Experiencing Voices in Electroacoustic Music

Andreas Bergsland

PhD Thesis
Department of Music
NTNU
2010

ISBN 978-82-471-2132-0 (electronic)
ISBN 978-82-471-2130-6 (printed)

Abstract

This dissertation presents a framework for describing and understanding the experience of voices in acousmatic electroacoustic music and related genres. The framework is developed with a phenomenological basis, where the author's own listening experience has been the main object of study. One component of the framework has been to group aspects that potentially can be attended to into *experiential domains* based on some common feature, relationship or function. Four vocal experiential domains related to the voice are presented along with three domains not directly related to the voice. For each of these domains, a set of concepts are introduced allowing for qualification and description of features of the experience. The second component of the framework, the *maximal-minimal model*, is partly described through these domains. This model presents *maximal* and *minimal* voice as loosely defined poles constituting end points on a continuum on which experienced voices can be localized. Here, *maximal* voice, which parallels the informative and clearly articulated speaking voice dominant in the radio medium, is described as the converging fulfillment of seven *premises*. These premises are seen as partly interconnected conditions related to particular aspects or features of the experience of voice. At the other end of the continuum, *minimal* voice is defined as a boundary zone between voice and non-voice, a zone which is related to the negative fulfilment of the seven premises. A number of *factors* are presented that potentially can affect an evaluation of experiences according to the premises, along with musical excerpts that exemplifies different evaluation categories along the continuum. Finally, the two frameworks are applied in an evaluation and description of the author's experience of Paul Lansky's *Six Fantasies on a Poem by Thomas Campion*.

Acknowledgments

There are a number of people who I wish to thank when this PhD project has now been realized into a finished thesis.

First of all, I would like to thank my supervisor, Carl Haakon Waadeland, for his support, encouragement, guidance and helpful comments throughout the work with this project, which have been immensely important over the 6 years that I have been working with it. Secondly, I would like to thank Rolf-Inge Godøy, my co-supervisor at the University of Oslo, for valuable feedback and for guiding me towards relevant research, especially in the starting phase of the project.

Furthermore, I want to thank the Faculty of Humanities at NTNU, which funded me with a 4 year fellowship and thus made it possible for me to engage in the project as a full-time job. The Faculty was kind enough also to grant me 2 months of additional support for finishing the thesis. Many thanks go to the Department of Music and my colleagues there, both for helping me with practical matters and for including me in the collegial community at the Department. I especially want to thank Tone Åse, Trond Engum and Øyvind Brandtsegg for being great colleagues, always helpful and interested.

In the first two years of my fellowship, I was a part of the interdisciplinary project *Aesthetic Technologies* at NTNU. The project had regular meetings where relevant issues were discussed, and it also hosted several international conferences, where I was allowed to present my work. This also resulted in the publication of two book chapters in two of the publications by the project. I want to thank all the participants in the project, but I especially want to thank Bodil Børset, with whom I had many stimulating and interesting discussions.

In 2005, I was accepted as a visiting researcher at the Sound Processing and Control Lab (SPCL) at McGill University, Montréal. I want to thank fellow students and researchers at the lab for letting me in on many interesting topics and for giving helpful comments to my project. I especially want to thank ass. prof. Philippe Depalle, who gave me invaluable help with getting good results with LPC. In Montréal I got the chance to record the voice of Nancy Helmes imitating Hannah MacKey, and she therefore also deserved my thanks. I also want to thank Paul Lansky, who I met during his stay in Trondheim, for his interest in my work and helpful comments related to my SFM-instrument.

Several people have been so nice as to read portions of my thesis along the way, and give me valuable feedback. I would especially like to thank Dawn Behne, Kåre Bjørkøy and Karl Jacobsen. Frank Ekeberg has also been of invaluable help in the finishing stages of the thesis. He has proofread my manuscript, giving me a great discount on his services, something I am very grateful for.

Lastly, I want to thank my parents for their support all along the way. I also want to thank my lovely kids, Kristian, Anna and Maria, who have had to live with a dad who sometimes had to stay long hours at the office, and who was sometimes not fully present when he was home. But lastly, and most importantly, I want to thank my dearest Mirjam, who has made it possible for me to go through with this project. Without her love, support and understanding, I wouldn't have made it all the way.

Trondheim, april 2010. Andreas Bergsland

Table of contents

Part I - Experiential Domains

Part II - The Maximal-minimal Model

Part III - Lansky's *Six Fantasies*

1.0 Introduction

1.1 The topic of the thesis

Hearing Åke Parmerud's *Les Objets Obscures* (1991, on Parmerud, 1994) for the first time was a turning point in my musical life. After having listened most of my life to music that was *played* by musicians, this piece presented a sound universe that couldn't be understood by the same premises as those I was used to. This time, I could hear no instruments in the ordinary sense, and I remember that it both fascinated me and made me feel alienated. Instead, there were a lot of sounds reminiscent of what I then heard as non-musical events, along with a female voice that was speaking rather than singing. This voice somehow seemed to bind the piece together through the recurrent appearances in the beginning of each of the four movements, thus functioning as focal points or centres around which the other sounds were distributed. And many of the sounds appeared to be either synthesized or under a heavy influence of some technological manipulation, including the female voice at certain points. At the time, the experience was in many ways *enigmatic* – I did not feel that I quite understood how to make sense of the piece, and still it was deeply fascinating.[1] The combination of this fascination and the urge to understand something partly enigmatic has followed me through the work with this thesis, where I have wanted to come to terms with this and similar kinds of experiences.

After my groundbreaking experience with Parmerud's piece I have later come to learn that this piece of music belongs to a genre called acousmatic electroacoustic music.[2] In this genre, the music is solely intended to be played back on loudspeakers, preferably on high quality systems, and therefore has no correlates in a live performance – the pieces of music are essentially fixed and exist only on record. While I could occasionally encounter melodies, rhythms and harmonies in a more traditional sense in these kinds of pieces, what most of them would highlight and explore, were other things: Firstly, the rich but much more ineffable world of *timbre*; secondly, a wide range of extra-musical associations; and thirdly, the possibilities of using sound technology as a primary means of composing.

[1] Later, I came to learn that the piece was in fact intended to be a kind of riddle, where the listener is challenged to discover "hidden objects" through the hints given verbally by the female voice, and sonically by the manipulated sounds presented.

[2] I will present a more concise definition of this genre in section 1.3.

After having heard a number of other works in this genre over the years, I have found that I have been particularly fascinated by pieces with voice. Many of these pieces explore the wide expressive possibilities of the voice, the huge potential for creating meaning on many levels, and the way in which voices can be shaped and modified by technological means so that they turn into chimerical creatures, strange hybrids or abstract instruments. Having written a M.A. thesis about the role of voice and poetic text in a contemporary orchestral lied, these pieces triggered a further interest, which ultimately led to the working out of the current research project on the topic of voice in electroacoustic music. The questions that I would like to address in this project are:

- How do we experience and make sense of voices in electroacoustic music?
- How can such experiences be described and compared?
- What factors affect these experiences?

In trying to answer these questions I will investigate relationships between theoretical and empirical knowledge in relevant fields and my own listening experience. Due to the many complex issues involved in musical listening, the multitude of meaning layers potentially carried by the voice, the ambiguities associated with not being able to see the sound sources, as well as the many transformations often applied, I have chosen to reject a systematic, empirical (i.e. experimental) approach in favour of a more open exploratory study. In this study, I will combine and structure existing theories and knowledge about the listening experience in electroacoustic music, studies of how people react and respond to voice in general and electronically manipulated voice in particular, with an exploration of my own listening experience. Therefore, the project will be both *interdisciplinary* and *phenomenological* in nature. Hence, I will in many respects follow prominent currents in the theoretical tradition within electroacoustic music, with people like Pierre Schaeffer and Barry Truax (see e.g. Schaeffer, 2002; Truax, 2001). In addition, I would like to explore the possibilities of using sound processing tools to manipulate voices as it is done in the music, to be able to see how different acoustical and control parameters affect the listening experience. I will describe further details about my approach in section 1.4.

During this thesis I will introduce two central ideas:

1) Firstly, I will present a framework of *experiential domains*. In this framework, I will group aspects of what can at the outset be a more or less holistic experience into a smaller set of constituent domains. I will then show how these domains are related to each other; that there are interdependencies between them, and that they can be organized into a kind of structure. By developing this framework, I hope to gain a better understanding and overview of the processes involved in the listening experience, but also to make it easier to communicate and demonstrate these processes to others. The resulting framework might therefore appear much "neater" and simpler than the situation is in reality, where different aspects might be much more entangled in each other and where distinctions might be much more blurred. At the moment, though, it is more important to develop a structured framework which perhaps simplifies and formalizes matters to some degree rather than pointing at ambiguities, paradoxes, and blurred boundaries that might undermine the framework.

2) The second central idea in my framework is the model of *maximal and minimal voice*. This model sets up two poles or extremes as reference points against which the experience of different types of transformed or manipulated voices might be judged and compared, namely the *maximal* and the *minimal* voice. The maximal voice can briefly and preliminarily be described as a typical informative and neutral speaking voice, resembling in many ways public radio broadcast voices. At the other end, the *minimal* voice is usually highly manipulated and often quite abstract, and thus defines the zone between what *is* voice and what *is not* voice. The imagined space between these two extremes is thought of as a continuum extending from a central zone, defined by the maximal voice, towards a peripheral zone, defined by the minimal voice. This continuum is also mapped out and described in a more detailed manner by formulating a set of *premises*, which can be seen as partly interrelated dimensions with which different vocal expressions in electroacoustic music can be evaluated. Taken together, this model expands and refines the idea that we generally experience transformed or manipulated sounds *in relation to* one that is unmanipulated (see e.g. Smalley, 1993; Smalley, 1997: 111-112; Schaeffer, 2004: 78-79). Moreover, it has parallels with theoretical models that have tried to describe the continuum between the concrete and reference oriented on one side, and the abstract and sound quality oriented on the other side (ten Hoopen, 1992a; Young, 1996; Chion, 1988; Emmerson,

1986). And, it embraces models that have described vocal sounds in electroacoustic music in terms of how intelligible any verbal material is for listeners (e.g. Segnini & Ruviaro, 2005).

Part of developing a framework for understanding and assessing the experience of voices in electroacoustic music has been to choose, often with a certain disappointment, what areas one will have to exclude or treat only superficially for some reason. By choosing the model of the maximal and the minimal voice and seeing that in relation to the different experiential domains, I have had to treat certain areas sparingly or even exclude them in order to maintain focus and delimit the scope of the project. For instance, aspects of the experience dealing with higher-level structures in compositions, i.e. beyond the single sound event or phrase, have been given little attention. The same goes for descriptions of the more properly "musical" features, such as pitch, duration, loudness and timbre. Therefore, the presented evaluations based on the developed framework must be seen as non-comprehensive; they present only certain aspects of the experience that particularly pertain to this way of viewing it.

1.2 Background: Voice as a musical element

In order to put the topic of this thesis into a broader perspective, I will now look into how and why the voice has been used in musical settings in general, and as a sound source for electroacoustic music in particular.

There are many indications that human beings have had an inclination to use the voice for musical purposes at all times and in all cultures. As Levman shows in his review of theories on the evolution of music, there have been many attempts to root the evolutionary origin of music in human vocalizations and to the development of language (Levman, 1992). Being the "instrument" that is ready made and always available, the voice could easily be used everywhere, at all times, and with no special skills required. Even if musical practices and notions of music have varied from culture to culture and have been changing through different times, it is believed that musical vocalizations and song plays and have played an important role in all cultures (Mithen, 2005). And, according to Mathiesen, the relationship between song, magic, science and religion (and by extension, state ritual) was very strong in all known ancient cultures as e.g. Egypt, Greece, Rome, the Middle East, Mesopotamia, India and China (Mathiesen, 2006). As for the Western musical tradition, one usually traces its practices back to Gregorian chant.[3] And even though vocal music has had variable prominence through the history of Western music, song has been a part of our notion of music ever since and indeed still is.

The branch of music that emerged in the latter half of the 20th century with the new technologies that allowed for recording, editing, and playing back of sound also incorporated voice. Pierre Schaeffer and Pierre Henry's first important work of *Musique Concrète* was *Symphonie pour un Homme Seul* from 1950, a work in which different types of voices were prominent sound sources (Schaeffer, 1998). Despite Schaeffer's placement of reduced listening practices, i.e. intentionally disregarding any associations to symbolic meaning or a sound source, at the heart of Musique Concrète, his sketches for this composition as presented in *A la recherche d'une musique concrète* display ideas which are overtly programmatic (Schaeffer, 1952). Here, Schaeffer envisions a composition where it is not so much the meanings conveyed by the voice as speech he is interested in. Rather, by using vocal sounds like humming, laughter, breathing and screams, the actions, the interior states and the exterior localization of the human beings are put into focus, *in addition to* the purely sonic properties of the vocal sounds that appears to be Schaeffer's primary interest (Schaeffer, 1952: 56-67). In many ways, Schaeffer and Henry's piece can be seen as a part of a more general tendency to explore and expand the use of the voice in music beyond singing, a tendency which can be traced back to Schoenberg and his use of the so-called "Sprechstimme" in *Pierrot lunaire* in 1912 (Anhalt, 1984).[4] An important part of this tendency, the exploration of the non-verbal repertoire, was later taken even further, with Berio's *Visage* from 1962 being among the most striking examples (Berio & Maderna, 2006).

The new technologies and techniques for sound generation, analysis and manipulation which blossomed in the 1950s appears to have triggered an interest in exploring the contact points between human voice and electronically produced sound. In what has been judged by many as the first real masterpiece within the *Elektronische Musik* tradition, *Gesang der Jünglinge* (1956, on Stockhausen, 2001), Stockhausen set out to create a continuum between sung speech sounds ("Sprachlaute") and electronic sounds, while imposing discrete steps in the

[3] The influences from Ancient Greek and Roman music are often taken to be of a more theoretical kind.

[4] Anhalt points out, however, that the tradition of the musical *melodrama* starting with Rosseau's *Pygmalion* used spoken voice that was more or less correlated to the instrumental accompaniment, and that Schoenberg's creation was therefore not *ex nihilo* (Anhalt, 1984: 7-9).

continuum that allowed him to use serial techniques in the organization of the continuum in the musical structure (Stockhausen, 1992; Stockhausen, 1960; Stockhausen, 1958). For Stockhausen the acoustic structure of speech sounds, based on insights from phonetics, therefore provided means for creating musical structures. The tendency to impose phonetic structures onto musical ones can also be seen with important composers like Ligeti and Berio in works like *Novelle Aventures* and *Thema – Omaggio à Joyce*, respectively (Anhalt, 1984; Murphy, 1999).

Following *Gesang der Jünglinge*, there were several others in which the idea of a continuum between vocal sound and other kinds of sound, either purely synthetic or recorded, was explored.[5] One important tendency in this exploration seems to have been to create *hybrid sounds*, which can be recognized as voice and something that is explicitly not voice at the same time, and to create temporally articulated *transformation*s or metamorphoses, in which there is a continuous or step-wise gradual transition between a vocal sound and a sound that is markedly non-vocal. Two classic pieces from the computer music repertoire that include both hybrids and temporal transformations are Jonathan Harvey's *Mortuos Plango Vivos Voco* from 1980 (Various artists, 1990) and Trevor Wishart's *Vox V* from 1986 (Various artists, 1989). Wishart, in particular, has emphasized the metaphorical potential that lies in linking voice in this manner to sounds from other sources (Wishart, 1996: 165-167).

Another tendency that is evident in electroacoustic music is composers' exploration of the musical qualities of *spoken* vocal material in their compositions. Berio's *Thema – Omaggio à Joyce* (Berio & Maderna, 2006) was ground breaking in that respect, applying a read passage from the 11th chapter of Joyce's *Ulysses* as the only sound material for the piece, subsequently subjected to extensive manipulation. In this piece, Berio wanted to extend the musicality that he found inherent in Joyce's text into the realm of music by fragmenting, manipulating, superimposing and restructuring the recordings of the reading of Joyce's text so that it was finally was loosened from its semantic bindings altogether (Dreßen, 1982). Several electroacoustic pieces from the 1970s and 80s were composed by similarly using speech, and in particular readings of literary texts, as the sole sound material for the compositions. *Speech Songs* (Dodge, 1994) by Charles Dodge and *Six Fantasies on a Poem by Tomas Campion*

[5] Bruno Bossis sees this idea as lying at the core of *artificial vocality*, i.e. expressions where sounds produced artificially, either recorded, transformed or synthesized and mediated by loudspeakers, resembles vocal production to a larger or smaller degree (Bossis, 2004; Bossis, 2005: 283-292).

(Lansky, 1994a) by Paul Lansky, which will be discussed in detail in chapter 12, are perhaps among the best known. The technology that was used in both of these pieces allowed the composers to change the intonation and the articulatory features independently, something that both composers used to make the speaking voices "sing" (Dodge, 1989; Lansky, 1989). In the following years, several other computerized techniques for processing and synthesizing the voice, such as frequency modulation (FM), formant wave functions (FOFs) and phase vocoding, were used increasingly by electroacoustic composers in exploring their musical potential (Cook, 1996; Wishart, 1988; Georgaki, 1998).

The link to literature is something that has always been present in the Western musical tradition, at least if one regards the Bible as literature, and the exploration of the sound qualities in speech and language through sound recording and manipulation technologies has also been conducted from a literary vantage point. With roots in the experiments and ideas of Italian and Russian futurists and the Dada movement, poets in Europe and America started to experiment with tape recorders in the 50s and 60s (Battier, 2003; Katz, 2004: 108; Wendt, 1993). French sound poetry, which included people like Henri Chopin and François Dufrêne, as well as Swedish text-sound composition, were perhaps the most prominent groups in that respect. For many of the poets following this tendency, sound fixed onto a recording rather than writing became the most important medium of expression, and sound technology became one important factor in the development of sound poetry (Hultberg, 1994). Thus, the exploration of the sounding qualities in language through sound was conducted from both a literary and a musical vantage point, resulting in a grey zone between music and poetry and a significant portion of cross-fertilization between the art forms.

To sum up, the richness and variability in the huge range of sounds that can be produced by the voice, adhering to established practices of sound making or not, has attracted composers and sound oriented poets in the hunt for interesting source material for their compositions. The structures of speech and language have also caught the interest of composers and have provided means for musical structuring and cross-fertilization. Furthermore, the special range of significations that the voice can evoke has been an attractive field for electroacoustic composers to explore, with verbal, non-verbal and mimetic meanings, literal as well as metaphoric. Barrière, commenting on his work *Chrèode* (1984), a work with a primary reference to the human voice, therefore provides a pertinent conclusion of this section:[6]

> The reference to vocal material has therefore real mimetic value; on the one hand it provides schemes for organization of material which we may draw upon, transform, anamorphize (in other words, it offers the possibility of a grammar); on the other hand, it is a carrier of meaning, it speaks to us more intimately than any other reference, and it furnishes a real learning experience for the imagination and perceptions. (Barrière, 1984: 183)

1.3 Choice of material

Throughout this study I will make reference to a rather broad set of artistic expressions and my experiences of these. This makes up the primary material upon which my analysis and discussion will be based. Four criteria have been important in selecting this material:

1. The material should include vocal sound.
2. It should be a part of an artistic and aesthetic context/purpose.
3. Its primary form of existence should include sound fixed onto a medium.
4. The qualities of the sounds in and of themselves should be a part of the aesthetic function.

As the forthcoming discussion will show, the criterion of "vocal" or "vocality" is not straightforward, but I will leave the intricacies of this criterion until later. The criteria 2-4 combined gives relatively clear boundaries, as I see it. They will, for example, exclude audio documents or broadcasts that are primarily non-artistic in nature, where the *informative* aspect is the most important, like in interviews, talks and speeches.[7] Moreover, it will leave out audio-books and recordings of readings of traditional poetry, because these artistic expressions primarily exist in a written form. I will also avoid all kinds of artistic expressions that are primarily presented live, that is, as a part of meeting between performers and an audience, thus excluding both recordings of live acoustic performances and oral poetry. Instead, I will focus on

[6] This piece was composed at IRCAM using the *Chant* and *FORMES* programs for synthesizing vocal-like sounds.
[7] These kinds of vocal sounds are, however, often appropriated into a musical context, often in fragmented, restructured or manipulated forms.

artistic forms of expression fixed onto a medium where this is considered to be the *artwork in itself*. That is, this artwork may be presented before an audience in concerts, but such a presentation is not considered essential for the reception of the artwork, but rather enhancing and amplifying its inherent qualities through playback on high-quality sound systems, often with possibilities for control of spatial parameters. Thereby, the *acousmatic* branch of *electroacoustic music*, as defined by Emmerson and Smalley, is an appropriate label for most of the material I have chosen in this thesis: "Music in which electronic technology, now primarily computer-based, is used to access, generate, explore and configure sound materials, and in which loudspeakers are the prime medium of transmission […] Acousmatic music is intended for loudspeaker listening and exists only in recorded form (tape, compact disc, computer storage)" (Emmerson & Smalley, 2009). I would also like to add to this definition, which is primarily focused on the way that technology is applied in the composition process and the medium on which it is presented, that I have focused on works that are what Leigh Landy calls *sound-based* rather than *note-based*, i.e. an "art form in which the sound, that is, not the musical note is a basic unit" (Landy, 2007: 17). In many cases this also means that I am considering works that by some may be said to fall outside of what is regularly considered to be "music". For example, I include several works regarded as *text-sound compositions*, a genre born from the grey zones between music and poetry as discussed in the previous section, and defined thusly by Sten Hanson: "Text-sound composition is a mixed art form standing right in the middle of poetry and music; a poem in which the speech sounds and the voice itself play equally important roles – or a bigger – role than the signification of the words, or a piece of music where the human voice and the music of language itself form the basis for the composed sound" (Sten Hanson, cited in Hultberg, 1994: 69, my translation).

In addition to the core material that I have delineated in this section, I will make occasional references to works outside these boundaries. The reason for this is simply that these works can enlighten the discussion of a particular topic in a particularly pertinent way. Therefore, I have also included works from genres like *soundscape composition*, *radiophonics* and *audio-visual art*. The criterion of pertinence to the discussion has also been an important criterion for the choice of all the other works *within* the defined boundaries.

1.4 Method

I will now delineate the central methodological strategies that I intend to apply when approaching the central questions in the thesis as presented in section 1.1. I will follow three main strategies: 1) *Interdisciplinarity*, 2) *phenomenology* and 3) *analysis by synthesis*. These will be described in separate sections in the following.

1.4.1 Interdisciplinarity

The voice is a universal phenomenon which has triggered scientific interest in a wide range of scientific fields from voice acoustics to linguistics and music perception. And within many of these fields, one has investigated vocal expressions similar to those used in electroacoustic compositions and the related genres as delineated in section 1.3 above. This first and foremost applies to the use of different kinds of manipulated and synthesized voices, which has flourished within research on the voice. Moreover, it is also significant that both my material and most research of voice and its perception operate with pre-recorded sounds that are mediated over loudspeakers. Therefore, an investigation of research literature from a wide range of fields can be of value in my study. Here are the research fields that I will take into consideration:

- Music performance (singing) studies
- Music perception
- Auditory perception
- Voice acoustics
- Voice physiology
- Voice perception – identification, recognition of speaker/voice features
- Speech perception – linguistics / phonetics
- Research on voice and emotions
- Research on synthetic voice and its perception
- Literary theory
- Cognitive science – research on categorization, metaphors in cognition,
- Media (radio) research

- Research on presence in virtual realities
- Information theory
- Neuroscience

Godøy has written that "the interdisciplinary scenario is a risky one, the chances of falling into a no-man's land, both as author and reader are very real" (Godøy, 1997: 34). Naturally enough, I have no pretentions of having comprehensive knowledge of any of these fields – they would each of them require years of study. Nor will I attempt to give a thorough account of any of these areas of study, which would fill several dissertations. What I hope for, however, is that I can be able to combine insights from different fields dealing with everything from low-level acoustic features to higher-level aspects related to meaning, so as to form a broad theoretical basis against which aspects of my own experience of voice can be held.

In addition to the problems of not being a specialist in all of the above areas, one faces challenges related to what Godøy calls the "hidden implications" within each field, i.e. underlying but not explicitly stated theoretical and perhaps ideological foundations that may underlie a field or a discipline (*ibid.*: 35). As I see it, research paradigms relying on the experimental method can be regarded as carrying such "hidden implications" that one needs to be conscious of. As I have gotten to know the fields above, it appears that in the majority of cases, the study of perception and cognition of voice, speech and music uses an experimental method which sets up a situation for a listener that is in many ways different from the ordinary daily activities of listening to a piece of music or listening to voice and speech in communication situations:[8] Firstly, the situation of being observed and "measured" might in itself affect listeners. The same goes for the environment of a laboratory, which may appear sterile and alienating to some. What is more, in many studies the stimuli used are radically simplified and/or taken out of context, so as to allow for measurements of the effect of a single or a small set of variables, resulting in stimuli that are highly artificial and which only vary in minute details. As I see it, this makes it necessary to complement experimentally oriented research with research that has a higher ecological validity, i.e. that deal with sounds and events more similar to those that are experienced in the real-world.

[8] As Carterette and Kendall notes, this is often a question of giving a privilege to reliability before validity because of the complexity of situations with higher ecological validity and the lack of methods that can satisfactorily deal with such situations (Carterette & Kendall, 1995: 3-4).

The interdisciplinary approach is well integrated into some fields. For example, in some areas of general auditory perception the knowledge of voice and music perception is high (e.g. Bregman, 1990; McAdams, 1984). In such fields, one can therefore see many examples of a similar way of using an approach which tries to establish knowledge about relationships between a wide range of physical and experiential phenomena on the basis of a wide range of methods and stimuli. I find Albert Bregman's book *Auditory Scene Analysis* an excellent example in that regard (Bregman, 1990).

1.4.2 A phenomenological approach

An important component of my method will be an investigation of my own listening process. The reliance on a subjective listening experience has a well established tradition within the studies of electroacoustic music, mainly owing to Pierre Schaeffer and his treatise *Traité des objets musiceaux* from 1966 (Schaeffer, 2002). Here, Schaeffer stresses the lack of direct and linear correlations between perceptual phenomena and the physical properties of sounds, "suggesting a psychological distortion of physical 'reality' and demonstrates that perception cannot be reduced to physical measurement" (Schaeffer, cited in Chion, 1983: 24, translated by John Dack). For Schaeffer, the main focus of study is on the experiential more than the physical issues of sound, and in his treaty he aims at charting the different aspects of the musical object.[9] Thus, he shows clear affiliation with a phenomenological tradition, particularly with Edmund Husserl, whom he also acknowledges explicitly: "For years […] we have been doing phenomenology without realizing it […] it is only after the event that we recognized Edmund Husserl's heroically rigorous definition the concept of the object postulated in our research" (Schaeffer cited in Chion, 1983: 32, translated by John Dack).

Similarly, my approach in this project can be seen as having a phenomenological component, since I attempt to investigate my own listening process through internal subjective inspection. Moreover, the aim of seeing how the experience is formed by my own previous experiences, knowledge and predilections and how I orient my attention and consciousness during listening is similar to other approaches that are explicitly phenomenological (Ferrara, 1984). More specifically, reflecting on my own experiences and the act of experiencing my

[9] The musical object is a sub-class of the sound object (*objet sonore*), which is an object constituted by consciousness rather than a material one (see Chion, 1983: 34-35)

approach resemble what Husserl called a *phenomenological reduction* (Føllesdal, 1989: 302-303).

Here, I have to comment shortly on the similarities and differences between this type of phenomenological reduction and the way that Schaeffer applies this term. Schaeffer associates phenomenological reduction with what he calls *reduced listening*, i.e. “putting in brackets” or suspending the question of the origin and semantic associations of the sound object, while focusing on the sound for its own sake. As Schaeffer, I am not interested in the true origin of the sound and the semantic associations that any of the sound making individuals *intended to communicate*. Rather, I am interested in the sound sources and any other semantic association *as they are experienced*. That is, I will suspend the question of *actual* sources and intended meaning: In listening to recorded sound any sources and causes are ultimately *virtual* – they are only constituted through the act of listening.[10] This is not to say, however, that the actual sources and the intended meaning do not play a part in listening: Both can certainly affect the listeners’ experience if they have any knowledge of them. Here, however, they are not the object of study.

The main difference between my own and Schaeffer’s version of the phenomenological reduction is that where he put the question of sound sources, causes and associations “in brackets”, I want to include all those aspects in my investigation. Thereby, Schaeffer’s and my own version of the phenomenological reduction through listening differ in the number of aspects that are included in the process, where Schaeffer is more restrictive than I, but where we share the same analytical approach to the subjective listening experience.

The phenomenological concept of *intentionality* can also be seen as central to my study, i.e. even if I will often use other related terms like “attention”. For Husserl, intentionality designates a “directedness” of consciousness towards an object, where “object” has to be understood in its widest sense, including living beings, actions, processes as well as physical objects (Føllesdal, 1989). Moreover, the term also implies that our consciousness adds or fills out some aspects in additions to those that are perceptually given through the experience. For example, when we look at a chair, we can see it only from one side, but we still experience the chair as one whole unit with many sides. In this thesis, the concept of intentionality is closely related to the *experiential domains* mentioned above, since these domains embrace a set of aspects towards which one’s consciousness (or attention) can be directed.

[10] See the discussion in section 2.6.3.

There are also points in which my approach perhaps might be considered less phenomenological. Ferrara states that “a distinctive phenomenological tactic is that, rather than manipulate a work through a formal grid of analytical questions or positions, one responds to questions posed by the work (Ferrara, 1984: 356). By setting up a framework for describing and evaluating my experience, it can be seen as shaped by a kind of “grid” or system. Thoresen also distinguishes an *open* kind of reduced listening, not oriented towards any particular features, and a more specialized, categorizing way of listening, the latter represented by Schaeffer’s typo-morphological framework (Thoresen, 2007b: 132). As Thoresen warns, this latter approach might quickly impose conceptual prejudices on perceptual givens. But, as he notes, this might be the disadvantage of any attempt to codify aural phenomena (*ibid.*). The open approach, on its part, has the disadvantage that the experience can be difficult to communicate through words. Even if I have often practiced a more open kind of listening for numerous pieces during the work with this project, especially in the early stages, the focus on issues related to voice in electroacoustic music has admittedly lead to a more selective type of listening, where the listening process has been affected by the theoretical framework. Thus, I have chosen to emphasize the communicability and consistency before openness, hoping that some of the issues that I have experienced in the early and more open exploration of material have carried over into the theoretical framework.

A potential criticism against the delineated approach might be that it renders results that are difficult to validate or repudiate and that it might delimit the potential relevance for other people. However, it is not my aim to focus on the personal aspects of my experience for their own sake in this thesis. I will rather aim at showing how particular aspects of an experience rely on different sets of previous experiences, of abstract knowledge, practical skills, cultural codes and conventions, and thereby to highlight the *relativity* of the experience. Through this relativity, however, I hope to delineate a framework that is *intersubjective* more than subjective in nature, because many preconditions can be assumed to be shared. For example, one can assume that the great majority of healthy individuals living in groups use speech to communicate, even if the codes that are used and the significations that are referred to will be valid for a smaller group of people. So, even if I possess a unique combination of previous experiences, knowledge and skills, many of these will be shared among a larger or smaller group of people and therefore also be intersubjective. The *degree of generality*, whether it is shared for a smaller or larger group of people, will naturally differ and it will always represent a certain degree of speculation to estimate how large this group is. Such questions will not be addressed to a large extent, however, since they would have to be based on extensive cross-cultural studies within fields such as physiology and psychology and thus be far beyond the reach of this investigation.

1.4.3 Analysis by synthesis

In the course of this dissertation I will also apply the principle of "learning by doing", often conceptualized as *analysis by synthesis* (see e.g. Risset & Wessel, 1999). The idea is that by making sounds with the same or equivalent means and techniques that are used in the production of the sounding artworks that I study, I will attain an enhanced general knowledge of the relationship between sound sources, the technological processes involved and aspects of the listening experience.

Moreover, another possibility this method opens is what Godøy calls the *epistemology of simulations* (Godøy, 1997: 295-296). By this he refers to the possibility of creating variants of a sound event or object where a certain trait or aspect is changed while others remain the same, and then observing and comparing the effect that this has on the experience. This strategy will be relevant to adopt in some cases where it would be useful to demonstrate the relationship between a certain control parameter, and perhaps also an acoustical parameter, and aspects of the listening experience.

I will mainly adopt analysis by synthesis as strategy in the exploration of Paul Lansky's *Six Fantasies*. By making a model that can produce sounds that resemble those in Lansky's piece, I hope to be able to investigate how sound and control parameters are related to the different experiential domains and the premises of the model of the maximal and minimal voice.

1.5 Outline

This dissertation is divided into three main parts:

1. In the first part, which comprises chapter two and three, I will delineate a set of *experiential domains* where each domain embraces a group of aspects that tend to be

experienced as belonging together, either on the basis of being related to a more or less unified and coherent sound source, behavior, semantic or aesthetic function. Here, I will try to apply and combine theories from the field of electroacoustic music studies and theories of voice and speech as a basis for the differentiation of such domains. I will then argue how these experiential domains can be organized hierarchically by referring to what is considered as the "materiality" of each of the domains and how it produces meaning for a listener. I will start by delineating a general set of experiential domains which might apply to all kinds of electroacoustic works (chapter 2), and then look into the experiential domains specifically related to the voice (chapter 3).

2. In the second part of the thesis, I will develop the mentioned theoretical framework of the *maximal* and the *minimal voice*. These concepts have been borrowed from the literary theorists Donald Wesling and Tadeusz Slawek (Wesling & Slawek, 1995), but as I will show, similar ideas can be found in studies of radiophonics and electroacoustic works. Then, I will try to break these concepts into seven *premises*, which are equivalent of Lakoff's cognitive models (Lakoff, 1987). The maximal-minimal model will also be discussed in the light of categorization. In chapters 5 to 11, I will elaborate on each of the premises in turn. I will try to map out what factors that can potentially have an effect on the particular aspect of experience dealt with by each premise, and show how the premises can be used to delineate criteria for making evaluations of segments of music according to the maximal-minimal continuum. On the basis of these criteria, I will then evaluate a few musical examples to illustrate and exemplify the evaluative side of the framework. In this evaluation, I will also make reference to the mentioned factors by suggesting how they have affected my listening experience in each case. This process will hopefully not only function as an argument in the construction of the framework, but also shed some light on the experiences of each of the pieces used as examples.

3. In the last part of the thesis, which comprises chapter 12, I will apply the evaluative framework delineated in the previous chapters on Lansky's *Six Fantasies on a poem by Thomas Campion*. To reduce the size of the material for evaluation, I have chosen shorter excerpts from the six movements in the piece, and each of these excerpts is then

described with reference to the experiential domains and evaluated in relation to the criteria for each of the seven premises in my framework. The evaluations are then interpreted in two different graphical forms, one privileging the temporal evolution of the evaluations, the other giving a better overview of the relationship between the evaluations of the different premises. As a part of this evaluation, I will also look at how the different evaluations could have been different if certain parameters were set or behaved in a different way so as to further illustrate the potential of the analytical framework. To do this, I will synthesize examples that approximate how the piece hypothetically could have been, and how this might have influenced the evaluations. Arguing that the excerpts are representative for a certain configuration of aspects relevant for the evaluations and certain ranges within which the aspects vary, I will then compare the evaluations for each of the six movements to see if this can give some interesting perspectives on the piece as a whole. These evaluations will also be compared with what other scholars have written about Lansky's piece. All in all, I hope that this will demonstrate how the evaluative side of the framework can be applied, as well as giving insight into how my previous experiences and knowledge along with the sounding structures of the piece have influenced the listening experience of *Six Fantasies*.

Chion has written that "part of the research into the sound object consists in defining new hearing intentions which groups of researchers can agree upon, with the help of a new vocabulary" (Chion, 1983: 30, translation by John Dack). I hope that this thesis will define a set of "new hearing intentions", new things to listen for in compositions with voice, in addition to constituting a novel analytical framework.

Part I

Experiential Domains

2.0 Experiential domains

2.1 Introduction

In trying to understand listening experiences involving voice in electroacoustic music and associated artistic expressions, finding suitable terms for describing different aspects of such experiences is an important step, grouping and structuring terms and aspects in relation to each other is another. Before turning to the aspects that are specific to the voice, however, it is necessary to delineate a framework that can account for how the special conditions of the acousmatic electroacoustic work might structure the experience. This includes seeing the possibilities of focusing on the traces of technological and compositional processes in the work, the sounding qualities of the voice itself, as well as the spatial layout and environment that the voice appears to be situated in. The main focus of this chapter will be to establish an understanding of these conditions and to be able to describe aspects that are related to them. The central concept in this respect is that of the *experiential domain*, to which I will now turn.

2.2 Experiential domains

As a first step in this process I would like to propose the term *experiential domain* to designate a number of aspects or properties of an experience that we tend to *group* together for certain reasons. I have chosen a term which refers to the more general phenomenon of experience rather than *listening*, even though what I will be studying is essentially experiencing *through* listening. This is because I would like to emphasize that what we experience often makes reference to other modalities as well – indeed to how we relate to the world in general. Thus, even though my study uses listening as method, I feel that by using the word *experiential domain* I embrace processes that are not particular to listening, but deal with more general ways of perceiving and making sense of the world. This is in line also with the phenomenological approach which I delineated in section 1.4.2. Don Ihde, who has attempted a phenomenological study of listening and voice, underscores that listening is indeed an activity of global character:

> […] through concentrating on auditory experience, a reevaluation of all the 'senses' is implied. For the first gain of phenomenology in regard to sensory experience is a recovery and reappreciation of the fullness and

> richness of the global character of experience. The very notion of an auditory dimension is problematic for phenomenology. (Ihde, 1976: 21)

What is common to all the experiential domains that I will introduce in this (and the next) chapter(s), is that they point to certain properties or aspects of a certain event or object that can be put in the centre of our attention, either actively or passively. We can choose actively to focus on one particular aspect of a sound, or we can be drawn more passively to it because its inherent properties are particularly salient, relevant or interesting to us. They embrace what one in phenomenological terminology would call products of *intentionality* – of directing one's intention towards something so as to constitute an object of consciousness (Føllesdal, 1989: 295). In my framework, an experiential domain will represent a class or group of experiences of such objects based on some common feature, relationship or function.

The grouping of the different aspects into these experiential domains has a lot in common with many theories of listening. Many authors have divided listening into different modes, types, levels, relationships or behaviors according to a similar set of criteria (see e.g. Schaeffer, 2002; Norman, 1996; Smalley, 1992; Bayle, 1989; Delalande, 1998). What I will do in the following is to ground and substantiate my differentiation of experiential domains on the basis of theories of listening and the particular conditions that apply for the object of study in this thesis.

The three domains that will be the focus of this chapter are what I will refer to as the *non-vocal domains*: The domain of sound qualities and structures (**SQS-domain**), the domain of technology, composition and mediation (**TCM-domain**) and the domain of space and environment (**SE-domain**). These domains are all considered to have *intrinsic* orientation, i.e. they embrace aspects that are experienced as directly related to the musical work. These stand in contrast to those that have *extrinsic* orientation, which include aspects related to the particular conditions of presentation and listening; the acoustical features of the room, the sound system applied, etc. In addition, the mentioned three domains share a dominantly *outward* orientation rather than an inward one – i.e. intentionality is directed outwards towards the incoming sensory information from the outside world rather than towards one's own bodily and mental response to this information. The *inward* orientation, on its part, implies focus towards what I refer to as the experiential domain of *body and mind*.

The inclusion of these three domains and the exclusion of the extrinsic and the body and mind domains is not made because I do not consider the latter two important, because they indeed are. The sounding result of playing an acousmatic work will always be a product of the equipment on which it is played back in interaction with the space it is played in, and we will we will always have our bodies and emotions with us in the experience, even if it need not necessarily be in the conscious focus of our attention. Nevertheless, the choice has simply been necessary to delimit the focus and scope of the dissertation.

In the following, therefore, I will start by discussing domains with inward and extrinsic orientation summarily, so as to be able to provide the basis for a more comprehensive overview and discussion of interrelationships that I will go into after the presentation of the domains in focus. What I will then do, is to try to ground and substantiate the mentioned three domains, and to specify the most relevant areas of experience that can be subsumed within each of them. Subsequently, I will present a model of how the experiential domains can be seen in relation to each other, also taking the distinctions and boundaries between inward and outward orientation, and extrinsic and intrinsic domains into account. Lastly, I will discuss how the distinctions between the different domains to some degree correspond to *ontological levels*, and how different levels of virtuality and reality can be played out in an acousmatic work.

2.3 Experiential domains with inward and extrinsic orientation

2.3.1 Body and mind domain

In many situations during listening, one's attention gets turned to one's inner world of bodily sensations or emotions more than the external properties we receive information about through our senses. Rather than focusing on what is going on in the music, we turn to the way the music make us feel – if the music gives us shivers down our spines, if it makes us relaxed and blissful, or if it evokes in us a strong urge to move. In so far as these things become the object of conscious attention, and it surely often does, such situations exemplifies the *internal* world of body and mind as an experiential domain in its own right.

That such a focus is pertinent in listening to electroacoustic music is evident from theoretical accounts as well as more empirically directed studies of listening modes or behaviours directed at body, emotions and self. For instance, Smalley's theoretical notion of the *reflexive* subject-object listening relationship is described as "subject-centred and is concerned with basic emotional responses to the object of perception" (Smalley, 1992: 520). François Delalande, basing his conclusions on qualitative interviews of listeners, describes the *empathic* listening behaviour as directed more towards the body than toward emotions (Delalande, 1998). This behaviour is characterised by attentiveness to sensations and the physiological products of the sound, where listeners speak of sounds as if they have been subjected to the movements implied by the sounds themselves.[11] Elisabeth Anderson, who has taken Delalande's approach with investigating actual listening behaviours further, has inferred a listening behaviour from a set of interviews that she calls *self orientation*, which comprises emotional as well as physiological responses, given the label of *sensation* in her framework (Anderson, 2007: 24-25). In addition, she notes that many listeners' responses are relatively neutral emotionally, something which she does not directly relate to physiological reactions. Rather, these responses, which she groups under the label *evaluation*, are more contemplative and intellectual, but personal in nature, and thereby still directed *inwards* toward the listener rather than focused outwards toward the object. Thereby, Anderson ends up with three subgroups of the listening behaviour of self orientation: 1. Sensation, 2. Emotion, and 3.Evaluation. In also taking evaluation into account, one sees that the link to the phenomenological approach I am taking in this project becomes apparent, because in doing a phenomenological study, I will indeed have to put my own listening into focus. That is, it may be difficult, or even impossible to direct one's attention toward what goes on in the music *simultaneously* as one "monitors" this experience. Rather, it will be a question of going back and forth between involved listening to the music and retrospectively making the experience into an object that one's attention can be directed toward. Therefore, Anderson's subsection *evaluation*, will indeed be one that will be in focus, but as a method rather than as an object of study on its own.

To conclude this section, I just want to emphasize again that I do not consider body and mind as unimportant or irrelevant aspects of the experience by excluding them from consideration in the further development of the framework, but merely that I have chosen to focus on the experiential domains that are intrinsic, i.e. experienced as pertaining to the acousmatic work in question. Moreover, body and mind will be implicit in some parts of the framework described in the next chapter, where I will link the perception of vocal gestures to the

[11] Another listening behaviour that is also related to the physiological sensations afforded by the music is what Delalande calls *immersed listening*, in which the opposition between the inner and outer is erased, and the music is experienced as surrounding and immersing the whole body.

unconscious “simulation” of production of equivalent gestures during perception when certain conditions are fulfilled. However, my main interest will still be the conscious reactions during listening rather than any involuntary response pattern.

2.3.2 The extrinsic domain

What I have called the extrinsic domain embraces all those aspects of the experience of acousmatic electroacoustic music which are heard as not being directly a product of the work, but which are linked to sources and causes extrinsic to it: the space that the music is heard in, the technology used in the presentation of the work, any aspects that can be attributed to human or automatic agents imposing some sort of control or influence upon the sound distribution or playback, or any sounds that do not come from the loudspeakers. Thus, this domain embraces directing attention toward what Chion calls *external space* (Chion, 1991) and what Smalley and Ekeberg calls *listening space*, which for all three writers has both a purely acoustical as well as a technological side (Ekeberg, 2002; Smalley, 2007). In Ekeberg’s words “this comprises the physical listening environment with all its acoustical characteristics, the type and arrangement of the sound system as well as the listening position relative to the loudspeakers and the physical boundaries of the listening environment” (Ekeberg, 2002: 19).

As for the acoustical properties of the room, with surfaces, objects and listeners that reflect, diffract and absorb, they may all contribute to the experience, but both diffraction and absorption will usually not be something that one can actually attend to during listening. It might be easier, then, to focus on the reverberant properties of a room, especially if reverberation times are long and if certain frequencies of the room or objects are reinforced so as to create resonances or more irregular vibrations.

As for the technological side, it includes in principle everything from the playback device to any other devices that in any way affects or modifies the sound (e.g. equalizers, filters, delay, converters), loudspeakers, power supply and even cables. In practice, however, one rarely focuses on such aspects during music listening unless there are explicit audible artifacts or faults such as humming, buzzing, clicks, distortion, feedback or other kinds of resonances that are clearly *not* a part of the composition. In some cases one might also react to the frequency characteristics of a sound system if it deviates markedly from what we usually listen to or what we take to be the “standard” quality of reproduction. In such cases, the act of *comparing* with one’s internalized mental templates will be important in the experience. Moreover, the sound system can also be under the influence of external devices that we clearly recognize as such – typical these days are cell phones, whose transmitting signals are received and amplified by the systems, something which clearly can cause a marked distraction and irritation.

When it comes to the influence of human or automatic systems for sound distribution, it will usually be integrated more seamlessly into the whole experience, so that it will be more difficult to assign single attributes or properties of the sound to either the musical work or the sound distribution system. If one knows a piece very well and have heard it being played in numerous different spaces by numerous interpreters, however, one might be able to attribute properties to the contribution of the diffusionist or diffusion system.

In addition to the factors included by Chion, Smalley and Ekeberg, I also want to add sounds from the listening *environment* to the extrinsic domain. Such sounds may direct our intentional focus toward them, for instance when a member of the audience at a concert coughs, or when somebody living upstairs from your domestic listening space draws his or her chair out from under the table so that one can hear it. Since acousmatic electroacoustic music as a rule is fixed to a recording medium and reproduced through loudspeakers, other sounds than those emanating from the loudspeakers will be regarded as extrinsic and in most cases regarded as unwanted distractions.[12] This is not to deny that in rare cases, environmental extrinsic sounds can also engage in fortunate interaction with the intrinsic sounds so as to be enjoyed and attended to in their own right.

2.4 The domain of sound qualities and structures (SQS)

Listening with a focus on the qualities or properties of the sound “in itself” is closely related to the practice of *reduced listening*, introduced by Pierre Schaeffer (Schaeffer, 2002: 270-272; Chion, 1983: 33-34). According to Schaeffer, when engaging in reduced listening, the listener is to get rid of all associations, references to sources, causes or symbols that the sound possibly

[12] One cannot avoid mentioning this without also mentioning John Cage’s radical inclusion of “unintentional” sounds into the spheres of musical experience with *4’33’’*. In this piece the lack of sound being made by the performer was to direct the attention of the audience towards the musical qualities of any sounds produced unintentionally in and in the proximity of the listening environment.

could evoke in favor of the *qualities of the sound in themselves* (Chion, 1983: 33). In other words, it is the material qualities of the sound rather than whatever the sounds may *refer to* that is the object of the intention of the listener in reduced listening, as is the case in the qualities/structure-orientation in my framework.[13] Moreover, for Schaeffer, reduced listening was an approach that allowed the listener to qualify and describe a sound object according to a set of what he calls typological and morphological criteria, many of which I will discuss further below. For now it will suffice to say that such qualities might for instance be the shaping of the sound in time – its attack, sustain and decay – the brightness of the timbre, the fluctuations in pitch, and so forth.

Many authors have used the term *abstract* or *abstraction* for a type of listening focus that resembles Schaeffer's reduced listening, and for situations in which no known source or cause can be perceived. Such terms appears to imply an attitude that abstracts qualities from more concrete objects and thereby also seems opposed to the more "concrete" and "direct" focus" toward things and events (see e.g. ten Hoopen, 1992a; Young, 1996; Chion, 1988).[14] Other terms used in electroacoustic theory that have related or equivalent meaning are *aural discourse* (Emmerson, 1986) and *spectromorphological* (Smalley, 1986; Smalley, 1993; Smalley, 1997). And, in many cases these terms are opposed to source/cause-related aspects, some of which have been labeled *mimetic* (Emmerson, 1986), *source-bonded* (Smalley, 1993; Smalley, 1997) and *reality* (Young, 1996), and for ten Hoopen (1992) and Young (1996) the oppositions are seen as poles or ends on a continuum between one and the other.

The term *abstract* also provides a link to the more structural properties of sound, which I have located in the same experiential domain as sound qualities in my framework, since focus on structural and formal disposition in the musical unfolding appear to offer a similarly abstract set of properties that one might attend to. For example, a formal disposition with two similar parts interposed by a different part (ABA) is applicable to all kinds of content, in the same way as a

[13] I want to emphasize, however, that I do not necessarily see the listening methods associated with reduced listening, i.e. especially cutting out or de-contextualizing a portion of the sound and repeating it continuously, as implied in the framework of experiential domains. In part II of this dissertation I will discuss several factors that may contribute to directing attention to different experiential domains.

[14] Schaeffer does not use the term *abstraction* in exactly the similar way, seeing *comprehending* as an abstract listening mode, and after an important crux in the argument of his *Traitè des Objets Musiceaux* he even sees the mode called *listening* (*écouter* – source/cause-oriented) as abstract (Chion, 1983: 39, trans. John Dack).

sound with a noisy sound spectrum can be produced by a number of different sources.[15] In short, the point with regarding these aspects together, which comprise everything from large-scale structures in a piece to variations in spectrum or pitch on a micro-level, is that both deal with properties that are abstract in the sense of being removed from the sources and causes related to the production of the sound as well as any other kind of signification or referential meaning carried by the sound.

While the presented definition of reduced listening suggests that reduced listening is incompatible with and opposed to a focus on referential aspects, there are writers that suggests that the boundary between these two is gradual rather than distinct. For instance, ten Hoopen and Young among others has proposed that terms like *abstraction* (sound quality focus) and *reality* (source/cause focus) constitute two ends of a continuum (ten Hoopen, 1992a; Young, 1996). Several examples can support such a claim.

Firstly, certain qualities of sounds, for example, seem difficult to describe without using source-related concepts. In Schaeffer's morphological framework, for instance, he distinguishes between three types of *allure*, which can be described as characteristic fluctuations in the sustainment of sounds, namely *mechanical*, *living* and *natural* (Schaeffer, 2002: 557-559). While *allure* is a morphological criterion which should describe abstract qualities of the sound, these concepts clearly refer to sources and causes.[16] Thus, it seems like certain qualities are difficult to describe without any reference to sources and causes.

Secondly, the inherent ambiguity of many sound sources in electroacoustic music can make the focal boundary between sources/causes and qualities/structure difficult to define. For many sounds, our attribution of sources or causes can be very vague and ambiguous, and what we are left with might be close to abstract descriptions, for example as when we can state that something is an "*impact* sound". In such a case, the notion of the impact is naturally rooted in an orientation toward objects (sources) and their behavior (causes), but the lack of any further

[15] The linkage between sound qualities and larger scale structures also parallels Smalley's spectromorphological framework, which covers everything from small scale properties to large scale structures: "In my spectromorphological approach, the concepts of gesture and texture, motion and growth processes, behaviour, structural functions, spectral space and density, and space and spatiomorphology maybe applied to smaller or larger time-spans which maybe at lower or higher levels of structure" (Smalley, 1997: 114). Thus, in Smalley's framework one can operate on several levels with the same concepts, several of which would correspond to what would have been labeled typological or morphological criteria in e.g. Schaeffer's framework.

[16] As Chion sees it, however, this does not go contrary to the rules of reduced listening, but rather attests to an interdependency between source/cause-oriented and reduced listening (Chion, 1983: 160, trans. John Dack).

qualification of the situation makes the event appear somewhat abstract.[17] The acousmatic situation together with electronic processing can also make the sources/causes completely unrecognizable so that only the *behavior* of the sound appears to be familiar. For instance, when we hear a sequence of short sounds where the time interval gets progressively shorter and shorter, one might recognize the event as something that is *bouncing*. Again, the lack of a definable object and context might make the sound event as a whole appear a little abstract, even if the behavior is concrete enough.

Lastly, there is also one last issue considering the discussed boundary when it comes to perceived *structures* in the music, namely the perception of temporal structures where the different sections can primarily be distinguished on the basis of source/cause properties. For example, if one hears three identical melodic phrases, two of them being sung by a different singer than the third, one might perceive the structure as AAB. In such a case, the temporal structure would still be an abstract one, but the contrasts on which the units are based are not based on the sounding shapes, but on source properties.[18] One can therefore argue that this will imply attending to both abstract structure and source properties at once.

Even if cases such as those discussed above suggest that the dividing line between the source/cause-oriented domains and the quality/structure-orientation is far from clear and that they might engage in interaction, I still want to maintain the distinction between the two in my framework. Firstly, I do not see any principle reasons why the existence of a grey zone between these two orientations does not exclude the possibility of maintaining two categories, since in the majority of cases it will not be difficult to decide whether certain properties should be assigned to one or the other. Secondly, I will deal with some of the problematic issues that are touched upon in the forthcoming explication of the max-min framework in the second section of this dissertation. In that way, I hope to be able to account for many of the ambiguous situations mentioned above.

[17] This situation resembles what Smalley calls *third-order surrogacy*, which is “where a gesture is inferred or imagined in the music. The nature of the spectromorphology makes us unsure about the reality of either the source or the cause, or both. We may not be sure about how the sound was made to behave as it does, what the sounding material might be, or perhaps about the energy-motion trajectory involved” (Smalley, 1997: 112)

[18] Emmerson covers such a possibility in his language-grid. This would correspond to category seven in his grid; mimetic discourse and abstract syntax (Emmerson, 1986).

I will now turn to **SQS-domain** to present concepts that can define and qualify the pertinent aspects further. I have chosen to rely largely on Schaeffer's typo-morphological framework, mainly as it is described in Chion's *Guide des objets sonores*, Thoresen's adaptation of this framework, and Smalley's related, but extended spectromorphological framework (Schaeffer, 2002; Chion, 1983; Thoresen, 2007b; Smalley, 1997). However, I will be eclectic in my choice of terminology in accordance with the level of detail that I find pertinent for the context of this thesis.

The notion of a basic distinction in a time domain (horizontal) and a spectral domain (vertical) is both intuitive and well-established in many descriptive frameworks. This distinction does not deny the fact that the experienced spectral content of a sound is interdependent on its articulation in time, but merely delineates two different areas toward which one can direct one's attention. Here, I have chosen to use Lasse Thoresen's concepts in his adapted version of Schaeffer's typo-morphology, namely *sound spectrum* and *energy articulation* (Thoresen, 2007a).[19] I will now discuss a few basic distinctions for these two "dimensions" in turn, and where it is appropriate I will try to exemplify terms and distinctions with references to vocal sounds.

2.4.1 Sound spectrum

2.4.1.1 Spectral type: The pitch–noise continuum

One basic distinction for the sound spectrum deals with what Schaeffer refers to as *mass* and what Thoresen calls *spectral type*. This refers to the difference between *pitched*, *inharmonic* and *noise based* sounds.[20] The sounds of language are among the most immediate examples of the first and last of these types, with voiced vowels being typical example of the first, and unvoiced consonants, for instance [s] and [h], typical examples of the last. Between these categories one can delineate several continua according to the following four criteria; *saturation*, *behavior*, *nodality* and *harmonicity*:

[19] In my view, Schaeffer's concepts of "mass" and "facture" and Dack's translation of these as precisely "mass" and "facture" are less intuitive than Thoresen's (Chion, 1983, translation by John Dack).

[20] I prefer the term "noise", because I feel that it is more intuitive and more precise than Schaeffer's term "complex". The category "inharmonic" corresponds to Thoresen's "dystonic" and to some degree, Schaeffer's term "cannelé" (channelled) (Chion, 1983: 148-149).

1) **Saturation:** (Smalley, 1997: 120; Thoresen, 2007b: 136) One can imagine a continuum from pitched sounds, via sounds that are still pitched but increasingly "fill out" the spectrum until the spectrum is so saturated that it will be heard as noise.[21] One can only envision hitting an increasingly greater number of neighboring keys on an organ to imagine how such a continuum could be realized – the higher number of keys hit, the noisier the sound. Consequently, sounds that consist of several components, either in the form of chords or any combination of different kinds of components, can be understood as an intermediary between pitched and noise based sounds.
2) **Behavior:** (Smalley, 1997: 120) If a number of pitched sounds with a non-regular or quasi random behavior are superimposed, the result tends to sound noisy. One needs just to think of the noise produced by a crowd of people attending an indoor concert during an intermission, especially if heard at a distance. When the crowd behaves more uniformly, as when an audience sighs, cheers or laughs, the sound appears much less noisy. Thus, the uniformity and regularity in behavior will also be a factor for the experience of saturated sounds as noisy.

3) **Nodality:** (Smalley, 1997: 120; Chion, 1983: 146-148) If one starts out from the noise end of the continuum, one can get a gradually increasing sense of pitched quality if certain parts of the spectrum are made more prominent than the rest – what is usually referred to as *colored* noise or *nodal* sounds. The sibilant speech sounds [s] and [ʃ] are both examples of nodal sounds. The narrower these spectral parts are, however, the more spectral "colouring" will turn into a pitched percept. If one filters white noise with gradually decreasing bandwidth, or if one blows a steady air stream while protruding and rounding the lips and then narrows the opening more and more until it becomes a whistling sound, one can get a sense of how a continuum between a nodal and a pitched sound can be realized.

[21] Cf. Smalley's concept of *saturate noise* (Smalley, 1997: 120)

4) **Harmonicity:** (Smalley, 1997: 120-121; Thoresen, 2007b: 136)[22] An additional factor that is of importance for the perception of pitch is the relationship between the different spectral components. If the components are harmonically related to each other, i.e. the frequencies are all integer multiples of the same fundamental, they will contribute to the experience of a single pitch. If, on the other side, they are inharmonically related, i.e. the frequencies are not related to the same fundamental, the sense of a definite pitch will be weaker or not present at all, or the partials can be heard as separate components. Typical sounds with inharmonic spectral components are bells and metallic resonances.

If one sets up these criteria along with their counterpart, one will have several terms to use for describing the continuum between pitched and noise based sounds:[23]

sparse	↔	saturated
harmonic	↔	inharmonic
uniform behavior	↔	irregular behavior
nodal spectrum	↔	flat spectrum

2.4.1.2 Other spectral criteria

Other spectral criteria that I want to include here are *stratification*, *spectral brightness*, *width and density* and *variability*:

- **Stratification:** We have already touched upon another distinction that can be useful in describing sound spectra, namely whether they are heard as consisting of one or several components. I have chosen to adopt Thoresen's term *stratified* for the latter, and refer to the former as *unified* sounds (Thoresen, 2007b: 134). For stratified sounds, like chords and inharmonic sounds, it is possible to focus either on the sound as a whole, or on some of the components of which it is composed.[24]

[22] Thoresen's term for an inharmonic quality is "dystonic".

[23] Both Chion, interpreting Schaeffer, Smalley and Thoresen have constructed diagrams which depict the continuum from pitched sounds (or equivalent terms) to noise including all or several of these criteria (Chion, 1983: 147; Thoresen, 2007b: 136; Smalley, 1997: 120). However, I find that these are either complicating matters or restricting the continuum to only some of the many possibilities that lie in combining these factors.

[24] For Schaeffer, the distinction between focusing on the whole sound or on its part is designated with the concept pair *external* and *internal morphology* (Schaeffer, 2002: 464).

One can also imagine even more complex situations with sounds consisting of groups of spectral components, thus consisting of several levels of components.[25]

- **Spectral brightness:** (Thoresen, 2007b: 136) I want to include Thoresen's notion of *spectral brightness*, which is a sound quality that is highly salient and which is also intuitively labeled with an opposition like *bright – dark* .[26] Again, the sounds of speech can provide some pertinent examples; [s] is brighter than [ʃ] and [i] is brighter than [ɔ].
- **Spectral width and density:** Spectral *width* (Thoresen, 2007b: 136) and *density* (Smalley, 1997: 121)[27] designate whether a sound covers a wide or narrow band of the spectrum, and whether the components in the band are densely or sparsely put together. Thus, the *saturated–sparse* continuum mentioned in the discussion of the pitch–noise continuum is really one of spectral density.
- **Variability:**[28] (Thoresen, 2007b: 133-134) Even if one should intuitively assign the question of variations in sound spectrum to the temporal unfolding, and thereby energy articulation of the sound, the distinction between *stable* and *variable* sounds is included under the dimension of sound spectrum in Thoresen's schaefferian approach. Most vocal sounds, especially in running speech, tend to have variable spectrum as well as pitch. Still, notes sung with stable pitches, stable vowels and without vibrato occur frequently in musical contexts, especially in styles and genres where vibrato is not a part of the stylistic vocabulary.

2.4.2 Energy articulation

Now, let's turn to *energy articulation*, i.e. to criteria that describe how a sound unfolds or is "articulated" in time. Here, it seems adequate to distinguish between different levels of detail; a *course* and a *fine*.

[25] The question of simultaneous components in a sound will be further discussed in a sound source oriented setting in chapter 11 on stream integration.

[26] According to Thoresen, this quality is not included in Schaeffer's typo-morphological framework (Thoresen, 2007b: 136)

[27] It is not necessary to include Smalley's further qualifications of the density criterion in this context.

[28] Even if the term *variability* seems to embrace a huge number of possible sub-categories, I will not present any such terms at this point, but rather depend on ad hoc terminology when a closer description is required.

2.4.2.1 Coarse level

On the coarsest level, which corresponds to the typology level in Schaeffer, there is a basic distinction between the *impulse*, i.e. a very short sound, the *sustained* sound and the *iterated* sound, which consists of a sequence of relatively rapid repetitions. In addition to these three main types, Thoresen has also included two others that extend the basic typology.[29] The first of these is the schaefferian notion of *composite* sounds, that is, "macro" sounds which consist of a smaller or larger group of more elementary sounds (Chion, 1983: 140-141; Thoresen, 2007b:134). Consequently, these sounds can be perceived at two levels, as was the case with *stratified* sounds; one global, focusing on all sounds in their totality; and one local, focusing on distinct single elements. In speech one may find sounds that can be perceived as composite. A single word or a group of words will often be perceived as a unity, based for example on pauses that distinguish it from neighboring sounds or an arch shaped intonation contour which constitutes a kind of wholeness. Within these macro units, however, one can distinguish between separate phonemes as well as syllables, all with their distinct qualities.

The second type that Thoresen includes in his extended typology is *accumulated* sounds, which consist of a larger amount of micro-sounds organized in an irregular manner. Typical examples are raindrops on a tin roof or pebbles pouring out of a bucket into another bucket. It is perhaps difficult to imagine vocal sounds as accumulations, but within electroacoustic music accumulations consisting of short vocal sounds organized in "clouds" of micro-events can be found in several works, such as Trevor Wishart's *Tongues of Fire* (1993, on Wishart, 2000b), Daniel Teruggi's *Fugitives voix* (1997, on Teruggi, 2000) and Alejandro Viñao's *Go* (1981, on Viñao, 1994).

This leaves us with the following criteria, here illustrated in **figure 2.2** with Thoresen's graphical objects, which illustrate the different energy articulations pertinently:

[29] Thoresen has also added *stratified* and *vacillating* objects as main types in his expanded typological diagram, but I will not to include these types at in the discussion at this point. In my view, *stratified* sounds should rather be included as a type in the sound spectrum dimension rather than the energy articulation dimension. The type of vacillating sounds is perhaps more marginal and "eccentric" than the other types, and I have therefore chosen not to comment it in the text.

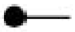
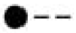
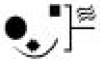
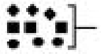
Sustained	Impulse	Iterative	Composite	Accumulated
				

Figure 2.1: Coarse level criteria of energy articulation with graphical symbols from (Thoresen, 2007b: 134)

2.4.2.2 Detail level

Moving down to a more detailed level, there is a number of different criteria in schaefferian morphology that describe the sounds' energy articulation, of which I have chosen to focus on a few. My terms for these criteria are *dynamic profile*, *intonation curve*, *fluctuation* and *grain*.[30]

1) **Dynamic profile:** (Thoresen, 2007b: 137-138; Chion, 1983: 154-158)[31] *Dynamic profile* refers to the shape or profile of the intensity that characterizes a sound, including its *attack*, *sustainment* and *ending*, although it is not always possible to separate sounds into consequent phases. The attack and ending portions of a sound can quite simply be described as having different degrees of *abruptness*, from the *abrupt* to the slowly swelling or *gradual*. For vocal sounds, stop consonants in initial positions would probably have the most abrupt attack, whereas vowel onsets would be more gradual. Stop consonants in ending positions would probably also be felt as more abrupt than vowels in the same position, especially if the release portion of the stop is "swallowed" or non-audible.
    a. **Characteristics of particular phases of the sound:** As Thoresen suggests, it might also be possible to say something about the spectral brightness of the three mentioned phases of the dynamic profile, in particular the attack phase, for which the presence of bright transients will be a salient feature (Thoresen, 2007b:138).[32] This seems indeed pertinent for vocal sounds, since many stop consonants in initial positions can be distinguished from the brightness of the

[30] I have chosen the term *fluctuation* since I find the English terms offered by Thoresen (gait) and Dack in his translation of Schaeffer (allure) unnecessarily opaque. Both Thoresen and Chion in his explication of the schaefferian terminology, however, use the term "characteristic fluctuation" when they try to explain their terminology. Here, Chion's paraphrase of Schaeffer: "*Allure* describes the oscillation, the characteristic fluctuation in the sustainment of certain sound objects, instrumental or vocal vibrato being examples" (Chion, 1983, translated by John Dack: 158, my underlining).

[31] See also Smalley's discussion of spectromorphological archetypes (Smalley, 1997: 113).

[32] In a seminal study of the multidimensional perceptual scaling of musical timbre, the presence of high-frequency energy in the attack portion of the sound was found to be one of three dimensions which most pertinently distinguished 16 music instrument tones (Grey, 1977).

attack; for example the attack in [ta] would seem brighter than the attack of [ga]. For the attack and ending phases it goes that the more gradual or soft the attack or ending is, the less salient this phase will be and the more difficult it will be to distinguish it from the sustained portion.[33]

b. **Characteristics of the sustained portion of the sound:** In so far as the sustained portion can be distinguished, it can be described in terms of several aspects, of which overall *shape/direction* (ascending, descending, arch-shaped, inverse arch etc.) *regularity*, *level* (from *ppp* to *fff*) and *abruptness of changes* are among the most important.[34]

c. **Correlation between spectral and dynamic profiles:** I will also mention, as Smalley does, how the spectrum usually changes along with dynamics so that the louder the sound, the brighter and richer the spectrum (Smalley, 1997: 113). This goes for most vocal sounds as well as instrumental sounds. For electronic sounds, however, there is no necessary relationship between the two, so that it will depend on the dispositions of the composer.

2) **Intonation curve:** (Chion, 1983: 162-164) The mentioned aspects of *shape*/*direction*, *regularity*, *level* and *abruptness of changes* would also apply to what Schaeffer would call *melodic profile*, which I have chosen to call *intonation curve*.[35] Intonation curve appears to be a particularly salient criterion for vocal sounds, since many such sounds, both in speech and in other vocal expressions have a pitch variation that is characteristic.

3) **Fluctuations:** (Chion, 1983: 158-162; Thoresen, 2007b: 138-139) *Fluctuations* can be found in the pitch, dynamic and spectral dimensions of a sound and will principally be related to the sustained portion. Moreover, one can further qualify these

[33] For a note played on a xylophone with a wooden mallet, for instance, one would have a relatively marked and bright attack portion, which then would lead into a gradually decaying phase that would last until silent – and this decaying phase would be impossible to distinguish from a sustained phase. Such *excitation – resonance* sounds are abundant both in everyday interaction between objects and in music.

[34] Here I am using Schaeffer, Thoresen and Smalley rather freely, even if all these notions can be traced to parts of these frameworks.

[35] Schaeffer's criteria for qualifying melodic profile are different from mine, but they cover roughly the same properties (see Chion, 1983: 162-164).

types of fluctuations in terms of *velocity*, (size of the) *deviation*, and *regularity*, that is, whether the fluctuations are periodic, random/irregular or something in between.

4) **Grain:** (Chion, 1983: 152-154; Thoresen, 2007b: 140) The *grain* of a sound, according to Schaeffer's framework, primarily pertains to the sustained portion of the sound and refers to the "overall qualitative perception of a large number of small irregularities of detail affecting the 'surface' of the object" (Schaeffer, cited in Chion, 1983, translation by John Dack: 152). As Chion explains, grain can be regarded as the microstructure of a sound which, by analogy to a material surface, can be either smooth or coarse. There are several types of vocal sounds that can have a granular quality. Good examples are 'creaky' voices, in which the glottal pulses are audible as irregular grains, the rolled [r] (as in Spanish) and the uvular [ʀ] (as in French).

   Whereas Schaeffer proposes several classes and genres of grain, I have chosen to apply Thoresen's simplified version of the framework. Thoresen restricts the range of parameters to two main ones; *coarseness* (from fine to coarse) and *velocity* (from slow to fast), and three optional ones; *sound spectrum* (to the extent that it differs from the spectrum of the 'carrier' sound), *weight* (how prominent the grain is in relation to the 'carrier' sound) and *register placement* (in which register the grain can be found).

In addition to the mentioned criteria for describing qualities and structures of vocal sound, it might also be pertinent to open for the inclusion of terms based in a more traditional Western musical idiom. Some works, among them Lansky's *Six Fantasies*, which will be scrutinized in chapter 12, focus on pitch and durational structures rooted in Western tonal systems, and in such cases the well established tonal, harmonic and rhythmic frameworks will probably provide well suited means of description. I will presuppose, however, that the basic concepts of this system are known to the reader, and I will therefore not elaborate further on it here. More complex concepts will be explained if necessary.

2.4.3 Structural features

Focusing on the structural features of a piece of music implies experiencing the music as constituted by objects, sections or parts on different levels and perceiving different types of relationships and structures between these objects. Such a focus corresponds in many ways to what Delalande describes as *taxonomic* listening behavior, which manifests itself in the listener's tendency "to distinguish sufficiently large morphological units such as sections or chains and to make a mental list of them; to qualify these, but just enough to distinguish them from each other; to notice how these units are arranged in relation to one another; to try and memorize all this data" (Delalande, 1998: 26-27). In the following, I will focus on the distinction of units, i.e. *segmentation* and relational arrangement, i.e. *structure*. Finally, I will briefly note how memorization can impose constraints on the comprehension of structural features.

2.4.3.1 Segmentation

Segmentation of the sonic continuum into smaller units, sections or streams with more or less defined boundaries is something that is essential in making sense of a piece of music (Godøy, 1997). Such segmentation can be vertical as well as horizontal, hence resulting in either temporally delimited "objects" or spectrally delimited "streams" or layers.[36] Features that tend to provide strong cues for temporal segmentation are discontinuities in the sonic flux, repetitions of shorter or longer segments and arch-shaped contours in pitch, tempo and dynamics. The result of a segmentation process is that by demarcating the boundaries, the segments are constituted as such.

Boundaries between units can have different strengths, depending on the different cues and of the degree of convergence between them, i.e. whether several cues indicate segmentation at the same point (Roy, 2003: 207).[37] Thus, while some segments may be markedly separated

[36] There appears to be a lot of mechanisms that are involved in the processes of segmenting a continuous sonic flux into units or sections, and many of these appear to be innate, while some are dependent on acculturation related to e.g. a particular language, genre or style. The Gestalt laws, largely taken to be universal, can account for many of the ways that we tend to form groups of elements and distinctions between these, so can the mechanisms of auditory grouping (see e.g. Roy, 2003: 204-210; Bregman, 1990). If one for example hears an isochronal sequence of subsequent notes, all the the same pitch, and then after some time this pitch changes, it would give us a strong cue for making a segmentation at that point. In this case, the Gestalt *law of similarity* could be seen as asserting its influence. However, I will leave the discussion of Gestalt principles and auditory grouping/streaming until later. The point here is not to show what mechanisms are involved, but rather to focus on the conscious outcome of these processes if they are given any attention.

[37] In Thoresen's terminology this can be seen in the different types of *field demarcation* (Thoresen, 1985:66-68).

from each other, other segments may be distinguished more subtly. The segments themselves can therefore also seem to “stand out” to different degrees. Without going further into empirical research on the topic at this stage, it will for now suffice to postulate that the strength of the demarcation appears to be related to the potential of attracting attention.[38] That a continuous flux of sound is parsed into segments need not imply, however, that the segments are actually put into the focus of attention and cognitively processed any further so as to form the basis of consciousness of some kind of musical structure.

2.4.3.2 Structure

Seeing units in relation to each other constituting some kind of *structure* is the crucial issue in Delalande’s taxonomic listening. One important class of relationships involves the assessment of the degree of similarity between two or more objects, segments or streams. One can experience two or more segments or streams to have different degrees of similarity – from maximal similarity when they are *identical*, to gradually less similarity when they are *variations* or *transformations* of each other, to when they so little similarity that they are *contrasted*.[39] Relationships based on the evaluation of similarities are involved when recognizing that a section of music is structured as, for example, A B A or A B A’ C A’’, with the [’] signs denoting variations. Additionally, there are a potentially huge number of relationships that one can focus on, and to give an account of these would require too much space than can be afforded in this context. I will restrict myself to mentioning a few relevant ones:[40]

- **Position:** Segments are attributed with a certain position in the piece or in a sub-section of the piece (beginning, middle, end).

- **Parts:** Segments are experienced as consisting of *parts*, i.e. other segments.

[38] This will be discussed in detail in chapter 10 on the *feature salience* premise.

[39] See e.g. McAdams & Matzkin, 2003; McAdams et al., 2004 or Deliège, 2001 for accounts of the role of similarity in music perception and how it relates to concepts like variation, transformation and repetition. The former of these also discusses, albeit somewhat superficially, transformation in electroacoustic music.

[40] Most of them are equivalent to those found in Roy’s functional and paradigmatic analysis and Thoresen’s auditive analysis of musical structures (Roy, 2003: 257-391; Thoresen, 1985).

- **Temporal orientation**: Segments are oriented towards what will happen next (forward orientation), what happens in the present moment (presence oriented), or what has just happened (backward oriented) (Thoresen, 1985: 68-69).

- **Foreground – background**: Different layers take on function as foreground, middleground or background relative to each other (Thoresen, 1985: 78-89).

- **Temporal stretching or compression** (Roy, 2003: 294)

- **Fragmentation** (Roy, 2003: 292)

- **Function**: Segments can take on different functions relative to each other, for example *call – response*. Furthermore, one segment can take on a certain function relative to those surrounding it, for example *interruption*, *transition*, *goal-point* (Roy, 1998; Thoresen, 1985: 71-72).

Most of the different structures mentioned above can all occur on different levels, from the smallest level of detail to the largest parts (e.g. movements) of works. Thoresen, for example, operates with four levels of *time fields*, which is his term for musical units that are perceivable for the listener: *Object*, *phrase*, *sentence* and *form*, listed from the lowest to the highest level (Thoresen, 1985). In this dissertation, however, I will keep my main focus on more local relationships, thus making larger scale structural relationships on the *form* level less relevant.

While this opens up for a complex interplay of different structures on different levels, there will most certainly be constraints on how much of this that actually can be attended to at a time. For instance, it seems that if the structures on the different levels are ordered in a *hierarchical* fashion, it seems to greatly facilitate mental processing and memorization.[41] And, since memory is constrained, one will either have to organize the segments effectively (which means in most cases, hierarchically) and/or delimit the number of segments that one tries to

[41] However, as Smalley notes electroacoustic music does not seem to have the hierarchical consistence that can be found in classical music: “Undoubtedly there are structural levels, but they do not need to remain consistent in number throughout a work, and a single level does not need to run permanently through the whole span of a work. For example, one might detect three or four levels in one part of a work and fewer or more in another part; one section of a work might comprise a neat hierarchy of small, unit-groupings, while another section might be a much larger, indivisible, higher-level whole” (Smalley, 1997: 114).

memorize, which can either mean that one restricts one's attention to a certain sub-section or that one focuses on the structures on the higher levels, which are per definition fewer.

This presentation of the sound qualities and structures must be considered as nothing but a rough outline, but still it contains a set of terms that can prove useful in describing the spectral and morphological qualities of vocal sounds, and to some extent the more abstract structures they enter into.

2.5 The domain of technology, composition and mediation (TCM)

Let's recall the encyclopaedic definition of electroacoustic: "Music in which electronic technology, now primarily computer-based, is used to access, generate, explore and configure sound materials, and in which loudspeakers are the prime medium of transmission" (Emmerson & Smalley, 2009). What this definition indicates is that the role of technology appears to be an implicit part of electroacoustic music, at least for the process of *making* and *presenting* it. That the technologies involved in the production of this kind of music along with the actions and choices of the composer can also be audible for the listener should be no surprise, and in some works aspects related to these technologies even seem to draw quite a lot of attention during listening. Consequently, I have included what I call *technology, composition and mediation*, abbreviated **TCM**, in the framework of experiential domains. In the following, I would like to give a short account of how such aspects have been treated in theory, and then try to identify some aspects that can further qualify this domain.

The view of the "stamp" of technology on electronic and electroacoustic music appears to have been ambivalent in theory as in practice. Electronic music was as early as in 1955 criticized by Adorno as being monotonous and "chemically pure", also noting that "every tone is stamped by the interposition of the equipment" (Adorno, 2002: 194). Later, Mike Vaughan could provide the diagnosis that "the perception of specific studio tasks in operation is generally thought to detract from the appreciation of the process as music", and that the "specific nature of the studio" therefore ideally should be transparent (Vaughan, 1994: 116). Also, a writer like Denis Smalley has viewed the technology involved in music making as something that should be ignored, as something that might "block" true musical meaning, and as something that ideally should be transparent (Smalley, 1997: 108-109).

Similar views of technology contending that it is something extrinsic to music and merely a neutral tool for realizing composers' musical ideas, have been met with scepticism and critique elsewhere. For example, Palombini addressed Tod Machover's view of technology as a neutral tool by quoting Heidegger: "We are delivered over to technology in the worst possible way when we regard it as neutral; for this conception of it, to which today we particularly like to do homage, makes us utterly blind to the essence of technology" (Heidegger, 1954 quoted in Palombini, 1998: 35). Other writers have advocated the view that technology is something that should be included in the musical discourse. Simon Emmerson states that "technology itself, may become a reference field, drawn attention to as a crucial signifier; the acousmatic veil torn down and the transparent means of production and dissemination become the subject of the discourse" (Emmerson, 2000: 205). Emmerson's viewpoint seems supported by Chion, who sees the bringing into light the means and technical devices employed for mediation and recording as an important narrative level in concrete music (Chion, 2005: 93). Simultaneously, he criticizes the 'naturalist' or 'immédiatiste' tendency in electroacoustic music, which suppresses the conditions of how the work has been made, while aiming at an imitation of the development of events and natural phenomena in real time.[42] The opposing tendency, which Chion affiliates with his own musical production, is the 'médiatiste' tendency, in which one doesn't seek to suppress or forget that the sound is mediated, but rather brings it out explicitly. The 'médiatiste' tendency is also associated with the early works by *musique concrète* composers such as Pierre Schaeffer, Pierre Henry, François Bayle and Alain Savouret, where montage is a central feature in the compositions.[43]

Another factor that can contribute to directing attention towards technology not mentioned by Chion is that of earlier experience with sound production and processing technologies. One should assume that the more a listener is familiar with the use of such technologies, the more technology-related aspects he or she could potentially focus on when listening to the music. This is supported by the results of Landy and Weale's 'intention and reception' project, where the listeners who were "experienced" or "highly experienced" in terms of music technology were the ones that primarily used technical terms in their descriptions of

[42] The critical view of the naturalist tendency can also be found in an earlier book by Chion, where he links the naturalist "ideology" to regression (Chion, 1991: 69).

[43] Smalley uses the parallel terms "naturalist" and "interventionist" in a later article to describe works where either the "composer's hand" is transparent or apparent to the listener (Smalley, 2007: 54).

electroacoustic pieces (Weale, 2005). Thus, when technical processes used to create and process sounds are familiar, they provide listeners with something that they will attend to and which can be recalled after the listening session – in short, they correspond to a "something to hold on to factor" as they are applied in Landy and Weale's project.[44]

However, it's not only the material components of technology that are audible, but also the actions and choices of the composer in the composition process. Indeed, technology should be understood as also embracing human *uses* of the material components in technology, according to Ihde (Ihde, 1993: 47). And, in electroacoustic music, as in most fields, machines and electronic systems will have to be put to use by somebody in order to do something at all, so that they will ultimately project some degree of human intention. This is also recognized by Chion reflecting on "audible montage", i.e. where a foreseeable sonorous movement is explicitly and suddenly interrupted. Here, he notes how this affirms for the listener both the materiality of the mediation and the presence of a "will to organisation" (Chion, 1991: 69). Thus, the focus on technology can also include the sense of some composing intelligence which has interacted with the technology in making the sound. Such a composing intelligence would correspond to what Emmerson calls *psychological presence*, which he sees as embracing the notions of "Will, Choice and Intention" (Emmerson, 2007: 23). As Emmerson shows through his argumentation, the relationship between composer and technology can be far from as simple as it was perceived in the case of "audible montage": The composer might automate parts of the composition process, use models or data drawn from observation of the human or non-human world, which are then mapped onto musical parameters, use mathematical algorithms or models, or even leave most of the decisions to chance operations (*ibid.*: chp.2) – all such practices may make it very difficult for the listener to map out any stages or actions in the composition process. The listeners' knowledge of such practices might also influence his or her experience of technology and the composing intelligence.

Certain aspects of technology seem to be less under the control of the composer than others. Specifically, one has seen that during the history of electroacoustic music, the different technological paradigms that composers have worked under have imposed constraints in terms of the methods and equipment available, so that for example the overall noise level of the composition has been beyond the control of the composer. Therefore, there might be aspects of a

[44] See Landy, 2006 and Weale, 2006 for an overview of the project and its results.

composition that are clearly products of technology, but where the link to “Will, Choice and Intention” is very weak. In other cases, when the musical parameters in a certain section of music appears to be under the control of a machine, the presence of the composing intelligence might seem more distant or perhaps to reside on another level than in the case of audible montage. Consequently, the human and the technological components might appear to have different degrees of pertinence in different cases, and their interrelationship can vary in complexity, something which can be reflected in the listener’s ability to perceive the role played by these components and the relationships between them.

After having grounded my inclusion of the **TCM-domain** in several theoretical positions, I would now like to qualify and specify this experiential domain further. Firstly, I will look at how the compositional process can be distinguished into several phases that can be audible for a listener. Then, I will introduce two terms that can further qualify several of these processes, namely *explicit* and *implicit transformation*. Subsequently, I will suggest how the *degree of specificity* for technological aspects can be used in the qualification and description. Lastly, I will shortly discuss how *technological connotations*, *imitation*, *behavior* and *effects of technology* can be useful terms for the same purpose.

2.5.1 Compositional process

It might be possible for a listener under some conditions to identify different parts or stages in the compositional process, especially if these conform to what Emmerson suggests is the ‘typical’ approach of *process-compare-combine* found in the ‘classic mix’ (Emmerson, 2007: 27). While the ‘compare’ part of the process would probably be difficult to trace in a finished composition, the ‘process’ and ‘combine’ stages might be possible to distinguish on the basis of listening, even if in many cases it can be highly ambiguous.[45] In addition to ‘process’ and ‘combine’, which I prefer to call ‘organize’, I would like to add ‘record’ and ‘synthesize’ as two additional stages, since they can in a similar manner be recognized as more or less distinct parts

[45] The ‘process’ stage will usually deal with one sound at a time, thus operating on a more local level, while the ‘combination’ stage will operate on a more global level, often organizing the processed sounds temporally and perhaps also spatially. However, this might not always apply, since in many cases processing can be heard as operating on a global level and organization as operating on a local level of a composition.

of the compositional process. Both these parts of the process can generate sound material that may be subsequently processed and/or organized, and Emmerson's 'typical' approach can thereby be extended to something like this: *record/synthesize-process-organize*. While all these parts of the process might be individually recognized in certain cases, more often the distinctions between them will be ambiguous and blurry.[46]

- **Recording:** Here, I will include everything in the sound that points toward the recording process for the listener. This can for instance be noises associated with the microphone, including its response to different types of sound and the way it is suspended or held, e.g. 'pops' from plosive consonants, wind noise, noise from hands holding the microphone, etc. Audible noise from the machinery of the recording device might also point to the recording process.
- **Synthesis:** This includes everything in a sound that points toward the process of creating a sound with the means of electronic/computer technology without using pre-recorded material as a basis. However, in the case of vocal synthesis, as in any case of imitative synthesis, the synthesis process is *modelled* on the human voice, thus giving the synthesis process a basis in a set of abstracted characteristics rather than a concrete pre-recorded sound.[47]
- **Processing:** Everything in a sound that points to the process of manipulating or processing a sound with the means of electronic/computer technology with the basis in a pre-recorded sound. Examples of processing are filtering, playback speed or sample rate manipulation, reverse playback, time-stretching and compressing, etc. Analysis-synthesis techniques tend to blur the distinction between synthesis and processing, but can also be recognized in their own right.
- **Organization:** Everything in a piece of music that points to the process of organizing recorded, synthesized and processed sound into a composition. This can refer to the

[46] As Bossis notes, the distinctions between processing and synthesis are not straightforward to make on the basis of technical criteria, especially for analysis/resynthesis techniques like PSOLA and LPC. See Bossis, 2005: 94.

[47] In the case of the voice, the categories of synthesis and processing correspond to Bruno Bossis' two categories "a voice as model" (une voix comme modèle") and "models of the voice" ("les modèles de la voix"), respectively He defines these categories in the following way: "Le première catégorie regroupe l'ensemble des modèles qui reposent sur des processus abstraits ne nécessitant pas une source sonore réelle tandis que la seconde met en jeu la voix naturelle comme un modèle concret" (Bossis 2005: 101).

temporal organization of the sounds, adjustments of the intensity level of sounds, repeating sounds with or without looping etc. Simple techniques like cutting, panning and fading might be assigned to the organization process, especially if they affect several distinct sounds simultaneously. If they affect only one single sound, they might just as well be assigned to the synthesis or processing processes.

2.5.2 Implicit/explicit transformations

Sounds based on pre-recorded material in electroacoustic music may project being transformed to different degrees, where the transformation is relative to the listener's mental template of the unmediated sound event, which can be either previously heard or not.[48] For some listeners a transformation may be experienced as implicit, often showing an inclination to erase its transformational character, while others may be experienced as explicit. While sounds tend to be experienced as either implicitly or explicitly transformed, one can also have ambiguous cases, where it is difficult to assign it to one or the other.

- **Implicit transformations** embrace the sounding traces of the necessary but often unnoticed prerequisites of electroacoustic music, such as the microphones, recording media, loudspeakers, studio facilities, noise reduction systems, compression and coding formats, etc.[49] All these aspects, which mainly make up the *mediation* component of this experiential domain, have their unique characteristics in terms of frequency response, phase modification, noise type and noise level, etc. Usually these aspects appear passive and unchanging and tend not to attract attention unless different configurations are

[48] Smalley refers to the base upon which a transformation is made as the *base identity*, and distinguishes between an *intrinsic* and an *extrinsic* base identity, where the former is "a specific sound with a specific site" and the latter is "a generic category of sound with no specific site" (Smalley, 1993: 285).

[49] Schaeffer also uses the term 'transformation' for the effects implied by the recording process when one can compare it to attending to the "original" event. For example, he notes that when listening to a recording, one can often hear a reverberation which was unnoticed at the original event. Moreover, the recording may also contain sounds that one didn't notice at the original event: background noises, the coughing of a neighbor, etc. He also mentions that the recording or playback devices in themselves can mark the recording, thus giving it a certain "signature" (Schaeffer, 2002: 77-79). Similarly, Jean ChristopheThomas, argues that recording sound with a microphone is already a 'first manipulation': "La première manipulation (qui en est à peine une, et pourtant radicale) sera celle qui consiste, et dès la "prise" de son, à grossir le détail du paysage extrait : exploration de l'inconnu qui est au coeur microscopique des phénomènes. Le microphone est immédiatement un microscope, apte à susciter les secrets de sons parfois très familiers" (Thomas, 2006).

contrasted with each other. Although these aspects *can* be the result of the composer's choices, they are often *not* experienced as closely related to "Will, Choice and Intention". More often implicit transformation may refer to:

- technological paradigms such as analogue/digital, magnetic tape/gramophone, etc.
- different types of equipment with different functionality, like telephones, dictaphones and radios
- historical or cultural contexts associated with certain technologies and their use within a specific period, institution, studio, etc.

- **Explicit transformations** are experienced as traces of a preceding process of interaction between a composer, sound processing technology and pre-recorded sound.[50] The properties of the sounds appear *changed* due to precisely this process, and the changes will often also display temporal variations, thus projecting a more active process than implicit transformations.[51] Therefore, explicit transformations tend to attract more attention than the implicit ones. The degree to which the technological components of the process are recognized will depend on the listener's experience with and knowledge of sound processing technology, identification and recognition of the properties of the pre-recorded sound, the complexity of the processes involved and the degree to which these have been widely used earlier. Listeners may often experience different degrees of transformation, e.g. from modest/mild/gentle to heavy/strong.[52]

2.5.3 Degrees of specificity

Different aspects related to technology, composition and mediation might be recognized with different degrees of specificity. Recognizing the 'technological paradigms' of the analogue versus the digital might be situated at the lowest level of specification, while recognizing the

[50] See Smalley, 1993 for an account of transformations in electroacoustic music. However, Smalley does not emphasize the role of technology for transformation.

[51] Thus, since explicit transformation requires an assessment of similarity between the present sound and another (absent) sound, either presented elsewhere in the piece or heard or existing as a mental template, the distinctions between focus on structure (**SQS**-domain) and on technology (**TCM**-domain) might be blurry in this case (cf. section 2.4.3.2 on *structure*).

[52] Cf. Smalley's notion of transformational distance (Smalley, 1993: 287).

mentioned processes of recording, synthesis, processing and organization in a musical work takes the specification one step further. All these processes might then be further specified according to the following three criteria:

- **Techniques and algorithms** can be recognized with different degrees of specificity. Less specific categories can be ‘close miking’ (recording), granular techniques (synthesis, processing), (processing/synthesis), modulation techniques (processing/synthesis), filtering (processing), etc., while more specific categories can be multiple-carrier frequency modulation synthesis, FOF synthesis, etc.
- **Hardware/software** may be recognized with different degrees of specificity, usually depending on highly experienced ears and/or hardware or software which provides the sounds with a marked “stamp”.
- **Parameter settings** for one of the two points above could potentially be recognized by somebody who has had extensive experience in using the device/equipment/technique in question and implies a high degree of specificity.[53]

2.5.4 Other aspects

- **Technological connotations:** Any aspects of technology might refer to a wide range of historical and cultural phenomena that is somehow associated with the technology in question. This might be historical period, country/region, institution, studio, compositional “school”, other cultural, commercial or artistic practices or concrete expressions, etc. (Cisinsky, 2007). Certain classes of sound might also connote functional characteristics of the technology in question, e.g. those that are associated with malfunction, error and failure, like clicks, contact flickering noise, distortion/clipping, feedback, metallic “ringing” (see e.g. Cascone, 2000).

[53] For example, in a course on digital signal processing at the Norwegian University of Science and Technology (NTNU) taught by Øyvind Brandtsegg, the students were asked in their final exam to specify what algorithms were used and roughly how the parameters were set for a set of different sounds. The work with the same techniques throughout the preceding semester had prepared them for this task.

- **Imitation:** One might experience that technology imitates acoustic sounds or phenomena, other instruments or sound producing technology. For instance, a *Mellotron* might imitate a choir singing a sustained chord, the *Chant* program may imitate an operatic singing voice, and a reverberation or echo unit or algorithm might imitate acoustical reverberation and echo.[54] In some cases, such imitations might even be imitated, as when a digital piece of software imitates the *Mellotron* playing sustained chords or an old plate reverberation unit. Thus, several layers of technology with different degrees of virtuality might be experienced. I will discuss related issues in section 2.8.3.

- **Behaviour:** Certain aspects of a segment of music might be assigned to a technological agent/machine due to characteristic behaviour. For example, a sound with static pitch and timbre, or with random pitch variations with high speed and high precision in timing and pitch, might be assigned to a machine or alternatively as having "machine like" behaviour.[55] In cases where different aspects of the same sound might be attributed to different sources, a sound may be experienced as hybrid or ambiguous.

- **Effects of technology:** In some cases the *effects* of technology just as much as the technology involved are recognized, i.e. certain defined characteristics of a sound are recognized as being affected by technology, even though one cannot specify what means are used to achieve these characteristics. For example, a sound may be experienced as fragmented, transposed, time-stretched or compressed, even if one cannot recognize what exact means are responsible for it.[56] These effects thus have some degree of abstraction,

[54] Artificial echo units rarely function in a true imitative manner. Natural echoes normally require relatively long distances (more than about 16 meters) to produce an echo, which gives considerable loss of high-frequency content in the echoed sound, something which is rarely modeled in echo units.

[55] As Schafer notes, "the flat continuous line in sound is an artificial construction. Like the flat line in space, it is rarely found in nature" (Schafer, 1994: 78). Truax points to *precision* as a factor that will create associations with machinery, the unnatural and the inhuman (Truax, 2001: 75). I will discuss further what factors that might contribute to machine-like behavior in chapter 7 on naturalness.

[56] Most of the types of transformations of morphology and facture described by Roy can be regarded as effects of technology without a precise reference to the technology involved (Roy, 2003: 292-294).

thus making them belong in a grey area between the **TCM-domain** and the **SQS-domain**.

Before I move on to the discussion of the experiential domain of space and environment I will just clarify that all notions of technology relates to technology *as experienced*. In most cases, what we perceive when we are listening will not necessarily give us precise and reliable information about how the sounds have been made. Still, there is a link between perception and generation exactly because the generative processes *tend to* produce results that are related.[57] Although a listener cannot with certainty deduce from listening what lies behind a particular work in terms of production, she or he can make qualified assumptions drawing on acquired knowledge of the links between different aspects of sound production and the sounds that are made from these – exactly based on the tendency that certain ways of production *tend* to mark sound in a particular way. Thus, what will be of great importance for the listener's experience of the different aspects related to sound production in electroacoustic works is the competence that the listener has acquired of these issues on a broader basis. Therefore, it is also a possibility that the sound that is the product of an electronic manipulation may be of a kind that it conveys no sense of transformation or "deformation" at all. In this way, it is not the process of electronic manipulation in itself that decides whether a sound *is heard as* a transformation of another sound – it depends both on the characteristics of the recorded sound that is being manipulated, and on the kind of manipulation that is applied to that sound. And, lastly, a sound may be heard as a transformation, even if no manipulation whatsoever has been applied. Indeed, this was the case in my own listening of the beginning of Jean-Claude Risset's *Sud* (1985, on Risset, 1987). Here, I took the almost "phasing" quality of the sounds of sea waves against rocks as proof of some manipulation, but could later on read Risset's own account of the production processes of the piece and find out that this sound was presented as recorded without any processing.[58] This shows that it is can be difficult to separate the transformations that the process of recording

[57] Cf. what Wishart has written on the composition of *Tongues of Fire* (1995, on Wishart, 2000b): "Sounds originating from very different generative processes may have significant audible similarities, while sounds from the input and output of the same process may have no notable audible differences at all. In practice, however, the generative processes I use […] *tend to produce sounds that are perceptually related* to each other in some way" (Wishart, 2000a: 23, my italics).

[58] Found in the interactive analysis of *Sud* at URL: http://www.ac-rennes.fr/pedagogie/musique/dswmedia/shock.html, retrieved 14/06/2009.

imposes on the experience and the transformations that are caused by deliberate electronic processing of recorded sound.

2.6 The domain of space and environment (SE)

Don Ihde, who has been investigating voice from a phenomenological point of view, has written that “sounds are ‘first’ experienced as sounds *of* things” (Ihde, 1976: 59). This is taken to indicate that we tend to intuitively and primarily experience sound with a reference to “things” – where “things” are taken in the extended sense; including living beings, material objects and substances as well as events and activities. Such a view largely corresponds to ecological accounts of listening, which tend to emphasize how humans are particularly sensitive to sound producing events or sources in everyday listening (see e.g. Gaver, 1993; Coward & Stevens, 2004). The view that the experience of “things” is valid and valuable also within the context of electroacoustic music seems now quite uncontroversial in theoretical writings. Aspects related to sound sources and environments in electroacoustic music has been conceptualized and theorized under labels such as “sound symbols and landscapes” (Wishart, 1986) and “indicative fields” (Smalley, 1992). Other theoreticians have conceived terms to designate a particular listening mode or behavior occurring when listeners attend to sound producing sources or sound causes as for instance “referential listening” (Norman, 1996), “causal listening” (Chion, 1994)[59], “figurative listening behavior” (Delalande, 1998) or simply as “event perception”, a term adopted from ecological perception by Luke Windsor (Windsor, 1995).[60]

We experience things, as well as beings and events, as occupying and being situated in some kind of spatial setting, and as localized within some kind of environment (Gaver, 1993: 7-9). When it comes to sound, there are many cues that give us information about the spatial setting and what kind of environment we are dealing with: reverberation properties, how far any perceived sound sources appear to be situated from us, overall level of ambient noise, and so forth. These aspects usually take on a role merely as background or contextual framing of whatever resides in the foreground of our attention. Nevertheless, I will group the aspects related

[59] Chion’s discussion of different listening types is taken from a book about “sound on screen”, but is sufficiently general to also apply to this discussion.

[60] This is recognized by many of these theoreticians as a rather primitive type of listening, since it has been of great importance for human survival throughout evolution – avoiding dangerous predators, localizing prey, or identifying possible candidates for mating. See e.g. Wishart, 1996, p.129, Schaeffer, 2002, p.120.

to space and environment into a separate experiential domain, because they can in some circumstances constitute the focus of our attention.

In most cases, however, the relationship between sound sources ("things") and space/ environment is one of interdependence: "Things" can make spaces "articulated" through reverberation, and the meanings and identities of both sound sources and environment will, at least partly, define each other and contribute to each other's recognition. In the following, I will substantiate this interdependence further, before I draw up some more specific aspects that can be assigned to this experiential domain.

2.6.1 Interdependencies

Studies on space in electroacoustic music over the last decade or so have pointed out ways in which sound sources or objects and the space/environment they inhabit are interdependent, constitute meaningful interrelationships with each other and affect each other's recognition (see. e.g. Smalley, 2007; Barrett, 2003; Ekeberg, 2002). I have condensed some of these theoretical issues in the following three points:

1. *Sound sources (or sound-objects) can play a part in constituting space.* "Sounds in general, and source-bonded sounds in particular, […] carry space with them – they are space-bearers" (Smalley, 2007: 38). This can happen in several ways:
    - *Through allusion:* When one hears a voice and imagines the vocal source as a person, this person has to be situated spatially, and preferably in some space that persons are known to inhabit – space is evoked through *allusion*, in Barrett's terms (Barrett, 2003).
    - *Through distance:* If a voice appears to be coming from a long distance, this distance, if no other clues to space are given, will define that space.
2. *Through motion:* By traveling through space, a sound source can also define a set of continuous spatial locations. This also highlights the fact that one often needs a certain amount of temporal exposure to sound to mentally build a spatial configuration. Such a

vectorial space, which Smalley calls it, can ultimately define both the trajectory of the sound source as well as the limits of the space (Smalley, 2007). For instance, if one hears a voice that starts out relatively close and then gradually seems to disappear into the distance until it is no longer present, this space will be defined by the whole range of this motion with the maximal distance defining the acoustic *horizon*.

3. *Perceived space results from interaction between sound-sources and space:* In many cases the configuration of a certain space, for instance its size and the nature and structure of any enclosing substances and objects, can only be heard through its effect on the sounds from objects or beings, mainly as reflection (echoes, reverberation) or absorption (spectral changes). The perceived sound will therefore be marked by the *interaction* between (active) sound source and the (passive) space surrounding it.
4. *Space/context can affect sound source recognition:* Whereas some sound sources (like the human voice in most cases) can usually be recognized regardless of context or apparent spatial location, other sound sources are dependent on a particular context or spatial disposition to be recognized.[61] Trevor Wishart mentions an example from the composition of *Red Bird* where a recording of a fly buzzing made so that the fly was stationary positioned close to the microphone in itself did not render a convincing source image of a fly. Only after attenuating the sound, applying a low-pass filter and panning so as to artificially create spatial motion did the sound appear convincingly as a fly (Wishart, 1996: 151).

2.6.2 Qualification of the SE-domain

From all the properties that might be used for qualification and description of spatial and environmental issues, I have chosen only a few that I have found relevant for this project. Firstly, I will discuss different *space frames*, as explicated in Emmerson's framework (Emmerson, 2007, chp.4), but expanded with Hall's categories of *proxemics* (Hall, 1979; Ekeberg, 2002). Secondly, I will look at properties related to reflection and absorption. Lastly, I will present some

[61] Wishart's terms for these two situations are *intrinsic* and *contextual recognition*, respectively (Wishart, 1996: 150).

properties related to context and environment, which I have tried to "boil" down to three different aspects of *setting*, namely indoor/outdoor, social and geographical setting.

2.6.2.1 Space frames

Emmerson's terminology of *space frames* provides a way of specifying different levels of distance when applied to acousmatic music.[62] Emmerson operates with four space frames, namely *landscape*, *arena*, *stage* and *event*, where the former two are defined as *field* and the latter two as *local*. While *landscape* is defined by the acoustic horizon, i.e. the outer limits of listening, *arena* is not too rigorously defined by Emmerson, but the term itself suggests a socially confined frame within which (amplified and largely one-way) communication can take place.[63] *Stage*, in turn, indicates the immediate space frame around the sounding *event*: "'Stage' suggests an area of clear perception from which we receive detailed, information-rich signals and to which we devote maximum attention" (Emmerson, 2007: 99). The space frame of the *event*, in turn, would be difficult to separate from any object and being in the space. And, as Emmerson notes, it is difficult also to draw distinct boundaries between the frames, especially in acousmatic music. Rather, "the frames essentially float in an unpredictable listening landscape in which any one frame can seamlessly move into the foreground or retreat into the distance of our perception" (*ibid.*: 100).

Social space frames: I want to extend Emmerson's space frames with a set of *social* space frames because such frames prove particularly relevant when dealing with an instrument of human communication such as the human voice. In particular, *distance* has important social implications, especially for speech and vocal communication, and these seem even more fine-grained than the frames in Emmerson allow to distinguish. The distance at which a speaker is perceived to be localized has social implications that are reflected in both *how* the voice is used and *what* the speaker talks about. In reviewing Hall's work on *proxemics*, i.e. "the study of man's perception and use of space", Ekeberg points to the links between a) distance, b) mode of

[62] Emmerson's terms *local* and *field* were originally developed for live-electronic music, but in later articles and his book *Living Electronic Music* he extends and refines these notions and opens up for listener-defined space frames and application on acousmatic works (Emmerson, 1998; Emmerson, 2007: chp.4).

[63] This corresponds roughly to Blesser and Salter's concept of the *acoustic arena*, which is described as "a region where listeners are part of a community that shares an ability to hear a sonic event. (Blesser & Salter, 2007: 22).

voice, and c) the subject matter of speech, a review that can be summarized in the following table (Ekeberg, 2002: 113):

Distance label	Approx. distance	Mode of voice	Subject matter
Intimate	0 – 0.5m	Very low	Secret or confidential
Personal	0.5 – 1.25m	Moderate	Personal
Social	1.25 – 3.5m	Speech	Non-personal, may be public information
Public	3.5m -	Raised / amplified	Addressed to group

Table 2.1: Summary of Ekeberg's review of Hall's categories of proxemics.

The different distances in the table can then be viewed as a set of space frames which can be seen in relation to Emmerson's. First of all, there seems to be some overlap. The *public* category, where one has to use raised or amplified voice in addressing a group, thereby seems to correspond to the space frame of the *arena*. Moreover, the *social* distance appears to cover the range of the *stage*, since performers will necessarily engage in an interplay that can be seen as an analogue to social intercourse. The *intimate* and the *personal* distances, however, are both shorter than what the space frame of the *stage* implies, and that I therefore see them as contained within it. It must be added, however, that while adopting his terms, I will still disregard Hall's relatively strict definition in terms of distance, and rather consider the boundaries between these frames as relative, fluid and variable, as Emmerson did. Consequently, I end up with the following configuration of (combined) frames:

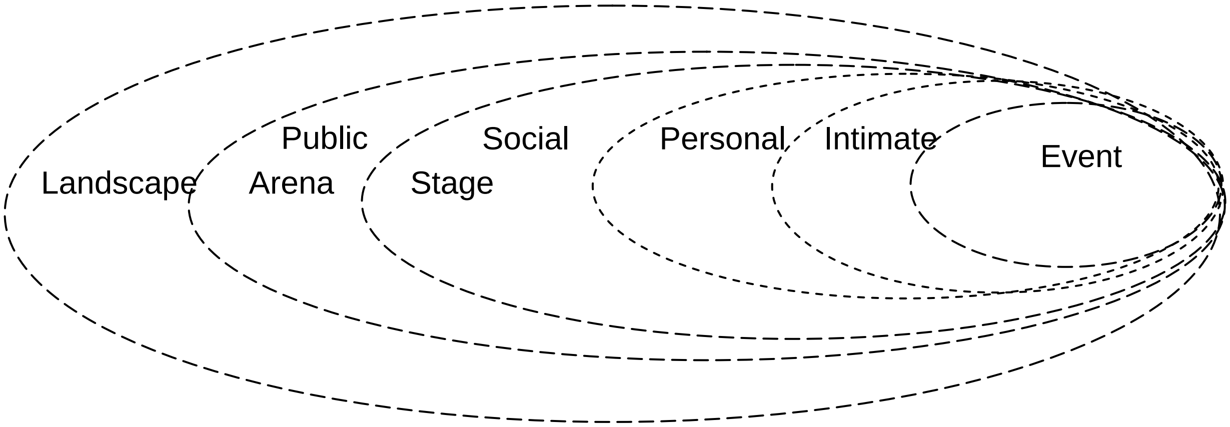


Figure 2.2: Model of space frames combined from Emmerson (2007) and Hall (the latter as presented in Ekeberg, 2002).

2.6.2.2 Spatial properties – reflection and absorption

Spaces might be experienced as more or less enclosed and as containing objects and/or substances at different distances and directions, dominantly due to *reverberation* and *absorption* characteristics and their relationship to direct sound. Pertinent characteristic features are:

- **Distance:** Experienced distance can be expressed in terms of the space frames just discussed. Important cues for experienced distance are: 1) proportion of direct sound versus reverberated sound (more direct sound indicates closer), 2) loudness (softer sounds tend to be farther away) and 3) high frequency content (duller sounds tend to be farther away) (Roads et al., 1996; 458, 462-463; Pierce, 1999). Since loudness and high frequency content are properties that differ also due to other factors, for instance the applied energy in making the sound, the assessment of distance will depend on the listener's familiarity with the characteristics or invariants of a particular class of sound sources, particularly if there is little reverberation present. Manipulation of the three mentioned cues is frequently used to simulate distance in electroacoustic music.
- **Direction:** Direction cues include both azimuth (vertical) and zenith (horizontal), but the latter will be less relevant in this context, and I will therefore not consider it here. The cues for azimuth are highly dependent on the number of speakers and their configuration, but since I will be exclusively dealing with stereo versions of electroacoustic works, I will not consider multi-channel spatialization. In a stereo system, the azimuth of a sound can be described as lying on the continuum between hard left and hard right with centre position in the middle. Manipulation of direction in electroacoustic works in stereo is usually done with *panning*, i.e. adjusting the sound level balance between the two speakers (Roads et al., 1996: 458-461).
- **Size/volume**: Long reverberation time with little high frequency content is usually experienced as indicating a large space, whereas short reverberation time with a great deal of high frequency content will be associated with a small space (see e.g. Blesser & Salter, 2007: 21; Emmerson, 2007: 21).

- **Materials/substances**: To some degree, reverberation characteristics can reveal something about the materials of the reflecting surfaces. Concrete or stone surfaces, for instance, tend to have another and perhaps more massive sonic imprint than wood.
- **Ambiguous configurations:** Lack of reverberation altogether has the potential of creating an ambiguous situation since it can indicate both an enclosed space where the enclosing materials are highly absorbent (as in an anechoic chamber or a studio where reverberation is minimized or where 'close miking' prevents recording of any reflections) as well as a "free field", i.e. with open space in all directions.
- **Present objects/substances:** Absorbing/reflecting objects and substances not a part of an enclosure will affect the frequency characteristics, reverberation time and amount. Whether this can be experienced in a recording is perhaps doubtful. Hypothetically, a listener with high awareness of the spatial characteristics of sound combined with recording and playback facilities that can realistically render such features might still possibly be able to attribute features of the sound to absorbing or reflecting objects in an acousmatic listening situation.

2.6.2.3 Environmental/contextual attributes

Attributes belonging to the space frames of the *arena* or the *landscape* can be defined in terms of different *settings*, which provides in some respect a context or environment within which the sounds of objects and beings are situated, and therefore provide important cues for how the sounds of the objects and beings are to be interpreted. There are potentially a huge number of attributes that could be included here, but I have chosen to include three that I find particularly important:

- **Indoor/outdoor setting:** In many cases, listeners can assign an acousmatic sound or a soundscape to an indoor or an outdoor environment. Cues can be found in the a) types of sounds that might be present, b) the presence of ambient noise (the noise level is usually higher outdoors than indoors), c) reverberation times (which are usually much shorter outdoor, at least in natural environments) and d) the degree to which sound transmission and reflection characteristics are static, which is more often the case for indoor than for outdoor sounds (Blesser & Salter, 2007: 341). Indoor/outdoor setting also has consequences for the social setting.

- **Social setting:** Properties of the experienced environment might be assigned to a more or less specified social setting. Such a social setting can be associated with a certain type of activity (e.g. religious worship, shopping, sports, education, play, social interaction), a certain indoor or outdoor setting, certain segments of the social structure (e.g. rich/poor, upper class/middle class, academics/"blue collar" workers, etc.) and perhaps a certain range of emotions. Such settings can also be defined in terms of the number and configuration of human beings or other living sound-making creatures that appear to be present; are there a few individuals or a big crowd, are they located at a confined location or more freely scattered around; do they appear to be stationary or move around, etc. Moreover, the social setting can be defined in terms of the degree to which it is associated with social interaction and which formal level this interaction takes place within (cf. Hall's categories above; *intimate, personal, social, public*).

- **Geographical setting:** Depending on background and experience, a listener might recognize a number of different features related to geographical setting from acousmatic sound with different degrees of specificity. Most people would probably easily distinguish urban, rural or natural settings, and many would probably also be able to specify further whether a sound is set in a metropolis rather than a small town, in the forest rather than at the coastline, in the jungle rather than at sea, etc. For listeners with the appropriate experience, certain cues can perhaps also give some indication of climate zone, for example sounds associated with snow (e.g. sounds of skiing, footsteps in snow, making snowballs) or tropical forests, whereas other cues can indicate localization in a particular area of the world (e.g. by identifying particular birds or animals that are characteristic of a certain area).

The discussed spatial and environmental properties are rooted in our ecologically grounded presuppositions of spaces and environments that are crucial in how we experience our surroundings in everyday life, but are also products of learning and experience within natural and cultural environments. These properties and their relationships within the external world also provide central frames of reference when we listen to acousmatic electroacoustic works (Windsor, 1995; Wishart, 1986; Clarke, 2005a). However, spaces and environments in electroacoustic music are necessarily deprived of all the multisensory cues of everyday life, and they therefore tend to be much more ambiguous and fleeting. Moreover, the abundance of synthesized and processed sounds and the possibility for superimposing and organizing them temporally in different ways often create spaces and environments that are far from what one can experience in everyday life. Often, one can experience imaginary spaces, gradually changing spaces, multiple spaces that can be nested into each other and environments that appear paradoxical because they contain sounds that point to objects or beings that can't co-exist in the same space in real life.[64] For example, the sound of a car engine in an indoor domestic setting would probably be experienced as "incompatible" or "paradoxical". Issues of ambiguity and imaginary, incompatible and paradoxical relationships must include all spaces, objects and beings in a sounding "scene". Therefore, it is necessary not only to be able to give a description of the individual aspects of spaces, environments and the sources and causes that are "inhabiting" them, but also the *relationships* between them. The result will in many cases be outside the frames of our daily experiences, thus transporting us into the realm of the imaginary or virtual. In the final section of this chapter, I will go further into a discussion of the virtual/imaginary as well as the relationships to the external world. Before I do that, however, I want to present some views of the relationships between the domains discussed up to now.

2.7 The relationship between the experiential domains

After presenting the experiential domains I will now introduce a general schematic representation of all the different domains and how they are related to each other, including those that were outside the focus of this thesis. **Figure 2.3** gives an overview of the experiential domains and how I see them as related to one another.

[64] This is discussed by both Emmerson and Wishart (Emmerson, 2007: 99-101; Wishart, 1996: 146-148).

In this figure, the basic ground or "canvas" coloured in pink in the figure is constituted by the body and mind-domain, reflecting that our body and mind constitute the basis upon which all our experiences are made and that our body and mind can also be subjected to intentional focus, e.g. in directing our attention towards for example our breathing cycle, heartbeats or any other bodily or mental reactions that come as a response to external stimuli.

Superimposed upon this canvas is a "plateau" containing everything in the experience that is external to the body and mind – in other words the canvas and the plateau thereby represent *inwards* and *outwards* focus (towards the outer world) respectively, something which is indicated with the double-headed arrow at the bottom-left of the figure.

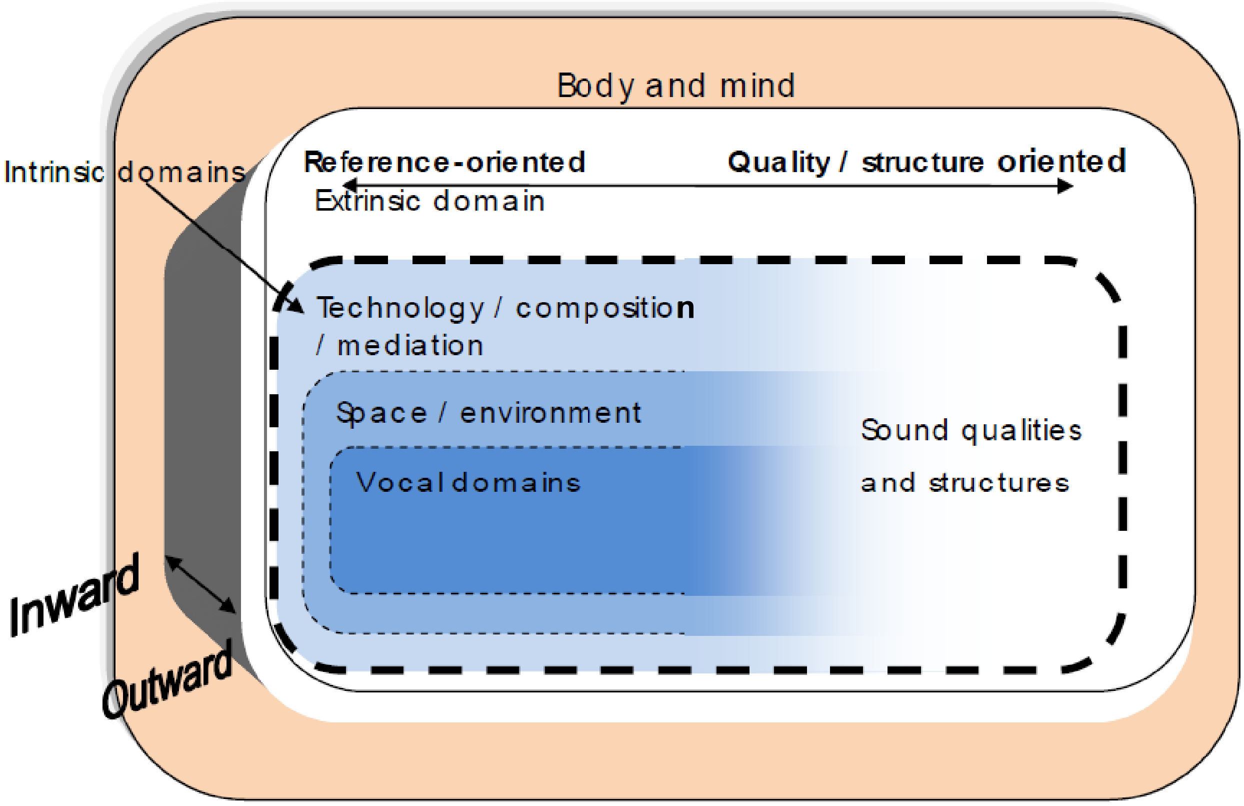


Figure 2.3: Schematic representation of experiential domains for acousmatic electroacoustic music listening. For a full explanation, see the text.

The plateau is, on its part, divided into two areas: the left half represents experiential domains that show an orientation towards references "outside" the sound, like sources, causes or symbolic signification; the right half represents a domain which represents an orientation towards more abstract qualities and structures, comprising the **SQS-domain**, but which may also include a focus on abstract properties belonging to the extrinsic domain, such as the spectral characteristics of the reverberation of a certain listening space. As one can see from the horizontal double arrow and the graded colouring from left to right, one can regard this distinction as a continuum, something I will discuss further below. The plateau is further divided into different areas in different shadings of blue, representing the other experiential domains; the technology/ composition/mediation (**TCM)** domain, the space/environment (**SE**) domain and what will be the main focus in the forthcoming chapter, namely the vocal domains, which embraces everything that is experienced as emanating from the vocal persona. The thick dashed line in the figure represents the discussed distinction between aspects of the experience that are considered to be *intrinsic* to the musical work and those that are *extrinsic* to it. The dashed style of the line between those areas nevertheless marks that in many cases it is not straightforward to assign a property or an event to either the extrinsic or the intrinsic domains.

The area of the intrinsic domains, which are those that will be considered in this thesis, embraces three areas delimited by rounded rectangles, where one is enclosed within another. These areas are clearly distinguishable on the left half of the area, where they are separated by dashed lines and given different shadings of blue. In the right half, however, the areas fade off into each other more and more the further right in the figure one moves. This is supposed to represent that even if the experiential domain on the right does not at the outset deal with referential properties like sources and causes, properties and structures of the sound can be abstracted to different degrees, where some are closer to the source/cause conceptions than others.

The placement of the three reference-oriented intrinsic domains within each other is done because it somehow corresponds to how it is experienced. To begin with, since the vocal persona will necessarily have to be located in some more or less defined space or environment, the placement of the vocal domains within the **SE-domain** should not be too difficult to attest to. As for the placement of the **SE-domain** (and thereby also the vocal domains) within the **TCM-domain**, it is motivated by conventions and practices within electroacoustic music implying that the composition is a product of actions of the composer in interaction with different kinds of technology, and thereby that any experienced sources, causes or other kinds of referential meaning would have to be "contained" within this frame of reference. Even if a composition features something that appears to be merely the “unmediated” or “unmanipulated” sound of a certain environment, this environment will always be a product of acts and choices by a composer and his/her use of the technologies of recording. A recording is always profoundly and implicitly marked by the fact that it is captured through a microphone, fixed onto a medium and played back from a speaker, usually in a different space. Whether the sounds are a product of mixing, editing or manipulation may vary, even if this is generally the case for electroacoustic music. The listener will therefore always in a sense have to listen *through* the technologically mediated and composed character of the sound, in order to experience spaces, environments, objects or beings within it.

As I have already implied, the boundaries between domains cannot be clearly defined in all cases. As for the (thin) dashed line between the **SE-** and the vocal domains, it indicates that it is not always straightforward to distinguish between what is a part of the environment and what pertains to objects and beings, such as a vocal persona, in it – indeed the environment can consist of beings as well as objects; the distinction will be dependent on both the salience with which a certain entity is perceived by the listener, whether the entities are experienced directly or indirectly (e.g. by their absorption or reverberant properties) and the focus of the listener. As for the boundaries between the **TCM-domain** and the two others it contains, it is not totally clear cut. For example, artificial reverberation might give the listener a sense of an enclosed space of a certain size and with certain characteristics, but it might also be attributed to a particular kind of reverberation unit. The same properties might therefore be heard as belonging to the **TCM** or the **SE**-domain depending on the focus or point-of-view. There can also be cases where the boundaries between the **TCM-** and the vocal domains are blurred. This especially goes for imitative synthesized sounds, which can at the same time evoke the use of a more or less specific technology involved in production, the virtual or imitated vocal persona, and the space the vocal sound is allowed to resonate in. I found it hard to give this lack of defined boundaries a separate graphical representation, however.

With the graphical layout of **figure 2.3** I have tried to imply that the reference-oriented and the sound quality/structure-oriented areas of the figure are at once separate (the dashed lines around the source/cause oriented domains only covers the left half of the figure) and continuous (with the continuous double-headed arrow and grading of the colours). This layout is motivated by what I discussed in the introduction of section 2.4, namely that on one side the dividing line between these orientations is difficult to define clearly, but on the other side, that in the majority of cases, one can still straightforwardly attribute one's experiential focus as being either one or the other.

While **figure 2.3** thus gives a fairly good overview of the experiential domains and the way they relate to each other, I think it is important to emphasize that it represents a simplification, and that in many cases the situation is far from as simple as what can be read from the figure. The possibilities of having several spaces or layers of technology within each other, for instance, will imply more complex nesting of domains within each other. Such cases will be discussed in more detail in the following section on virtuality and the relationship to the real.

2.8 Virtuality and the relationship to the "real"

The imaginary or virtual status of experienced sound sources and environments in acousmatic electroacoustic music is perhaps most immediately perceived when sources and environments take on properties that go beyond everyday experience. However, as many theoreticians of electroacoustic music have commented and discussed, all experienced sound sources in acousmatic recorded sound have a virtual or imaginary status, since they are all products of imagination – the sources are not really there (Chion, 1991: 25; Wishart, 1996: 136; ten Hoopen, 1992a: 122). Rather, the only *actual* sound sources the listener experiences in acousmatic electroacoustic pieces are the loudspeakers in combination with the reverberation of the room – both belonging to what I have called the *extrinsic* domain in this framework. This kind of virtuality seems paradoxical since one in some cases can experience sound as having an element of "actuality", "reality" or "documentary". As John Young comments: "The transference of real-world sound material into electroacoustic music signifies a shift in context from actual experience to the 'virtual' while at the same time the idea of sound documentary is invoked" (Young, 1996, p.77). On this background, I will suggest that while the virtual or imaginary status for the experience of all recorded sound is an important *epistemological* insight, the terms of *virtuality* and *realism* can be valuable concepts in the description of spaces, environments and the sources and causes that are in them. I will begin with a discussion of the latter.

2.8.1 The documentary component

The documentary status of recorded sound is not surprising, considering the long history it has had as historical document. After all, sound recording began as a process of letting actual sounding events interact with the recording technology so as to produce a *trace* that can act as a link to the (historical) events for the listener, and from early on sound recording devices have been used as scientific instruments for documentation and archival purposes in several scientific fields.[65] In some cases sound recordings can even be taken as proof of "what actually happened" or "what was actually said".[66] The view of recorded sound as documentary material of people, places, culture, animals, cities and the like can also be found in the practice and theory of electroacoustic music, for example in so-called *soundscape composition* that has been developed during the last 30-40 years.[67] In this genre, recordings of environmental sound have been used both for documentation as well as artistic purposes.[68] The use of historical recordings in electroacoustic music is another example of recorded sound that can be experienced as having documentary status.[69] In such pieces the historical significance of the recognized events or people appears to put an emphasis on the link to the actual event or person, and perhaps also the historical period in which the recording was made. Thus, such recordings can have potential for evoking broad historical and cultural connotations.[70]

The degree to which the sense of documentary is evoked for the listener can vary greatly. This depends on the listener's recognition process in hearing the sound, a process which may or may not give a correct judgment of the sound in question, depending on the knowledge and background of the listener and the degree to which cues in the sound allow recognition.

[65] According to Jonathan Sterne, there are accounts of American ethnologists possessing phonograph cylinders of music from several Native American tribes with the intention of historical documentation and preservation as early as 1889-90 (Sterne, 2003: 315). As early as 1899 the first academic archive of sound recordings was founded, the *Phonogramm Archiv* of the *Ostereichissche Akademie der Wissenschaften*.

[66] As can happen when recording of emergency phone calls are used for evidence in a trial.

[67] This artistic practice grew out of the concern for preserving and improving our sonic environment, especially as expressed by the *World Soundscape Project*, an organization founded in the late 1960's by R. Murray Schafer. See e.g. Schafer, 1994 or Truax, 2001.

[68] Examples of the former can be found in the episode entitled *The music of Horns and Whistles* from the CD *The Vancouver Soundscape* (Schafer et al., 1996).

[69] A good example of use of sound material is found in Wilhelm Zobl's piece *Andere die Welt, Sie braucht es* (1973, on Various_artists, 1988) where several political speeches, among them Richard Nixon's speech after having won the presidential election in 1968, are presented in what appears as a close to intact form. Other examples can be found in Trevor Wishart's *Two Women* (1998, on Wishart, 2000b), which consists of four parts in which public statements from relatively well known voices, namely Princess Diana and Margaret Thatcher .

[70] Certain characteristics, usually the ones that signal "old" recording equipment, might even reinforce the impression that something is "historical".

However, I will not go further on a general discussion at this point, since what is the most important for now is that the degree to which something is experienced as “historical” or “documentary” adds to the terms with which sound sources and environments can be described and qualified. The veridical status of these descriptions is not an issue here.[71]

2.8.2 Degrees of realism

While the documentary aspect may be important when listening to certain sounds, in other cases sound sources, environments and spaces might be experienced as overt constructions. Despite such an experience of constructedness, there is a sense in which we can evaluate how “real”, or perhaps “realistic”, they appear to us. This sense of realism is clearly different from the sense of “documentary” discussed above, since experiencing something as realistic does not preclude the recognition of virtuality.[72] As long as the resemblance with the experiences of the external, physical world is maintained to a certain degree, we might be willing to accept the sounding environment and its spatial features as realistic, despite our recognition that they are clearly a construction. For instance, if the distance to a sound source is simulated by maintaining the relationships between high frequency roll-off, amplitude and the degree of reverberation added in the mix, we might still perceive the situation as fairly realistic, even though we recognize the characteristics of artificial reverberation. And, we might judge the *degree* of realism by considering how the situation corresponds to our experience of spatial properties of sound.

Not only spatial characteristics, but also individual sound sources and their behaviors might be evaluated in terms of realism, even if we are not in a position to assess the documentary status of the sound or have recognized it as a simulation. And, in some cases simulated sounds might give us a more realistic image of a particular source than a high-quality recording of the same source. As Wishart notes,

> […] the recreation of the effect ‘fire’ by purely auditory means, can simply fail to evoke the power of the multi-media image of fire. In this particular case, where we are restricted to the medium of sound, the use of studio fabrication (such as the recording of crinkled cellophane and it subsequent speed-changing,

[71] I will take up the discussion of recognition with special reference to the voice in chapter 3.
[72] In many ways, this resembles how Wishart uses the term “real” and “realistic” in his discussion of the disposition of sound-objects in space. Wishart here exemplifies how different kinds of hypothetical editing and mixing procedures can entail different combinations of real or unreal objects and spaces.

> filtering and mixing with other sources) provide an aural image which is more acceptable than the real thing. (Wishart, 1996: 138)

In this case, the experience becomes actually more "true" for the listener in the version fabricated in the studio than an actual recording of fire. Still, it has to be commented that the opposite can also be the case – a simulation can appear highly unreal(istic), and a recording of animals and birds in a forest can be perceived as highly real(istic).

The last point that I want to mention about the evaluation of realism is that it is dependent on the particular range of knowledge and experience of the listener. As both Wishart and Chion note, for somebody that has specialist knowledge of birds and can recognize species from their song, an auditory image of a forest might be experienced as unrealistic if it contains birds that don't co-exist together in nature (Wishart, 1996: 146; Chion, 1994: 108-109).[73] For listeners without this knowledge, the auditory image may still be perfectly realistic. All in all, realism can be a useful term in the description of spaces, environments and the sound sources that inhabit them, as well as the interrelationships between environments and sources, bearing the relativity of the concept in mind . For the voice as sound source, however, I will go into a separate and more thorough discussion of aspects closely related to these in chapter 7.

2.8.3 Ontological levels

When discussing terms like virtuality, documentary and realism one also has to deal with the fact that in many cases there appears to be several levels of reality one can refer to. Such levels can introduce new sets of experiential domains similar to the ones I have already drawn up *within* the original ones, thus complicating the picture considerably compared to **figure 2.3**. I would like to refer to these as *ontological levels* since they refer to a certain frame of existence or reality within which objects and beings can be situated.

For electroacoustic music there are, as I see it, mainly three ways multiple ontological levels can emerge; 1) by imitation, 2) by insertion of mediating technologies within an already established space frame and 3) by verbal means. I will go further into how the voice can constitute multiple ontological levels in the following chapter (section 3.4.3), where especially 1)

[73] Chion's example deals with television, but the point would be the same as with acousmatic sound.

and 3) are central. At this point I will only present some general ideas that can serve as a useful background to the discussion of the voice.

2.8.3.1 Imitation

In the case of imitative synthesized sounds, the experience of multiple ontological levels may vary greatly. A highly conventionalized imitation might, for example, not be experienced as an imitation at all. An electric piano, for example, is nowadays heard as an instrument in its own right, even if it was once conceived as an imitation. On the other side, a non-conventional imitation is probably recognized as having (at least) two separate source identities, one being the *imitator* and the other being the *imitated*. The *imitator* is then experienced as in some way "responsible" for the *imitated*, as its "originator" or "maker". Even though it may remain more or less hidden at times, it will have to be recognized (but not necessarily identified) at some point for the situation to appear at all as an imitation. The *imitated* would then appear as a "product" of the first, in a way forming the "surface" part of the two source identities, necessarily evoking some sense of presence for the imitation to be effective. The degree of similarity with the target of the imitation would have to be sufficient for the listener to evoke this sense of presence. Therefore, the example with the Mellotron imitating choir singing in section 2.5.4 would probably not be very effective in constituting two source identities with corresponding realities, since the vocal sounds would likely not evoke sufficient presence for the listener – it would probably be heard only as a "Mellotron using choir samples".[74]

In a segment of the first movement of Jean-Claude Risset's *Sud* (ca.0:34-1:40), the situation is different (1985, on Risset, 1987). Here, a kind of forest environment is introduced, but it is soon evident that several of the "birds" are synthetically produced, i.e. that they are *imitated* birds. Still, despite the recognition of the "true" origin of these sounds as a "computer" or "synthesizer", the sounds still present themselves as "birds" situated within some forest-like environment. In other words, the birds clearly have a virtual status, but they are still experienced as constituting a "reality" of some kind. The "reality" of the sound generating systems, however, is much more hidden in this case, but can nevertheless be recognized and assigned with a

[74] Imitations that have been conventionalized, as the "strings" and "brass" sounds of non-sample-based synthesizers, which would now hardly be associated with string instruments at all, might be called *implicit* imitations, while those that are recognized as imitation might be called *explicit* imitations.

"reality" that is in a sense more "real" than the birds. Thus, one can relate the experience to two different ontological levels. And, in this case as with other imitations of this kind, a focus on the *imitator* would be assigned to the **TCM-domain**, whereas the *imitated* would be assigned to the **SE-domain** (or possibly to any objects or beings within an environment, like birds, cars, flies, etc.).

2.8.3.2 Nested media and spaces

Different ontological levels might also arise from situations where one experiences media *within* media, i.e. where one can recognize a medium such as a radio, television, dictaphone, telephone, CD-player etc. situated within a space that is itself mediated. One will thereby have a situation in which the "world" that is played out "inside" the medium, for example in a radio play, an interview, a band playing on a record, would be situated within the "world" or space in which the medium is situated.[75] One would thereby have two sets of space and/or time frames, which, depending on the degree of interaction between them, would be more or less detached from each other.[76] Listening to a hypothetical musical work in which one seemed to be present at one end of a telephone call, thus hearing one 'unmediated' and one 'mediated' voice, for example, would project two different spaces, one within the other, but where the participants in the conversation would share a common time frame and a "social space". The two ontological levels would thereby be partly connected. If the mediated sound played back within a space is recognized as a *recording*, an additional distance is introduced, namely time. This would imply a more accented dissociation between the two ontological levels.

In a section of Yves Daoust's *Mi Bémol* (1990, on Daoust, 1998: 1:08-1:35) one can find an example that illustrates multiple ontological levels: In this section one can recognize (two?) typical news reports as they would sound if coming from television sets or radio localized in an indoor setting (a kitchen?). This would indicate a time frame that would be approximately the same for the "world" of the news report and the space (the kitchen?) where it appears to be presented (news reports often contain both live and pre-recorded material). The space frames of the two, however, would be dissociated from each other, and even more so than in the telephone

[75] This parallels the distinction between what Chion calls *initial sources* and *terminal sources*, where the latter is the mediated source within a narrative that is itself mediated (Chion, 1994: 77).

[76] This is clearly related to the three "acousmatic dislocations" of *time*, *space* and *mechanical causality* (see e.g. Emmerson, 2007:91).

example since there would be no interaction between them – the news report is typically one-way communication.[77]

In any case, the ways in which the ontological levels would relate to each other, the degree to which they would be detached from each other and the way these levels would be related to the experiential domains would all be pertinent criteria in the description of the context in which a voice is experienced. And, as it will be further shown in later chapters, the voice can indeed engage in the play with different space or time frames through imitation and projection of realities by verbal means.

2.9 Chapter summary

In this chapter, I have introduced the concept of *experiential domain* to designate a group of aspects that we potentially might direct our attention towards during listening to acousmatic electroacoustic music, where the aspects have a common function, feature or relationship that intuitively bind them together. Three experiential non-vocal domains were presented: The domain of sound qualities and structures (**SQS-domain**), the domain of technology, composition and mediation (**TCM-domain**), and the domain of space and environment (**SE-domain**). These domains were seen as providing a frame in which the vocal persona and the related vocal experiential domains, were situated and contained. Each of the domains was linked to theories of electroacoustic music, thereby supporting the distinctions between the domains. Moreover, I presented several concepts in relation to the domains to allow for qualification of pertinent aspects belonging to each of them. The three domains were also seen in relation to each other and to other experiential domains outside the focus of this dissertation, namely the *body and mind* and the *extrinsic* domains. Lastly, the discussion of the three domains in focus was expanded with the notions of *ontological levels*, designating levels of reality projected through the acousmatic work. The ontological levels were also related to the concepts of *virtuality*, *documentary* and *realism*, which were proposed as additional qualifiers for the aspects of the discussed domains.

[77] In so far as the situations described designate multiple levels of reality, it can be compared to the situation in narrative fictional literature. The narrators of narrative fiction can be distanced from the characters involved in the plot in different degrees and in different ways, most importantly in this context, in time and space. See e.g. Genette, 1980.

3.0 Experiential domains of the acousmatic voice

3.1 Introduction

After having delineated a general framework of experiential domains and explicating three non-vocal domains in the last chapter, it is now time to go more specifically into features related to the voice. I will continue to use the concept of the experiential domain in grouping aspects that share a certain function, feature or relationship. In that way, I hope to disentangle some of the multidimensional capacities of the voice so as to establish a structured vocabulary for describing vocal sound that can provide a basis for the further development of the theoretical framework in the following part of the thesis.

Before I begin the discussion of what I will refer to as the *vocal* experiential domains, I feel that it is important to emphasize that the voice is not just a sound source/cause like any other. There is little doubt that the voice has a special status for human beings across all cultures, being the primal carrier of verbal communication, a very important one for non-verbal communication, and of course one of the primary "instruments" of musical expression. This special status is also mirrored in a perceptual sensitivity to vocal sounds, and to any meaning that these sounds may convey, be it linguistic, identity related or affective. That this sensitivity is apparent at a very early stage of our development, shows its fundamental importance.[78] The special status of vocal sounds also seems to be reflected in our brain structure, where certain areas and mechanisms appear to be especially dedicated to processing vocal sound.[79] Generally, humans with normal hearing have small problems in recognizing a voice in different acoustical settings, be it at a distance, in different enclosed and open spaces, or mediated through loudspeakers.

But also for listeners of electroacoustic music, the voice tends to afford "something to hold on to", i.e. something one can use to make sense of and appreciate a particular work. Through their studies of listener responses to electroacoustic music, Landy and Weale found that the voice was indeed a factor that the listeners held on to and applied in their

[78] E.g., tests with heart rate measurements of foetuses have suggested that even before we are born, we can recognize our mother's voice among other voices (Kisilevsky et al., 2003). There are also indications that newborns tend to prefer their mother's voice over others' and that infants grant a special status to speech sounds before other types of sound (DeCasper & Fifer, 1980; Vouloumanos & Werker, 2004).

[79] Within neurological research one has attained results that points towards voice sensitive areas and mechanisms in the brain, i.e. areas and mechanisms that are especially devoted to processing vocal sounds (Belin et al., 2000; Levy et al., 2001; Gunji et al., 2003; Levy et al., 2003).

interpretations and descriptions of a work (Landy, 1994; Weale, 2006). Whereas Landy and Weale did not assess whether the voice was a "hold on to factor" that was either more or less important than other factors, some composer-listeners have noted the special status of the sound of the voice in electroacoustic works. For instance, Smalley suggests that the occurrence of the voice in an electroacoustic piece will entail a perceptual shift when introduced:

> "The moment a voice is perceived in a sounding context the listener's ear is drawn to it and interpretation shifts to focus on the unseen human presence, trying to decode the meaning of its utterances and the relationship to the person to the sounding environment […] a human presence where previously there was none changes everything" (Smalley, 1992: 541)

Thus, a voice not only represents an interesting sound source in terms of its sonic repertoire, ranging from singing via speech to non-sense noise making, but implies the introduction of a "human presence", notions of utterance, expression and meaning, and some kind of relationship to an environment surrounding this human presence. Thus, Smalley's statement would make "vocal" electroacoustic music display something highly comparable with the *vococentric* character as Michel Chion diagnoses for cinema listening. Chion states that for a cinema audience "*there are voices, and then everything else*" (Chion, 1999: 5), implying that when a human voice is present, it sets up a hierarchy of perception with the voice at the centre. Moreover, Chion writes that the "*presence of the human voice structures the sonic space containing it* […] If a human voice is a part of [a sonic space], the ear is inevitably carried toward it, picking it out, and structuring the perception of the whole around it. The ear attempts to analyze the sound in order to extract meaning from it – as one peels and sqeezes a fruit – and always tries to *localize* and if possible *identify* the voice" (*loc.cit.*).[80] For Chion and Smalley, then, sounds perceived to be vocal would have a special status for listeners also when the voices are integrated within a work of electroacoustic music – instigating an image of human presence, a search for identity of this presence, its localization in an environment, and to the possibility of utterance and meaning.[81]

[80] Here, Chion paraphrases Christiane Sacco.

[81] This kind of human presence would be an invisible one within the acousmatic branch of electroacoustic music, thus constituting something very close to what Chion labels *acousmêtre* within cinema (Chion, 1999: 21)

3.2 Vocal experiential domains

The vocal experiential domains have already been set in relation to the general framework of experiential domains presented in the last chapter, as represented in **figure 2.3**. In this figure, the vocal domains were located on the side of the figure that was reference-orientated, and the voice can be said to be reference-oriented in (at least) three respects:

1) In pointing to sound producing gestures or actions, i.e. *causes* involving the vocal apparatus of a human being
2) In referring to a sound *source*, namely the (virtual) person or character that we will in most cases imply from a vocal sound – a character I would from now on like to refer to as the *vocal persona*, to adopt Edward T. Cone’s term (Cone, 1974: chp.1)
3) By having the possibility to refer to other things than itself through the use of verbal and non-verbal codes

This leads us to the four vocal experiential domains that I will focus on in this chapter, namely:

- **The vocal gestures-domain (VG-domain):** Embraces features that are directly related to the bodily terms of sound production, involving all the physiological organs and the movement and behavior of these during vocalizations.

- **The identity-domain (ID-domain):** Includes quasi-permanent features that are inferred from the **VG-domain** as well as the other two domains below, such as gender, age, personality, ethnicity and socio-economical background.

- **The affective domain (AF-domain):** This domain comprises features related to affective/emotional states such as emotions, moods, affect bursts and interpersonal stances. Even if affective states might be expressed through verbal means, I will confine this domain to those features that can be directly inferred from the **VG-domain**.

- **The linguistic domains (LI-domain):** Refers to verbal features uttered through vocal gestures (**VG-domain**) and any meanings they may convey.

These four domains will be the main focus of this chapter. One could perhaps have added other domains to these four to deal with issues related to things such as social or group interaction. However, I have chosen not to deal with these issues to delimit the scope of the dissertation, and since it seems that the projection of interaction between individuals is used relatively sparsely in electroacoustic music.

Even though one can see from the descriptions of the four domains above that they are interdependent, there are both practical and theoretical reasons for considering them separately: Practically, because it will be easier to discuss these aspects separately rather than all at once, and theoretically, because correlates of many of these domains have been studied and regarded in separation. For example, the **linguistic-**, **affective-** and the **identity-domains** seem to correspond quite well to the three semiotic layers of spoken communication often applied in linguistics, namely the *linguistic*, *paralinguistic* and the *extralinguistic* layers, respectively (see e.g. Laver, 2003).[82] Of these layers, the paralinguistic is often defined wider than just embracing the affective and the emotional: According to Laver, the paralinguistic layer helps to "emphasize key aspects of verbal information, to imply the speaker's pragmatic intent, to demarcate rhythmic and intonational units of spoken language, to manage the cooperative time-sharing of the role of speaker, to indicate the momentary affective and emotional condition of the speaker, and much else" (*ibid.*: 414). Thus, one sees that an important part of the paralinguistic layer is to engage in an interplay with the linguistic, supporting, complementing, and enriching it as well as making the verbal more intelligible. However, I choose *not* to focus so much on these aspects of the paralinguistic in this context, but constrain myself to considering its affective information.[83] Within brain research, one has also found indications that the brain processes the information related to the mentioned domains separately, at least initially, so that it is only integrated into more unified representations at higher and more abstract levels of semantic processing. For example, Belin and colleagues see the processing of voice by the brain as involving three partly dissociated and independent processes of perception and cognition, namely *speech information*, *affect information* and *identity information*, all corresponding to domains included in my framework (Belin et al., 2004).[84]

Lastly, the criteria used in distinguishing between the four domains is partly related to two of Charles Sanders Peirce's three categories of signs, namely the *index* and the *symbol*:

[82] In my view, however, the term *extralinguistic* has the unfortunate property of indicating that such definitions are external to linguistic cues, when studies point to the importance of lexical, grammatical and realizational aspects of spoken language in defining regional affiliation, group belongingness as well as social class (Laver & Trudgill, 1979).

[83] I will discuss issues related to *prosody* in section 3.6.3 that can be seen as partly paralinguistic.

[84] The model of Belin and his co-workers is in its turn based on Bruce and Young's model of face perception, and the latter model is partly included in the former, regarding face and voice perception as engaging in multimodal interaction contributing to a more holistic person perception (Bruce & Young, 1986).

Firstly, both the **VG-** and the **ID-domains** deal with features that function as *indexical* signs, i.e. “sign[s] with a direct existential connection with its object” (Fiske, 1990: 47). Both these domains have a direct connection with its object by being the source and the cause of the sound. The vocal **VG-domain** might be said to be even more direct than the **ID-domain**, though. This is because the **ID-domain** also embraces features that point beyond vocal production, such as when vocal age provide features for making an estimate of the whole body size, or when vocal gender has implications of other bodily signs of gender that are not related to the voice.[85] As for the **LI-domain**, it deals with the use of arbitrary or convention-based codes and therefore take on the function of *symbolical* signs, which are signs “whose connection with its object is a matter of convention, agreement, or rule” (*ibid.*: 48). The **AF-domain**, on its part, might be situated between the indexical and the symbolical, since it relies both on conventional codes and involuntary patterns of bodily/neural reactions which are universally recognizable.

Since all of the features in the domains I have discussed will have to relate to vocal gestures in some way or another, it seems appropriate to start with a presentation of the **VG-domain**.

3.3 The domain of vocal gestures (VG)

The domain of vocal gestures can be referred to as the *how*-domain of the four, encompassing the experience of the vocal apparatus in action. When our attention is focused on sound producing actions and gestures, we are engaged in a source-cause experience of the most direct kind: We hear the vocal apparatus and the body within which it is functioning in an interactive gestural play; we hear vocal folds in action: lips, teeth and tongue moving around, lungs exhaling and inhaling, and the jaw opening or closing. We can, for instance, when hearing somebody whisper, then sing softly and lastly shout, experience the changes in *how* the sound is produced: The mode of vocalization shifts from letting only air through the larynx to the engagement of the vocal folds in vibration, first softly and then more powerfully;

[85] Of these two domains, one might also see an *iconic* component in the **VG-domain** that is less present in the **ID-domain**. In the former, one can see that there exist certain similarity relationships between the properties of the sound and properties of the production. Such similarity relationships are more difficult to find for the **ID-domain**. For instance, there seems largely to be a correlation between the perceived sound level and the energy applied in making it (see e.g. Traunmüller & Eriksson, 2000). This relationship might not apply when comparing different types of phonation, however, since certain phonation types require more energy (subglottal pressure) for producing sound compared to others. E.g. so-called *flow phonation* will be a lot more effective in terms of needed subglottal pressure relative to sound level than so-called *pressed* phonation. See the discussion of voice quality/phonation types in section 3.3.5 for details.

the jaw that starts out slightly opened, and then fully opens.[86] The domain of vocal gestures can therefore be seen as the *bodily materiality* in experience, where materiality is understood in Chion's sense as "the sound's details that cause us to 'feel' the material conditions of the sound source, and refer to the concrete production of the sound's production" (Chion, 1994: 114). Thereby, this domain parallels Roland Barthes' famous "grain" of the voice; "the materiality of the body speaking its mother tongue" (Barthes, 1991: 270).

3.3.1 Perception involves imitation

Johan Sundberg has stated: "Probably in our imagination we project the voice timbre we hear from our own voice organ, and we analyze how phonation would feel under these imagined conditions. Then we describe the timbre via this imagined phonation" (Sundberg, 1987: 157). Hence, the way we can relate what we hear to *how* we can make similar sounds ourselves can be reflected in the way that we *describe* vocal sounds: A voice sounds "tense" or "relaxed" because we know how it would feel to produce it. And, some vocal sounds are "guttural", whereas others are "nasal" because we can immediately associate them with the areas of the vocal organs that vibrate and resonate.

Indeed, the ability to engage in *imitation* of other people's vocal production appears to be a part of our pre-programmed disposition in how we relate to voices of others from very early on, and it is also crucial for spoken language acquisition. Recent research has indicated that, even without explicit imitation, the imitative process is so much a part of our perceptual system that it is going on continuously as a form of simultaneous mental simulation that is triggered by stimuli from all sense modalities. Quite recently one has detected a class of neurons that are active not only in the motor execution of action, but also when perceiving or imagining the same actions, namely the so called *mirror neurons*.[87] Much of the research on these neurons have strongly suggested that when one perceives an action being executed, or even imagines that one executes an action, one will simultaneously *simulate* the action in a manner often described as *embodied simulation*; "a specific mechanism by means of which our brain/body system models its interactions with the world" (Gallese, 2006: 16).[88]

[86] Fonagy was able to demonstrate this in an experiment in which the listeners were able to "hear" the shape of the lip opening (Fonagy, 1967 reported in Sundberg, 1987: 157).

[87] These neurons were originally detected in macaque monkeys, but have later also been found to exert a similar functions in humans (Gallese et al., 1996; Rizzolatti et al., 2001).

[88] There are also theoreticians that investigate the possibilities for relating more abstract structures to embodied simulation. Godøy has proposed that musical perception has a motor-mimetic component that is not only related to sound producing movements, but in general to more abstract properties of sound (Godøy, 2006). If Godøy is

That this kind of imitation or simulation is automatic to us does not mean, however, that it allows us to know *where* a certain sound is produced or *how* different anatomical systems behave while making it. For instance, one can know *how* to produce the "gravel" voice of Louis Armstrong without knowing the name of the articulators.[89] And, one might know how to make a "creaky" voice, but one might not know that this is due to the vocal folds being strongly adducted and having weak longitudinal tension. So, to actually *know* the location or the name of the organs involved and how they behave in vocal production appear to be a more mediated way of experiencing a vocal sound than engaging in imitation, tacitly or aloud. In this context, the knowledge of the means involved in production can prove particularly important, since it provides a link to a vocabulary that will be necessary when attempting to describe the experience of aspects of vocal sound. Therefore, I will give a description of the most important features involved in different modes of vocal production below. Before I do this, however, I will discuss how *familiarity* of different features can have an effect on the degree of embodied simulation and thereby indirectly how this can affect the specificity of description of the vocal gestures involved.

3.3.2 Familiarity affects degree of imitation

There seems to be certain constraints that determine the degree to which actions are mentally simulated due to the fact that the *familiarity* of the actions appears to play an important role. This is supported by a recent neuroimaging study that found indications that the action or motor repertoire of the observer affects the degree of activation in the motor system (Buccino et al., 2004). Whereas actions belonging to the motor repertoire of the observer were mapped onto the motor-system of the observer, those that didn't were instead mapped and categorized based on visual properties. Similar results have been found when comparing observations of dance movements done by skilled dancers with those of untrained observers, and when piano music has been played for pianists versus non-pianists: The expertise in the motor activities correlated in both cases with the activation of corresponding areas of the brain involved in executing the motor movements (Haueisen & Knosche, 2001; Calvo-Merino et al., 2005;

right, the link between perception, imagery and action can be related to a more abstract form of gestural involvement, where specific sound sources and particular action patterns are less important than a general and abstract form of gestural activation that can be realized as all kinds of bodily gestural expressions, overt as covert. In other words, one can speak of both *motor* and *source equivalence* at this abstracted level. This can for instance be observed in so-called sound-tracing studies, which show that listeners display similarities in the way that they depict or "trace" the properties of a sound graphically (Jensenius, 2007: 82-92).

[89] Trevor Wishart suggests that this kind of voice is produced by a sub-glottal vibration in the windpipe (trachea) (Wishart, 1996:264).

Calvo-Merino et al., 2006). Moreover, a few studies have reported of a higher degree of neurological activation of motor areas during observation of actions made by humans compared to actions made by artificial agents, indicating either that mental simulation will be stronger for human than artificial agents (Press et al., 2005), or that it will happen for human actions only (Tai et al., 2004).

Several comparable studies of speech perception, that thereby have a more direct relation to vocal gestures, have detected activation of areas of the brain that are active both during perception of speech and during speaking (Fadiga et al., 2002; Hickok et al., 2003; Watkins et al., 2003; Watkins & Paus, 2004; Wilson et al., 2004; Skipper et al., 2005). It has been suggested that speech gestures rather than speech sounds therefore are the perceptual primitives of speech (Liberman & Whalen, 2000), and that neurologically speech processing therefore has a pre-lexical *how* processing pathway (Scott & Wise, 2004). The finding that the familiarity of actions will constrain the degree of mental simulation has also found support for the special case of speech. In one study the researchers found that when comparing the perceptions of words known to the listeners to quasi-words outside the listeners' vocabulary, the former showed a much stronger activation than the latter, indicating that the familiarity of the words affected activation (Fadiga et al., 2002).[90] This suggests that the general tendency for mental simulation to depend on the familiarity of actions can also be applied to speech and other vocalizations. Consequently, one can hypothesize that phonemes outside one's own repertoire, unfamiliar linguistic units, unfamiliar modes of vocalization (e.g. over-tone singing for somebody unfamiliar with this) and properties of sound that are caused by the technology of manipulation or synthesis, would probably give less activation than listening to somebody speaking in one's own language. For example, when vocal sounds are manipulated, this will in most cases render results that are impossible for a human being to execute, even if one's skills in vocal imitation are high. E.g. ring-modulating a speech signal with a sine wave with a frequency of 150 Hz, the spectrum of the resulting sound would have an inharmonic quality that would most likely be impossible to imitate. Or, by increasing playback speed of a speech recording to a certain point, one would get high-pitched speech with a fast rate and a "thin" quality that would be very difficult or impossible to imitate.[91]

[90] The authors could not say with certainty whether the effect of increased activation for words over quasi-words would be due to the familiarity of the words or whether the presence of a (known) meaning could have facilitated the activation of the speech motor centers.

[91] Some people, however, have extraordinary imitative skills, and can imitate sounds that seem impossible to imitate for others. The American actor and comedian, Michael Winslow, is according to Wikipedia, "known as the 'Man of 10,000 Sound Effects' for his ability to make realistic sound effects using only his voice" (Wikipedia contributors, 2007a).

Taken together, this suggests that the engagement in embodied simulation will be much stronger for familiar vocal sounds and patterns than unfamiliar ones. Except for trained singers, who have a privileged knowledge of vocal organs and their actions in singing, speech in one's own language and dialect will most likely induce the highest level of embodied vocal simulation. This seems to support a theoretical model in which speech in a listener's own language constitutes a privileged point of reference in listening, and which therefore also can allow for the richest and most specific descriptions if coupled with knowledge of the sound producing organs and actions involved. Therefore, I would now like to give an account of the anatomical systems and organs along with a description of their function and manner of operating for different kinds of vocal production.

3.3.3 The three vocal subsystems

In vocal production, three major anatomical subsystems with different functions are involved:

- **The respiratory system** includes the lungs, the chest wall and the diaphragm, and its function is to create and maintain an airflow with a certain pressure through the glottis and the vocal tract (cf. the definition of the articulatory system below).

- **The phonatory system** includes the vocal folds and the muscles for controlling its configurations located in the larynx tube (see **fig 3.1**). Its function is to generate sound by letting sound pass through the larynx so as to generate vibrations in the vocal folds or simply pass the airflow through it without engaging in vibrations.

- **The articulatory system** includes the vocal tract, i.e. the tube running from just above the larynx all the way to the mouth, the nasal cavity, and the *articulators*, which are movable structures including the tongue, lips, jaw and velum. Its function is to shape and modify the sound generated by the vocal folds by changing the length and the shape of the vocal tract, to obstruct the air flow at different locations, and to generate different kinds of oscillations and/or noise.

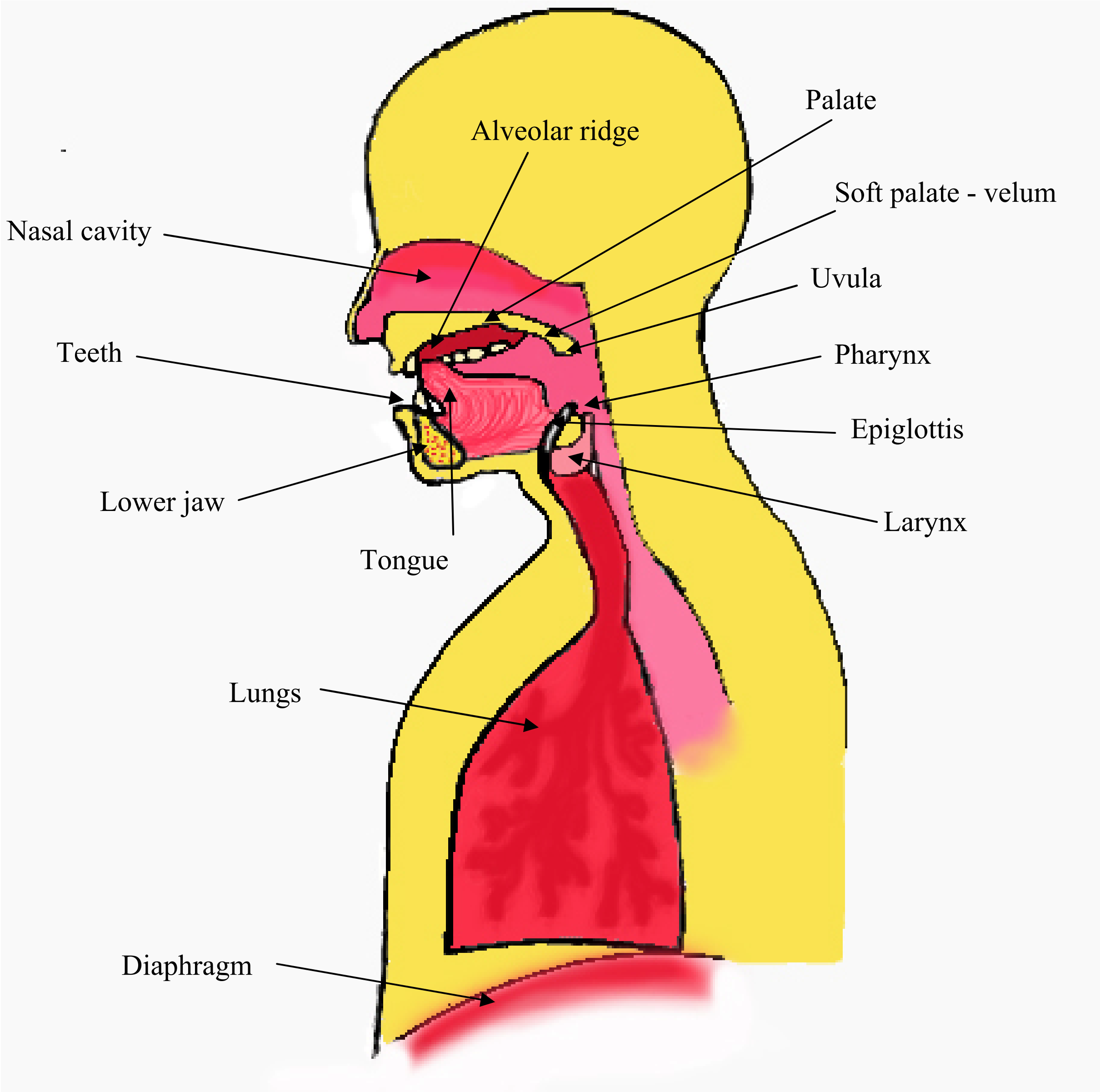


Figure 3.1: Schematic drawing of the physiological components involved in vocalizations.

The location of the different components in the three systems can be seen in the schematic depiction in **figure 3.1**.

The functioning of these systems is not only of interest for understanding the physiology of vocal production. The distinction is also easy to grasp in terms of both production and perception, and it can thereby provide the first crude stage in an experientially based description of vocal sound: We can identify the respiratory system in action when we hear inhalations and exhalations, and we can recognize whether the lungs provide the other systems with a high or a low airflow. As for the phonatory system, we can easily tell when the vocal folds produce pitched vibrations or not, and whether the pitch is high or low. We also easily identify when the articulatory system is at work in its shaping and modification of the sound, or when it is the source of noise or vibrations. Hence, we are so familiar with the basic functions of these systems that we in most cases can identify their individual contributions.

This way of partitioning voice is not only pertinent in a physiological and a perceptual perspective – it also provides a very basic scheme for physical/mathematical modeling of the voice, at least partly. This basic scheme is often referred to as a *source-filter model*, where the contribution of the phonatory system is often referred to as *source*, and the contribution of the articulatory system is called *filter* (Sundberg, 1987: 10). This model is interesting in the context of my thesis, since a number of processing techniques in electroacoustic music can be interpreted as source-filter models, e.g. the LPC-technique, which will be central in the discussion of Paul Lansky's music. The source-filter model can be traced back to early mechanical devices for producing speech sound, like those of von Kempelen (1791) and Johannes Müller (1848), and it was later put in practical use in Homer Dudley's *Voder* and *Vocoder* (Lieberman, 1984: 115; Mattingly, 1974: 2453-54; Dudley, 1939).[92] In his *Acoustic Theory of Speech Production*, a modern version of the source-filter theory, Gunnar Fant postulated that speech production could be theoretically quite well modeled by seeing the contributions of source and filter as independent. According to this theory, speech output can be calculated as a linear product of the spectra of the source and the transfer function of the filter (Fant, 1960).[93] The theory also had the practical benefit of facilitating synthesis of vocal sound, because the linearity of the theory meant that the source sound could be synthesized on its own and the filter could subsequently be applied to the source sound. Even if it has been shown that the source-filter theory has to be corrected for energy losses in the vocal tract, time-variant features and interaction with the respiratory system (Kent & Read, 2002), the source-filter theory provides a useful and simple model for understanding the acoustics of the voice as well as its application in synthesis and processing of vocal sound.

I would now like to go on to discuss each of the subsystems involved in vocal production to try to establish a vocabulary for a more detailed description of vocal gestures.

[92] According to Lieberman (1984), it was Johannes Müller who developed the source-filter theory.

[93] However, the filter part really has two components in the acoustic theory of speech production; the transfer function of the vocal tract and the radiation characteristics of the lips, which can be seen as a simple high-pass filter with a constant rise in the spectrum of 6dB per octave up in frequency. For voiced phonation the spectrum of the laryngeal source is usually presented in an idealized form, with a harmonic spectrum in which the amplitude of the partials is declining with 12dB per octave up in frequency.

3.3.4 Respiration

The respiratory system creates the airflow that passes through the phonatory (larynx) and the articulatory (the vocal tract) systems, and in that respect it can be considered as the main "motor" or provider of energy for voice production. Since breathing will always pass through the vocal tract, it will always be spectrally coloured by it, and thereby never be heard independently or "directly". For phonated sounds the effects of the respiratory system will be heard even more indirectly, namely as changes in loudness, spectrum, and (partly) pitch of the phonation component – and this component will inevitably in turn be spectrally "shaped" by the vocal tract.

3.3.4.1 Breathing

Breathing, along with pulse, body temperature and blood pressure, is one of the vital signs – in other words, breathing is a direct indication of life.[94] We need to breathe to provide a constant refilling of oxygen for our cells, and situations that imply hindrances for breathing are potentially life-threatening. It can be observed externally by the movements of the chest/rib cage, which expand and contract rhythmically, usually at least every 5 seconds, depending on the activity level (Sundberg, 1987: 33). Breathing can also be audible, but this depends on the intensity of breathing as well as the distance to the listener – we usually have to be at a personal distance to hear the breath of somebody at rest, while we can hear a person at much larger distance when he or she is breathing heavily, for example during or after demanding physical activities. Relevant experiential descriptors of breathing sounds are therefore 1) rate of the inhalation/exhalation cycle, 2) periodicity of the inhalation/exhalation cycle, and 3) loudness. We can also often identify "activities" or emotional states from breath sounds only, such as sleep, sexual arousal and terror. Thus, breathing can provide a close link between the **VG-** and the **AF-domains**.

Breathing is an obligatory companion to all kinds of vocalizations. It creates necessary pauses of shorter or longer duration in between vocal phrases, usually in the form of inhalations, audibly or silently. Experientially, the frequency of breathing pauses can be a relevant aspect to include in a description, since we automatically will anticipate a breathing pause in a vocal phrase. Our anticipation will probably grow in strength the more time it has been since the last breath intake was heard. Therefore, operatic singers singing sustained notes

[94] It is not surprising, therefore, that breath has been used as a metaphor for a life-giving force, as when the biblical God breathes life into "the dust of the ground" so that it becomes a living being (Genesis 2:7).

of long duration can create a kind of dramatic tension related to how long they are able to hold the note before taking a breath.[95] Phrases that are artificially made too long might also seem unnatural because one will usually anticipate a breathing pause. Moreover, when listening to artificial voices that don't breathe, listeners can find it more unnatural of precisely that reason (Whalen & Hoequist, 1993).

3.3.4.2 The role of respiration in vocalizations

The main role of the respiratory system in producing vocal sound is to provide what is usually referred to as *subglottal pressure*, i.e. air pressure that is created below the area called the glottis (the vocal folds and the space between them, located in the middle of the larynx, see **figure 3.1**). Several muscles contribute to creating this pressure, among them the *diaphragm* (see **figure 3.1**) and the muscles in the abdominal wall. Subglottal pressure has been shown to have a major impact on the amplitude of the vocal sound. Thus, increased pressure is correlated with louder sounds, although phonatory parameters are also important (Ladefoged & McKinney, 1963; Plant & Younger, 2000). Fundamental frequency of the vibration of the vocal folds also tend to rise with subglottal pressure, especially for singing in the highest part of the pitch range (Sundberg, 1987: 36-37).

Studies reported by Sundberg indicate that the relationship between subglottal pressure and experienced loudness in singing is relatively straightforward, with about 9 dB increase for a doubling of subglottal pressure (*ibid.*: 39). However, for acousmatic music it makes little sense to use loudness as a parameter in the qualification of a vocal sound other than in a relative sense, since loudness will depend on the settings of playback sound level. Still, we seem to be able to judge intuitively how much force or effort has gone into the production of a vocal sound. Rather than referring to subglottal pressure, which is a physiological parameter, the term *vocal effort* seems to be a useful term for the experienced "force" applied in vocalization. This would be especially relevant for speech, for which well-established dynamic categories like *pianissimo* and *forte* are not applicable. Vocal effort is considered an important factor for how loud we experience a vocal sound to be (Eriksson & Traunmüller, 2002). Effort is immediately recognized by the ear so that even if two vocal sounds have equal sound pressure level (SPL), one can easily distinguish between different degrees of vocal effort, for instance when distinguishing between a shouting versus a normal voice. We

[95] The dramatic tension is also clearly related to the performer displaying his/her skills in vocal production efficiency in terms of the amounts of air used in phonation and of vital capacity of the lungs.

probably make use of several spectral cues that can tell us something about the effort applied; in its simplest terms the spectral brightness generally increases with higher effort.[96] For speech, the duration of vowels also appears to be prolonged compared to consonants with increased effort, and the overall pitch level of an utterance tends to rise (Andersson et al., 1996). Lastly, studies of vocal effort has also found results that link vocal effort to the perception of communication distance, in fact to such a degree that vocal effort has been defined relative to it: "vocal effort is the quantity that ordinary speakers vary when they adapt their speech to the demands of an increased or decreased communication distance" (Traunmüller & Eriksson, 2000: 3438). Thereby, it is easy to see the link to Hall's correlation of distance with "mode of voice". I want to add to this, however, that in a musical context increased vocal effort might not indicate an increase in experienced physical distance *per se*, but rather function as a *projection of* psychological/social distance.

The last qualifier that I want to add for respiration in vocalization is simply the distinction between *inhalation* and *exhalation*, which I assume can be recognized in some cases, while it may be difficult in others. While exhalation is the "natural" mode of vocalization, there are a number of vocalizations that can also be performed while inhaling, including phonated vocalizations. For instance, even if I had not tried it before, I was able to sing a short phrase with inhaled phonation, meaning that the subglottal pressure will have to be an under-pressure instead of over-pressure. However, the phonation was clearly marked by the reversal of the pressure, creating a more "breathy" quality, and I was only able to sing for a short time before the pressure ended. Trevor Wishart, who has systematically explored the musical potential of the human vocal repertoire, describes and demonstrates a number of possibilities for creating inhaled sounds: "By varying the tension of (I believe) the larynx and the filtering of the oral cavity [during inhalation] a great number of sounds can be produced: from pure tones [...] which may be outside the normal range, click trains [...], sub-harmonics [...], more complex multiphonics [...] to complex and unstable oscillations (usually produced at the end of a long in-breath when the pressure inwards is difficult to maintain) (Wishart, 1996: 275). As Wishart also notes, inhaling voice is more unstable than exhaling, most probably due to problems with control (Wishart, 1996). However, when listening to Wishart's sounds, I must admit that I could not recognize that they were produced by inhalation. It might be, therefore, that sounds have to be explicitly "marked" as inhalations or preceded and

[96] Acoustically, increased vocal effort has been found to imply 1) increased sound level (SPL), 2) raised first formant, F1, and 3) increase in the amplitude of the higher partials relative to the fundamental/spectral tilt (Traunmüller & Eriksson, 2000; Liénard & Di Benedetto, 1999).

followed by exhalations to make it easy to identify the direction of the airstream. However, I will assume that for “normal” vocalizations like speech and song, it would be relatively easy to distinguish inhaled vocalizations from exhaled vocalizations.

3.3.5 Phonation

The phonatory system is centered around the glottis and the vocal folds in the larynx. In addition to providing the physiological function of preventing airflow to the lungs so as to make the respiratory system a foundation for pushing e.g. during child-birth, the vocal folds can produce sound. The sounds produced at the glottis can be either *voiced* or *unvoiced*. In the first case the vocal folds engage in periodic vibration, producing a pitched sound corresponding to the rate of vibration. In the second case, air is merely let through the glottis so as to produce turbulent noise, as in whispering. Whispering, even though it does not involve vibrations in the vocal folds, is still usually regarded as one type of *phonation*. Since it seems to offer the greatest range of variation, I will describe only features related to voiced phonation in the following.

3.3.5.1 Voiced phonation

Pitch: The perceived pitch of a vocal sound is usually closely correlated with the fundamental frequency (f0) of the vibrations of the vocal folds. In addition to being affected by the subglottal pressure, f0 is mainly controlled by varying the tension and the length of the vocal folds (Tempelaars, 1996: 43). There are potentially many properties related to pitch that can be relevant in this context. Since pitch by itself can be experienced as abstracted from its production and related to the musical system, e.g. intervals, arpeggio, melody, glissando etc., I will consider such aspects to reside within the **SQS-domain**. Here, I will rather present some aspects that relate more directly to vocal production, namely *range*, *register*, *fluctuations* and *intonation/stability*.

- **Range:** Range can be described either in itself as being either wide or narrow, or in terms of upper and lower limits. The range of pitches that can be produced by a voice is dependent on several factors. Among them are the following:

- **The length of the vocal folds:** This is the main explanation for the difference in pitch produced by men, women and children (Titze, 1989; Kent & Read, 2002: 191).[97]
- **Phonation type:** By using falsetto phonation (see below) the pitch range can be extended upwards.
- **Training** can usually extend the pitch range both upwards and downwards.
- **Individual factors:** Different voices have different ranges. In Western classical singing voice categories like soprano, alto, tenor, baritone and bass, have vocal range as one of the significant characteristics (Handel, 1989: 166; Erickson, 2003).[98]
- **Relative location:** As listeners, we can usually hear in what part of an individual's vocal range a certain vocalization is located, that is, whether it is in the middle, comfortable register or significantly above or below this part of the range.[99]

- **Register:** Relative location is partly related to *register*, which designates ranges of the voice, particularly in singing, where the phonatory quality sounds similar and where it feels more or less the same to produce a sound (Sundberg, 1987: 49-51). In classical singing one distinguishes between several registers, which will be beyond the scope of this dissertation to go into detail about. In general, the difference between *falsetto* and *modal* registers is perhaps the most salient, especially for the male voice. This will be discussed further in the following section (3.3.5.2).

- **Fluctuations:** Fluctuations can in principle be described by the same parameters as those presented in section 2.5.2.2 on the detail level of the **SQS-domain**, namely *velocity*, (size of the) *deviation* and *regularity*. In song or song-like expressions, regular fluctuations in the form of *vibrato* are often an important part of the expression. The *bel canto* vibrato, as well as fluctuations in most natural sound sources, tends to involve not just pitch, but intensity and spectrum as well, even if the pitch fluctuations are by far the most salient in the experience (Verfaille et al., 2005). Within the speech sciences, irregular fluctuations of the voice are often labeled *jitter* (fluctuations in f0) and *shimmer* (fluctuations in intensity) (see e.g. Aoki & Ifukube,

[97] See also the discussion of the identity domain in section 3.4 below.

[98] Formant frequencies are also taken to be important in voice classification in classical singing, suggesting that vocal tract length also plays a role, see e.g. Cleveland, 1977.

[99] This has experimental support, see Honorof & Whalen, 2005. .

1996). Such fluctuations might also point to reduced health or old age (Feinberg, 2004), or affective states like fear. Indirectly, one can therefore be directed both toward features of the **AF-** and the **ID-domain** through this parameter.

- **Intonation/stability:** Deviations from norms and standards when it comes to pitch structures, such as in intervals, melodies and harmonies, may draw attention towards the physical apparatus of production, i.e. the **VG-domain**. This might happen, for example, if pitches are not properly intonated, or if the pitches of sustained notes are not stable.

3.3.5.2 Phonation type

The vocal folds can produce different types of phonation, which have several characteristic features both as to how they are produced and how they sound.[100] Here, I will consider the most common phonation types used in speech and singing:

Modal voice: The "normal" type of voiced phonation used in speech is usually referred to as *modal* voice. It is produced by applying a moderate tension of the vocal folds so that the folds are vibrating efficiently with minimal irregularities and without audible friction (Laver, 1980: 111).

Flow phonation: Modal voice appears to be not too far from what Sundberg refers to as *flow phonation*, which is characterized as being an effective type of vocalization with less subglottal pressure, and vibrations with higher amplitude. Therefore, this kind of phonation is the one usually applied in classical singing (Sundberg, 1987: 80).

Harsh/pressed voice: If the vocal folds are tightly adducted so that a high sub-glottal pressure has to be applied, the result is often called *harsh* voice (Laver, 1980: 126-132) or *pressed* phonation (Sundberg, 1987).[101] The result is usually a "raspy", "rough", "metallic" or

[100] Phonation type is also closely related to the concept of *voice quality*, which nevertheless tend to be used to denote the auditory quality of the vocal sound which is dependent both on phonation type and supralaryngeal features (Trask, 1996: 381).

[101] This kind of phonation is much less efficient because it is produced with a high sub-glottal air pressure combined with a high degree of pressure and tension adducting the vocal folds.

"raucous" voice quality, with a certain degree of aperiodic noise caused by irregularities in the vibrations (Laver, 1980: 127).

Breathy voice: If the vocal folds are only partially adducted so that parts of the airflow can pass freely, and that only parts of the vocal folds vibrate, the result is often labeled *breathy* voice (Sundberg, 1987: 80; Laver, 1980: 132-135). The degree to which breath is allowed to pass through the vocal folds will correspond to the degree of breathiness or noisiness in phonation.

Creaky voice: *Creaky* voice, or *vocal fry*, is a type of phonation in which the vocal folds vibrate with a very slow vibration rate driven by a very low sub-glottal air pressure. The vibrations can be so slow that almost single vibrations can be heard, and the sense of pitch disappears altogether (Laver, 1980: 122-126). Creaky voice is sometimes also referred to as the *pulse register* (Sundberg, 1987: 50). This phonation type is also characterized by irregularities in the vibration rate. Some vocal performers, e.g. the Norwegian Sidsel Endresen, are capable of creating single glottal pulses, and even create a continuum from single pulses to audible notes.

Falsetto: Falsetto is characterized by higher fundamental frequency and breathier quality than the modal voice, the latter due to air leakage through the glottis during phonation (Laver, 1980: 118-120). In singing, it can also occur under the labels of *loft* register (Hollien, 1974).

Other qualities: Features like *hoarseness* or *roughness* can also be valuable in describing phonation (Parsa & Jamieson, 2001). For such qualities, the presence of extraneous noise, jitter and sub-harmonics can be a part of the picture (Omori et al., 1997). These qualities might direct attention towards the bodily production of the voice, and can also be used intentionally with an artistic purpose.[102] In a *diplophonic* voice the two vocal folds will vibrate independently with slightly different frequencies, causing a kind of "double" phonation with inharmonic qualities (Gerratt et al., 1988).

[102] Trevor Wishart describes how he can produce sub-harmonics intentionally (Wishart, 1996:265).

3.3.6 Articulation

The articulatory part of vocal production is very complex, since it involves several organs of very different configuration. The articulatory system is responsible for *resonances* as well as *sound production* by vibration, turbulence (e.g. as in frication sounds) or release of pressure. The ways in which the organs can shape, alter and produce sound are so many that it would be impossible within this context to give a comprehensive presentation. Therefore, I would only like to provide some very basic notions that can be useful in the description of articulation from a listening perspective.

3.3.6.1 Resonances

The *resonances* created when the sound from the respiratory and phonatory systems enter through the vocal tract have a major influence in determining the experienced quality of the vocal sound. Firstly, the *length* of the vocal tract is important since it determines the frequency range of the resonant frequencies – the shorter the vocal tract, the higher the frequencies of these resonances (Kent & Read, 2002: 28-32; Fant, 1960). The placement of the resonant frequencies is also a valuable indication of the size of the person, and vocal tract length is therefore an important property in judgment of the identity of the voice (cf. section 3.4). However, the length of the vocal tract can also be varied by lowering or elevating the larynx, something which also affects the resonant frequencies and the perceived vocal timbre (Sundberg, 1987: 113-115). Secondly, the positioning of the different articulators (see **figure 3.1**) at different locations can create many resonant configurations with distinct perceptual qualities. Many of the configuration possibilities for the articulators are utilized in speech, in particular the vowel sounds, and this provides a link between the vocal gestures domain and the **LI-domain**. Our ears have a particular sensitivity to the distinction between such sounds, especially those from our own language (cf. section 3.3.2). However, we can to some extent also recognize the placement and configuration of the articulators independent of linguistic content; we can in particular recognize the *tongue*'s placement, and the *lips* rounding or retracting. Again, familiarity is an issue, meaning that we can both reproduce and describe familiar resonances more easily and precisely.

The resonant qualities of the vocal tract will cause some parts of frequency spectrum to be strengthened in amplitude, whereas other parts are attenuated. The strengthened parts of the spectrum can be described in terms of resonant peaks in the frequency spectrum, in acoustics usually referred to as *formants*. These formants have been shown to have a close relationship to the perceptually distinct character of many speech sounds, and vowels in particular. Most vowels can be described and modeled quite well by the frequencies of the first two formants, which together are frequently seen as forming a two-dimensional chart, often referred to as the *vowel space*, in which the whole palette of vowels can be placed.[103] As we can see from **figure 3.2** this vowel space is delimited by the upper and lower boundaries for the possible formant values, which differ among men, women and children.

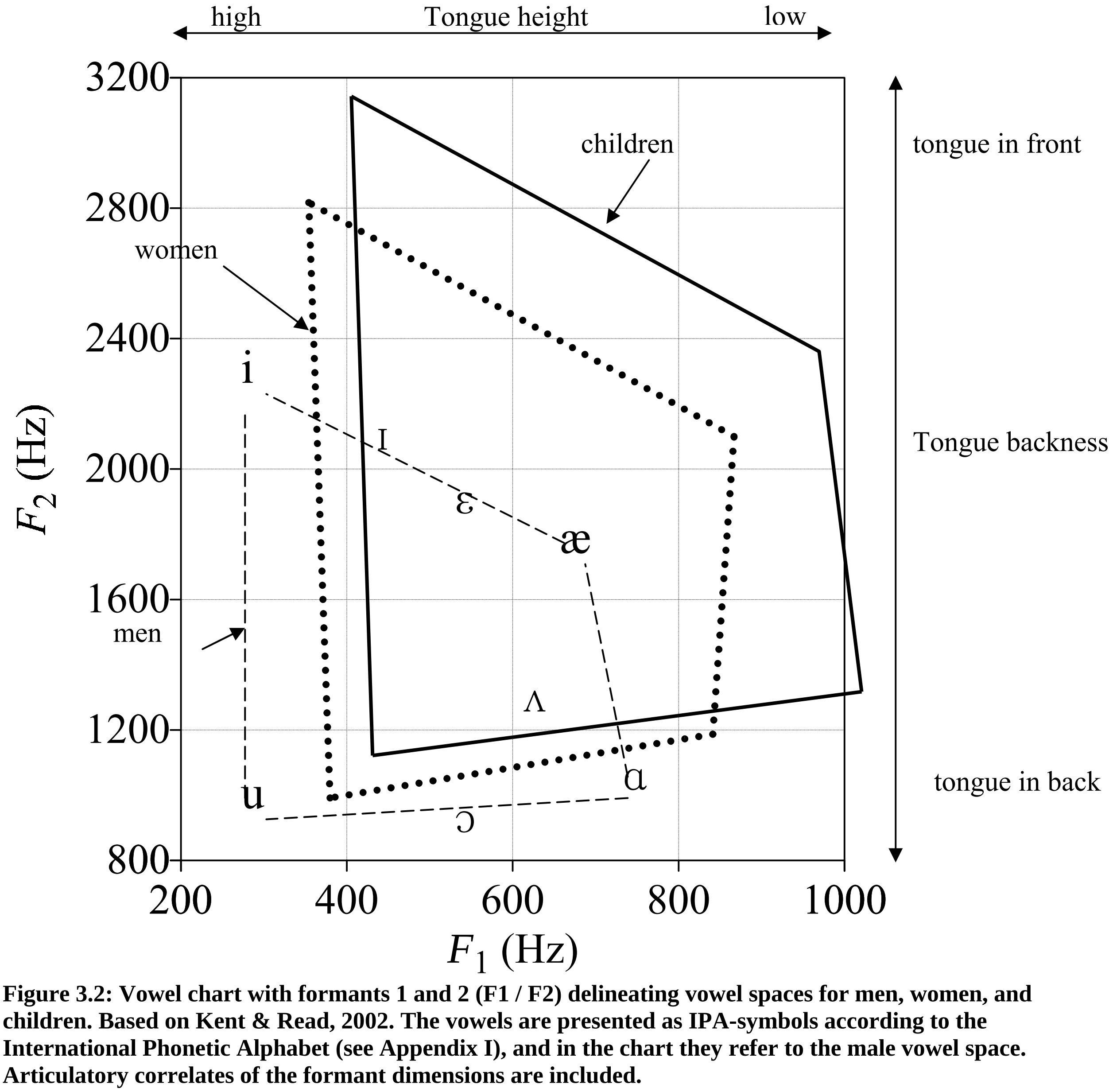


Figure 3.2: Vowel chart with formants 1 and 2 (F1 / F2) delineating vowel spaces for men, women, and children. Based on Kent & Read, 2002. The vowels are presented as IPA-symbols according to the International Phonetic Alphabet (see Appendix I), and in the chart they refer to the male vowel space. Articulatory correlates of the formant dimensions are included.

[103] Still, it is important to emphasize that a model based solely on the first two formant frequencies is a simplification. Vowels also possess many dynamic qualities, and when considering vowel sounds in the context of other speech sounds it can be necessary to consider e.g. the trajectories of formant frequencies, formant amplitudes, and fundamental frequency (see e.g. Kent & Read, 2002: 108-110).

What is perhaps the most interesting thing in this context is that there is a correlation between the two first formants and two articulatory dimensions. Whereas the first formant correlates negatively with *tongue height* during the vowel production, the second formant correlate with *tongue advancement* or "backness", i.e. whether the tongue is placed in the back or the front of the mouth during articulation (Kent & Read, 2002). *Lip rounding* is also a factor that creates perceptually salient distinctions in the vowel qualities, e.g. between [i] and [y] (rounded) and [e] and [ø] (rounded). Since lip-rounding lowers both formants, it cannot be mapped as easily to the vowel space, however.[104]

The last thing to mention about the possibilities for creating resonances with the articulators in the vocal tract, is the optional use of the *nasal cavity* as a resonator, such as in the nasal consonants ([m], [n]) and nasal vowels (e.g. [õ]). When lowering the velum so that the nasal cavity resonates, one will regularly find that nasalization has the acoustical results of adding a set of nasal formants and *antiformants* (negative peaks in the frequency spectrum) to the oral formants, which are not perhaps as perceptually salient as the formants, but which still contribute in making nasal sounds recognizable as such in most situations (Kent & Read, 2002: 171-177).

3.3.6.2 Articulators as sound producers

As for the sound producing action of the articulators, there are a number of different ways this can happen, and here I will only very briefly mention a few. The most important ways are by stopping the airstream temporarily and then releasing it again in *stops* and *clicks*, by creating *constrictions* that create frication noise, and by producing *oscillation*. These will be considered in turn below. Again, which articulators are used (cf. **figure 3.1**), where they are *located* and how they are *configured* relative to each other (e.g. tongue pointed towards the teeth or bent backwards towards the palate, upper teeth against lower lip, etc.) are aspects we can use to describe these sounds. And as previously stated, those sounds that we are familiar with through language are those we will be most likely to recognize and give a precise description of.

[104] Ladefoged suggests that lip rounding is correlated negatively with the second formant, whereas Kent & Read (2002) maintains that lip rounding will cause a lowering of all formants (Ladefoged, 2005: 41-42).

Stops and clicks: In the stop sounds, the airflow is stopped completely by the constriction (lips against each other as in [b] and [p], tongue against different parts of the palate or velum as in [d] and [k]), and then subsequently released, so as to create a burst of air and energy with a distinctive character depending on the *place of articulation* and presence and timing of any *voiced phonation*. Such bursts of energy can also be found when other forms of tensions or energy build-ups are released, such as clicking or smacking with the tongue.[105]

Fricatives: In the *fricative* sounds, the constriction is not complete, so that air passes through the narrow constriction, thus creating frication noise, e.g. [f] and [s]. Many of these sounds have a characteristic noise spectrum with great portions of the energy located in a relatively high frequency range.

Oscillations: The articulators can be used to create periodic oscillations or trills of different kinds, such as in the rolled and uvular /r/'s, ([r] and [ʀ] in IPA notation, as in Spanish and French, respectively). Many of these oscillations are in the sub-audio range, but with extra effort such vibrations can even reach the audible pitch range (Wishart, 1996: 263-264). This is dependent on the *regularity* of the oscillations, however. For instance, it is not difficult to create an audible tone by speeding up the rate of oscillations for an [r] sound, while it seems impossible to get anything but noise from speeding up the oscillations for the [ʀ].

***

Taken together, the phonatory and articulatory functions, powered by the respiratory airflow, can produce an amazingly wide range of sounds. Many of the effects of different types of phonation, resonances (filtering), and articulatory sound production are often combined into a palette which forms the basis of the speech sounds of the languages of the world. But, as Trevor Wishart has both argued and shown through sound examples, the possibilities for human sound production greatly exceeds speech sounds. For instance, Wishart demonstrated how two and sometimes three different sources of periodic oscillation can be active simultaneously, how tongue and mouth positions that usually are associated with different speech sounds can be combined, and how parts of the articulators other than those used in speech production can be made to oscillate, both in sub-audio and audio frequency ranges

[105] Some of these sounds are even used as consonants in sub-Saharan African languages.

(Wishart, 1996: 263-286). Whether such complex sounds can be recognized and precisely described by listeners that are not practiced vocal performers themselves, however, is less likely.

3.4 The identity domain (ID)

The experiential domain of *identity* can be regarded as the *who*-domain within this framework, since it deals with the recognition, identification, categorization and discrimination of the identity of the vocal persona by the listener.[106] This can be a question of recognizing a voice as a person that is familiar to us to some degree, of categorizing an unfamiliar voice in relation to features such as age, gender, size, ethnicity, social belongingness and so forth, or identifying two vocal phrases as being related to either the same or different identities. In this section, I will try to show how such features are partly dependent on the listeners' background and knowledge, how they can be related to acoustical properties and properties of vocal production, and how these features may be affected by manipulations.

3.4.1 Categorization of voice identity

An important part of the identification of a voice will be to *classify* it according to a set of categories, so as to *define* the identity of the vocal persona. Such categories can refer to a wide range of features such as gender, age, personality, ethnicity, size and other quasi-permanent bodily characteristics, pathological conditions, socio-economic and geographical belongingness, regional accent and dialect. Attributes related to profession and social, economic or political function might also be relevant in certain cases. These are all features which are mostly *not* voice specific, and which one can apply on a general basis in daily social interaction, but which can *potentially* be recognized from the voice, depending on the availability of cues and the experience and background of the listener.

The listener's background will be crucial for recognizing culture-specific features, whereas some features will largely be universally recognized. Thus, the identity categories

[106] The terms "identification" and "recognition" seem to be used interchangeably in many studies referring to the ability to assign a category or a name to a voice. However, McAdams notes that these terms are used in a stricter sense in some areas of research. Here, "recognition" refers to the process of making a match between what is currently heard and something that has been heard in the past, but which needs not be defined. "Identification", on the other hand, comprises the process of naming or labeling an event (McAdams, 1993: 148).

can be divided into two groups depending on whether the listener will need culture-specific knowledge to be able to recognize them, or not:

1. **Universal** or cross-cultural categories include gender, age, size and common pathological conditions such as hoarseness, sore throat or a running nose.
2. **Culture-specific** categories such as regional/social belongingness will depend on more specific kinds of knowledge the listener has acquired, for example what language he/she speaks/understands well, what part of a region a speaker comes from, what kind of social group the speaker can be associated with and so on.

Admittedly, universal categories might also have culture-specific features. There are for instance many gender cues that are socio-culturally rooted (see Günzburger, 1995 for a short review). For features like personality and attractiveness, there will probably also be a combination of universal and culture-specific issues involved, even if I have no support of that claim. Anyway, for all categories that might involve cultural-specific factors, this can be a source of differentiation between listeners when it comes to recognition. Since many of the voices applied in electroacoustic compositions use languages other than my own (Norwegian), and since my knowledge of regional differences and social relationships in other countries attributed from voice is limited, I will not have access to as many cues as native speakers. In some cases it might be possible to remedy this by studying relevant cultural aspects, but in most cases one will have to settle with recognizing that the cultural relativity will favor listeners differently from piece to piece.

By identifying *acoustic* cues that potentially can affect culture-specific categories, one can make some assumptions regarding the degree to which they are affected by different kinds of electronic manipulations. Even if the categories mentioned above might be sufficient to describe the identities of voices in electroacoustic music, it might also be interesting to link some of the most central categories to relevant research in the area. In **table 3.1** I have gathered results and conclusions from a number of studies on voice identification, recognition and discrimination of gender, age, attractiveness, size, personality, ethnicity, pathological conditions/health and regional belongingness/origin. In the table, the identity categories are presented in relation to acoustical and vocal production-related properties, something which allows for a more comprehensive understanding of the interrelationships involved. One might

Identity feature	Effects on classification, recognition/identification and discrimination/similarity evaluation	Related acoustic and production features involved
Gender	· Gender is a highly salient cue in classification. [1] · Gender can be evaluated as a binary (male/female) or a continuous property (feminine/masculine). [2] · Identification of gender can be made with high reliability. [3] · Identification of gender can be made from whispered sounds, but with lowered recognition rate than voiced vocalizations. [4] · Identification of gender can be made from filtered vocalizations, but with lower recognition rate than unfiltered sounds. [3] · Gender can to some extent be recognized from sine-wave replicas of sentences (see also sect.3.6.4). [5] · Gender identification can be made from children's voices. [6] · Synthesized voices can be distinguished based on gender cues. [7] · Changes in playback speed can affect gender judgments. [8]	· F0 is generally about 1.7 times higher for women than for men. [9] · The upper and lower limits for F0 for singing voices differ with about an octave or higher between female and male voice types. [10] · F0 st.dev. is higher for women than for men. [11] Higher F0 variability => more feminine sounding. [12] · Sharper gradients in ΔF0 for women than men. More monotone => more masculine. [11] · Longer apparent vocal tract is judged more masculine. [13] · Formant frequencies are about 20% higher for women than for men. Formant amplitudes are lower for women than for men. [14] · Wider formant bandwidths for women than for men. [15] · Steeper spectral slope (=less high frequency content) for women than for men. [15] · H1-A3 (Source spectral tilt) is lower for women than men. [16] · Women's voices have generally lower SPL than men. [9] · Higher noise content (more breathy voice quality) in voiced segments for women than for men. [17]
Attractiveness	· Vocal and visual attractiveness are correlated for women. [18] · Men: Lowered F0 by manipulation were rated more attractive by women than unmanipulated, and raised F0 by manipulation were rated less attractive. [19]	· Men: Lower F0 is judged more attractive. [19], [20] · Women: Higher F0 + higher formants is judged more attractive. [18]
Age	· Age is a highly salient characteristic of the voice. [7] · Listeners can estimate age reasonably well from voice. [21] · Listeners can estimate age from whispered sounds better than chance. [21] · Changes in playback speed can affect age judgments. [8]	· There is a decrease of F0 from about 400Hz to about 220Hz between birth and 12 years old. There are no significant gender differences. [9], [6] · F0 for boys drop from about 220Hz to about 110Hz between 12 and 16 years old. [6] · Gender differences for formant frequencies develop around 7-8 years old and increase from puberty. [9] · Women: Decrease of F0 after menopause. [22] · Men: Decrease of F0 between 20 and 45, increase after 45. [21] · St.dev. of F0 and intensity increases with age. [21] · High jitter/vocal tremor and shimmer => high rating of perceived age. [23] · Vibrato rate decreases and vibrato extent increases for professional singers between approx. 20 and 60 years old. [24] · Speaking rate tends to be slower for children compared to adults [25], and for children it tends to increase with age up to a point. [26] · Speaking rate decreases with perceived (old) age. [27] · Decrease of formant frequencies with increased age from about 1000Hz(F1) and 3000Hz(F2) for infants to 500Hz(F1) and 1000Hz(F2) for adult males.[9] Decrease with (old) age has also been observed. [28]
Size	· Body size and vocal tract length are correlated [29], but for adults the correlation is only weak. [30]	· F0 and spectral envelope interact in speaker size judgments: => higher F0 and spectral envelope higher in frequency => perceived size is smaller. [31]

Ethnicity	· Afro-American and Caucasian voices can be distinguished at better than chance levels. [32] · Filtering can affect ethnicity judgments more than gender judgments. [33]	
Regional belongingness/ origin	· Speakers are more easily recognized when they speak a language which the listeners have good familiarity with.[34] · Regional dialect can over-shadow other speaker cues in voice identification. [1] · The degree of resolution in classification of regional origin based on dialect depends on the listener's knowledge and experience of the dialect regional areas. [1] · Similarity judgments of dialect depends on knowledge and experience with dialects. [1]	· F0 mean and range, intonation patterns, temporal patterns, and formant frequencies can be significant factor in identifying dialects and accents. [35]
Pathology/ health	· Health rating and age have been found to be negatively correlated. [36]	· High jitter and shimmer => low rating of perceived health. [36] · Low Harmonic to Noise Ratio (HNR) => low vocal function [37] and physical fitness. [38]
Personality		· Extroverts and emotionally stable speakers tend to have louder voices (greater effort) and greater dynamic range than introverts and emotionally unstable individuals. [39] · Extroverts tend to have larger higher mean F0, wider F0 range, and faster speech rate than introverts. [40] · For rate manipulated voices, faster rates are judged as less benevolent and slower rates are judged as less competent. [41]

Table 3.1: Reported results/conclusions from studies or reviews of studies of different identity related features. Reported results and conclusions refer to studies on speech or speech segments unless otherwise indicated. [1]: Eriksson, 2007; Murry & Singh, 1980, [2]: Ko et al., 2006; Nass & Brave, 2005; Wolfe et al., 1990, [3]: Lass et al., 1976, [4]: Schwartz & Rine, 1968; Lass et al., 1976; Ingemann, 1968;Schwartz, 1968, [5]: Fellowes et al., 1997, [6]: Perry et al., 2001, [7]: Nass & Brave, 2005, [8]: Chiba & Kajiyama, 1958, [9]: Kent & Read, 2002, [10]: Handel, 1989, [11]: Günzburger, 1995, [12]: Wolfe et al., 1990; Perry et al., 2001; Smith, 1979; Ko et al., 2006, [13]: Feinberg et al., 2005; Coleman, 1976, [14]: Wu & Childers, 1991, [15]: Childers & Wu, 1991 [16]: Hanson & Chuang, 1999, [17]: Karlsson, 1992; Klatt & Klatt, 1990, [18]: Collins & Missing, 2003, [19]: Feinberg et al., 2005, [20]: Collins, 2000, [21]: Linville, 1996, [22]: Ferrand, 2002; Brown et al., 1996; Xue & Deliyski, 2001; Winkler et al., 2003; Linville, 1996, [23]: Feinberg, 2004; Brückl & Sendlmeier, 2003; Debruyne & Decoster, 1999, [24]: Sundberg et al., 1998, [25]: Kent & Forner, 1980 [26]: Walker et al., 1992; Sturm & Seery, 2007 [27]: [1] Whereas some studies indicate that speaking rate generally decreases with old age (Shrivastav et al., 2003; Schötz, 2003; Schötz, 2004; Ptacek & Sander, 1966), other have found that this is only for reading, indicating a decline in cognitive processing (Winkler et al., 2003; Brückl & Sendlmeier, 2003), [28]: Debruyne & Decoster, 1999 [29]: Fitch & Giedd, 1999; Perry et al., 2001, [30]: González, 2004, [31]: Smith & Patterson, 2005; Smith et al., 2005 [32]: Trent, 1995; Walton & Orlikoff, 1994 [33]: Lass, 1980 [34]: Goggin et al., 1991, [35]: Foreman, 1999; Kent & Read, 2002 [36]: Feinberg, 2004 [37]: Ferrand, 2002, [38]: Ramig & Ringel, 1983, [39]: Trouvain et al., 2006; Scherer, 1979; Scherer, 1978, [40]: Nass & Lee, 2001, [41]: Brown et al., 1973

in particular note how spectral envelope (often only measured as formant frequencies) and f0 are closely related to the categories *gender*, *age* and *size*. One can observe, for example, that women have f0 about 70% above that of men, and that the spectral envelope for women lies about 20% above that of men on average. One also sees that for children, the percentage difference in these two parameters relative to men is even higher, depending on age. These two parameters are also correlated with perceived body size, with higher parameter values rendering the experience of smaller persons, even though one must keep in mind that this is largely due to the correlation between age and size before adulthood. Electronic manipulations that affect these parameters, such as varying playback speed, can influence judgments of both age and gender, as one can see from the table. Since the percentage difference of f0 and spectral envelope between male, female and children's voices are not equal, manipulations of playback speed tend to give unnatural results, however (Chiba & Kajiyama, 1958: 181). I will leave the discussion of details of relationships between other identity features, acoustic and physiological features and sound processing techniques until later chapters, where they might prove relevant for the discussion of a particular listening experience.

3.4.2 Voice familiarity

For familiar voices, a listener will have to mentally *match* the voice that he or she hears at the moment with the memory of one heard before, possibly attaching a name and a face to it, and any other associations that this particular voice might evoke. Such processes are often referred to as *recognition* or *identification*, where the former is often taken to specifically designate the matching to an existing mental trace or representation, and the latter to the verbal *naming* of the sound source and activation of other semantic structures (McAdams, 1993). Voice familiarity has been shown to affect the following:

- **Voice identification:** The rate of correct identifications is highest for highly familiar voices (Schmidt-Nielsen & Stern, 1985; Yarmey et al., 2001).
- **Multimodal associations:** Familiarity with a person can affect the range of associations made and the degree in which other modalities are evoked (von Kriegstein et al., 2005). For example, the voice of a person that we have an intimate

> relationship with can evoke visual, tactile and olfactory associations within us, whereas it will be more difficult to evoke such associations from unfamiliar voices.[107]

For instance, after having met the British composer and writer Katharine Norman personally at a conference after hearing her presentation, it became possible for me to recognize the voice in the piece *Losing It (Insomnia remix)* as belonging to the composer herself, something which made me experience her voice as much richer in its associations – I could to some degree imagine her speaking.

My familiarity of Katharine Norman is naturally of a different kind than for people I interact with on a daily basis. Some people are clearly more familiar than others, and familiarity can therefore imply more than being simply a question of either-or, which it usually is in voice recognition studies involving familiarity as a parameter. Rather, I will suggest that we can speak of different degrees or modes of familiarity based on several parameters:

1) **Frequency:** How often and for how long we encounter somebody?
2) **Range of expression:** Do we know somebody's entire range of expressions (calm, agitated, polite, argumentative, laughing, crying, singing, imitating, etc.), or do we only know them through one single type of expression?
3) **Memorability:** To what degree are persons memorable to us, e.g. by being distinct from others, having an emotional impact on us, being present at a memorable occasion, etc.?
4) **Distance:** What distance we usually interact with somebody; intimate, personal, social; do we interact with somebody through mediated experience or face-to-face?
5) **Multimodality:** Do we know somebody by their voices only, are we familiar with somebody's visual attributes, and if so, from still or moving images, or face-to-face interaction, do we know how somebody smells, how it feels to touch them, etc.?

Even if I am personally acquainted with several composers and other persons whose voice has appeared in acousmatic music, it is only for a limited number of pieces that I have personally met the people whose voices are in the music. I presume that this will be the case for the majority of listeners. I would guess that typically, listeners encounter voices that are

[107] In the research literature different kinds of *cross-modal associations* have been studied that demonstrate how stimuli within one modality might evoke imagery or associations within other modalities (Gilbert et al., 1996; Amedi et al., 2002; von Kriegstein & Giraud, 2006; von Kriegstein et al., 2006). In extreme cases this can result in *synaesthesia*, i.e. an involuntary perceptual experience of a cross-modal association (Popova, 2005).

either a) not previously heard in any other context than one particular work in question, b) previously heard in other works, usually by one and the same composer,[108] or c) belonging to somebody that is well-known from mediated expressions like television, Internet and radio.[109] Compared to voices that we are familiar with from daily social interaction, this will usually imply a lower degree of familiarity, especially concerning points 2), 4) and 5) above.

3.4.2.1 Cues in recognition/identification

Laver and Trudgill point out that there are *physiological* factors as well as *social* and *psychological* factors that are important in providing cues for voice identification and recognition in speech (Laver & Trudgill, 1979):

- **Physiological features** include most of the features discussed in section 3.3: The length and shape of the vocal tract, dimensions of lips, tongue, nasal cavity, pharynx and jaw, dental characteristics and geometry of laryngeal structures, dimensions and mass of the vocal folds and respiratory volume (Laver & Trudgill, 1979: 7).[110] These anatomical features will impose invariant limitations for each speaker's possible ranges of fundamental frequency, amplitude, and spectral configurations.[111]
- **Habitual features** are acquired through social interaction or as a result of habits, including what is often referred to as *voice settings*, which can include habitual adjustments of the vocal tract and larynx, as well as habitual choices of a comfortable range within the physiological limits (*ibid.*: 14). These features are both learnable and imitable and often function as social and cultural markers. Habitual features can also be a result of psychological disposition or personality (cf. **table 3.2** below).
- **Affective features:** If articulatory features that signal affective states are a part of a speaker's habit, they can be interpreted as markers of personality (*ibid.*: 17). In English the habit of using harsh phonation might for example indicate an aggressive personality.

[108] Examples are the voice of Hannah MacKay in several of Paul Lansky's works and the voice of Cathy Berberian in the works of Luciano Berio.

[109] E.g. the voice of Ronald Reagan in *OUT* by Alain Thibault (1985, on Thibault, 1990) or Ernesto "Che" Guevara on Sten Hanson's *Che* (1968, on Various_artists, 2006c).

[110] Several studies using whispering have confirmed the intuitive idea that voices can be recognized solely based on the resonant qualities of the vocal tract (Pollack et al., 1954; Yarmey et al., 2001).

[111] This is not to say, however, that training cannot affect any of these parameters for all kinds of vocalizations, but such training can only affect these parameters up to a point limited by physiological features. Studies with singers have shown that both fundamental frequency range and sound pressure level were significantly extended during singing training over four semesters (Mendes et al., 2003).

- **Linguistic features:** Linguistic units have been found to affect voice identity judgments (Remez et al., 1997), and the success of voice imitators in impersonating a particular speaker have been found to rely in part on verbal semantic issues (Zetterholm et al., 2002). The articulatory realization of linguistic units is the basis for speakers' *accents*, which can serve as a marker both for regional belongingness and often social class (Laver & Trudgill, 1979: 17). Grammatical and lexical features (see section 3.6.1.3) can also function as social markers (*ibid.*: 22-26).
- **Acoustical features:** All the above mentioned features can be related to acoustic parameters, such as:
    - Mean fundamental frequency (F0) [112]
    - F0 contour [32]
    - Spectral envelope / formants[113]

As one might notice from the points above, some cues are quite similar to those of the **VG-domain** discussed in section 3.3. Moreover, there are also cues here that are associated with the **AF-** and the **LI-domain**. Thus, the cues of the **ID-domain** have a basis in other domains, but here it is the quasi-invariant features that are of interest rather than the dynamic and constantly changing ones. Thus, the variations within a certain time-span can give rise to an estimation of invariant properties that can be used in identification. Even if some identity features might be recognized only from short snippets, longer stimuli with more cues present (for instance whole sentences instead of single phonemes or syllables) will therefore generally increase the rate of identification (see e.g. Bricker & Pruzansky, 1966; Pollack et al., 1954; Compton, 1963; Schweinberger et al., 1997).

As for the relationship between the different types of cues, it seems less likely that the same cues were equally important in the recognition of all voices. The lack of conclusive results in many early studies on identification made Handel propose that identification is not tied to a single cue, but rather that "any single cue will afford some accuracy, and any single cue when combined with another cue will improve identification" (Handel, 1989: 222). Other studies have supported conclusions similar to those of Handel in suggesting that different acoustic cues are used for different voices, and that different manipulations of vocal parameters affected single voices differently (Van Lancker et al., 1985a; Van Lancker et al.,

[112] See van Dommelen, 1990.

[113] Lavner and colleagues concluded that when they substituted vocal tract characteristics of a speaker with artificially derived characteristics, recognition dropped dramatically, and compared to manipulations of glottal source characteristics (F0 or glottal wave form), the effects were much more pronounced (Lavner et al., 2000).

1985b; Lavner et al., 2000; Kuwabara & Sagisaka, 1995; Schmidt-Nielsen & Stern, 1985).

Handel also attributes much of the inconsistencies to the sets of voices used in each experiment, i.e. the *context* against which each voice is compared (Handel, 1989: 255). In comparing the voices in one set the cues that are salient for identification might be different from the ones that will be salient in another set of voices. The cues will therefore vary according to the salient *differences* between the voices in the set. A very low pitched voice will for instance stand out among high pitched voices, and a female voice will stand out among male voices, and so on.

If the conclusions of these studies are adequate, it would be difficult to make predictions about the influence of one single parameter on recognition and identification, because it would depend on the characteristics of the voice in question. This will also prevent me from saying something general about how different kinds of electronic manipulation will affect the recognition of voices unless the salient cues for one particular speaker could be identified beforehand. For instance, if the intonation pattern of the speaker appears very characteristic, a manipulation that would replace the intonation curve with a stable pitch will probably reduce recognition dramatically. But even in such cases, the influence of other cues would be difficult to predict, because it is not evident what set of cues for a particular speaker that will be salient for one listener in one particular context. However, this does not mean that the indications of the research in this domain is not useful in this study, on the contrary – it can give us some clue of the probability with which a certain manipulation might affect recognition, and about what cues might be important when we recognize a voice as being the same or different. Therefore, I have included a table (**table 3.2)** in which the conclusions and results of several studies are provided in a compact form, both with regard to different vocal production features or identity features (upper part of table) and to different electronic manipulations (lower part of table). Many of the studies reviewed have applied different kinds of manipulations in order to study the influence of different parameters, and several of these manipulations are parallel to many of those that can be heard in electroacoustic music. Indeed, reverse playback, filtering, rate alterations with or without pitch or spectral shifting are a part of the "standard" repertoire of manipulation techniques for electroacoustic composers. And, as we see from the table, most of these kinds of manipulations will indeed affect identification in a negative way. I have also included some factors that are related to vocal production, such as changes in voice quality or register, since these have been shown to have a significant negative effect on recognition.

Vocal production / other factors	
· **Pitch change (register) in singing**	· Great differences in pitch (an octave or more) between subsequent tones lead to a decrease in same-different judgments regarding vocal source . [1]
· **Vocal disguise/changes in voice quality**	· Different vocal disguises reduce the ability to discriminate voices to different degrees depending on the disguise. [2] · Whispering reduces identification. [3] · Use of falsetto reduces identification to below chance. [4]
· **Regional dialect**	· Change in dialect/accent made listeners fail to identify a speaker. [5]
· **Imitation of other speakers**	· Can be effective in disguising speaker identity. [6] · Can cause listeners to mistake different speakers for each other. [6]
· **Ideolect***	· Individual voices can be recognized solely on the basis of time-varying phonemic features, i.e. from sine-wave speech stripped of features like f0 and voice quality. [8]
· **Gender** · **Age**	· Are highly salient features for voice recognition. [20]
Electronic manipulations	
· **Filtering**	· Low-pass filtering reduces identification depending on the range of frequencies allowed through the filter. [9] · High-pass filtering reduces identification depending on the range of frequencies allowed through the filter. [10]
· **Rate alterations, no pitch shift**	· Time compression generally reduces identification, but different voices are affected differently. [11] · Time expansion generally reduces identification, but different voices are affected differently. [12]
· **Reverse playback**	· Generally reduces identification for familiar voices [13], but different voices are affected to different degrees. [14]
· **Spectral envelope shifts**	· Reduces identification. Upward and downward shifts reduced identification to about half. [15] A few % change can affect voice individuality radically. [16] · Identification was reduced when formants were individually shifted, but more for F3 and F4 than F1 and F2. [15]
· **F0 shifts**	· F0 shifts generally reduced identification, but more for increased F0 than decreased, and with great variability for different voices. [15] · Up to 50% change without affecting voice individuality. [16] · In speaker recognition tasks, listeners were highly sensitive to shifts in F0 for low- or high-pitched voices, but not for voices with intermediate pitch. [17]
· **Sine wave analogues**	· Speakers can to some degree be recognized from sine-wave replicas of naturally produced sentences. [18]
· **LPC rendering**	· Reduces recognition rate. [19]

Table 3.2 Findings regarding factors affecting voice identification and recognition not linked to a particular identity feature. [1]: Erickson et al., 2001, Handel & Erickson, 2001. [2]: Reich & Duke, 1979, [3]: Pollack et al., 1954; Yarmey et al., 2001, [4]: Wagner & Köster, 1999, [5]: Eriksson, 2007. [6] Sullivan & Schlichting, 1998, [7]: Sullivan & Schlichting, 1998, [8]: Fellowes et al., 1997; Remez et al., 1997, [9]: Pollack et al., 1954 Compton, 1963, [10]: Pollack et al., 1954; Compton, 1963, [11]: Van Lancker et al., 1985b, [12] : Van Lancker et al., 1985b, [13]: Van Lancker et al., 1985a; Bricker & Pruzansky, 1966; Sheffert et al., 2002, [14]: Van Lancker et al., 1985a, [15]: Lavner et al., 2000, [16]: Kuwabara, 1996, [17]: van Dommelen, 1990, [18]: Remez et al., 1997, [19]: Schmidt-Nielsen & Stern, 1985, [20]: Eriksson, 2007. * Ideolect is a language variety that is unique for one individual.

Since this table mainly presents research on identification and recognition from spoken voice, it might only hint at relationships affecting the recognition and identification of voices in electroacoustic music, which comprise all types of vocal expressions. An identification study using sung voice provides results that in my opinion suggest a very straightforward general principle as important in recognition and identification from voice, namely that the greater similarity between sounds that are to be identified or recognized and sounds one have experienced earlier, the greater is the chance that one will identify or recognize the sound. In this study, the listener subjects were asked which of three notes were sung by a different person (Erickson & Perry, 2003). The researchers found that the note farthest in pitch tended to be judged as the different voice, suggesting that similarity in salient features such as pitch is important. If the researchers included as much as six notes comprising a wider range of pitches, the task became easier, indicating that it was important to hear how a note sounded like in a particular pitch range to make a correct identification. On this background, I will hypothesize that if one has only heard the voice of somebody singing before, the chances that this voice will be recognized and identified will be greatest when we hear that person singing rather than speaking. And, if we have only heard a filtered version of a voice earlier, the chances for recognition and identification are greatest when we hear a filtered version once more.

3.4.3 A note on imitation and ontological levels

In traditional vocal forms of expression, like song or storytelling, the identities of the voices can function merely as a neutral contextual frame which is projected only as a neutral "instrument" for performing a musical or a narrative content. However, identities can also be played out more actively through *imitations* as when one single voice creates different vocal personas during one and the same performance. For instance, the famous Schubert Lied *Erlkönig* invites the singer to take on several different vocal personas during the performance, the narrator, the father, the son and the Erlking, who all have their "lines" in the lyrics (Cone, 1974: chp.1). Thus, one will in such a situation thereby have vocal personas on several ontological levels in accordance with the earlier discussion: on the level of the singer and on the level of the characters.[114]

For the voices in electroacoustic music, one might identify many of the same things – one might have "neutral" voices whose identity and function remains constant throughout a piece, and one might find imitations which create vocal personas on different ontological levels. However, in addition to vocally induced imitations, one will also frequently experience that electronic manipulations can contribute to projecting different types of vocal personas.

[114] One can also claim to see that the narrator, with his/her distancing to the events happening to the characters, constitutes a separate ontological level.

These identities might even change gradually from one to the other in a kind of transforming metamorphosis, or they might be superimposed on top of each other (see Landy, 1993; Smalley, 1993; Wishart, 1996: 155-159). Hence, the identities experienced in electroacoustic music might appear to have an even greater *plasticity* than more traditional vocal expressions.

3.5 The affective domain (AF)

The affective experiential domain embraces a range of phenomena including attitudes, interpersonal stances, moods, emotions and preferences which can be inferred from vocal sound. All these phenomena may in principle be encountered in acousmatic electroacoustic music, and I will present a set of parameters that may distinguish them from one another below. However, I have chosen mainly to focus on emotions, simply because they are the most salient expression of affective states, having both high intensity and short duration, and also because they change rapidly. Emotions therefore seem easier to detect and describe than more diffuse affective states like moods. Emotions also exert a major influence in directing our attention and assigning cognitive resources to a sound. A vocal outburst with intense emotional content will probably catch our attention more easily than an emotionally neutral utterance. Research in the cognitive sciences has suggested that emotionally charged information will be prioritized and will receive privileged access to attention and awareness, especially for negative and threat-related emotions (Vuilleumier, 2005).[115] However, I will not address questions regarding the listener's emotional response to music. I am primarily concerned with the emotional aspects that the listener will experience as belonging to or conveyed by the vocal persona – even if such aspects can in many cases be mirrored by the listener's responses.

Traditionally, emotions have been regarded as an important component in music practice and experience, something which has been reflected in musical thought of earlier centuries (e.g. the *Affektenlehre* ascribing musical properties to affects). Furthermore, a number of writings have regarded emotions as central for musicological study (see Cook & Dibben, 2001 for an overview), and a growing body of empirical research with a number of different approaches attests to the relevance of emotions for music (Juslin & Sloboda, 2001) .

[115] Emotional vocalizations has been also been found to increase activation in the voice-sensitive regions in the brain compared to neutral vocalizations (Grandjean et al., 2005).

From my review of the literature, however, it appears that relatively little interest has been taken in the role of emotions in electroacoustic musical theory and practice.[116] Whether this is a reflection of a lack of interest on the part of theoreticians or composers is hard to say. Nevertheless, vocalizations that appear strongly charged with emotions do sometimes appear in electroacoustic music, for example in works such as Takemitsu's *Vocalism AI* (1956, on Takemitsu, 2004), Berio's *Visage* (1961, on Berio & Maderna, 2006) and Wishart's *Red Bird* (1977, on Wishart, 1992). Such examples are in themselves sufficient reason to include the affective domain in this context. Yet, it might appear paradoxical that emotions can be experienced from voices that are subjected to electronic manipulations, modifying, reducing or extending voices with the use of "cold" technology. I will try to show in this section that the technology and manipulations can alter the cues that the decoding of emotions depend on, but not necessarily so as to completely neutralize it; it may increase the chances of confusions or decrease the chances of recognizing the emotive content altogether, but it may also transform or even emphasize emotional content. In my argumentation I will draw heavily on empirical studies on voice and emotion, due to the scarcity of theory related to electroacoustic music and the fact that a great number of the manipulation techniques that are applied in studies on voice and emotions are similar to many of those found in electroacoustic works with voice.

3.5.1 Categories of affective states

According to Scherer and Zentner, phenomena such as moods, attitudes, emotions, personality traits, preferences and interpersonal stances are all regarded as different types of affective states which are distinguished by a set of parameters, or "design features" as they are called (Scherer & Zentner, 2001: 362-363). These design features might also constitute a means for describing affective states in electroacoustic music. The ones that seem most relevant for this study are:

- **intensity**
- **duration**

[116] Causton's discussion of Berio's *Visage* is one exception (Causton, 1995). Wishart also touches upon the topic of emotion in his discussion of human utterance (Wishart, 1996: chp.11). Delalande's analysis of empathic listening behaviour appears to approach aspects related to emotions in its focus on listeners' conceptualization of bodily sensations during listening, but emotive aspects are not a part of his analysis (Delalande, 1998).

- **rapidity of change**[117]

Scherer and Zentner then classify the different affective states according to these design features:[118]

- **Emotions** (e.g. angry, sad, joyful, fearful, proud, elated, ashamed, desperate): High intensity, short duration, very high rapidity of change
- **Preferences** (e.g. like, dislike): Low intensity, medium duration, very low rapidity of change
- **Mood** (e.g. cheerful, gloomy, irritable, depressed, buoyant): Medium intensity, long duration, high rapidity of change
- **Interpersonal stances** (e.g. distant, cold, warm, supportive, contemptuous): Medium intensity, medium duration, very high rapidity of change
- **Attitudes** (e.g. liking, loving, hating, valuing, desiring): Medium intensity, long duration, low rapidity of change.

The examples given for each of these types also serve as guidelines for possible descriptors of vocal utterances. The number of possible categories is naturally huge, and I therefore don't want to list all possible affective qualifiers that belong to each type. What can be useful, however, is to classify affective states according to a limited set of categories or dimensions. The so-called dimensional view of emotions might be subject to some controversy, but it nevertheless appears to provide a useful framework for classification, and can thereby also be useful qualifiers for the **AF-domain**. These dimensions are (Scherer, 2003, Sloboda & Juslin, 2001 and Pereira, 2000):[119]

- **Arousal** – indicates the degree of activity. E.g. are "angry" and "delighted" taken to be higher in arousal than "sad" and "serene".

[117] The other design features used by Scherer and Zentner are synchronization, event focus, appraisal elicitation and behavioral impact (Scherer & Zentner, 2001: 363).

[118] I have considered that since *personality traits* deal with quasi-invariant features, these are considered to be a part of the **ID-domain**, even if it clearly borders onto the **AF-domain**.

[119] The first two of these dimensions are most commonly included in dimensional models according to Sloboda & Juslin, 2001.

- **Valence** – indicates whether an affective state is felt as positive or negative. “Sad” and “distressed” are considered to be negative, whereas “excited” and “satisfied” are positive.
- **Power** - indicates potency or control and distinguishes emotions initiated by the subject from those elicited by the environment. E.g. “contempt” and “hot anger” are taken to be high in power, whereas “fear” and “sadness” are low.

3.5.2 Recognition of emotions from voice

In numerous studies, the human abilities for recognizing emotions from the voice has generally been found to be quite good – only slightly lower than from the face – even if certain emotions are more difficult to recognize vocally than others.[120] And since these studies generally have used voices presented without visual stimuli through loudspeakers, the conditions are not fundamentally different from what is the case in electroacoustic music – both clearly deal with acousmatic, loudspeaker-mediated listening. Studies have indicated that recognition of emotions is partly dependent on cues that are cross-cultural and partly on culture/language specific cues.[121] A possible explanation for these results is that the involuntary physiological changes that underlie emotional arousal that markedly affect vocal production, so-called *push effects*, are universally recognized in the same way as facial expressions, like the smile, are recognized across cultures. The rules regulating emotional display that are linked to more voluntary processes, so-called *pull effects*, might be more cultural-specific, being bound up in cultural conventions such as display rules, social relationships and situational context.[122]

[120] The recognition of emotion from standardized voice samples, using actor portrayals, has attained an average of about 60% over more than 30 studies according to Scherer, 2003, which is about five to six times higher than chance level. However, the results for disgust and shame showed a low degree of recognition in Banse & Scherer, 1996, 15% and 22%, respectively. As for singing, eleven professional singers sang phrases from songs, and their task was to express different moods: joy, sadness, neutral anger and fear. In a listening test around 80% of the emotions were correctly identified, with the exception of joy, which were identified by 56% (Kotlyar & Morosov, 1976). According to Scherer, facial expression studies generally report an average accuracy percentage of around 75% (Scherer, 2003).

[121] Scherer and colleagues found that listeners across cultures and languages could recognize emotions from content-free speech with relatively high accuracy, indicating that universal rules play a part in inferring emotions from voice (Scherer et al., 2001). Still, in the same study there was a tendency that the linguistic distance or difference between speakers and listeners was reflected in the results, so that speaker of languages with Germanic origin (German, Dutch and English – the actors portraying the vocal emotions were German radio actors) had higher accuracy than listeners native in Romance languages (Italian, French, Spanish), who in their turn had higher accuracy than Indonesian listeners.

[122] See Kappas et al., 1991 for an explication of the push and pull effects on vocal emotion.

Acoustic parameters	Fear	Joy/elation/happiness	Rage / hot anger	Sadness	Tenderness
F0 mean	+ [1], ++ [2], + [3]	+ [1], + [2], + [3]	+ [1], ++ [2], + [3]	- / 0 [1], (-) [2], - [3]	- [3]
F0 range	+ [1], + [2]	+ [1], + [2]	+ [1], + [2]	- [1], (-) [2]	
F0 variability	+ [1], - [3]	+ [1], + [3]	+ [1], + [3]	- [3]	- [3]
F0 changes / contour	Rising [3]	Smooth, upward inflections [2], rising [3]	Abrupt on stressed syllables [2], rising [3]	Downward inflections [2], falling [3]	falling [3]
Jitter	+ [1]	+ [3]*	+ ? [3]*	- [3]*	
Tempo/rate	+ [1], + [2], + [3]	+ [1], + / - [2], + [3]	+ [1], (+) [2], + [3]	- [1], (-) [2], - [3]	- [3]
Mean intensity	0 [2], - (except in panic) [3]	+ [1], + [2], 0 [3]	+ [1], + [2], + [3]	- [1], - [2], -[3]	- [3]
Intensity variability	+ [3]		+ [3]	- [3]	- [3]
High freq.energy	+ [1], - [3]	+ [1], 0 / + [3]	+ [3]	- [3]	- [3]
Proportion of pauses	- [3]	- [3]	- [3]	+ [3]	+ [3]
Presicion of articulation	+ [2]	0 [2], + [3]	Tense [2], + [3]	- [1], - [2], - [3]	- [3]
First formant freq/bandwidth	- / wide [3]	+ / narrow [3]	+ / narrow [3]	- / wide [3]	
Voice quality	Tense [1], Irregular voicing [2],	Tense [1], Breathy, blaring [2],	Breathy, chest tone [2], Tense [3]	Lax [1], Resonant [2], Lax [3]	
Microstructure	Irregular [3]	Regular [3]	Irregular [3]	Irregular [3]	Regular [3]
Voice onset / tone attack		Fast [3]	Fast [3]	Slow [3]	Slow [3]
Staccato/Legato	S [3]	S [3]	S [3]	L [3]	L [3]
Vibrato magnitude/rate	Small/fast [3]	Fast [3]	Large/fast [3]	Small/slow [3]	Small/slow [3]
Singer's formant †		+	+	-	-

Table 3.3: The table shows reported results from three reviews of research on emotions in voice and music. All indications are relative to an emotionally neutral voice. 0 = unchanged / medium value + = increase/higher value, - = decrease / lower value, + / - = increase or decrease. Differentiations in degree: (-) = slight decrease, - = decrease, -- = strong decrease, ? = less certain. [1]: Review of 39 studies by Scherer (1986), [2]: Review of research literature on vocal emotions by Murray & Arnott, 1993, [3]: Review of 104 studies of vocal expression and 41 studies of music performance by Juslin & Laukka, 2003. The parameters of this study are common to voice and music except *jitter, proportion of pauses, precision of articulation*, and *first formant*, which are specific to vocal studies, and *staccato/legato* and *vibrato* which are specific to music. *The authors consider the results on jitter to be only preliminary.

To get a gross overview of the acoustic and voice production-related characteristics of different emotions, I have included a table (**table 3.3)** which summarizes results from three review articles (Murray & Arnott, 1993; Scherer, 1986; Juslin & Laukka, 2003). Since there are considerable differences in what emotions are studied in each case, I have included only the five most studied emotions in the table. The similarity in values between the data from the different reviews indicates quite strongly that cues for emotion are rather consistent across individual variations, something which also confirms the cross-cultural similarities mentioned above. Therefore, it seems likely that the chances of identifying emotions from vocal utterances in electroacoustic music independently of cultural background are relatively good, at least in principle. In practice, however, while the vocal material in empirical studies usually are produced with the sole intention of using it in the study, the voices in electroacoustic music are produced with an artistic intent which might not aim for the utterance of emotions that can be unequivocally recognized.[123] Moreover, stylistic conventions that vocalizations may rely on may forefront the culturally dependent facets of emotion while reducing the universally recognized ones. Lastly, as I will look more into in section 3.5.4, electronic and computer processing might make recognition more difficult by removing or distorting salient cues, but it may also impose vocalizations with an emotional content which was not there prior to processing.

3.5.3 Emotions and abstract qualities

There seem to be parallels in how emotions are inferred from the human voice and in non-vocal music. Based on the main tendencies in the reviewed studies in **table 3.3**, Juslin and Laukka concluded that emotion-specific patterns of acoustic cues can be used to communicate discrete emotions in both vocal as well as musical expression of emotion (Juslin & Laukka, 2003). Not only did they find strong indications of similar patterns of decoding accuracy, but they could also see relatively clear tendencies of similarity between emotion-specific patterns of acoustic cues in music performance and vocal expression.[124] This was interpreted by the

[123] Wishart is one of a few exceptions that I know of. In making *Red Bird* he directed the vocalists to apply different types of affective coloring of a number of phonemes, syllables, words and phrases “in order to project such implications into the phonemic objects themselves” (Wishart, 1996: 298)

[124] However, the authors recognize that their study include only one of several kinds of emotional signification, namely that which is motivated by a structural similarity between music and vocal features. Thus, emotions elicited by expectancy or arbitrary association are not considered in their study. Another fact that inhibits the generalization of the conclusions is that the range of musical expressions is restricted to Western genres such as classical music, jazz, European folk music and opera.

authors as indications of an *iconic* and universal component of musical emotion, grounded in the commonalities of human biology, as opposed to emotional components rooted in conventions, such as when harmonic modes (major-minor) are used to indicate "happy" or "sad". If there is something to this claim, it implies that more abstract sound qualities and structures can carry emotional content in the same manner as vocally conveyed emotion, regardless of sound source identity. Thus, if a voice is heavily manipulated so as to attain characteristics foreign to the voice, it still can carry emotional meaning through e.g. its pitch contours, its intensities and its rhythmic configurations. In other words, a sound can be on the verge of being heard as non-vocal while still conveying emotions.

What is relatively evident from **table 3.3** is that several of the emotions seem to have very similar sets of values for the different acoustic properties. For instance, one can see that very little distinguishes the values for tenderness and sadness, and the same is the case with the values for fear, joy, and anger. And, if the cues have a high degree of similarity, they will probably not provide a basis for distinguishing between them. It seems therefore that this branch of research so far has not been able to account for the relatively high recognition rates for vocal emotions. Scherer has argued that similarities between the mentioned groups of emotions are caused by similar levels of *arousal*, because high- or low- arousal emotions can be characterized by either high or low values for F0 mean, range, variability, intensity and tempo/rate, and several studies have indicated that the acoustic cues related to arousal have the largest effects in signaling emotion (Scherer, 1986: 144; Scherer, 1995).[125] The results from experiments have suggested that these acoustic parameters vary in a continuous rather than categorical manner causing gradual changes in the arousal level, and even if there are some cases of interaction between verbal structures and specific (categorical) intonation patterns, the effects of these are relatively small (Ladd et al., 1985; Scherer et al., 1984; Banziger & Scherer, 2005).

3.5.4 Emotions from cue manipulation or synthesis

Cue manipulation or synthesis is an approach that has been frequently taken to find the appropriate cues that characterize the different emotions. In several studies of vocal emotion researchers have used manipulations that parallel those applied by electroacoustic composers,

[125] Scherer also comments that many studies suffer from the fact that one single emotion can differ in the degree to which it is subdued by the speaker, so that it has to be specified whether emotional labels such as "joy" refers to the more aroused "elation" or the more subdued "happiness", or whether "anger" refers to the "hot" (aroused) or "cold" (subdued) version (Scherer, 1995).

evaluating how either the deletion or modification of acoustical cues from the voice, or synthesizing a voice from scratch, can affect the attribution of different emotions. It can therefore be worth taking a look at research applying such a methodology. For instance, Mozziconacci's doctoral dissertation applied resynthesized versions of emotionally neutral sentences, i.e. sentences which were spoken with no particular emotion, which were manipulated by changing the pitch level, pitch range, intonation pattern, intonation declination, and speech rate using the Time-Domain Pitch-Synchronous OverLap-and-Add algorithm (TD-PSOLA) (Mozziconacci, 1998). This technique allows for a manipulation of duration and fundamental frequency while retaining a high degree of naturalness, at least if the deviations from the original F0 are not too great. By letting listeners evaluate different manipulations systematically, she was able to obtain a set of optimal values for six different emotions, all relative to an emotionally neutral sentence. These values are given in **table 3.4** below. In addition to the absolute values given by Mozziconacci, I have rendered two of the values as ratios, so that a pitch value of 2.00 will imply that the pitch is twice as high in frequency, i.e. one octave higher, compared to that of the neutral sentence.[126]

	F0 Range		**Speech rate**		**Mean F0**	
	Rel.	**Abs.**	**Rel.**	**Abs.**	**Rel.**	**Abs.**
		st.				Hz
neutral	1.00	5	1.00	-	1.00	94.0
joy	2.00	10	1.20	-	2.62	246.2
boredom	0.80	4	0.67	-	1.00	93.8
anger	2.00	10	1.27	-	1.97	185.6
sadness	1.40	7	0.78	-	1.71	160.3
fear	1.60	8	1.12	-	2.95	277.0
indignation	2.00	10	0.85	-	2.97	279.0

Table 3.4: Ratios corresponding to the optimal parameter values for six emotions in resynthesized sentences found by Mozziconacci (1998). In addition to the absolute (Abs.) values given by Mozziconacci for F0 range (in semitones, st.) and mean F0 (in Hz), I have added relative (Rel.) values for these parameters in the columns with grey color. No absolute values were given for speech rate by Mozziconacci.

What this can indicate in this context is that manipulations of intonation parameters of a spoken sentence can indeed make the listener experience emotional aspects of the voice that was never intended by the speaker. Thus, rather than imposing technologically "coldness" onto the voice, manipulations here make the voice more emotionally intense. This opens up

[126] The different intonation patterns used in her study are not included in the table, since it is based on a system derived from Dutch spoken language.

for a discussion on emotional implications of voice transformation also in electroacoustic music where comparable source-filter techniques are applied, such as in Lansky's *Six Fantasies*, that will be subjected to study in the chapter twelve.

It is also evident from several studies that many forms of manipulations mainly remove cues to emotion recognition, so that they are more poorly recognized or more often confused with other emotions. For instance, low-pass filtering will remove most of the spectral cues as well as attenuate perceived loudness, reverse playback or random splicing will destroy temporal cues like f0 contour and rate, and destroy any emotive content that is a result of an interplay with the linguistic domain. For example, Scherer and colleagues low-pass filtered sentences that had been found to have emotional content, and when comparing recognition rates, they found that the manipulated and unmanipulated sentences were only correlated for one out of nine emotions (Scherer et al., 1984). Similarly, Friend and Farrar could report that a speech-content masking method that consisted in having natural sentences mimicked with non-sense syllables, called *reiterant speech*, scored higher on listeners' ranking of emotional content (angry, excited, happy) than random splicing and low-pass filtering (Friend & Farrar, 1994).

There are also a number of studies that have investigated the degree to which synthetic voices can carry affective content. While the results are variable, it seems that many methods are capable of producing synthetic voice with recognizable emotions, even if the recognition rates are usually lower than for natural voice (Schröder, 2001; Burkhardt & Sendlmeier, 2000; Gobl & Ní Chasaide, 2003; Cahn, 1990).

All in all, these studies show that affective content can be recognized from synthesized or manipulated vocal sound, often applying techniques that are similar to ones that are used in composition of electroacoustic music. Some techniques, however, might affect cues so that this becomes increasingly difficult and the chances of confusion increase, whereas other techniques seemed able to "apply" emotions synthetically, where there was none before. To identify and describe affective content in electroacoustic music using the categories discussed in section 3.5.1 should therefore not be an impossible task, even if the chances for ambiguities and vagueness are still present.

3.6 The linguistic domain (LI)

Verbal elements constitute an important aspect of many electroacoustic works. Many of the works that have voice in them also use language, thereby generally encouraging the listener to include comprehension and interpretation of the verbal material into the listening process. Composers might support this indirectly by also including written versions of the verbal elements in CD booklets and concert programmes.[127] For some genres, such as text-sound composition, the reference to verbal material is even included in the name of the genre and the composition of text and sonic structures are seen as integrated (Hultberg, 1994).

The linguistic domain includes everything in an experience that is related to language in the listening experience, including verbal elements and structures as well as the meanings conveyed by these elements and structures. Even if the field of language perception and comprehension is very complex and the amount of research literature is huge, I will single out a few terms and levels of description that I consider important for this experiential domain. I will organize the discussion according to a distinction between 1) higher-level knowledge or learning, and 2) lower-level perceptual processes that are more directly dependent on the nature and condition of the verbal material. These two approaches are often seen as having different directions; while the first goes *top-down*, from higher to lower level processes, the other goes *bottom-up* (see e.g. Davis & Johnsrude, 2007a). I will start by looking at top-down influence on processing and the way it involves the listener's competence and knowledge.

3.6.1 Top-down: Listener competence and knowledge

To successfully process and comprehend verbal structures there has to be a degree of concordance between listener's linguistic and contextual competence and the cues available to a listener. This requires a *shared* linguistic code between sender and receiver, as well as shared knowledge of how the linguistic code can be applied in different situations to attain different goals relating to different frames of reference (see e.g. Fiske, 1990). When engaging in a conversation with somebody with the same social and regional background and affiliation as yourself, e.g. one's next door neighbour or brother, you will usually be able to understand almost all of what he or she is saying, because the codes of communication will be shared. You speak the same language, the same dialect, probably shared "slang" words, have the same sense of formality and informality, and so on. This thereby enables exchange of speech acts that are meaningfully related to a *shared context*, that is, the time, place and situation you are both in at the time. Moreover, in such a situation one can also support and extend the verbal channel of communication by using non-verbal means such as gestures, body posture, facial expressions, pointing, mimicking and so forth. Lastly, face-to-face communication is

[127] See e.g. Vande Gorne, 1998; Lejeune, 2000; and Dhomont, 1996.

basically two-way; it is open for interaction, for possibilities of clarifying what is not completely understood, for perceiving each other's response to every word and phrase.

In acousmatic music, when "faced" with the vocal persona, however, the listener will be deprived of all visual cues, the possibilities for interaction, as well as the immediate, direct context of shared place and time compared to daily face-to-face communication. One will instead be left with verbal material that has to be interpreted from listening only, on the basis of wider or more general frames of reference, and with codes that are generally less specific, that is, it cannot refer to the precise time, location and situation which normally are shared between interlocutors.[128] When engaging in interpretation of verbal material in acousmatic music, this will usually imply that the degree of ambiguity is higher.

The interpretation of verbal material in electroacoustic music will no less than in daily conversations draw on the listener's earlier experiences and knowledge, both related specifically to any verbal codes involved, and to general knowledge of the world. For example, when listening to Sten Hanson's text-sound piece *Che*, it is my knowledge of one of the central leaders of the Cuban revolution, Ernesto "Che" Guevara, that enables me to link the vocal persona speaking in Spanish throughout the piece to this historical person (1968, on Various artists, 2006c). Moreover, it is my poor knowledge of Spanish that hinders me from understanding everything that he says, but that still allows me to recognize a few words like "repression", "sobrevivir" (survive), "fáciles" (easy), "dificiles" (difficult), and "guerrilla" and to interpret what is said as a political speech.

When earlier experience, learning or memory are involved in the perceptual and cognitive processing of sound, such as in this case, one often speaks of a *top-down* influence or processing, as opposed to the automatic and pre-attentive *bottom-up* influence exerted by the basic auditory processing of the incoming stimuli (Davis & Johnsrude, 2007a). The *top-down* influence on the perception of speech can be observed within several levels at work in speech perception; the *phonetic*, the *lexical*, the *grammatical*, the *syntactic* and the *semantic*. On all these levels, one's competence will influence what one perceives, in that it will impose constraints on the range of possible variations on what sounds can be identified and distinguished (phonetic), what possible words can be included in a sentence (lexical), what orders the words can be placed in (syntactic), what grammatical forms they will appear in

[128] In another perspective, one can claim that the codes of a piece of electroacoustic music, like every work of art, can be *more* specific than what is the case in face-to-face communication, in the sense that it establishes its own autonomous "code" or rules that is only valid within the work. For instance, if sound event *a* is repeatedly associated with sound event *b*, *a* can subsequently be taken to "stand for" *b*. In semiotic terms, there has been established an *arbitrary* association between two signs, that subsequently makes one become the signifier (*a*) and the other (*b*), the signified.

(grammatical), and the resultant meaning that the combination of words can evoke in a listener (semantic). By imposing restrictions on the possible interpretations on the different levels, top-down influence will generate a basis for making predictions about what is to come, in that certain units or combinations of units more are more likely to occur than others. As I see it, these levels can also serve as levels of description for verbal material in electroacoustic works, and I will therefore discuss them in the following with a special focus on how top-down processing influences each of these levels.

3.6.1.1 Phonetic level

At the *phonetic* level, I will use the established framework for describing *phonemes*, the minimal sounding units of speech, namely the International Phonetic Alphabet (IPA). I will assume that most readers are familiar with the basic signs of this system, and I will therefore not present the alphabet here. For readers not acquainted with IPA, I have included a nomenclature of the symbols used in this dissertation in Appendix A. The phonetic alphabet is categorized based on the vocal gestures used in sound production. Hence, place of articulation and type of articulatory gesture distinguish phonemes into categories like *bilabial* (with both lips), *labiodentals* (with lip and teeth) and *alveolar* (articulated at the alveolar ridge) as well as *stops*, *trills* and *fricatives*. This underlines the link between the phonetic level and the **VG-domain**.

Top-down influence can often affect how single speech sounds are interpreted and how the distinctions between the different speech sounds are made. One’s language background can in some cases determine how one can differentiate speech sounds, so that faced with foreign speech sounds one can risk that different phonemes can be grouped into one single phonetic category. For example, native speakers of Japanese tend to have problems distinguishing the English rhotic [r] from the glide [l] (Logan et al., 1991). A related phenomenon that has given rise to considerable debate and study is *categorical perception*. This phenomenon, which will be further discussed in the following chapter, is often taken as indicative of the mentioned top-down influence (see e.g. Harnad, 1987). Experiments with synthesized phonemes and syllables have shown that one single acoustic variable can vary gradually and continuously, but that somewhere along this gradual change a sudden shift will occur between one phoneme category and the next, demonstrating the boundaries of learned categories.[129] Such category boundaries can differ among languages. For instance, stop consonants are distinguished on the basis of differences in voice onset time (VOT) which differ among many languages (being either pre-voiced, having a short time-lag or a long time-lag) (Kessinger & Blumstein, 1997; Handel, 1989: 289). All in all, this shows that one has to see the phonetic categories in relation to one's own language competence.

3.6.1.2 Lexical level

The lexical level deals with *lexemes*, which can roughly be described as embracing all the different forms that a single word can take on. For example, both "throw", "threw" and "thrown" are all form of the lexeme THROW. I will not make a point of the linguistic issues related to the *lexeme* term, however, and I will therefore in most cases simply refer to the lexical units as *words*. In contrast to phonemes, which have no meaning in themselves, words represent a link to the semantic level by having through convention a defined range of significations to which it can potentially refer.

Listeners' *lexical* knowledge can influence perception and comprehension. For instance, even if phonemes are missing, degraded or ambiguous in a word, they can still be interpreted on the basis of the lexicon of a language. An example of such a "lexical" effect is the so-called "Ganong-effect", which refers to how an ambiguously sounding consonant will be perceived differently depending on which lexical context it is placed within (Ganong, 1980). A phoneme which is ambiguous between a [g] and a [k] will for instance be perceived as [g] when followed by –ift, and [k] when followed by –iss, because of the lexical influence of the words "gift" and "kiss" (and because there are no words like "giss" and "kift" in the English lexicon). A similar effect is reported by Norris and colleagues, who found that when listeners had been exposed to an ambiguous consonant in a context which favoured interpretation of one or the other consonant, the interpretation that was lexically biased by each group persisted also when the ambiguous consonants were presented in isolation (Norris et al., 2003).

[129] Examples of categorical perception include investigation of second formant transitions and its effect on the perception of stop consonants. In a classic study by Liberman and colleagues, gradual formant transitions caused abrupt shifts from perception of [b] to [d] and finally to a [g] (Liberman et al., 1957). Voice onset time (VOT) in word-initial stops is also a classical example of categorical perception. If the voicing (phonation) starts between 0-20ms after the stop burst, it is perceived as a voiced consonant [b], and if it starts between 40-60 ms, it is perceived as an unvoiced consonant [p] by English speakers. However, for Thai speakers there are three different categories of VOT in the same range (Abramson & Lisker, 1967; Abramson & Lisker, 1968).

Lexical competence can also aid the crucial process of *segmenting* the sound stream into words and sentences. Continuous speech often tends to include few pauses and other unambiguous acoustic cues for the listener as to where word boundaries need to be placed, especially in rapid conversational speech. Thus, speech perception needs some way to parse the continuous stream into a sequence of words, and one of the ways this is done is by reference to lexical constraints.[130] There are indications in recent research that such top-down cues tend to dominate, and that the use of bottom-up, acoustic, and statistical segmentation is confined to situations where the lexical cues are minimized, masked or distorted (Davis & Johnsrude, 2007a).

3.6.1.3 Syntactic and grammatical level

The syntactic level deals with the organization of words into larger units, typically sentences, whereas the grammatical level refers to how the words are specified and function relative to grammatical forms and categories, such as tense (which designate the time something occurs), person (which distinguishes between speaker, addressee or others) or number (singular or plural). Verbal aspects at the syntactic and grammatical levels are integrated in the structure of language, and will rarely attract any attention unless they are ambiguous or deviate from established rules or conventions.

Linguistic competence on the s*yntactic* and *grammatical* level is important in aiding perception and comprehension. This is because there are constraints on which grammatical classes that can follow each other in a language, and by knowing these constraints one narrows down which classes of words can follow each other. Pickering and Garrod have reviewed results that show how syntactical and grammatical rules impose predictabilities in a sequence, causing disruptions or anomaly effects if they are violated. If they are obeyed, though, it ensures speed and ease in processing (Pickering & Garrod, 2007). For instance, because an adjective is very likely to be followed by a noun (at least in English and Norwegian), a listener can predict with relatively high certainty that a noun will follow when an adjective is encountered.

[130] In the TRACE model of speech perception, a model of word segmentation is presented where lexical effects are the main contributors to finding word boundaries, i.e. without considering factors such as stress, syllabification, and juncture cues (McClelland & Elman, 1986).

3.6.1.4 Semantic level

The *semantic* level of the **LI-domain** refers to the meaningful outcome of natural language processing – involving all the mentioned linguistic levels, but where attention is directed towards interpreting verbal structures so as to construct meaning from them. In other words, when we focus on *what* the vocal persona says, we focus on the semantic level of the **LI-domain**. This level is thereby distinct from focusing on any of the levels discussed above *for their own sake*, which would imply a kind of abstraction process. Thus, this level is firmly located in the reference-oriented part (left) of the **LI-domain** in **figure 2.3**, whereas a focus on the linguistic structures in themselves would imply a location further towards the right of the figure. Since I will distinguish this reference oriented mode of intentionality from any other ways of attending to linguistic structures, and since I will be referring to it a lot in this thesis, I will give the semantic level of the **LI-domain** its own abbreviation, namely **LI/sem**.

There are a lot of issues at the semantic level that can influence and guide perception and comprehension in a top-down manner. Our knowledge of the world can impose constraints that form the basis of predictions that help us narrow down the possibilities of what will come next, something that will ease perception and comprehension. Such constraints can for example delimit the lexical decision to specific instances that belong to one single category. In the sentence starting with "Peter felt very hungry so he went to a restaurant and ordered a ….", it is likely that the last word will belong to the category that can be described as "edible objects that can be bought in restaurants". If the last word is degraded or ambiguous, so as only to render the ending [-εΙκ] audible, semantic cues can provide likely candidates for a plausible interpretation; in this case, it is a lot more likely that the last word will be "steak", rather than "lake" or "break". However, since semantic constraints are bound up in many factors, such as types of discourse (formal, informal, everyday, poetic), culture, beliefs, and situation, the workings of semantic influence on perception are very complex. Semantic influence also works on many levels; from within a single sentence to longer stretches such as narratives, speeches or song lyrics. Such constraints are often simply referred to as "context", at least within linguistics.[131] Its importance for speech perception is often illustrated by showing that two sentences can sound identical, and the only way one can

[131] It seems that it is easier to make use of context effects in the language which is our mother tongue. Van Wijngaarden and colleagues studied differences in intelligibility for natives and non-natives, and found that non-natives required 1-7dB better speech-to-noise ratio to obtain 50% sentence intelligibility than native listeners. They explained this as largely pertaining to the non-native's less effective use of context effects, especially semantic redundancy (van Wijngaarden et al., 2002).

tell them apart is through contextual cues, such as in the sentences “How to wreck a nice beach” and “How to recognize speech”.[132]

Interpretation of the semantic level is not only dependent on the mentioned levels and the verbal context, but the whole context that we associate with the vocal persona. The **ID-domain** is often very important in linking the verbal semantic content to a person with certain properties, the **AF-domain** will give information about the affective state of the vocal persona and the **SE-domain** might give us some clues about the environment in which the vocal persona is situated so as to aid interpretation of anything that the verbal material refers to. The **TCM-** and **SQS-domains** might even influence the semantic level: The former might affect the interpretation of the verbal content if hiss and pops accompany the voice, signalling that it is recorded many years ago; the latter might affect the semantic content through relations of sonic similarity, just as when two or more words are linked by rhyme or alliteration (repetition of initial consonant sound). Thereby, the wide range of semantic meanings that can be interpreted from the contextualized verbal utterance within the context of an electroacoustic piece can comprise much more than the **LI-domain** in itself. Still, all the other domains will have to take a *contextual* or background function, i.e. they cannot be at the centre of attention, since this will imply that we will have to de-engage the verbal processing for a while. The many questions related to the levels of attention, of redundancy or lack of coherent information in different domains, of different degrees of complexity, and of learning, are all important here, but all these questions will have to wait until the discussion in the premise chapters, first of all chapters 5 and 6.

At this point, I will merely exemplify how the linguistic meaning, i.e. the **LI/sem** domain, can relate to the meanings conveyed by other domains. Sometimes, the referential aspects of several domains may indeed overlap – as when a vocal persona states that “I am old and feel miserable” in a sad voice that is recognized as being that of an elderly person. At other times, however, the referential aspects of the vocal persona may seem almost totally distinct compared to those constituted verbally. For example, if a very neutral and disengaged voice reads a story involving statements of several fictive characters, the statement “I am old and feel miserable” can mean something that doesn’t relate to the vocal persona of the reading voice at all – it clearly belongs to the fictive character.

[132] In cross-linguistic studies, one has found some empirical support for the proposition that some languages depend more on context than others, meaning that in some languages situation, tone of voice, gestures, and other non-verbal cues will have a stronger influence in communication than in others (Kitayama & Ishii, 2002; Ishii et al., 2003). For example, Japanese and Spanish are seen as high-context languages, whereas English and Germanic languages are seen as low-context.

3.6.2 Bottom-up: Perception of acoustical cues

Even though high-level linguistic competence and contextual knowledge plays an important role in perceiving and comprehending linguistic structures, low-level perceptual processing that is more directly linked to the nature and condition of the incoming stimuli will naturally also play an important part – in other words what I have referred to as *bottom-up* influence. From early on, researchers were aware of the influence of top-down processes. In a set of pioneering studies, Harvey Fletcher wanted to focus on bottom-up processes by testing the accuracy in which non-sense syllables or words were identified. This was defined as *articulation* (Fletcher, 1922). Later studies have tended to vary more in terms of rigorousness in excluding contextual information, studying *comprehension*, i.e. understanding and reproducing the semantic content of verbal structures, usually sentences, and *intelligibility*, a term somewhat intermediate in terms of context influence, often referring to the ability to correctly identify or reproduce words, or in some cases sentences (Arons, 1993; Allen, 1994; Foulke & Sticht, 1969; Pisoni et al., 1985).

Recent research has indicated that top-down and bottom-up processing are interactive (Davis & Johnsrude, 2007a; Pickering & Garrod, 2007). According to this view, the redundancies in linguistic structures, both from a top-down and a bottom-up perspective will enable the listener to make several kinds of predictions based on multiple cues on multiple levels, and it is the degree of correspondence between these cues and the input stimuli that will determine the success of the identification of the verbal structures. Pickering and Garrod have presented a model of this interaction, in which production constraints on the phonetic, syntactic and semantic levels act as feed-forward predictive filters which affects the bottom-up processing continuously. Depending on the degree of noise, these predictions are weighed against the bottom-up analysis of the incoming stimuli in a mechanism analogue to a Kalman filter:[133] "If the prediction is strong and the input is noisy, there is low Kalman gain (strong top-down influence on interpretation); if the prediction is poor and the input clear, there is high Kalman gain (strong bottom-up influence)" (Pickering & Garrod, 2007: 108).

[133] Welch and Bishop have described the Kalman filter in this manner:"The Kalman filter is a set of mathematical equations that provides an efficient computational (recursive) means to estimate the state of a process, in a way that minimizes the mean of the squared error. The filter is very powerful in several aspects: it supports estimations of past, present, and even future states, and it can do so even when the precise nature of the modelled system is unknown" (Welch & Bishop, 2004).

Two issues are of particular importance for the bottom up influence:[134]

1. **Auditory stream integration** refers to the process of linking auditory units to each other so that they form a continuous event for the listener. This process is described and explained by Bregman in his theory of *Auditory Scene Analysis* (Bregman, 1990) and will be subjected to a more detailed discussion in chapter 11. At this point, it is sufficient to state that this process, i.e. forming a continuous auditory stream, appears to be a necessary condition for perceiving ordinary speech. Without forming a single auditory stream, it will not be possible to process the single phonemes in the right sequential order. For example, if pre-recorded words produced within the context of different sentences are spliced together into one and the same sequence, listeners find it difficult to comprehend it as an intelligible sentence, and will often hear only detached words coming from different speakers (Bregman, 1990: 543). Also, if large discontinuities in pitch are introduced in a sequence of vowels that otherwise is spectrally continuous, listeners are likely to hear them as separate streams causing the vowels to be heard as separate syllables. In normal speech, there will usually be continuities in frequency, spectrum and spatial position that will ensure that speech is held together as one single stream. This is partly due to *coarticulation*, that is, the articulatory interaction between subsequent phonemes that may or may not be audible.[135] E.g. in producing the word *stew* [stu], most speakers will start the rounding of the lips during the [s], even if it is only required for the [u]. Thus, the [u] influences the pronunciation of the [s]. In pronouncing the [s] in *stay*, for instance, no lip rounding will be observed (Kent & Read, 2002: 223-226). Hence, one can interpret this as a form of articulatory redundancy that can also be of help in the bottom-up processing.

2. **Salience of acoustical cues:** Depending on the degree of redundancy provided by different kinds of top-down constraints and contextual cues, there has to be a minimum of acoustical cues present for the listener. In cases where context can provide listeners with minimal aid in linguistic perception, bottom-up cues will dominate. Examples of such situations can be when perceiving foreign names or other

[134] Both these issues can be linked to the coupling between speech perception and speech production discussed above (sect.3.3.1, see also Davis & Johnsrude, 2007a and Pickering & Garrod, 2007).

[135] See Farnetani, 1999 for a review of coarticulation.

words, when the level of ambient noise is high, or when communicating over a low-quality communication system. In such cases, the features and structures working on many of the earlier mentioned levels need to be clearly articulated and differentiated so as to minimize ambiguities. What is often referred to as “clear” or “highly intelligible” speech can give some cues as to how such differentiations can be facilitated from the articulation and acoustic point of view (Kent & Read, 2002: 227-228): Clear speech is (1) slower, has longer pauses, and some speech sounds are lengthened, (2) reduced forms of vowels and consonants are often avoided, (3) it has greater intensity, especially for stop consonants. Highly intelligible speech has (1) wide F0 range, (2) relatively expanded vowel space, (3) substantial F1 variation, (4) precise articulation of point vowels [i], [a], [u], and (4) high precision of intra-segmental timing. I will elaborate further on the salience of vocal features in chapter 10.

3.6.3 A note on prosody

Before I go on to discuss the effect of having degraded or ambiguous verbal cues, I would like to take a brief look at the effects of prosody on the processing of verbal material. Prosody, which usually is taken to embrace intonational parameters like pitch, loudness and duration, as well as rhythm, tempo, stress and boundary cues like pauses, changes in duration or adjustment, plays an important part in the processing of linguistic structures in speech. According to an extensive review by Cutler and colleagues, prosody is very important in creating structural cues in speech; it helps listeners perform lexical segmentation efficiently, it gives important clues to syntactic and grammatical structuring and semantic continuity that can even overshadow some of these when in conflict. What is more, it can affect semantic issues; prosody can provide the cues to distinguish a question from a statement, it can provide clues to interpret ambiguous sentences, and it can highlight or emphasize semantically important words, bringing e.g. new or contrasting information (Cutler et al., 1997). If prosody is disrupted, this can also lead to deteriorated processing. For instance, experiments have showed that intelligibility can be negatively influenced by flattening, inverting, exaggerating or reversing the fundamental frequency contour of a speech segment by artificial means (Laures & Bunton, 2003; Hillenbrand, 2003; Schlauch et al., 2005).

It is difficult to assign the prosodic features to any one of the domains or levels in my framework. On one side, they have a clearly linguistic function, but on the other side the “musicality” and bodily materiality of the cues might locate them on the verge between the **SQS-**, **AF-** and the **VG-domains**. Even if prosodic features are thereby not a part of the **LI-domain** *per se*, it is important to be aware of their effect on perception and comprehension of verbal structures.

3.6.4 Degraded or ambiguous verbal material

If some of the acoustic verbal cues like the ones just mentioned are degraded, ambiguous or altogether missing, perception has to rely on other bottom-up cues and top-down influence in making a plausible interpretation. In many of the empirical studies of perception and comprehension of verbal material already mentioned, different kinds of electronic manipulations were used, such as filtering, changes in playback speed , manipulations of pitch or frequency spectrum, time-stretching and time-compression. Since many of these types of manipulation parallel those used by electroacoustic composers, I have compiled results and conclusions from a number of studies of speech perception in a table below (**table 3.5)**.

Taken together, the reviewed studies indicate that speech perception is affected by many of the mentioned manipulation techniques, but still surprisingly resistant to electronic manipulation, even in cases where relatively little contextual redundancy is afforded to the listeners. Moreover, one can see that different categories of speech sounds are often affected in different ways; vowels, for example, are affected more than consonants in some cases, in other cases it is the other way around.

In some studies of the perception of verbal content in sounds that retain only a minimum of acoustical cues, such as sine-wave speech and noise-vocoded speech, authors have found strong indications of the top-down influence in speech. (Remez et al., 1981; Davis et al., 2005; Davis & Johnsrude, 2007a):[136] These studies have generally indicated that:

1) given the information that the manipulated sounds are in fact speech, listeners have been able to decode the verbal content

[136] Sine-wave speech is produced by analyzing the formants of a speech signal, and using the time-varying tracks of the first three formants to control sine-waves, which subsequently can be presented simultaneously or in isolation. Noise-vocoded speech, on the other hand, is produced by filtering the speech signal into separate filter bands. The amplitude of the signal from these filter bands are extracted and smoothed, so as to remove the most rapid amplitude variations in the signal. Then, the signals from the different bands are used to modulate a wide-band noise signal within the range of each filter band, and lastly these noise modulated signals are combined into the final noise-vocoded speech sequence.

Electronic manipulation	Effect on speech perception
Reverse playback	· No priming effect on lexical decision task. [1]
	· Words and syllables could not be recognized from reversed sentences. [3]
· reverse playback of local segments keeping overall order of segments intact	· Increasing the length of segments reduces intelligibility. Perfect intelligibility at segment lengths of 50ms, 50% at 130ms. [2]
Filtering **· High pass**	· HP: Articulation* 33% with cutoff frequency (CF) at 2150Hz, 78% with 1250Hz CF. Vowels are affected more than stops. [5]
· Low-pass	· HP: Vowel recognition dropped to 5% when filter allowed only formants above the third formant. [12]
	· LP: Articulation* 40% with CF 1000Hz, 75% with CF 1950Hz, stops are affected more than vowels, fricatives are affected most, especially [s] and [t] with an articulation of 40% and 66% with CF 3000Hz, respectively. Women's voices are affected more than men's. [5]
· Band-pass	· BP: Everyday English sentences can be recognized with 82-98% accuracy with center frequencies of 1500, 2100, and 3000Hz (approx.12-14% bandwidth). [4]
· Time-varying filters	· Filtersweep: Vowels (a,e,i) produced by a woman recognized with above 90% accuracy in an unmanipulated condition, were slightly less recognized (70-86%) when filtered with a time-varying band-pass filter (filtersweep) varying the centre frequency along a sinusoid with a Q of 0.6, low point of 640Hz and and top point of 8000Hz at a rate of 2Hz. With a sharper filter (Q=0.1), faster sweeping rate (3.5Hz), and shifted high and low points (500 and 3500Hz) recognition was poorer for /a/ and /e/ but well for /i/. [11]
Speed changed playback	· Changes below 10% have very little effect on articulation*. Above 10% speed change articulation decreases. Decreasing playback speed has a greater effect on articulation than increasing it. [5]
	· Vowels (a,e,i) produced by a woman recognized with above 90% accuracy in an unmanipulated condition, were poorly recognized (0-36% correct) when played with speed factors of 0.5 or 2.0. [11]
	· Perceived vowel qualities change when playback speed changes. Vowels produced by men, women and children are affected differently, and different vowels are affected differently. [7]
	· Vowels are degraded more than consonants. [15]
· decreased speed	· A speed factor of 0.63 (60% decrease) gave an articulation* of 40%. [5]
· increased speed	· A speed factor of 1.54 (54% increase) gave an articulation* of 56%. [5]
	· The correct recognition of short high redundancy phrases only slightly decreases with a speed factor of 1.5 (50% increase).[6]
Pitch independent changes of rate **· Time-compression by periodic sampling and discarding**	· A restricted vocabulary of 50 monosyllables, in which portions up to 80% of sampled periods of sampling intervals between 10 and 80ms had been discarded, retained intelligibility of above 90% for listeners familiar with the vocabulary. [13]
· increased rate	· For listeners trained in listening to speech at 325-475wpm intelligibility did not drop significantly for speech at double rate. [16]
	· In speeding up sentences (PSOLA) subjects' ability to repeat the spoken sentences correctly dropped to 50% when the speech rate was about 12-13 syllables/sec, which roughly corresponds to a rate factor between 2.7 (male) and 3.5 (female). [17]
· decreased rate	· Slowed down sentences (decreased rate) using a pitch synchronous time-domain algorithm were recognized slightly poorer than those with original speech rate. [14]
Spectrum independent pitch changes	· Vowel recognition did not drop below 70% for glottal pulse rate (GPR) values from 5 to 640Hz when spectral envelope was kept constant. [10]

Spectral envelope shifts **· envelope shifted up** **· envelope shifted down**	· Vowel recognition is affected negatively by spectral envelope shifting. Vowels produced by children are affected more by upwards shifting and less by downwards shifting compared to adults. [8] · The recognition threshold (at 50% recognition) for simultaneous manipulations of glottal pulse rate (GPR) and spectral envelope ratio (SER) was different between vowels; /e/ was most resistant against manipulations, /o/ least resistant. [10] · Vowels produced by men are affected less than those produced by women and children when shifted with a ratio of 2.0 (a shift corresponding to a doubling of formant frequencies). [8] · Vowel identification was better when fundamental frequency was shifted upwards in the same direction as spectral envelope, than when fundamental frequency was kept constant.[9] [10]
· with constant F0	· Vowel identification stayed above 50% using a constant f0 of 80Hz and spectral envelopes shifted up and down with ratios from 0.55 to 2.8.[10]
Distortion **· Input overload**	· Using vacuum tube amplification, an input overload up to 15dB above overload point reduced articulation* from 79 to 77%, with values above this articulation drops off rapidly. [5]
Vocoding (source-filter decomposition) **· noise vocoded speech**	· Recognition of words in sentences showed good performance (>85% recognition) when rendered through an analysis source-filter analysis-resynthesis system (filter banks) with a noise source using four filter bands. [18] Other studies report that high [19] or medium-to-high (75% intelligibility) [20] speech recognition performance was obtained with three [19] and seven [20] bands of modulated noise. · Significant effects of learning on speech recognition from noise-vocoded speech are reported. [21]
· hybrid sounds	· The filter component of hybrid speech (source and filter components from different speakers) were recognized well (80%) when using eight filter bands. The speech component from the source part was recognized in just above 20% for 1 filter band, then decreasing to below 10% from 3 filter bands and more. [18]
Sine-wave speech	· When told to listen for a comprehensible sentence, listeners performed well in a transcription task. [22]

Table 3.5 : The effect of electronic manipulations on speech perception. References are [1]: Kreiner et al., 2003 [2]: Saberi & Perrott, 1999, [3]: Newbrook & Curtain, 1998, [4]: Stickney & Assmann, 2001, [5]: Fletcher, 1922, [6]: Klumpp & Webster, 1961, [7]: Chiba & Kajiyama, 1958, [8]: Assmann & Nearey, 2003, [9]: Assmann et al., 2002, [10]: Smith et al., 2005, [11] Unpublished sound recognition experiment by this author, [12]: Peterson & Lehiste, 1959, [13]: Fairbanks & Kodman, 1957, [14]: Nejime & Moore, 1998, [15]: Foulke & Sticht, 1969, [16]: Orr et al., 1965, [17]: Versfeld & Dreschler, 2002, [18]: Smith et al., 2002, [19]: Shannon et al., 1995. [20]: Davis & Johnsrude, 2003, [21]: Davis et al., 2005, [22]: Remez et al., 1981. * The term *articulation* refers to the probability of correctly identifying non-sense speech sounds, here given as a percentage value (Allen, 1994).

2) there is a learning effect, so that when presented several times, the manipulated sounds were easier to identify in terms of verbal content
3) presenting a written transcript or un-manipulated vocal sounds before the corresponding manipulated ones, increased the training effect

Davis and Johnsrude also linked their findings on noise-vocoded speech to what they call a process of perceptual "retuning" facing degraded or manipulated cues – i.e. an adaptation process to novel and changing linguistic environments directed by top-down influence towards lower-level processing (Davis & Johnsrude, 2007a: 142). This "retuning" was not seen as specific to manipulated or synthesized speech, however. Similar kinds of perceptual retuning could be encountered in natural forms of variation like speakers with unfamiliar regional or foreign accents.[137] This has also been supported in research on speaker identity, where it has been shown that speaker identity will affect phoneme classification, word classification, recognition and memory, and sentence recognition (Eriksson, 2007). Eriksson concludes a review of this kind of research with the following statement: "Thus it can be concluded that the processing of voice information is an integral part of speech perception and that although the processes pertaining to linguistic content and speaker recognizing are regionally separated they overlap and influence each other" (*ibid.*: 8). This once more underline the interdependence of the vocal experiential domains.

Taken together, we have seen how several studies demonstrate the effects of learning through gradually "tuning in" to what a sound conveys, as well as show how external information (as written versions of the linguistic contents) can guide listening in a certain (top-down) way so that listeners can actually perceive a spoken message that was not perceived before this information was given. One might easily imagine that both these phenomena can occur when listening to a piece of electroacoustic music with heavily manipulated voice. Therefore, one should assume that both listening to an electronically processed voice several times as well as studying liner notes to a composition which may give away some of the verbal material may assist in interpreting the verbal content.

On my part, I have observed the effect of learning while listening to Michel Decoust's *Interphone* (1977, on Various artists, 1989). About 1:43 into the piece, a heavily filtered spoken voice loop enters, i.e. it is not recognizable as a voice until much later. Then, during

[137] One could probably also include pathological voices to this list. Moreover, there are indications that similar processes are active when encountering new speakers in general, what is normally referred to as *speaker normalization* (Nusbaum & Morin, 1992; Sheffert et al., 2002).

the next minute or so, the filter gradually widens so that more and more frequency components are allowed, and finally one can recognize the sound of a speaking woman. However, I found that after having listened to this transition a lot of times, I could hear both the vocal origin and the verbal content from the heavily filtered source at the beginning of the transition, thus demonstrating how the memory of the voice loop exerted a strong top-down influence on my listening.[138] Thus, *repeated exposure* clearly had an effect on intelligibility. I would guess that many listeners also have experienced the positive effect on text intelligibility that reading a written version of verbal material included in electroacoustic pieces prior to listening can have.

This indicates that when discussing speech intelligibility and comprehension in the context of an electroacoustic piece, one has to consider (1) the effects of repetitions and re-listening, (2) the time needed to "tune in" to the particularities of a new kind of manipulation, a new voice or a new accent (3) any earlier experience with one particular type of manipulation, (4) in cases where the effect of manipulation is time-varying, to what degree perceptual "retuning" is needed to retain intelligibility, and (5) to what degree the verbal contents is known from beforehand. These issues will be further discussed in chapter 9.

3.7 A note on utterance mode

Many of the issues that have been addressed in this chapter have been related mainly to speech, something which reflects the fact that speech is the most important point of reference in my theoretical framework. This will become even clearer in the forthcoming chapter, where I will introduce the maximal-minimal framework. However, to have some guidelines for describing other types of vocal expressions, I would like to include the notion of *utterance modes* as applied by Istvan Anhalt (Anhalt, 1984: 214-216). Utterance modes can be thought of as patterns or configurations of voice involving a multitude of features, belonging to all the vocal domains, plus a number that belongs in the **SQS-domain**. While some of these utterance modes, such as *bel canto* singing, have been conventionalized, others appear to have been used only occasionally. Anhalt illustrates the concept of utterance modes by presenting a multitude of examples from the contemporary vocal music literature, many of them

[138] This can be seen as being in accordance with research on the effects of *repetition* on speech processing. For example, studies have confirmed that 1) repetitions of speech segments tend to improve intelligibility if something is not perfectly intelligible from the outset (Pollack, 1959), and 2) that the perceived clarity of speech in noise is substantially improved by stimulus repetition (Jacoby et al., 1988), particularly if the same talker produces both first and subsequent presentations (Goldinger et al., 1999; Davis & Johnsrude, 2007b).

employing extended vocal techniques. For many of these, intermediary modes between speech and song have been applied, like Ligeti's *Speech-song with fixed pitches* and Haubenstock-Ramati's *chanting*, which is defined as a middle stage between song and speech. Anhalt's list of factors that underlie many of these examples serve as a useful guide to such intermediary modes, in addition to opening up for other modes outside of this continuum (Anhalt, 1984: 214):

- existence or absence of a recognizable tuning system
- relative vowel lengths
- statistical vowel-consonant ratio per unit time over a certain duration
- degree of pitch stability
- vocal resonance patterns
- patterns of voice quality

Based on this list, conventional singing would have 1) a recognizable tuning system, 2) long vowels, 3) a high vowel-consonant ratio, 4) a high degree of pitch stability, 5) a high frequency resonance that makes the voice "stand out" more (singer's formant), and less distinct resonances for many vowels, and 6) stable voice quality; whereas speech would have 1) no tuning system, 2) shorter vowels, 3) lower vowel-consonant ratio, 4) pitch variability, 5) more distinct resonances, and 6) perhaps somewhat less stable voice quality. Using these factors, one can also imagine many intermediary modes, as well as other modes outside the imagined continuum between speech and song. One can also recognize from this list several features that have been dealt with in relation to the experiential domains already discussed.

Peter Stacey's notion of *vocal styles* can also be useful to consider here (Stacey, 1989b: 21; Stacey, 1989a: 27).[139] In a more structured manner than Anhalt, Stacey draws up a continuum between "lexically dominated" and "musically dominated" styles of vocal production. On this continuum, the conventionalized patterns of speech are located on one end. Stacey introduces styles that imply gradually more "musically dominated" styles, with a gradual application of quantized ("prescribed") rhythms and pitches, i.e. pitches located in a tuning system. Schoenberg's *Sprechgesang*, described by Stacey as "a compromise between speech and music, using the unstriated pitch contours of language but touching the cardinal points of a prescribed melody" (Stacey, 1989b: 16), serves as an intermediate style here.

[139] Stacey deals with issues related to the relationship between score and performer that I do not find relevant for this framework, however.

Finally, Stacey reaches two categories of conventional singing, namely, *syllabic* and *melismatic*, where the former refers to a style where the syllables correspond to the musical notes and where the latter designates a mode of singing which applies several notes per syllable. The vocal style of *melismatic singing* is thereby the most "musically dominated" conventional style. However, Stacey adds two other categories beyond melismatic singing in the direction of the "musically dominated". In these styles voice is used as pure sound, one in which voiced sounds dominate and the other where percussive (noisy) sounds dominate.

With Anhalt's and Stacey's terms, one sees that one has moved on to a level of description that deals with the configuration of several aspects together into a "mode" or a "style". The highly conventionalized modes of *speech* and *singing* were important reference points for both of them, defining together a set of intermediary as well as more extreme modes. Moreover, for Anhalt the different modes he found in contemporary vocal music translated into a set of factors. In some regards, this way of seeing several factors together and using that as a basis for setting up intermediary configurations between two reference points anticipates the maximal-minimal framework that I will present in the next chapter. Here, I will consider a number of features in relation to each other, and speech will here be the central point of reference.

3.8 Chapter conclusions

In this chapter, I have presented four experiential domains which are all intimately linked to the voice and what I called the *vocal persona*. The vocal domains are seen as situated at the core of the general framework of experiential domains in acousmatic electroacoustic music presented in the previous chapter. The division into the four domains of vocal gestures (**VG-domain**), identity of the vocal persona (**ID-domain**), affective expression (**AF-domain**) and verbal material (**LI-domain**) was motivated by the idea that a voice carries meaning in multiple layers and with reference to different aspects. For each of the domains, I have tried to establish a basic understanding of the types of meaning that can be inferred by the listener, of the most important factors that contribute in constituting this meaning, including factors related to the listener and his or her background, as well as acoustical cues, and how different types of electronic manipulations can affect the factors. By doing this, I have also introduced several concepts related to each of the domains that might prove useful when trying to describe vocal sounds in electroacoustic music when setting up the theoretical framework of the maximal and minimal voice, which will be the focus of the following part of the thesis.

By introducing these domains I have not wanted to claim that the domains are clearly distinct and independent. Rather, I have shown that the domains are partly interdependent and that the boundaries between them are sometimes hard to delineate. What I would like to maintain, however, is that each of them *can* be put at the centre of attention, so as to highlight the meanings related to each of them.

Part II

The maximal-minimal model

4.0 The maximal – minimal model

In the two preceding chapters, I presented the idea that aspects of an experience of voice in electroacoustic music can be assigned to what I have called experiential domains based on whether the aspects were attributed to the same source/cause, or share a function, feature or relationship. Furthermore, I presented a set of concepts related to each of the domains that can be useful in the qualification and description of these domains and the aspects they subsume.

Still, what is hitherto missing from the framework constituted by these domains is a way to relate experiences to two important points of reference: The first of these is the communicating speaking voice, which we are so used to hearing and responding to, both in daily communication and through mediated experiences. The second point of reference is more specific to the field of electroacoustic music and related genres, namely the critical boundary between voice and not-voice. Both of these can be experienced in electroacoustic works, and both offer a means of evaluating all other experienced voices or voice-like sounds relative to them. It is on the basis of these two reference points, or perhaps more correctly, *reference zones* – since they can't be pinned down precisely, that I have developed the maximal-minimal framework.

In this chapter, I will present the basic ideas of this framework, beginning with an excursion to the literary theory that inspired it. Then, I will go on to look at some related theories dealing with radio and radiophonics, before I present the seven *premises* that together define the maximal voice. After showing how this framework is related to other theories of electroacoustic music, I will go on to present the idea of viewing the framework as a centre-periphery model, with the maximal voice located at the centre and the minimal voice at the periphery, which simultaneously constitutes the boundary towards not-voice. I will also argue that the maximal-minimal model has a structure similar to a so-called *cluster model*, a type of category described by Lakoff (Lakoff, 1987), and that the model can be seen as having two kinds of boundaries. Finally, I will give an outline of the structure and organization of the following chapters dealing with the seven premises of the model. In these chapters the link to the experiential domains will also become clear.

4.1 Borrowing from literary theory

The concepts of maximal and minimal voice is borrowed from the literary theorists Donald Wesling and Tadeusz Slawek and their use of these concepts in the book *Literary Voice - The Calling of Jonah* (Wesling & Slawek, 1995). In this book, Wesling and Slawek are concerned with notions such as subjectivity, intentionality, expressivity and presence, and the dismantling of these within postmodern literary practices. In their discussion of these issues, the authors introduce the notions of maximal and minimal voice as conceptual tools for the analysis of literary texts. The two concepts are presented as ends or extremes of literary voice, which the authors theorize mainly by analyzing texts that are seen as belonging to each of the extreme positions. Although it is difficult to find concise definitions of these terms in their book, especially the maximal voice, one can still get the main ideas from their analyses, which I will give a brief presentation of in the following.

In accordance with central theories within the post-structural line of literary theory, Wesling and Slawek take on a critical attitude towards concepts like authorship, meaning, selfhood, originality and self-presence, maintaining the metaphysical founding of such concepts.[140] Nevertheless, they seem to be interested in investigating how such notions have come to play a role with regard to the literary voice. One of the reasons for this seems to be that these ideas still seem to play a role in how people generally react to voices, both in literary texts, and as sound. Despite practices such as avant-garde writing and electronic networking, that seemingly defy the ideas of self-centred, self-contained and clearly bounded subjects, the authors maintain that "we still refer to intention and agency, making assumptions about the social and other placement of person, of voice" (Wesling & Slawek, 1995: 11). And, although these attributes have been eliminated from postmodern writing, the reader is still forced into a "furious hunt for continuity, individuality, centeredness, bounding outline, [and] social tones of voice" (*loc.cit*). Thus, for Wesling and Slawek, the reader plays a constitutive role in defining the voice, in constituting its presence, and anchoring it to a provenance, although there are several reasons why this is ultimately deceptive and illusory: Writing separates reader and writer in time, the boundaries between self and world or self and other cannot be defined clearly, one can never be fully present in what one says or writes, etc.. The constitutive role played by the reader in literature also seems to be filled by the listener when listening to singing voices: "Through voice in popular song, a new form of presence is constituted, a presence which is to be created by the listener, without whom the song cannot

[140] See e.g. Terry Eagleton's account of this type of criticism (Eagleton, 1983: ch.4).

exist and through which the listener makes a self into a speaking subject" (*ibid.*: 76). In many ways, Wesling and Slawek's description of the reader/listener and his/her "hunt for continuity, individuality, centeredness, bounding outline, [and] social tones of voice" matches some of Chion's ideas, presented in the previous chapter, of the listener and his/her structuring of the voice in perception : For Chion (and Sacco, whom he cited) perception tends to centre on the voice, "picking it out", thus separating it from the totality of sounds present, to "analyze the sound in order to extract meaning from it" and to "*localize* and if possible *identify* the voice" (Chion, 1999: 5). It seems then, that from a reception point of view, ideas of presence, identity, intentionality and meaning are relevant for our understanding of voice, both as writing and as sound, even though such ideas can ultimately be shown to be, at least from a philosophical point of view, metaphysical illusions.

When it comes to the concepts of *maximal* and *minimal* voice Wesling and Slawek do not present any clear-cut definitions, but rather chose to link them to a set of other related concepts. Especially with regards to the maximal voice they are not particularly explicit, and the closest they come to an explication of this term is that it "has highly overdetermined meanings – meaning grafted onto meaning - and is thus literary voice as such" (Wesling & Slawek, 1995: 10). Moreover, they refer to the maximal voice as a "full voice" (*ibid.*: ix). In their analyses of texts that apply this maximal voice, however, it is somehow easier to get a better grip on this notion, especially in the authors' analyses of "bardic voice".[141] The authors see the bardic voice as designed to counteract "the increasing erasure of the self from poetry" with a poetics that can be defined as "a longing for personal voice and the will to simulate the illusion of personal voice" (*ibid.*: 114). Thus, it must be seen as a projection of person and self, albeit an illusory one. The bardic is also linked to the tradition that, since Plato, has privileged speaking over writing, grounded in the view that the oral has a presence which writing has not: "The bardic is print culture's nostalgia for oral culture" (*ibid.*: 113). The bardic can therefore also be described as "voice as presence", encouraging belief in a

[141] This "bardic voice" is described as "a national presumption in literature", as "a mode of representation [which] is a myth of national memory", and as "a kind of writing that brings voice into the public sphere – a genre directly political and historical, in a hopeful if pitiable relation to existing power" (*ibid.*: 106). This kind of literary voice is mainly traced to a certain historical period covering parts of the eighteenth and nineteenth centuries, but as the authors note, can come to life again during the (re-)building of a nation. Here, the authors explicitly refer to the new nations rising after the fall of the "iron curtain" around 1989-90. See Wesling & Slawek, 1995: 106.

correspondence between person, self and voice as sound, where the rendering in print merely is seen as a layer of transparency.[142]

For the *minimal voice*, however, ideas of presence, self and person seem remote and a lot harder to detect. In many ways, minimal voice constitutes the main focus of the authors in their account of literary voice, and since it constitutes the opposite pole of the maximal, it therefore becomes easier also to get a clearer picture of both poles. First of all, the minimal voice is described as literary voice in which tensions, ambiguities, uncertainties related to meaning, to intentionality, to self and person come openly into play. This can happen in two different ways, according to Wesling and Slawek, in the form of two different modes of the minimal:

1) The first mode, is "voice as noise"; voice not yet becoming concept "where the vox confusa of animals or the body has not reached vox articulata but still has human meanings" (*ibid.*: 11). This is exemplified by the authors with "exclamation, birdsong, babble, phatic utterance, phonic material that seems on the way to being speech" (*ibid.*: 10).[143] Furthermore, this "vox confusa", will make it more difficult to identify a speaker and his/her socio-cultural linkage, because it is "operating on the borders between languages, circulating among various speeches, and denying easy national identification" (*ibid.*: 165). Thus, in this mode, a primitive communicative function expressed in a grosser, gestural manner will replace symbolic, culturally coded verbal meaning and thus simultaneously deprive a receiver from identifying parts of the speaker's identity linked up with conventional language.

2) The second mode of the minimal is "voice as non-sense", where "language is used to mime breakdown of language" (*ibid.*: 10), i.e. where syntactical and grammatical structures are violated, and where the play with the material properties of literature as writing are put in focus. The structures that normally bind a verbal message into a fairly coherent entity will in these cases be broken, resulting in open-endedness, fragmentation, ambiguity, loss of coherence and effacing of an expressive self and its intentions.

[142] With this as a basis, the authors then develop a critique of bardic voice and other types of "voice as presence", showing how they still are implicitly bound up in dialogic relationships between self and other, and always already ingrained with writing and literariness, which together undermine and decentres notions of person and self.

[143] Here, the authors refer both to the actual phenomena and the description of these in literary texts. Thus, it is not always clear if the authors deal with sound or ideas.

For the authors, these two modes of the minimal expose how voice is indeterminate, displaced, and dispersed in relation to notions as person and self. And taken together, Wesling and Slawek's theoretical framework incorporate several ideas that are applicable also for voice in electroacoustic music: The idea of a general continuum that embraces graded, not binary oppositions between the verbal and the non-verbal, presence and absence, transparency and materiality, coherence and incoherence of meaning, specificity and lack of specificity of identification, between centeredness and diversion, continuity and rupture, individuality and multiplicity, clear and blurry boundaries.

While Wesling and Slawek's theory is valuable as a starting point for my theoretical framework, it is important to emphasize the differences between the literature and electroacoustic music and how they are experienced. For example, whereas a literary text is usually constructed through language, or at least linguistic units, electroacoustic music with voice only optionally includes language. Moreover, the primary medium of literature is writing, whereas in electroacoustic music it is sound. Nevertheless, both art forms still seem to rely on a conception of voice as a form of *expression*, i.e. where *somebody* conveys some form of *meaning*. Moreover, even if both media represent a fundamental separation in time and space between the act of giving voice to something (either by writing or speaking, singing, or recording) and to the reception of this voice, the fact that a recorded voice often will have an immediate perceptual similarity with the un-mediated voice can't be underestimated. The medium of recording will *not* have to deal with the primary symbolic transformation that is implicit in all literature, something which has important consequences for the perceptual and cognitive processes involved.

Interestingly enough, Børset and Dyson have applied these concepts, or quite similar ideas, to voices in radio, something which shows that the concepts might be applicable for loudspeaker-mediated voice (Børset, 2006; Dyson, 1994). I will now discuss some of the issues raised by these two writers and simultaneously show how many of the ideas of Wesling and Slawek's can be transferred to the realm of sound. Furthermore, I hope to identify issues that have to be developed further in the specific context of voices in electroacoustic music.

4.2 Maximal and minimal voices in radio

The potential for a conceptual transference of the idea of a gradation from the maximal to the minimal voice from literature to loudspeaker mediated voice has already been noticed by

Bodil Børset, who in her analysis of Nathalie Sarraute *pièces radiophoniques* uses Wesling and Slawek's terms (Børset, 2006: 128-131).[144] For Børset, it has been important to see Sarraute's literary texts and the different radiophonic realizations of these texts in relation to each other. In her discussion she therefore comments both on aspects that are related to the pieces as literature and as sound. In the parts engaged in analysis of the acoustic realizations of the pieces, she focuses on aspects like:

- the strength of the link between character and voice
- individuation, i.e. the degree of separation between a voice, other voices and background noise
- voice familiarity
- intelligibility of the spoken words

Børset sees all these aspects as positively related to the maximal voice. That is, strong links between characters and a recognizable voice, clear separation of voices from other sounds, and good intelligibility are all features that according to her, characterize a maximal voice. For instance, she describes the maximal voice as "a voice which is clear ("tydelig") and has a large degree of individuation. It must be easy to hear what is said and relate this to a speaking who. Per definition, this kind of voice should be easily recognizable and markedly separate itself from other voices" (Børset, 2006 : 129, my translation).[145] Individuation is also seen as related to *recognition*, because in recognizing a familiar voice, e.g. a known actress or actor, the voice will be more easily distinguishable from other voices. Conversely, these issues are seen as negatively related to the minimal voice. As for the first point in the list above, Børset's discussion of problems with linking voices to characters clearly shows one way in which voice can be minimal, even if the term "minimal voice" is not explicitly brought up in the discussion. Here, Børset shows how in many of Sarraute's pieces one doesn't know the name of the characters that are linked with the different voices, often resulting in the experience of a chaotic and continuous flux of voices, rather than clearly distinguishable

[144] Børseth's thesis was part of the interdiciplinary project *Aesthetic technologies 1700-2000*, a project in which I also took part until its termination in 2006. Wesling and Slawek's notions have been discussed both formally and informally within the frames of this project, and thus has brought both of us to try to apply the terms on different genres of loudspeaker mediated sound.

[145] Norwegian original: "en stemme som er tydelig og som har stor grad av individuasjon. Det må være lett å høre hva som sies og knytte dette til et snakkende hvem. En slik stemme skal per definisjon kunne vær lett gjenkjennelig og skille seg markant fra andre stemmer".

interlocutors. As for the second point in the list, it is more explicitly linked with minimal voices when viewed negatively: "The minimal voices are those which […] recurrently merge into other voices or resound into background noise or silence" (*ibid.*: 130, my translation).[146] Two other issues that are positively related to minimal voice also come up during Børset's analysis:

- the "materiality" of sound
- the dissolution of sense and meaning

Here, she draws explicitly on Wesling and Slawek's two modes of the minimal, "voice as noise" and "voice as non-sense", analogously to each of the above points, respectively.

What Frances Dyson has written about radio voice corresponds in many ways to the ideas of Wesling and Slawek (Dyson, 1994). Rather than including a notion of a graded continuum between extremes, however, she focuses on the ideological and technological foundations of what she calls the "dominant" radio voice, a type of voice that has many parallels with the literary maximal voice. Dyson's notion of the radio voice is one that is usually associated with the more traditional broadcast voice emphasizing information rather than entertainment. She writes that "[generally], the dominant radio voice talks – its speech is clear, articulate, sometimes eloquent. Most of what it says is perceived by the listener as factual and informative […] It does not mumble or stutter, it pronounces full and meaningful sentences, it says something" (Dyson, 1994: 167). Hence, the "full and meaningful sentences" clearly parallels Wesling and Slawek's "overdetermined meanings – meaning grafted onto meaning". Moreover, we can recognize the point of clarity or intelligibility from Børset.

However, in addition to presenting factors that constitute the dominant radio voice, she also demonstrates ways in which these factors can be negated, and thereby it is possible to relate her discussion to the minimal voice as well. One such potential factor is the "bodily" aspects of the voice that can be apparent in e.g. trembling voices, throat cleansing, coughing, sneezing, panting, etc., in other words, what I referred to in section 3.3 as the bodily materiality in experience constituted by the **VG-domain**. It is therefore not surprising that Dyson describes the dominant radio voice as a disembodied voice; bodily sounds are by convention absent from it or edited out "because the body represents *noise*" – i.e. something

[146] Norwegian original: "De minimale stemmene er de som i stemmegradasjonen stadig går i ett med andre stemmer eller lyder tilbake i bakgrunnsstøyen eller stillheten".

that might clutter or hinder the conveying of *what* is said (*ibid*.: 178).[147] She then shows how this favoring of signification conveyed by words rather than the body is related to the view of the voice as an instrument of the soul rather than the product of a body, a view that can be traced back to Christian theology and early Greek philosophy, especially Aristotle, according to Dyson. The genealogy of the dominant radio voice therefore goes back to some of the cornerstones of Western thought.

Another potentially disruptive factor is the medium itself. Just like bodily noise has to be abandoned to retained fullness and unambiguousness of meaning, distortion and noise from the medium also has to be abandoned in the striving for 'true' reproduced sound. As Dyson expresses it, the reproduction of the voice has to be "without distortion and as ontologically identical as possible to the original" (*ibid*.:179).[148] Moreover, the dominant radio voice has to appear either as the only audible sound source, or as the foreground phenomenon with very few other sound sources interfering, so as to prevent anything that potentially could threaten the intelligibility of what is being said. Hence, Dyson claims that such a voice has to be *singular*, i.e. it must be controlled so as to appear *one at a time* (*ibid*.: 181). Neither environmental sounds nor sounds implicit in the recording process are allowed to come to the foreground, preferably by minimizing noise by using low-noise recording devices, "close miking", or soundproof studios to shut out interfering sounds.

Finally, Dyson also makes a point that the dominant radio voice is one of authority (*ibid*.: 180-181). Radio is a medium usually authorized by a broadcasting institution and technological systems of distribution, often controlled by institutions with power and money, ultimately deciding who is allowed on the air and who is not.[149] Thus, radio is itself a medium of power.[150] The authority is also apparent both in the mentioned abandonment of noise from bodies and mediating technology and the preference for male voices, a preference linked to views rooted in Western culture; the male as the more truthful and rational; the female as untruthful, irrational and hysterical. Listener expectations of truthfulness, continuous censorship and surveillance on what is being aired, and the broadcaster's risk of being assumed full responsibility for any statement – all this makes questions of power and

[147] However, as Katharine Norman notes, other radio voices may indeed use noises of the body and its actions in the studio to gain trustworthiness and to be convincing (Norman, 2004a:116)

[148] These ideas are also developed by Norman, who discusses radio art which apply *clean* and *unclean* voices, i.e. voices which are undisturbed/unaffected (clean) or disturbed/affected (unclean) by the technologies of recording and mediation (Norman, 2004a: 103,122).

[149] This is about to change however, with radio transmitted over the internet. Because of the internet's open and global character, it is more difficult to enforce laws of censorship on these kinds of broadcasts.

[150] The huge importance of radio in Hitler's way to power is maybe one of the best examples on this.

authority especially poignant in radio. This is not to say that issues of power and authority are irrelevant in electroacoustic music. Indeed, Hannah Bosma has shown in many of her articles that authorship in electroacoustic music is related to both power and gender issues in the disfavour of women (Bosma, 2003; Bosma, 1995). Indeed, the composer has been a dominantly male figure in Western music, often endowed with unquestionable authority and geniality, being the "true" origin of the music (rather than the instrumentalists and vocalists) and the one that authorizes the performers' choices. Even if such tendencies have been counteracted by ideas of openness, process and interactivity, especially in the thoughts and music of John Cage, electroacoustic music, especially the acousmatic vein, still relies heavily on the Western notion of the composer as the Master of music and sound – as a *phoniurge*, to use Chion's term (Chion, 1991: 35). Even though issues of power and authorship are all important, I choose not to discuss them any further in the following in order to delimit the scope of the dissertation.

To sum up, Dyson delineates a type of voice that:

- is a speaking voice
- is clearly articulated, without any mumbling or stuttering or other kinds of bodily "noise"
- is factual, informative and meaningful
- is without any interference from mediating technology
- appears in singular – one at a time
- is one of authority and power, and is usually male.

Together, Børset and Dyson, Wesling and Slawek have provided ideas that have been important for the framework that I am developing in the following. In particular, they provide a basis for my conception of the maximal voice. Moreover, the latter two's idea of a graded continuum is an important component in my framework. For electroacoustic music the main challenge with the minimal end of the continuum seems to be that it is a lot more diverse than the maximal end, bearing in mind the seemingly limitless possibilities of the technologies of processing and synthesis. In that respect, the two modes of the minimal, suggested by Wesling and Slawek, and retained by Børset, seem to contain too little room for differentiation and refinement. Lastly, the "voice-as-nonsense" mode of the minimal would probably not be very applicable for the majority of electroacoustic works.

4.3 Maximal voice as a set of premises

Retaining many of the ideas presented above, I will now formulate the maximal voice as a set of *premises*, as I choose to call them. The premises each express one particular aspect or feature of the maximal voice, similar to several of the points mentioned with regards to Dyson and Børset above. As in the previous chapters, I am still dealing with an experiential point of view, and define the premises from a listener's perspective rather than a production perspective. Each of these premises can be seen as conditions that can be fulfilled to different degrees, and when they are all fulfilled, the result is what I define as maximal voice. By considering different degrees of fulfilment, it is also implicit that the premises in themselves can be seen as continua running from the maximal to the minimal, in line with the idea of a graded continuum of Wesling and Slawek, only differentiated for several aspects.

The seven premises I include in my framework are:

1. **Linguistic-semantic focus of attention**: The semantic level within the linguistic domain receives sustained and maximal attention.
2. **Balanced information density**: the information density of the experiential domains is optimal for the processing/decoding of the LI-domain.
3. **Naturalness**: The sound has maximal resemblance with one produced by a human being and his/her vocal apparatus.
4. **Presence**: The listener experiences a sense of a shared "here and now" with a vocal persona.
5. **Clarity in meaning formation:** Meaning can be constructed from the voice with a high degree of clarity – also implying specificity, certainty and coherence
6. **Feature salience:** Vocal sounds and features "stand out" perceptually – for themselves and relative to other sounds and features.
7. **Stream integration**: The sound of the voice is integrated into one coherent and continuous sound stream (cf. auditory scene analysis).

It might be difficult at this point to see the links between some of these premises and the points from Børset and Dyson listed in section 4.2, maybe except numbers 1, 5 and 6. The premises will be extensively discussed and exemplified in the following chapters in turn, and instead of going into much detail at this point, I leave it to these chapters to provide the links to Wesling and Slawek, Børset and Dyson. Furthermore, some of the premises, such as that of *naturalness*, only seem meaningful when considering how they might *not* be fulfilled for the multitude of transformed vocal expressions in electroacoustic music. For example, it might seem superfluous to state that the maximal voice is natural unless one considers the possibilities of experiencing highly unnatural voices in the music. Thereby, even if these premises are meant to define the maximal voice in a positive sense, they are nevertheless conditioned by the possibility of partial or no fulfilment. And, when the premises are fulfilled only minimally, we are dealing with minimal voice. However, since minimal voice is a boundary phenomenon, its definition is more complicated, and I will return to the relationship between these premises and minimal voice below.

As one might notice from the seven premises stated, maximal voice describes something relatively far from electroacoustic music and sound art in general, which usually focuses on sound rather than meaning, and in which disruption of established meaning systems is sought rather than eschewed.[151] And in my experience, the maximal voice in its fullest sense is not encountered too often in the type of material that I focus on in this thesis. In contrast, in other sonic expressions as radio and audio-book genres like the interview, causeries, speeches and lectures, voices close to my notion of the maximal is the rule. One example that approach the maximal voice can be found in *Les objets obscures* (1991, on Parmerud, 1994) by Åke Parmerud (**sound example 4.1** from the second movement, 0:00-0:14). In this piece, a French female speaking voice is heard at several salient points in the piece. Several times this voice appears on its own, close and without any reverberation, devoid of ambient noise, and it speaks in an articulate and fluent manner, clearly intelligible, at least for listeners who understand French. Yet, on the semantic level, there is not so much clarity: The sentences are not always complete, and what is referred to remains mostly rather obscure since the woman is presenting a riddle, a riddle which hints at the "hidden objects" that can be heard in the piece, and that the title refers to:

> Le deuxième: Un paysage ambulant. Un déplacement perpétuel. Quelque chose qui frôle sans toucher. Un mouvement sans but. Un objet de repos.
> (The second : a landscape on legs. A constant moving. Something that touches without touching. A movement without goal. An object to rest in.)

Even if the answers to the riddles are presented in the fourth part of the piece, the general impression is that this is quite far from the meaningfulness of the informative and factual radio voice. Moreover, the voice is only present in short sections at a time and functions mostly as introduction to the more "musical" parts of the piece. This woman is therefore felt

[151] According to Eco, contemporary art is characterized by an oscillation between rejection of established order and its preservation (Eco, 1989: 60).

to retreat to an undefined absence during most of the piece, thus being only temporarily present for the listener.

One can see that even if this voice doesn't bear all of the characteristics described above, it still seem to draw on some of the central "resources" of the maximal voice, largely fulfilling premises such as a focus on attention (at least for those understanding French), information density, naturalness, presence, salience and stream integration. And, as I intend to show in the course of this dissertation, there are several other examples of how voices in electroacoustic music can draw on these resources.

4.4 Related theories of electroacoustic music

Even though the idea of the maximal-minimal framework has its origin in literary theory, there are several theories of electroacoustic music that are based upon similar ideas. By giving a brief account of some such theories in the following, I hope to show that many of the components of my framework are already presented as elements of other theoretical stances, both within electroacoustic music in general and acousmatic electroacoustic music with voice in particular. In other words, I am not introducing all my ideas from scratch, but rely on many accepted notions in the field. Therefore, the framework can be seen as an attempt of relating a number of established ideas to each other.

I have already mentioned theories which present the idea of a continuum between concepts like "abstraction" and "reality" (see section 2.4), i.e. where the qualities of the sound "in itself" are in focus on one end, and where the referential aspects related to sources and causes are in focus at the other end (e.g. ten Hoopen, 1992b; Young, 1996). One can indeed see some resemblance between this idea and the maximal-minimal continuum: For the *focus of attention* premise, a focus on abstract qualities, i.e. the **SQS-domain**, will clearly imply *not* to focus on the semantic level of the **LI-domain**, and as Børset mentioned, the "materiality of sound" is indeed related to the minimal voice. At the maximal end of the continuum, one can also argue that source/cause-related aspects are important, since a clear definition of the identity of the vocal persona is important for the fulfilment of the *clarity of meaning* premise. However, for the maximal voice, source/cause-aspects are only a part of the context or background – for the *focus of attention* premise, it is the semantic content of the verbal message (**LI-domain**) that is at the centre.

The theories that are specific for voice in electroacoustic music are naturally even more interesting. In Bruno Bossis' groundbreaking book *La voix et la machine* he introduces the concept of *artificial vocality*, a term that in its turn conjoins the concepts of "vocality" and "artificiality", in principle covering all aspects of vocal or vocal-like expressions in sound mediated by loudspeakers, but in practice focusing on voice in electroacoustic works (Bossis, 2005). By using the adjective "vocality" rather than speaking of the voice, Bossis opens up for seeing the phenomenon as a continuum rather than a question of either-or. In summing up the ideas presented in the book, Bossis emphasizes how artificial vocality is played out in a generalized continuum, where all musical parameters are continuous (*ibid.*: 288). More specifically, he draws up a continuum between the "frankly vocal" ('franchement vocal') and what is "not at all vocal" ('pas du tout vocal') (*loc.cit.*). Moreover, he argues that the detachment of the voice from bodily production, which characterizes artificial vocality, delineates a continuum between the human and the synthetic (*ibid.*: 289). All in all, the idea of the continuum is a very important one in Bossis discussion of artificial vocality, and it is not difficult to see the similarity between these ideas and the gradation between the maximal and the minimal fulfilment of the premise of *naturalness*, which as we have seen is related to the number of properties that can be attributed to the human vocal apparatus. Even if I do not explicitly set up "artificiality" or "synthetic" as antitheses of naturalness, as Bossis does, the premise of *naturalness* can also embrace this idea.

A much less developed theory, but still interesting in this context, is presented by Segnini and Ruviaro in their paper "Analysis of Electroacoustic Works with Music and Language Intersections" (Segnini & Ruviaro, 2005). Here, the authors break down the analysis into two dimensions, namely *intelligibility* and the listeners' judgment of the *speech-like* versus *music-like* features. These two dimensions are then set up as axes in a "music-language sonic space", as they call it, which I have made an adapted version of in **figure 4.1**. Here, we can see how "musicness"-"speechness" constitutes one axis and "unintelligible text"-"intelligible text" the other. The authors have made a selection of 6 electroacoustic pieces with voice (in addition to some vocal compositions and genres), which are then located in different locations or along different paths in this space. For instance, *Six Fantasies* by Paul Lansky, which is the object of study in chapter 12, is located along a diagonal path from

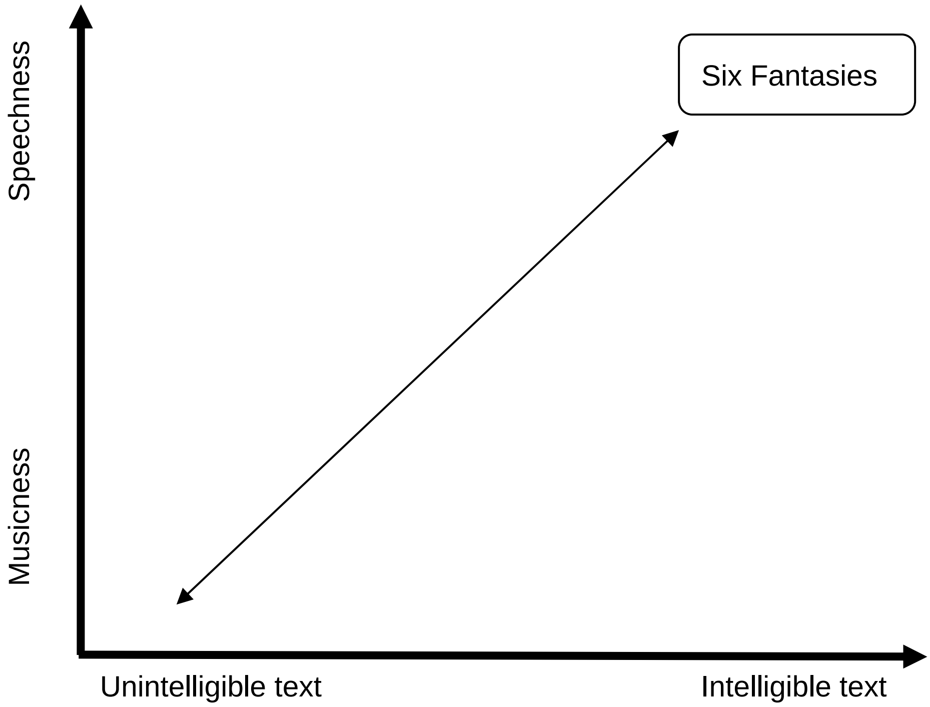


Figure 4.1: Music-language sonic space. Adapted from Segnini & Ruviaro, 2005.

bottom-left to top-right, indicating that it moves between “musicness”/”unintelligible text” and “speechness”/“intelligible text”.

There are several parallels between Segnini and Ruviaro’s way of analysing electroacoustic works with voice and language and my maximal-minimal framework. First of all, their use of the “intelligibility” dimension clearly matches one important facet of the *clarity of meaning* premise.[152] Their “speechness”-“musicness” dimension, on its part, is perhaps more difficult to relate directly to one of the premises, especially since “musicness” is not defined in detail.[153] Nevertheless, since the maximal voice is defined as speech, and since “musicness” would probably imply greater focus towards the **SQS-domain**, it can be seen as loosely related to the *focus of attention* premise. The parallel that still is most interesting in my view, is their use of a spatial model with two continuous dimensions. That the premises of my model can, at least in a relative and inexact manner, be treated as dimensions, will be clear from the forthcoming discussion in 4.6.

All in all, there are several overlapping ideas between existing theories of electroacoustic music and my framework. More generally, the idea of a continuum appears to be present in many theoretical frameworks, even if there is some variation in what aspects

[152] I argue in chapter 9 that intelligibility is far from a simple variable, however.

[153] Rather, this continuum has more in common with the continuum of *utterance modes* between speech and song discussed in section 3.7.

these continua are taken to represent.[154] As we have seen, there are also some theoretical positions that have presented continua which have similarities with some of my premises, in particular *focus of attention* and *naturalness*. Still, as far as I know, the combination of a greater number of premises/dimensions which are organized in one single model is unique to this framework. Another feature I believe to be unique is the view that the dimensions are structured into a centre with a surrounding periphery bounding onto what is not-voice, and that this structure parallels that of prototypical categories. These two features will be examined further in the following two sections.

4.5 Centre and periphery

When the experienced voices only partly fulfil the premises of the model, they will depart from the maximal, and at some point of negative fulfilment, what we might call *violation* of the premises, they will ultimately reach a state where they can be characterized as *minimal*.

The minimal voice is much more difficult to define than the maximal voice, however, because it comprises a wider range of possibilities and modes of expression. For the mentioned seven premises, there are simply more ways in which they can be violated than be fulfilled. Thus, it seems that the relationship between the maximal and the minimal can be described through the dichotomy narrow-broad. For the *clarity of meaning formation* premise, for instance, the maximal voice will be confined by conventions of speech and language. The minimal voice, on the other hand, isn't confined by any mode of voice at all – it can comprise speech, singing and vocal experimentation. E.g. in Western classical *bel canto* song, which often has a basis in a meaningful text, the text is often unintelligible due to melismatic passages and phrases in high registers that make pronunciation difficult. In artistic vocal expressions of the Avant-garde, within as well as outside of music, there are many examples of other kinds of voices that in different ways have presented voice with little clarity of meaning, e.g. by using non-sense phonetic texts or pseudo language (as in *Ursonata* by Schwitters (1922-32, on Schwitters, 1992) or *Nouvelles Aventures* by Ligeti (1962-65, on Ligeti, 2006)). And clearly, in electroacoustic works one can find a great many examples of types of electronic processing that will affect the clarity of meaning negatively; filtering, time stretching and compressing, granulation and distortion, many of which will be discussed in the following chapters. Similarly, for the *focus of attention* premise, the fulfilment of the

[154] As Godøy notes, the notion of musical space as in principle continuous along several axes is present in the writings of people like Boulez, Schaeffer and Xenakis (Godøy, 1997:192).

premise implies directing attention towards the semantic level of the **LI-domain**, whereas one can imagine that attention towards both the **SQS-**, **TCM-** and **SE-domains** all violate the premise. Again, therefore, the maximal voice appears more narrowly defined than the minimal.

In addition to representing one of two extremes in my framework, the minimal voice also represents the possibility of transgressions into what *is not voice* – it represents a boundary zone where the voice appears to be on the verge of turning into what is no longer a vocal sound. And, in electroacoustic music the boundary between what is voice and what is not voice is the subject of exploration in several works, among them Stockhausen's *Gesang der Jünglinge* (1956, on Stockhausen, 2001), Wishart's *Red Bird* (1977, on Wishart, 1992), *Mortuos Plango, Vivos Voco* by Jonathan Harvey (1980, on Various artists, 1990) and *Chant d'Ailleurs* (1992, on Viñao, 1994) by Alejandro Viñao. In these and several other pieces one can experience gradual transformations between vocal sounds and sounds that are clearly of a different origin.[155] In the very beginning of the latter work (**sound example 4.2**, 0:00-0:30), for instance, there is a sustained note which transforms gradually and continuously from a wind instrument with a rather ethnic and Eastern flavor into a singing voice. This happen rather slowly and in several stages, and at one point (I experience this around 0:20) the sound begins to take on the qualities of a voice, while still lacking greatly in naturalness compared to a real voice due to its static character. When the pitch fluctuations set in a little later, however, the sound becomes more natural, until it finally sounds just like a sung note. One can thereby experience that the sound gradually changes from non-voice into voice, and that, at one point, the sound passes through a boundary zone between the two. I will return to a more in-depth discussion of such boundary transgressions below.

One useful way of visualizing the discussed relationships between the maximal and the minimal voice is a circular centre-periphery model (see **figure 4.2**). Here, the maximal voice constitutes the centre and the minimal voice the periphery, which borders onto what is not voice.[156] Moreover, in this model the maximal voice is clearly more narrowly defined than the minimal, representing the multitude of ways that voice can depart into the minimal and ultimately into non-voice. The idea of a graded continuum between the maximal and the minimal is also retained in this model. Still, the model lacks the connection to the seven

[155] See e.g. Smalley, 1993; Wishart, 1996 and Landy, 1993 for accounts of transformations/metamorphoses in electroacoustic music.

[156] The idea of the maximal voice as a centre is not explicitly formulated by Wesling and Slawek, but can in my opinion be a consequence of their idea of a maximum pole or end in which departures towards the margins and the two minimal modes take off in different directions.

premises introduced above, and I will therefore look into ways of expanding it. Theories on categorization and prototypes appear to offer such a link, and consequently, I will look into these theories in the following section.

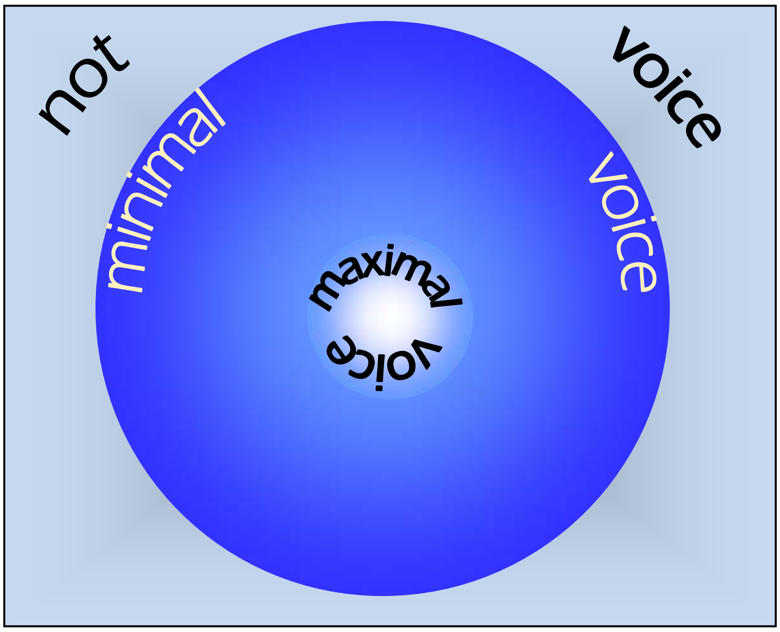


Figure 4.2: Centre-periphery model of maximal and minimal voice

4.6 Parallels with categorization models

Eleanor Rosch's and George Lakoff's writings on categorization and prototype theory present ideas that pose some interesting links to the framework I have hitherto delineated. I hope to show in the following that my model in many ways is structured like a particular type of category described by Lakoff, where the category in my specific case will be "experienced voice in electroacoustic music".

Prototype theory was formulated by Eleanor Rosch in the 1970s, but has several predecessors, both within philosophy, cognitive anthropology and social psychology (Rosch, 1975; Rosch & Mervis, 1975; Rosch, 1978).[157] Her theory was formulated in opposition to the "classical" view of categories, which saw them as rooted in the objective structures in the physical world. Within the classical view of categories, certain members of a category could *not* be seen as more central or typical to the category than others, since the categories in

[157] Lakoff mentions Ludvig Wittgenstein's concept of *family resemblances*, Brent Berlin and Paul Kay's work on color terms, and Roger Brown's study on basic-level categories as important predecessors for Rosch's prototype theory. See chapter 2 of Lakoff, 1987.

themselves would be determined by shared properties rooted in an external and objectively given world, thus endowing all members of a category equal status. However, through reviewing a series of earlier empirical studies and undertaking a set of new ones Rosch was able to find what she called "prototype effects", i.e. that some members of a category were taken to be more prototypical or better examples of the category than others. E.g. she found that some species of birds were thought of as better example of the category "bird" than others; whereas robins and sparrows were considered the best examples, owls and eagles were not so good examples, and penguins, emus and ostriches were considered worst examples of the category (Rosch, 1975).[158]

If we go a bit further into the claims of prototype theory and related theories, we can see that the relationship between the typical and the less typical members also resembles the relationship between maximal and minimal voice, if we regard the whole continuum between the poles as constituting the category "experienced voice in electroacoustic music". In the same way as minimal voice is defined in relation to maximal voice, in prototype theory non-prototypical members of a category are defined *in relation to* the prototypical. Again, in the words of Lakoff and Johnson: "We understand the nonprototypical chairs as being chairs, not just on their own terms, but by virtue *of their relation to a prototypical chair*" (Lakoff & Johnson, 1980, my italics). If we return to the premises that define the maximal voice, we can see that they at the same time define what the minimal voice *is not*. Thus, anything that would be considered as minimal voice would be, in a similar manner as in prototype theory, defined in relation to the maximal voice. In this respect, my model seems to fit the structure of categories in prototype theory.

Moreover, one can see that the spatial metaphor that is implied for the prototypical categories, with some members that are more central than others, in many ways resemble my model, in that they both can be thought of as graded continua between centre (the prototypical members) and periphery (the worst examples). This is evident from Lakoff's summary of the basic results of prototype theory: "Some categories, like tall man or red, are graded, they have inherent degrees of membership, fuzzy boundaries, and central members whose degree of membership (on a scale from zero to one) is one […] Other categories, like bird, have clear boundaries; but within those boundaries there are graded prototype effects – some category members are better examples of the category than others" (Lakoff, 1987: 56). Without

[158] Another example is referred to by Lakoff and Johnson, whose theories have been greatly influenced by prototype theory: "A prototypical chair, for us, has a well-defined back, seat, four legs, and (optionally) two armrests. But there are nonprototypical chairs as well: beanbag chairs, hanging chairs, swivel chairs, contour chairs, barber chairs, etc." (Lakoff & Johnson, 1980: 122).

addressing the question of the different kinds of boundaries mentioned here at this point, we see that in both cases there is a question of *graded prototype effects*.

As we saw above, I listed several constituent premises that together converged to define the maximal voice of my model. This resembles what Lakoff calls *cluster models*, which designate a source of prototype effects in his theory. Cluster models involve several *cognitive models*, which for Lakoff are mental constructs involved in the organization and structuring of knowledge and meaning, e.g. in forming categories (Lakoff, 1987: 74). These cognitive models will in some cases cluster together or converge to form categories that are psychologically more basic than the models taken individually, hence the term cluster models. When all the cognitive models in a cluster converge, it will then result in a more central or prototypical category member than when there are only just a few models clustering or no clustering at all, something which will result in more peripheral members. I will give a short presentation of Lakoff's example using the concept *mother* to clarify this. This concept normally involves a set of cognitive models that combine to form a cluster model. The models he mentions are:

> - The birth model: The person who gives birth is the *mother*. […]
> - The genetic model: The female who contributes the genetic material is the *mother*.
> - The nurturance model: The female adult who nurtures and raises a child is the *mother* of that child.
> - The marital mother: The wife of the father is the *mother*.
> - The genealogical model: The closest female ancestor is the *mother*. (*ibid.*: 74)

These models converge, then, to form the generally accepted conception of what a mother is. When they diverge, however, the resulting sub-categories can still be regarded as *mothers*, but in a more peripheral sense: *stepmother*, *surrogate mother*, *adoptive mother*, *foster mother*, *biological mother*, *donor mother*, etc. Hence, there is a set of less central subcategories that are understood as variants of a more central category. The category *mother* is therefore seen as having a *radial structure*. This also implies that the subcategories are not understood purely on their own terms, but based on their relationship to the central model (*ibid.*: 91).

Taken together, both the cluster models and the radial structure appear to have many similarities with the way in which the maximal – minimal framework was structured; the maximal voice was defined according to a set of premises, in many ways resembling the cognitive models of Lakoff; the minimal voice was defined *in relationship* to the maximal; and, the maximal was seen as more central than the minimal. One difference, however, is that whereas the subcategories in Lakoff's case were restricted to established, conventionalized categories, my model is in principle open to all kinds of combinations of fulfilment/violation of the premises.

Regarding the prototypicality of the maximal voice, however, it can be questioned in what respect and on what grounds maximal voice can be considered as prototypical or as a "best example" of the category "experienced voice in electroacoustic music" apart from sharing the graded centre-periphery structure. Is *speech*, i.e. the category that seems to have the best fit with my concept of the maximal voice, regarded as a better or more typical example of the voice than e.g. singing, laughing, crying or screaming, or even electronically processed voices? On one side, non-verbal vocal expressions such as laughing and crying are more typical on the basis that they are more universal compared to the culture specific expressions of speech and song. On the other side, electronically processed voices might seem more prototypical since processing is a characteristic mark of electroacoustic music. From a third point of view, speech could be seen as more prototypical in that it is probably the mode of vocalization that most people apply most often (maybe except professional singers). However, according to Rosch, questions regarding prototypicality are first of all verified on an empirical basis (See Rosch, 1978: 36). To attain an empirical basis comparable to the studies by Rosch and colleagues would require extensive experimental work that I regard to be beyond the scope of this thesis.

Rather than proceeding into speculation on the prototypicality of my notion of the maximal voice, I will settle with the fact that my model is not empirically based, but is a theoretical construction that first of all is thought to be a tool for understanding voice in electroacoustic music. Thus, even if I will draw on empirical research as well as other bodies of theory in the further development of the theory, its value will first of all be judged in the application in analysis and interpretation. As the discussion on prototypically hopefully has shown, there are considerable resemblances between my theoretical framework of maximal and minimal voice and that of Lakoff's *radial* categories and their structuring into *cluster models*. This resemblance also seems to reinforce the spatial representation in a circular centre-periphery earlier shown in **figure 4.2**. What is more interesting is that it seems to open up for an inclusion of the premises of the model into this picture, since the premises are seen as converging in the centre and diverging towards the periphery. When also taking into account that the premises can be regarded as graded continua running from the centre to the periphery, the spatial representation of the premises as kinds of dimensions having axes pointing outwards from the centre, almost suggests itself. Consequently, an expanded version of the centre-periphery model would be something like **figure 4.3**. However, I want to emphasize that despite the similarity with mathematical representations of an n-dimensional

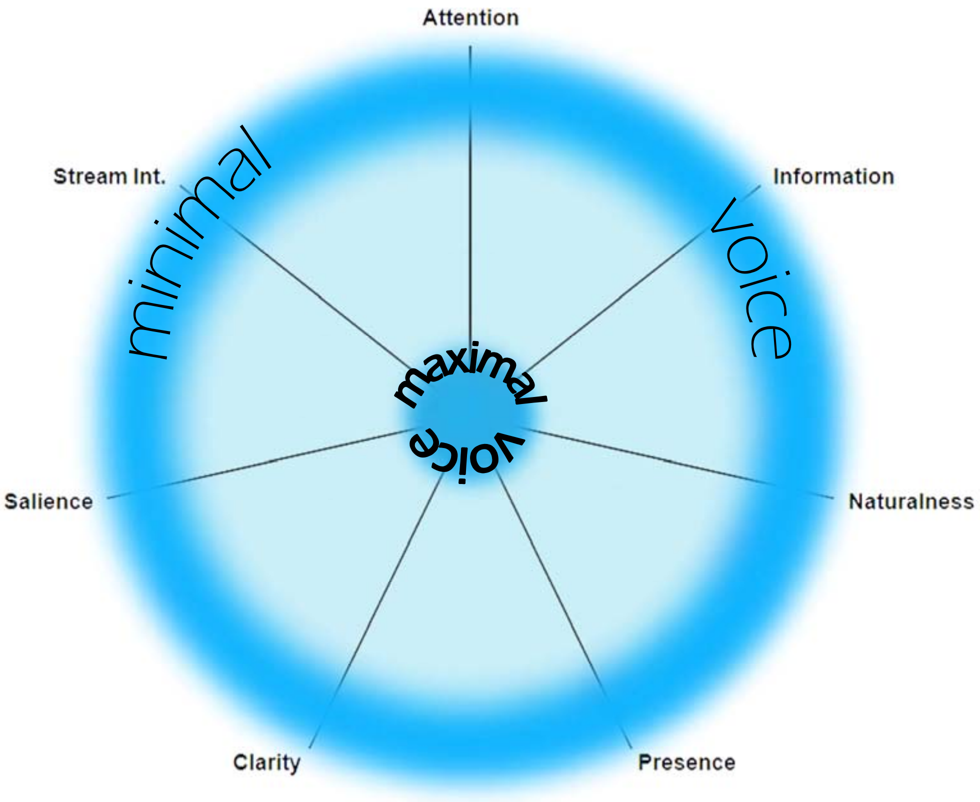


Figure 4.3: Extended centre-periphery representation of the maximal-minimal framework. The seven axes represent the seven premises introduced in section 4.3.

space, the axes in this case are not "true" dimensions in the sense of being orthogonal or independent.[159] Rather, I think of this representation as a system of axes in the sense proposed by Godøy (Godøy, 1997). Godøy regards the notion of axes as applicable in principle to anything that can be regarded as an aspect of music, and axes are seen as having both a hermeneutical role as a visualization and understanding of an aspect, and a role in generation/simulation (*ibid.*: 186-87). While the latter is not an issue in this context, Godøy's view of the generality of axial representation in representing aspects along a continuum seems to fit perfectly with my model, and his emphasis on the hermeneutical role that allows for

[159] A true dimensional model in the mathematical or physical sense would require full independence between the dimensions/axes of the model.

understanding and visualization is also very much in line with my intentions of using such a representation.

Perhaps the most important point of these axes is that they allow for *comparisons*, both between two or more different segments of music and between single segments and hypothetical values along an axis: It is by seeing musical segments, both actual and hypothetical ones *in relation to one another*, that the possibility of knowledge and understanding lies, according to Godøy (*ibid.*: 190). Thus, by assigning a segment of music to a value along one axis, one has the possibility of seeing this segment *in relation to* the other values on the continuum, for instance the maximum or the minimum, as well as the possibility of seeing it in relation to another segment of music evaluated with a value along the same axis. Hence, axes in Godøy's sense are tools for relational thought, rather than an absolute mapping of exact values where the assignment of a value along an axis for a musical segment constitutes an act of relative comparison. This value might be graded in different resolutions, from coarse (high-medium-low) to fine (1-20), depending on the possibilities for differentiation along an axis (*ibid.*: 146). In my framework, I will largely apply a relatively coarse resolution with the continuum into divided into five categories, of which two are the maximal and the minimal poles, one is at the intermediate position, and the remaining two are between the intermediate and the maximal and minimal poles. Thus, the five categories are:

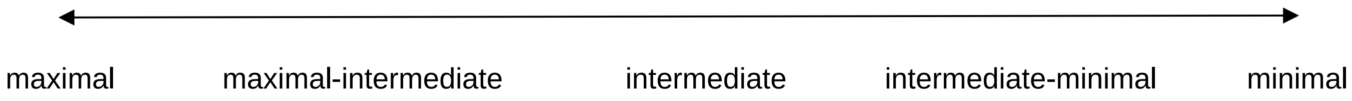


Figure 4.4: Five value categories along the maximal-minimal continuum.

In principle, however, one can choose other resolutions depending on the degree to which it is possible to make distinctions and comparisons. In the following seven chapters on the premises of the model, I will present evaluation criteria relative to the mentioned five categories with representative examples of the different steps along the continuum. As I will demonstrate in chapter 12, if these evaluations are made along all axes and mapped onto a representation such as the one in **figure 4.3**, it will give a representation that allows for comparisons among different musical segments for many aspects simultaneously, and give an interesting overview of the relationships among the premises for each segment of music.

One important difference between **figure 4.3** and **figure 4.2** that must be noted is that I have removed the “non-voice” label. This is done because not all of the premises can transcend from the minimal voice to non-voice, consequently making the questions of boundary transgressions more complex and more difficult than for the simple centre-periphery model. Another thing that adds to this complexity is that there is more than one way of conceiving of the boundary between voice and non-voice. I will therefore go deeper into the questions of boundaries in the following section, also relating that question to theories on categorization.

4.7 Boundaries of the voice

One central question regarding the boundary between voice and non-voice is whether a transition from voice to not-voice or vice versa, like I described for Viñao’s *Chant d’Ailleurs* above, is gradual or more abrupt. I want to start the discussion of this issue by turning to my own experience. On one side, I have experienced that such a transition has happened relatively abrupt, almost like a sudden revelation. This happened the first time I listened to a particular section of Decoust’s *Interphone* (1977, on Various artists, 1989, **sound example 4.3**, 1:43-3:00). Here, a pitched, repetitive sequence lasting about 10 seconds can be heard, and as commented in section 3.6.4, it was at first undefined in terms of source whereupon it gradually expanded in spectral width until it suddenly was apparent that I was listening to a voice. On the other hand, I have experienced when experimenting with LPC source-filter decomposition of vocal recordings, that when a buzz source is controlled by the fundamental frequency and amplitude extracted from a recorded speaking voice, it will take on the characteristics of speech, even if it is clearly a synthetic sound and not “proper” speech.[160] At this point, when a filter is gradually applied, the sound gets gradually more and more speech-like without the sudden “leap” from being not-voice to being voice.

As for the first case mentioned, this resembles what has been labelled *categorical perception*. Categorical perception refers to the perception of sensory phenomena into different categories, and is usually opposed to continuous perception, in which sensory phenomena are located along a gradual and smooth continuum. One classical example of categorical perception is in colour perception, where differences between e.g. reds and

[160] A buzz source is a complex, pitched sound, often with as much as 50 evenly spaced harmonics, in which the harmonics decay progressively towards the upper frequency ranges. This kind of sound is often used as source sound in source-filter synthesis because of its rich spectral content.

yellows look much smaller than equal-sized differences that cross the red/yellow boundary (Harnad, 1987: 535).[161] Hence, even if the frequency of the light changes at regular intervals, the perceived changes will be greater when a category boundary is crossed – gradual changes will be translated into discrete categories, as a kind of analogue-to-digital conversion. This can also be expressed in terms of an *either-or* logic: *Either* something is yellow *or* it is red. As for the second case, however, there is no sudden change between two categories, and consequently it can be described as having so called *fuzzy* boundaries, or as belonging to a *fuzzy set*.[162]

Before I can discuss how these boundaries relate to my model, however, it is necessary to relate the question of boundaries to the seven premises. As I see it, only some of them represent possibilities of entering into non-voice. Firstly, I will anticipate that it is first and foremost the *naturalness*, *presence*, and the *feature salience* premises that allow for transgressions of the boundary into non-voice. Minimal presence implies *absence*, and minimal feature salience as well as naturalness implies that no features of the sound are heard as belonging to a voice anymore – hence, all premises can enter into non-voice. The focus *of attention*, *information density*, *clarity of meaning* and *stream integration* premises, however, are more difficult to conceive of as transcending the boundary of the minimal voice into non-voice.

To clarify some issues of boundary transgressions, I will use the premise of *naturalness* in the discussion. This premise deals with the attribution of properties to a human vocal apparatus. However, experienced naturalness need not be an all-or-nothing matter, since one can often experience in electroacoustic music that only some properties originate from a voice, whereas others do not. For instance, it is possible to separate the *source* and *filter* components, described in section 3.3.3, so that one can keep only the information about fundamental frequency, amplitude variations, or the characteristics of a filter corresponding roughly to the articulatory organs. In that way, one can hear sounds that are perceived as partly originating from human vocal apparatus and partly from something else, implying that the categorical boundaries are fuzzy rather than all-or-nothing. Such fuzzy boundaries also fit better with experiences of sounds that are heard as having only faint resemblance to the voice,

[161] For overviews of research on categorical perception in auditory perception, speech and music see Handel, 1989, chapter 9 and McMullen & Saffran, 2004. An idealized depiction of categorical perception is also given in Harnad, 1987, p.55.

[162] Lotfi Zadeh devised a form of set theory to model graded categories which he called *fuzzy set theory* (Zadeh, 1965). The two mentioned types of category boundaries have also been expressed in mathematical *set theory* terms, where the categorical boundary case is named “classical” or “crisp” set.

e.g. in possessing a vowel-like quality, a speech-like intonation curve, etc., without any other properties common to vocal sound. Hence, seeing the maximal-minimal voice model as consisting of fuzzy boundaries, seems to fit well with my experience of a gradual transition from faintly voice-like to "just like" a human voice.

Such faintly voice-like sounds might not be categorized as voice at all. Consider the sound example from Lars-Gunnar Bodin's GYZO (1:00-1:25, **sound example 4.4**, from Bodin, 2006). Here, the single pitched sound in the foreground that enters after about 5 seconds is clearly a sound with the characteristics of electronic origin. Still, it has an intonation curve and a spectral evolution which resembles a speech gesture to some extent, almost so that one can transcribe it phonetically. Thus, this segment is "speech-like", or more generally "voice-like", while maintaining its source identity as an electronic sound, i.e. as *not* belonging to the category of "voice" since it is not perceived as emanating from a human vocal apparatus. Hence, one can see that "likeness" can be used to compare the properties of two things even if they belong to different source categories. Similarly, a chimpanzee can for example be regarded as having many human-like qualities, such as a human-like physiology with hands with human-like palms and five fingers, non-verbal communication with hugs, kisses and pats on the back, and so on. In the same manner, in referring to the abstracted properties of something, e.g. through transforming categories into adjectives like "humanness", or by referring to similarity, for example using "likeness" or "like", one can introduce fuzziness. As in the example of chimpanzee being included in the fuzzy category "human-like", "fuzzifying" a category can extend the range of members it can include.[163]

As for the experience that indicated categorical perception discussed above, it can be illustrated with an adapted version of Harnad's idealized representation given in **figure 4.5**. Here, the uni-dimensional change in the sound stimulus is given on the x-axis, and the response is given on the y-axis, in which we note the break at the boundary between the category "voice" and "not-voice", in the sense of being experienced as having a human vocal source or not. What Harnad calls "output of iconic representation" would in this case represent our perception of the gradual changes in the sound, whereas the output of the categorical perception obviously has a discontinuous change from one to the other. What

[163] This can also be compared to so-called *hedges* in linguistics, originally defined by Lakoff as "words whose meaning implicitly involves fuzziness – words whose job is to make things fuzzier or less fuzzy" (Lakoff, 1973: 471). By introducing the *hedge* "sort of", Lakoff introduced a fuzzy boundary for the category "bird", which in other cases could be regarded as having a clearly defined boundary (cf. the citation above on the category boundary of birds (Lakoff, 1987: 56). He could then construct a "birdiness hierarchy" in which not only prototypical birds and less typical birds were included, but also bats, which would probably not have been included in the category "birds" without the hedge.

Harnad calls categorical perception (CP) “boundary bias” or “boundary effect” does not represent the categorization in itself, but rather the perception of similarity as given on the

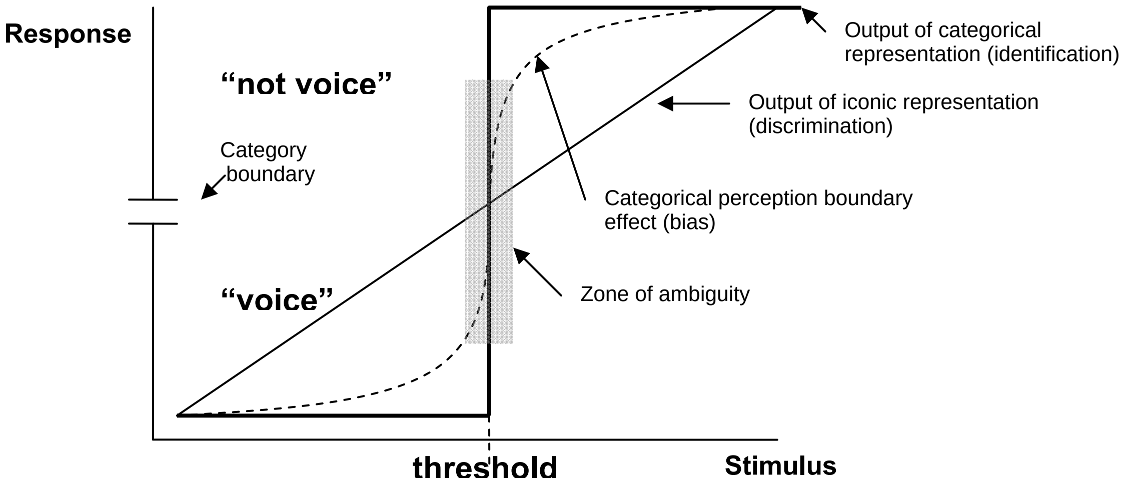

Response
“not voice”
Category boundary
“voice”
threshold
Stimulus
Output of categorical representation (identification)
Output of iconic representation (discrimination)
Categorical perception boundary effect (bias)
Zone of ambiguity

Figure 4.5: Adaptation of Harnad’s idealized depiction of categorical perception. Here, the input, the stimulus, is given on the x-axis, and the output or response is given on the y-axis. What is labelled the *iconic representation* is regarded as an analogue of the sensory input, proportional with changes in the stimulus. The categorical representation is represented with the bold line is an all-or-none function that determines the category boundary between the categories, here specified as “voice” vs. “not-voice”. The dotted line shows what Harnad calls the categorical perception (CP) boundary effect: the biasing influence of the category boundary on perceived similarity. The zone of ambiguity is not given by Harnad. (After Harnad, 1987: 555).

response axis so that similar values represent perceived similarities. As one can see, when tracing the dotted line from left to right, the values change only marginally at first, designating a relatively high degree of similarity. When it approaches the threshold, the values change dramatically, indicating a perceived difference rather than similarity, ultimately leading to the “leap” when passing category boundaries, such as the one described above in Decoust’s *Interphone*. However, one can imagine a zone around the threshold at the x-axis and at the category boundary at the y-axis in which the decision is more likely to be of a more random nature, thus representing what I called “zone of ambiguity” above (this is the grey square in **fig.4.5**). Despite the idealized nature of his depiction, Harnad admits this ambiguity when stating that there is “[in] reality […] always some variability at the boundary” (Harnad, 1987: 555). Hence, even when maintaining strict category boundaries, one has to consider that when approaching some category related threshold, there will be a zone of variability that affects categorization. In a temporal transition from one to the other in the form of a continuous process, this phase of ambiguity need not be felt that strongly because the temporal continuity of the process will entail a sense of constancy due to scene analysis principles of sequential integration: “The sound coming from [a single sound-producing event] tends to change continuously rather than abruptly in its property as the event unfolds” (Bregman, 1993:18-19). Thus, even when one perceives a change in the sound, the continuity of the change will imply a constant source, delaying, in a sense, the potential ambiguity imposed by the changes. When the changes have come to a point at which another source category can be recognized, this can be felt as quite an abrupt change resembling the idealized situation in **figure 4.5**.

To sum up my discussion on the boundaries of my model, which have been based on the premise of naturalness, we have two ways of regarding the category boundary, in which the second extends the range and “fuzzifies” the category relative to the first: (1) In a stricter sense based on sound source identification on an all-or-nothing basis; (2) In a fuzzy sense based on abstracted properties (likeness).[164] And, as I have argued, the fuzzy category will extend the boundaries in (1). Taken together, I end up with a dual notion of the category boundary. The model can thus be illustrated with a clear boundary contained within a fuzzy boundary that continues gradually into not-voice, as can be seen in **figure 4.6**. Here, the rightmost end of 1) indicates what is *just* beyond the minimal voice, and going left from this end, one would enter into the zone of the minimal. This zone continues into and through the zone of ambiguity in 1), whose leftmost end coincides with the end of the minimal. The fuzzy continuum is less relevant for the boundaries of my model, but it still represents a possibility for making comparisons regarding the degree to which a sound resembles a human vocal apparatus. Hence, it can be used in the evaluation of the naturalness premise, something I will elaborate on in section 7.3.

[164] If we regard Bruno Bossis’ definition of “vocality” that he applies in his study of artificial vocality, it seems to be based on properties or qualities and thus fit with this second way of defining categorical boundaries: “[…] la vocalité est la qualité vocale d’un évenement sonore. En consequence, cette qualité révèle une analogie perceptive avec la voix naturelle” (Bossis, 2005: 8)

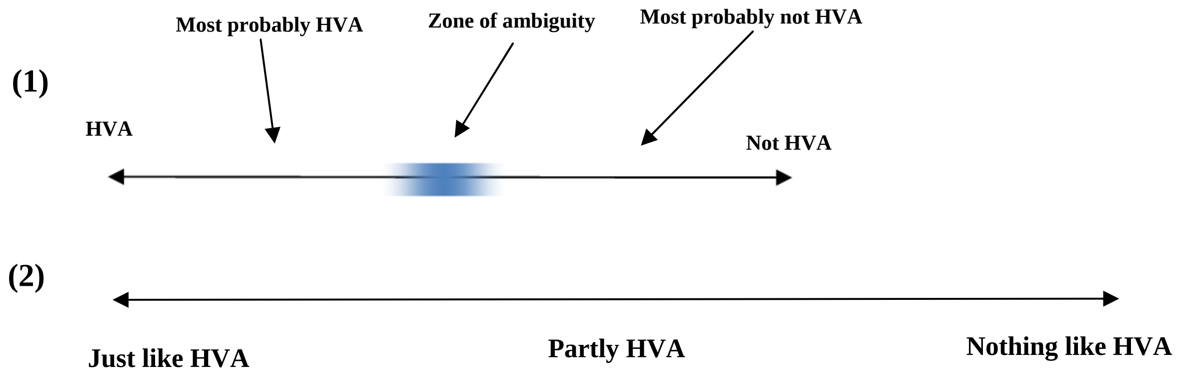


Figure 4.6: Dual notion of the boundary between voice and not voice. (1) represents the source related clear boundary with a zone of ambiguity between the end points, representing maximal certainty of category decision between *human vocal apparatus* (HVA) and not *human vocal apparatus* (not HVA). (2) represents the fuzzy boundaries related to abstracted properties ranging from maximal likeness to human vocal apparatus (just like HVA), via intermediate percentage values representing what is partly like a human voice (partly HVA) to minimal likeness (nothing like HVA).

4.8 Chapter conclusion and outlook on premise chapters

In this chapter, I have presented the maximal-minimal model of voices in electroacoustic music. This model was based on ideas presented by Wesling and Slawek, but also had similarities with theories of radio as well as theories of electroacoustic music. The main idea with the model was that the experience of voices in electroacoustic music can be regarded in relation to two zones of reference constituted by the maximal and the minimal voice, between which there is a continuum. These reference zones were defined according to a set of premises, where the maximal voice was defined by the convergence of the premises. The minimal voice was related to the premises in a negative manner, at the same time as it defined a boundary zone towards what is not voice. The two zones could be thought of as centre and periphery, respectively, and the premises could be represented as axes running from the centre towards the periphery. This structuring of the model was shown to have similarities with prototypical categories, especially those structured as cluster models. In the light of theories of clear and fuzzy categorization, the boundary zone constituted by the minimal voice was shown to have a dual structure, corresponding to these two types of categories.

The following seven chapters are structured around each of the premises of my framework. In the beginning of each chapter, I will begin with linking the premise to the framework in general, including seeing it in relation to the experiential domains. Then, I will give a theoretical account of phenomena, concepts or features that are at the core of each premise, thus hopefully substantiating and clarifying the foundations of the premises. Moreover, I hope to show through this account how many of the premises are linked to phenomena and theories of these phenomena that are described within several fields, many of which were listed in section 1.4.1. This will lead to the presentation of a set of factors that the premises are seen as depending on. When it comes to setting up criteria for how the premises can be used in an evaluation, it differs from premise to premise. For some premises, the criteria have been explicitly formulated, for others not. And, for some premises I have had to discuss these criteria extensively, while for others, they have been sufficiently precisely defined by the premises themselves. Anyway, the criteria for each of the premises have then been used in an evaluation of a number of sound examples extracted from electroacoustic works or pieces from the related genres mentioned in section 1.3, so as to demonstrate in practice how the evaluations can be made. For many of the premises, I have also compared this with an evaluation of the factors that were seen as potentially contributing to the evaluation of the premise. Taken together, the following seven chapters therefore give a thorough explication of premises of the framework, which ultimately will be applied together in the evaluation of *Six Fantasies* in chapter 12.

5.0 Focus of attention

> **Premise one of the max-min model:**
> *Linguistic-semantic focus of attention*: The semantic level of the linguistic domain receives sustained and maximal attention.

In this chapter, I give a theoretical account of the importance of attention for listening and link this to my concept of the maximal and minimal voice. Then I delineate a set of configurations for how attention can be directed towards different domains and combinations of domains, and see how this can be mapped onto the continuum between maximal and minimal voice. Finally, I exemplify some of these configurations in a discussion on excerpts of a few electroacoustic works.

Attention is crucial in listening. Sounds that receive our attention will likely make the most important contribution in our experience of these sounds. On the other hand, sounds that are not attended to will most likely leave only a faint imprint, or none at all. The current premise refers to how the maximal voice requires a maximal and sustained focus towards the semantic level of the **LI-domain**, which I have abbreviated **LI/sem** (cf. section 3.5.1.4). In other words, this premise implies that as a listener, one pays full attention to "what is said", rather than for example the quality of the voice in and of itself or the background ambience that signals the environmental localization of the speaker. This is clearly in line with Dyson's emphasis on the informational content, "what is said", of the radio voice (Dyson, 1994). The premise of focus of attention also implies that if the listener's attention is attracted to any of the other aspects of loudspeaker mediated voice than **LI/sem**, it will entail a displacement towards minimal voice. Additionally, I will see the directing of one's attention towards those domains that I have designated as "non-vocal" (**SQS-**, **TCM-** and **SE-domain**) as implying an evaluation more towards the minimal than if attention is directed at the remainder of the "vocal" domains (**LI-domain** except the semantic level, **AF-**, **ID-**, and **VG-domain**). But, before I present how this premise could be used in evaluating musical examples, I will need to clarify several aspects of attention and put forward a set of factors that potentially can affect this evaluation.

5.1 Attention – selection, divided attention and distraction

Let me start by introducing three important aspects of attention.

1) Attention implies *selecting* from the stimuli that reach us, either actively or passively, a smaller subset of the sensory stimulation. In theories of attention, this is often referred to as a kind of *filtering* (see e.g. Driver, 2001; Knudsen, 2007; Pashler, 1999).[165] This selection process is central for the maximal voice in that it favours the linguistic type of processing rather than, let's say, the range of intonation of the voice. Selection can be initiated by our will, i.e. by volition, which consciously guides attention towards certain aspects of the input for some reason. Alternatively, it can be a result of the properties of the sounds that *capture* one's attention without any conscious decision. These ways to direct attention at certain features of the incoming sounds are often referred to as *top-down* and *bottom-up* respectively (Knudsen, 2007), and studied under the labels of *selective attention* and *distraction* (or alternatively *attentional capture*).
2) Attention involves *applying our limited perceptual and cognitive resources* to whatever appears most interesting or useful in a situation. In other words, attention is a question of perceptual and cognitive economy, so that we use our available resources on what is most useful, interesting or relevant (Pashler, 1999; Jones & Yee, 1993).
3) Attention often results in a *facilitation of sensory processing* and a *more detailed awareness* of the objects of attention, something which is also mirrored in increased neuronal activity in the brain (Hugdahl et al., 2003; Bregman, 1990).[166]

In the context of the max-min model, it is especially the first of these points that will be further elaborated, since it is the *selection* of the domains of experience and potential *distractions* from these that will be the most crucial point in the evaluation process. The two other points still constitute important underlying ideas for the following discussion. There are three different configurations of attention in particular that will prove to have importance for the evaluation process related to the current premise, namely *selective attention*, *divided attention*, *distraction/attentional capture.* I will therefore give a short explanation of each of these concepts, in addition to discussing some relevant temporal aspects of attention.

[165] A highly influential model of attention was Broadbent's (1958) *filter theory*, also referred to as *early selection theory* that proposed that all properties of the stimuli reaching the senses were processed, but that mechanisms that functioned as a filter only allowed one stimuli to be processed further. (See Driver, 2001 or Pashler, 1999 for an account of Broadbent's theory).

[166] Knudsen attributes such an improvement in the quality of information to two mechanisms; a) the orientation of the sense organs, in the case of audition this will most often mean turning one's head towards the direction of the sound source, and b) mechanisms modulating the neural circuits that represent the information (Knudsen, 2007).

5.1.1 Selective attention

The term *selective attention* is usually reserved for the top-down initiated part of the selection process. Hence, we apply our selective attention whenever we consciously direct our attention towards one particular sound source, or one particular aspect of a sound. For example, when we listen to music, either at a concert event or at home, we will often have to filter out those sounds belonging to the extrinsic domain (cf. section 2.3.2), like noise from the audience, passing cars, hum from fluorescent lighting, noise from air-conditioning, etc., in order to be able to attend fully to the music. Furthermore, when deliberately directing our attention towards certain aspects of a sound, as when practicing *reduced listening* (cf. section 2.4), one will also apply selective attention to sonic qualities while filtering out any source related aspects.[167] In the current framework, the role of intention or *volition* is seen as a factor potentially affecting attention.

5.1.2 Divided attention

Even if there are apparent limitations as to how much of the incoming stimuli can be processed, there are certain situations where attention can be divided between different aspects of the input. Often, one can easily attend to several features of a sound at a time without giving any special precedence to something in particular. Especially, this can happen if each of the features does not demand too much perceptual and cognitive effort, something I will discuss further below in the section on processing load.

The discussion of divided attention has also been noted in writings about electroacoustic music. According to David Evan Jones, attaining a “dual attention focus” can be obtained by composers by drawing the listener’s attention either toward sound qualities, for sounds which at the outset draw the listener’s attention at phonetic information, or toward phonetic qualities, for sounds which draw attention to sonic properties at the outset (Jones, 1987: 143).

There appears to be some constraints on divided attention as to what aspects or tasks one can divide attention between. Experience from social gatherings, for example, tells us that it is difficult to follow what is said by two different speakers talking at the same time. The famous studies of the so-called “cocktail-party effect” by Cherry pointed in the same direction (Cherry, 1953). In his experiments listeners were instructed to attend to the speech presented

[167] The phenomenological reduction that reduced listening implies can also be seen as a filtering process, thus resembling filter theories of attention.

on one ear, while also being presented with another speech stream on the other ear. It turned out that in most cases listeners were not able to identify neither the language spoken, the individual words, nor the semantic content of the ear that was not attended to.

5.1.3 Distraction / attentional capture

In contrast to selective attention, *attentional capture* is generally taken to be passive, automatic and unaffected by expectancies or conscious decisions (Ruz & Lupiáñez, 2002).[168] A phenomenon often seen as closely related to attentional capture is that of *distraction*, which designates a situation where something draws the attention away from the current task or focus. For example, during larger portions of the writing of this dissertation, I have been located in an office at the floor just below several practice rooms for music performance students, while the room exactly above contained a grand piano. When students practiced on this grand piano I could hear everything they played. As one can guess, this posed problems of distraction in the writing process, since even if I was heavily engaged in the task of writing and thinking, the piano music from above regularly captured my attention, and thus distracted me from my primary task. This was especially a problem when the students played loud passages with lots of heavy accented chords. This caused me to "get out of it" a lot more often and having larger problems with coming "back on track" than when they played softer passages with a more even flow. Not surprisingly, therefore, *loudness* was an important factor as to the degree of distraction – the loud chords "stood out" more, i.e. they were more *salient* than the softer passages. As we will see in the in the next section, both *salience* and several other factors can be shown to have an effect on distraction.

5.1.4 Temporal aspects of attention

The temporal aspect of distraction must also be mentioned, especially since the premise states that the maximal voice receives *sustained* maximal attention, implying that attention remains maximally focused during a longer period of time. For this to take place, however, it is of essential importance that a listener, when first detecting a speaking voice, *initially* pays some attention to aspects regarding the identity of the speaker (**ID-domain**), the surroundings and

[168] This has some resemblance to what Truax calls *listening-in-readiness*, which he describes as a "kind of listening […] in which the attention is in readiness to receive significant information, but where the focus of one's attention is probably directed elsewhere (Truax, 2001: 22). Truax exemplifies this kind of listening with a mother that is woken up by her baby's cry at night while sleeping, but not by trucks or other noises.

the situation that he or she appears to be situated in (**SE-domain**), and the medium that he or she is heard through (**TCM-domain**). Such contextual aspects are indeed important for interpreting the linguistic content appropriately – "what is said" is clearly not the same when these aspects are different (Kappas et al., 1991: 203; Schirmer & Kotz, 2006; 25). Moreover, it also appears that a process of mapping out the range of possible speech gestures (**VG-domain**) is necessary as a framework for interpreting speech, and that once this is done for a certain speaker, the attentional demands on speech perception are reduced (Nusbaum & Morin, 1992). For a maximal voice, the mentioned aspects need only be attended to at the beginning of a session, however, since after this point there will be minimal change and thus minimal need to attend to them.[169] For example, the identification of the most important identity traits of a speaker like gender, age, social/regional belongingness and ethnicity will most often be done in only a few seconds after the encounter with the speaker (Nass & Brave, 2005). After this initial attending, there will be no further need to attend to these aspects, that is, as long as they stay relatively constant. Only when new voices enter do we again need to mobilize attentional resources to repeat the identification/recognition process for a short few moments.[170] Thus, the "sustained maximal attention" that the premise refers to will as a rule only follow the initial process of interpreting the contextual "framing" within which the vocal utterances should be interpreted.

Another temporal aspect of attention concerns the frequency with which distractions occur – the more often something interrupts our focus, the less attention will be devoted to the intended focus of our attention. A higher frequency of distractions will therefore also imply more distraction seen in a temporal perspective. In this framework I will refer to the joint effects of the *amount* of attention that a distracting sound will cause, the amount of *effort* applied in turning attention back to the original focus, and the frequency of distractions, as the *strength* of the distraction, because regarding these issues separately will lead to an unnecessary complication, in my view.

5.2 Factors potentially affecting attention

In the following, I describe several factors that potentially can contribute in guiding attention, inhibiting or provoking distraction and allowing for divided attention. These factors are

[169] Palmeri and colleagues also found that details about how a word is spoken by a specific speaker can be retained in memory for a certain period of time (Palmeri et al., 1993).

[170] See the discussion of the factor *novelty/change* in section 5.2 below.

volition, *processing load*, *same/different type of processing*, *salience*, *relevance/interest*, *novelty/change*, *unpredictability*, and *emotional salience*. The first of these, *volition*, is particularly related to selective attention, *processing load* will have relevance for all terms, and the remaining factors will primarily be related to attentional capture/distraction, but can have relevance for the two other theoretical constructs of attention as well.

- **Volition:** Implies an active intentional attitude that can consciously direct attention towards an event, feature, or object.[171]

- **Processing load:** The perceptual and cognitive processing load applied when attending to a sound segment or a feature of sound may affect the degree to which other sounds or other aspects may cause distraction or allow for divide attention (Wickens, 1991; Muller-Gass & Schröger, 2007; Berti & Schröger, 2003).[172] Generally, attending to something which demands a higher processing load, tends to inhibit distraction more than compared to a lower load. Lower loads will more often allow for both distraction and divided attention. This is related to the limited perceptual and cognitive resources mentioned in point 2) of section 5.1. For instance, it is known that short-time memory, which is believed to be a functional component of attention, can only hold a limited amount of information at a time (Berti & Schröger, 2003; Miller, 1956).[173] Processing load is therefore related to how many units that have to be retained in memory at a time, which in turn might be affected by issues such as speed of presentation and the ability to structure information into larger units. One might assume, therefore, that if the processing load of a speaking voice was lowered through reducing speaking rate or inserting more pauses, the chances that one would be able to process properties of any other sounds present, would increase, and if raised, that the changes would decrease.

[171] Several theories of listening relevant to electroacoustic music have underlined the active and conscious aspects of attention: Pierre Schaeffer's third mode of listening, *entendre*, deals with both *intention* and *selection*: "Entendre, here, according to its etymology, means showing *an intention to listen* [écouter], *choosing from what we hear* [ouïr] what particularly interests us, thus "determining" what we hear […] manifests an intention" (Chion, 1983: 25, translation from the EARS website, URL: http://www.ears.dmu.ac.uk/spip.php?rubrique218, retrieved July 14, 2008, my italics); And with Katharine Norman; "attention is the *state* of applying our mind to, in this case, sounds, intention is the *determination* – an active process – to 'stretch out' towards this state" (Norman, 1996); Barry Truax, on his part, has designated a mode of listening which he calls *listening-in-search* for which "the ability to focus on one sound to the exclusion of others […] is central to the listening process" (Truax, 2001: 22).

[172] Wickens uses the term *processing resources* instead of processing load.

[173] See Knudsen, 2007 for the role of working memory in attention.

This factor is also related to the premise of information density, which will be discussed in the following chapter.

- **Same/different type of processing:** The degree to which two aspects demand similar type of processing might affect the degree of distraction that one aspect asserts on the other or the possibility of divided attention between the two aspects. An aspect that demands the *same* type of processing tends to pose more distraction and inhibit divided attention more strongly than an aspect that demands a *different* type of processing. Hence, when listening to continuous speech, the most serious problems in terms of distraction would probably be posed by introducing another speech stream, something that Cherry's "cocktail-party" effect studies also showed (Cherry, 1953). Dividing attention between tasks that differ more from one another, however, is generally easier.[174] For instance, it appears that attention can be divided between verbal and melodic processing without too much difficulty.[175]

- **Salience**: Attentional capture is often linked to the concept of *salience*, which designates the degree to which properties of a sound will "stand out" for a listener, for instance by being markedly louder than other sounds, or clearly distinguished from other sounds in the temporal or spectral domain (Kayser et al., 2005; Sussman et al., 2003).[176] Salience is discussed in more detail in chapter 10.

- **Relevance/interest:** A factor that appears to be important in capturing the attention of listeners is the potential *relevance* that a sound has for a listener. The most typical example of relevant sounds are our own names, which in most cases will attract our attention or cause distraction if attending to other things (Wood & Cowan, 1995; Holeckovaa et al., 2006). There are also indications that concerns that have personal relevance and that stimuli that are associated with somebody's mood or special interests

[174] For example, Allport found that subjects were able to repeat speech at the same time as doing other unrelated tasks, such as recognizing pictures and sight-reading piano music, just as well as when they were performing one task at a time (Allport et al., 1972).

[175] Bonnel and colleagues found that the ability to detect incongruities in melody and lyrics in operatic songs was not significantly different when the tasks were done independently as when they were done simultaneously, something which indicates independent processing of melodic versus word processing (Bonnel et al., 2001).

[176] The importance of salience on distraction is also acknowledged in the empirical research literature, where studies have indicated that the salience of a distractor, sometimes referred to as the *strength* of distraction, usually understood as the degree of deviation of the distractor compared to the task-relevant stimuli, has correlated with the electrophysiological indications of distraction (Berti et al., 2004; Escera & Corral, 2007).

will have a greater tendency to evoke distraction, and therefore also will more easily attract attention than those with little or no relevance (Gilboa-Schechtman et al., 2000; Compton, 2003). Again, it is apparent how attention is dependent on factors that are individual and therefore might differ from listener to listener.

- **Novelty/change:** When aspects of a sound change sufficiently fast and to a large enough degree, it may lead to attentional capture. For instance, studies showed that when listening radio broadcasts with two voices, listeners exhibited an *orienting response*[177], i.e. they oriented their attention when there was a change from one voice to the other (Potter et al., 1998; Potter, 2000).[178] Also, it turned out that the listeners were less capable of recalling what was said in the few seconds *after* the onset of a new voice than before, suggesting that the change of voice mobilized parts of the listeners' attention so that the processing information from the verbal domain was delimited during this phase of orienting towards the new voice.[179] More generally, acoustic novelty and change tend to attract attention. In particular, the attack and decay phases of a sound tend to attract attention if they imply a major change. Escera and colleagues rapport a number of studies that support this, especially where onsets (attacks) follow a longer silent period, and when offsets (sudden decays) follow a continuous sound (Escera et al., 1998).

- **Unpredictability:** Unpredictable events are more likely to attract attention and/or cause distraction than predictable ones, even in cases where the prediction is based on another modality than the one in which the event itself is taking place (Sussman et al., 2003). Highly predictable events, on the other hand, might lead to a decline of, or in some cases a reorientation of, attention. This can be observed when attending to a looped spoken phrase. While the verbal content communicated by the voice might attract attention at first, subsequent repetitions of a short portion of tend to make listeners attend away from the semantic content, and towards the **SQS-domain** (Deutsch, 2003).

[177] In addition to orientating sensory receptors toward the stimulus, orienting response is associated with several physiological responses associated with attention, like slowing of heart rate and increase in skin conductance and temperature (Lang, 2000).

[178] The rate of change in the high-change condition in Potter's experiment was 0.17 Hz or higher (at least 20 voice-changes during a 2-minute segment).

[179] However, after the first few seconds of a new voice, the verbal content could be recalled better than before the change. It seemed therefore, that the orienting response induced by a voice-change would temporarily increase the information-processing capacity after an initial decrease.

Making referential aspects of a sound redundant by looping was also one strategy Schaeffer applied in attaining a focus towards the qualities of a sound in itself in reduced listening (Chion, 1983: 33).

- **Emotional salience:** Sounds conveying or invoking emotions, including voices, capture people's attention more easily (Compton, 2003; Vuilleumier, 2005). If a voice has a clear emotional content that is mediated verbally or non-verbally, especially for negative emotions like anger or fear, it will tend to be prioritized and receive privileged access to attention and awareness.

5.3 Evaluating the focus of attention premise

After having presented factors that potentially influence attention, it is now time to go deeper into the criteria for evaluating something that is based on the premise of attentional focus. This evaluation process is not straightforward, and raises many complicated methodological issues. The criteria for the maximal voice, maximal and sustained attention towards **LI/sem** ("what is said"), appear relatively clear. I have to stress, however, that maximal attention towards "what is said" does not imply that any of the other domains are completely unattended. More specifically, it seems impossible not to allocate any attention towards features related to the vocal domains, since these features inevitably are conveyed by the same sound source, namely the vocal persona. And, crucial aspects of communication are often carried by the **AF-domain**. The non-vocal domains, however, can remain close to unattended for the maximal voice. Therefore, "maximal attention" in this context, implies attending to the vocal domains in the background, but leaving the non-vocal domains unattended.

Compared to the maximal voice, it is much less evident how attention is configured for the minimal and the intermediate voice. This is to a large degree a product of the fact that two criteria have to be combined in the evaluation of the attention premise, namely what I will refer to as attention *locus* and *level*. Here, *locus* deals with what (group of) experiential domains receive attention from the listener, and *level* refer to different degrees of attention the listener is engaged in:

- **Locus of attention:** Here, I will deal with three different loci: 1) The semantic level of the **LI-domain** (**LI/sem**), 2) the vocal domains (**VG-**, **ID-**, **AF-** and **LI-domain**, except

LI/sem), and 3) the non-vocal domains, (**SE-**, **TCM-** and **SQS-domain**). In the evaluation, attending to 1) implies an evaluation that is more towards the maximal than 2), whereas attending to 3) implies an evaluation more towards the minimal than 2). This reflects that the maximal voice is a source bonded conception, closely related to the vocal persona. When attending to other sound sources (e.g. the **SE-domain**), more abstract properties (the **SQS-domain**) or properties related to another ontological level (processing or organization stages of **TCM-domain**), this represents a further detachment from the sound source of the vocal persona than when attending to features that are more immediately related to the vocal persona. Thereby, one can see a parallel with the reality – abstraction continuum discussed in section 2.4, since a focus on abstract properties (**SQS-domain**) is opposed to the source bonding of the maximal voice also in my model.

- **Level of attention:** The level of attention deals with how strongly attention is allocated to one of the three loci mentioned above, hence implying at the same time to what degree other loci can receive any attention. Hence, distraction and divided attention towards 2) or 3) will mean a displacement towards the minimal, since it implies that the attention level directed at **LI/sem** is lower.[180] Additionally, since there can be different degrees of distraction, the size of the dislocation will be dependent on the *strength* of the distraction, as it was defined in section 5.1.4 above.[181] For example, one can easily filter out a speaker's brief and discrete throat cleansing in between two phrases, but if such a throat cleansing is sufficiently loud, of long enough duration or does not fall between linguistic phrase boundaries, there is a great chance that they will attract some attention and therefore also impose an element of distraction.

In combination, the criteria of locus and level of attention allow for a number of different configurations, which can delineate different locations along the maximal-minimal continuum. To make it easier to apply these criteria consistently, I have arranged a number of different configurations in **table 5.1**. Here, the three loci are given individual columns, and the level of attention is then given for each of these loci for five steps along the continuum

[180] Again, Dyson's account of the radio voice has parallel claims in that she notes how trembling voices, throat cleansing, coughing, sneezing, panting, as well as aspects related to technology and mediation represent potential sources for interference for the verbal meanings conveyed by the voice (Dyson, 1994).

[181] I have to note, however, that the limits of tolerance for distractions can be relative to the listener, the speaker, social conventions and situation.

from the maximal to the minimal. For configurations involving distraction and divided attention, I have specified the strength of the distraction as well as which locus that is regarded as the distracter. It must be noted also that I have included the possibility that one locus can be distracted by an additional stream, i.e. something usually implying another sound source, where the distraction comes from the same kind of locus. For example, if the attention towards **LI/sem** of one vocal persona is distracted by the **LI/sem** of another vocal persona, this is seen as implying an evaluation as maximal-intermediate. Even if there might be possible configurations that are not in the table, I still regard it as covering the most central configurations, and as providing guidelines for how to apply the premise of focus of attention in evaluating vocal phrases in electroacoustic music. I will now exemplify how such evaluations can be made in the following section, in addition to showing how the evaluations are related to the factors discussed in section 5.2 above.

loci **evaluation**	**LI/sem (1)**	**Vocal domains (2)**	**Non-vocal domains (3)**
Maximal	Max attention	Background	Unattended / background
Max-int.	Divided		Unattended / background
	Intermediate strength distraction from other verbal stream (1)	Background	Unattended / background
	Intermediate strength distraction from (2)	Intermediate strength distracter	Unattended / background
	Low strength distraction from (3)	Unattended / background	Low strength distracter
Intermediate	Divided with (3)	Unattended / background	Divided with (1)
	Divided		
	Unattended / background	Max attention	Unattended / background
Int.-min	Unattended / background	Divided	
	Unattended / background	Intermediate distraction from other vocal stream (2)	Unattended / background
	Unattended / background	Intermediate strength distraction from (3)	Intermediate strength distracter
Minimal	Unattended / background	Unattended / background	Max attention

Table 5.1: This table shows several configurations for attending to the three different loci introduced in the text, namely (1) the linguistic level of the LI-domain (LI/sem), (2) the vocal domains, and (3), the non-vocal domains. The table shows the levels of attention for each of these loci (columns), from maximal attention to unattended, and the evaluation according to the premise in five discrete steps along the max-min continuum (rows). Distractions of different strengths and divided attention are also shown in the table.

5.4 Evaluation of musical examples

In the discussion of the premise of focus of attention I have chosen to discuss excerpts from five different electroacoustic works which can exemplify five different steps along the max-min continuum. I will start at the maximal end and work my way through the five examples, of which the last can be placed in the minimal end of the continuum.

5.4.1 Maximal: Åke Parmerud, *Grains of Voices*

I will start by taking a look at the opening of Åke Parmerud's *Grains of Voices* (1995, on Parmerud, 1997), which was the one excerpt I could find that came closest to making me pay maximal attention to the semantic level of the linguistic domains (**LI/sem**) (**sound example 5.1**, 0:00-0:23). Here, one encounters a deep, resonant male speaking voice speaking, or more precisely, reciting a text. The text, which is easily recognizable, is a re-written version of the opening of the biblical Genesis: "In the beginning when God created the heavens and the earth, the earth was a formless void and silence covered the face of the deep. While a wind from God swept over the face of the waters, then God said: 'Let there be sound!'".[182] The slight accent in the English pronunciation points towards a non-native speaker. Some artificial reverberation is added to the voice, and no other sounds can be heard in addition to the voice. Since my understanding of English is fairly good, my attention was dominantly directed at the **LI/sem**, at least at the first time of listening. Some listeners might, as I did in subsequent listening sessions, be drawn to the details of the accent so as to reveal the origin of the speaker. However, the accent is not particularly marked, and the nuances from which it can be recognized are quite subtle, so that its potential for distraction is very modest. We can relate the evaluation of this example to some of the factors introduced in section 5.2:

- **Salience:** Due to the lack of other potentially distracting sounds and the subtle processing that is added (reverberation), the voice is highly salient and the sole focus of attention.
- **Novelty/change:** There are *no* changes or novel aspects that could create distraction/ orienting response in the **ID-**, **SE-** or **TCM-domains** after the initial onset of the phrase.

[182] The re-writing is made on the basis of Genesis 1:1, where the word "darkness" is replaced with "silence" and "light" is replaced with "sound".

- **Processing load:** The processing load is within a range so that it is not particularly prone to distraction: The speech rate is relatively slow, with a generous amount of pauses in between word groups, but still not of such a length that one "falls off" or looses interest.
- **Relevance/interest:** With my personal and cultural background, I experienced the semantic issues in the recited text as having medium relevance for me. On one side, I am brought up in a Christian society and occasionally attend church services, but on the other side, I am an agnostic that does not believe in the Biblical version of Genesis. Being professionally engaged in issues of sound and music, however, the shift from a visual to an aural focus that the semantic content of the text here represents, could potentially have made it somewhat more relevant and interesting for me. As I remember my own first listening to this piece, however, I thought of the text as something of a "cliché".

Even though this example does not display maximal ratings for all of these factors, I still feel that it demonstrates how the maximal end of the current premise can be experienced. As I stated in the chapter that introduced the maximal-minimal model, it is rather rare to find examples of maximal voice in electroacoustic music, and the previous sound example shows pretty much how close one can get within a dominantly musical expression such as electroacoustic music.

5.4.2 Maximum to intermediate: Paul Lansky, *Things she carried*

The first movement of Paul Lansky's *Things She Carried* (1996, on Lansky, 1997) from the work carrying the same name, can exemplify the maximal-intermediate category. I have chosen an excerpt from around the middle of the movement as an example in this respect (**sound example 5.2**, 2:52-3:19). In this excerpt, one hears the voice of an adult woman, probably middle-aged, who speaks in a well articulated manner, easily comprehensible, even for me as a non-native speaker of English. Even if it is not too evident from the start, the semantic dimension that binds the spoken phrases together becomes clear relatively shortly: The text gives an account of the contents of somebody's purse, presumably a woman's, and through that account, the text simultaneously gives an indirect description of the owner of the purse (presumably an adult American woman). Hence, the most intuitive way to listen to this piece would be to focus on the semantic content of the mentioned account. As Katharine

Norman expresses it in her analysis of the narrative aspects of *Things She Carried*, there is an element of radio drama which takes place in the piece, and this drama is the most apparent aspect to focus on once the piece has been "framed" by introducing the speaker, the environment in which she speaks and the title that signals the semantic link between the listed items to come: "we can settle back as the action (even a monologue is active internal dialogue) unfolds before our ears and inner eye" (Norman, 2000: 219). However, the presentation of the voice through what sounds as ringing comb filters along with the accompaniment which sometimes becomes pretty loud, does not allow the listener to focus exclusively on the semantic dimension (as can be the case e.g. in a thrilling radio drama): The ringing filters along with the slight reverberation that is added as the piece proceeds both have implications for the experience of the surroundings of the female vocal persona (**SE-domain**); they hint at a process of manipulation (**TCM-domain**) and create interesting spectrally and harmonically changing textures (**SQS-domain**), all aspects that easily can demand some attention from the listener, as they did for me. Thus, the **LI/sem** will have some competition for attention from several non-vocal domains. To link this to the factors of the premise:

- **Relevance/interest:** The concept of creating a portrait of a woman in sound through an account of the contents of her handbag felt interesting to me. Being married for over a decade I have realized the importance of a woman's handbag, and what it can tell about its owner. This therefore was an incitement to attend closely to the semantic content.
- **Novelty/change + unpredictability:**
    - **LI-domain:** Apart from the phrase "three pens and two pencils", none of the verbal phrases in the excerpt can be heard before this point in the movement, and they are therefore "novel" at this point. Therefore, one does not risk "falling out" because of tedious repetitions.
    - **SE-domain:** There is a subtle change between the reverberant characteristics up to the point of the excerpt and what can be heard in it. For the first two phrases, "keys" and "calculator", one can notice reverberation that was not there earlier, but in the following phrases one can even hear that the reverberation is "coloured", especially when the drone in the background disappears. This change is quite unexpected, since reverberation characteristics

often remain unchanged. One might therefore pay some attention to this aspect at this point in the piece.

  - **SQS-domain:** From the phrase “ticket stubs” and onwards one can recognize that the reverberation is “coloured”, so as to comprise musical chords. These chords then change from phrase to phrase and might capture some attention from the listener.
  - **TCM-domain:** Corresponding to the issues mentioned for the **SE**- and **SQS**-domains, but focusing first and foremost on the technological aspects of these properties.

- **Salience:** The vocal sound is salient throughout the example, even compared to the sustained drone sound that can be heard in the first half of the example. The vocal sound is also much more salient than the coloured reverberation, which is rather subtle.

Taken together, there are two conflicting tendencies here: On one side, the semantic content is relevant and interesting and contains enough novel elements to maintain attention towards the **LI/sem**. On the other side, there are subtle changes that might draw slight attention to **SE**-, **SQS**- and **TCM-domain**s, which all are non-vocal domains. In my view, this qualifies as a mild to intermediate distraction by non-vocal domains, and according to **table 5.1**, then, it seems that this falls into the maximal-intermediate category.

5.4.3 Intermediate: Lars-Gunnar Bodin, *CYBO II*

In the next example, which is an excerpt from Lars-Gunnar Bodin’s text-sound piece *CYBO II* from 1967 (Various artists, 1992, **sound example 5.3**, 04:01-04:19), the **LI/sem** receives even tougher competition from the other loci. In this excerpt, we can hear five unaccompanied spoken verbal phrases, all separated by about a second of silence. All five phrases are very clearly articulated in what in my ears sounds like relatively standard, easily comprehensible, British English pronunciation. The phrases are apparently spoken by a female, but due to what I hear as electronic processing, the quality of the voice and the age that it implies *change* paradoxically during the course of the excerpt. In my view, this is difficult to neglect for a listener, and will probably attract his or her attention. The most important factor contributing to this guiding of attention in this case is precisely novelty/change:

- **Novelty / change:**
    - **ID-domain:** Here, the changes in the voice catch my attention, rather than the verbal contents. During all the verbal phrases but the third one, the voice changes in the course of each phrase, and even if I can recognize that these changes are due to processing, I still cannot help assigning them to the changing identity of the speaker. And in my ears, these changes can mainly be mapped onto the age/size continuum, starting out with a full-voiced middle-aged woman, the voice then changes to that of a younger woman, and in the final two words of the fourth phrase, we can hear the voice of a girl. Thus, even if the continuity in the sound tells me that this comes from one and the same source, the mental image of the identity still changes during the course of most parts of the excerpt. And this change will likely not pass unnoticed; rather the contrary, it will probably catch our attention for two reasons: 1.There is a process of change going on, implying new and relevant information. 2. The unexpected nature of the change: This kind of change has not occurred before within the work, and it is not common to hear those kinds of changes in everyday contexts either.
    - **TCM-domain:** Even if the changes in vocal identity in this case must be assigned to processing, the consciousness of the processing still tends to retreat to the background for me when I am listening, at least for the first four phrases. This may be due to the nature of the manipulation, which sounds like an increase in playback speed, possibly in three or four steps from slower to faster. These steps in themselves are much more abstract than the changes in the **ID-domain**, which correspond to much more differentiated identities. Therefore, the **TCM-domain** will provide less information for the first four phrases, in that the same change happens more or less each time. The last time, however, a new kind of processing enters, introducing a granular type of sound, something which implies new information and the potential attraction of attention.
- **Same/different type of processing:** One also has to address the question as to what degree paying attention to the identity changes will potentially exclude other domains

from our processing, especially the linguistic domain. In other words, how large a part of our attentional resources can be occupied by such identity changes, and what amount of cognitive resources are left for other kinds of processing? In this particular case, I do not experience that attention to the **ID-domain** necessarily will be of hindrance to comprehending the verbal message; the very high clarity of the articulation, the long pauses in between the phrases, the lack of any interfering sounds, the overall focus of semantic contents of the verbal phrases in the piece, and the semantic link (admittedly not crystal clear, but still present) between the phrases and the overall theme of the piece, here contributes so that attention is at least shared between the **LI-** and the **ID-domains**.[183] But I do experience that attention is *drawn* towards what is going on in the **ID-domain**, and that even if trying to focus exclusively on the verbal content of these phrases, the changes in identity are difficult to ignore.

To sum up and relate this explicitly to the attention guiding factors:

- **Novelty/change:** Changes in the **ID-** and **TCM-domains** attract attention.
- **Processing load:** Not very high for **LI-domain** due to pauses.
- **Salience:** (**LI-domain**) High clarity of verbal features, due to lack of interfering sounds.
- **Relevance/interest:** The semantic link between the excerpt and the preceding verbally mediated content retains an interest in the **LI/sem** .

Taken together, I would say that attention is shared or divided between the **TCM-**, **ID**, and **LI-domains** in this example, and that this is mainly due to the information present in each of these domains. From **table 5.1**, one sees that such a situation implies an evaluation as *intermediate*.

[183] As the title of the piece suggests, CYBO II deals with the relationship between humans and machines. Among other things, one of the voices in the piece gives a description of a person attached by electric chords to the back of his head to a kind of control box attached to his back. Many of the voices that can be heard in the piece use a highly detached language, so as to allude to a de-humanized presentation.

5.4.4 Intermediate to minimal: Jacques Lejeune, *Messe aux oiseaux*

In the opening of *Christe eleison* of Jacques Lejeune's *Messe aux oiseaux* from 1987 (Lejeune, 2000) there is a passage which can exemplify the intermediate-minimal category for the *focus of attention* premise (**sound example 5.4,** 0:00-1:04). Here, one can hear a texture of several superimposed voices, many of them with different qualities and apparent identities, in addition to two layers of non-vocal sound; one in the high-frequency region, and one in the low. All the voices that one can hear in this excerpt have one thing in common: they all say "Christe eleison". This Greek phrase is known by large numbers of people acculturated in the Christian tradition, since it has been used in the Liturgy of the Mass, as well as in numerous musical settings of that text over the centuries. I was not familiar with the *exact* meaning of the word "eleison" beforehand, however, something I expect is the case with the majority of Norwegians of my generation and younger.[184] The lack of exact semantic verbal meaning along with the extensive repetition (more than 50 times) of the phrase tend very soon to make all levels of the **LI-domain** redundant in this case, hence giving the listener no need to devote any attention to it at all. In other words, there is no change in the **LI/sem** during the whole excerpt. Instead, there are a lot of other things going on that the listener will probably attend to instead, and much of this can be related to the novelty/change factor:

- **Novelty/change:**
    - **VG-domain:** The vocal qualities and the vocal effort are changing constantly between whispering and modal voice and between the medium soft and the almost inaudible.
    - **AF-domain:** The slight differences in phrasing and intensity also indicate different levels of arousal of the speakers.
    - **TCM-domain:** Most of the voices appear manipulated in ways that constantly change. Even if there are recurrent types of manipulation (mostly speed manipulation, amplitude modulation, artificial reverberation or the combination of the three), it is not always easy to predict which of these will occur next (indicating also that the factor of **unpredictability** is high, something which also contributes to guiding attention towards this domain).

[184] I have later come to learn that it means "have mercy".

In addition to this, one could also possibly find several other aspects related to the **SE**-, **ID**- or **SQS-domains** that a listener could find interest in. In any case, attention is shared between the vocal and non-vocal domains, and the **LI/sem** is left unattended or in the background. This indicates, then, that this example is located in the *intermediate-minimal* category according to **table 5.1**.

5.4.5 Minimal: François Bayle, *Théatre d'Ombres – derriere d'image II*

The last example that I will discuss in this section is an excerpt from the second movement of the *derrière d'image* part of François Bayle's *Théatre d'Ombres*, composed in 1988 (Bayle, 1998). In the sound example (**sound example 5.5**, 0:11-1:05), we can hear a sequence in the piece which is dominated by a texture of short vocal sounds. Each individual sound consists of a quite neutrally uttered syllable, which I hear as either [bʌ] or [dʌm], and together these syllables constitute nothing but a play of sound to me, without any connection to specific words. I hear the texture as being a product of transposed versions of these syllable sounds by increasing playback speed (or the digital equivalent of changing sample rate), so as to produce a set of about 6-8 pitches within a span of approximately an octave, at least until the final few seconds where the range extends by several tones. Hence, since all sounds in the sequence are transposed versions of the same syllable, they share vocal identity as well as having been processed in the same manner.

The way in which the sounds are organized appears to be structured, but still with random or quasi-random elements: Through most of the excerpt, I hear the sounds divided into two groups, one dominating each channel, where the group to the left comprises the lowest part of the pitch range and the group to the right, the highest. A few times I think I can hear a third group, but I am not really sure of this. Rhythmically, I hear it as going back and forth between irregular and regular patterns, where there is a slight temporal dislocation between the channels. This dislocation also appears to vary, so that the two groups almost go in and out of synchronization a few times. In terms of pitches, there also seems to be a mixture of regularity and irregularity, with some patterns that are recurrent, but where there are elements that vary in a less predictable manner. As for the spatial layout of the texture, there is very little depth here – the syllables are very much "in the speaker". Still, the play between the right and the left channel brings some spatial interest, although the configuration with one group in each channel remains relatively stable.

The **SQS-domain** is what clearly is the focus of my attention during most of this excerpt. For me, the constant rhythmic play within as well as between the channels, with different degrees of temporal dislocation, combined with the variations in what pitches can be heard, is what attracts attention in this excerpt. The unpredictable nature in terms of rhythmic articulation and the sequence of pitches adds interest. The lack of connection to any lexical units, the neutral emotional content, the use of one single voice as well as one type of processing make the **LI/sem**, **AF-**, **ID-** and **TCM-domains** unlikely candidates for attracting attention beyond background processing. The artificial and highly repetitive organization of the syllables makes me hear the syllables as dominantly disembodied, thus disengaging them from the **VG-domain**. Still, I admit that I at times during listening have projected the sequence of syllables onto vocal apparatuses of two virtual vocal personas, despite the overt artificial character of the syllables, thus experiencing the syllables as *uttered* rather than *organized* by a composed intelligence.

All this can be related to some of the factors for this premise:

- **Novelty/change:**
    - **LI/sem, AF-**, **ID-**, **VG-** and **TCM-domains :** No change
    - **SQS-domain**: Constant changes in rhythmic organization, in temporal dislocation between the two groups of syllables, and of pitch patterning.
- **Unpredictability:**
    - **SQS-domain**: The organization of rhythmic and pitch patterns has elements of irregularity as well as regularity, which constantly produces new configurations that attract and hold attention.

Taken together, I experience that a non-vocal domain, the **SQS-domain**, is what is in focus most of the time for this excerpt. Thereby, it is evaluated as *minimal* according to **table 5.1**. Since I occasionally hear the syllables as uttered, i.e. as vocal gestures rather than sounds, the evaluation at these times will correspond to the *intermediate-minimal* category. The excerpt might therefore be categorized as dominantly minimal, with occasional variations towards the intermediate-minimal.

5.5 Chapter conclusions

In this chapter, I have presented the first premise of the max-min model of experience of loudspeaker mediated voice. The premise states that for the maximal voice, attention is maximal and sustained and directed at the semantic level of the linguistic domain, **LI/sem**, as presented in chapter 3. Through a discussion of several central issues of attention, it was argued that the premise implied selection as well as filtering out other aspects, and that distraction and divided attention were relevant phenomena. The attribution of attentional processes were further refined and presented as a set of factors that could potentially influence attention, namely *volition*, *processing load*, *same/different type of processing*, *salience*, *relevance/interest*, *novelty/change*, *emotional salience* and *unpredictability*. Two sets of criteria were then presented for evaluating the current premise: 1) *Locus of attention* referred to the grouping into (a) **LI/sem**, (b) the vocal, and (c) the non-vocal domains; 2) *Level of attention* referred to how strongly attention is allocated to each of the loci, and whether there were any instances of *distraction* or *divided attention*. These two sets of criteria were then combined into a graphical representation (**table.5.1**) which assigned different combinations of the two criteria to the max-min continuum delimited to five discrete categories. This representation was then used as a basis for the evaluation of five excerpts from electroacoustic works, so as to exemplify the five discrete categories. The above mentioned factors were linked to the evaluation in each case, showing how each of them had contributed during evaluative listening.

6.0 Information density

> **Premise two of the max-min model:**
> *Balanced information density*: The information density of the experiential domains is optimal for the processing/decoding of the **LI-domain**.

This premise is related to the concept of *information* and how this is used in information theory, and especially how this theory has been discussed in relation to artistic practices by Umberto Eco in his book, *The open work* (Eco, 1989). Accordingly, I take information to refer to *something that the listener does not already know* – in other words, that tells the listener *something new*, something that is *added* to one's existing knowledge (*ibid.*: 45). In line with this theory, the term information is also linked to *predictability*, since what we can predict with great certainty really tell us very little. Thereby, high predictability means little information and vice versa. Lastly, I will relate this premise to the limitations of cognitive processing, as they are expressed in Cognitive Load Theory (CLT) and theories of relational complexity (Paas et al., 2004; Halford et al., 1998). Consequently, I will see information as the result of interpreting or coding features of the external world into some kind of stable mental representation which is retained in memory for a period of time.

When I refer to information *density*, it is because I want to emphasize that we are dealing with the amount of information that a listener can infer during a certain period of time. Hence, if a listener gets lots of information in little time, information density is high, if the listener gets less information in more time, information density is lower. When I have formulated the premise so that information density should be *optimal* for the processing/decoding of the **LI-domain** for the maximal voice, this means that information density should neither be too high nor too low for processing the verbal features of the voice. I explain this more in detail in the following.

In the following, I discuss information density and predictability and what I mean by optimal processing. Then, I look at the minimal voice, which for this premise is divided into two separate continua; one towards a mode of the minimal where information density is too low, the other towards a mode where information density is too high. These are labelled the *reduced* and the *noise* mode of the minimal, respectively. Subsequently, I go on to look at a set of factors that can potentially affect information density, before I draw up some guidelines for evaluating the premise. Lastly, I discuss a number of excerpts that can exemplify different evaluations.

6.1.1 Optimal information density for the maximal voice

We have already seen that Dyson's notion of the radio voice, which in many ways corresponded to my notion of the maximal voice, is a source of information and meaning: "Most of what it says is perceived by the listener as factual and informative […] it produces full and meaningful sentences, it says something (Dyson, 1994: 167). To be "informative" in this sense implies that there has to be something new in *what* is said, in other words in the **LI/sem**. Still, there cannot be too much information presented during too short time either. If so, it will be hard to follow and structure all of it. Therefore, the information has to have some *redundancy*, i.e. it has to be reiterated or reinforced in some way, at the same time as some components have to constitute what is new. That information generally has to be *balanced* to be maximally meaningful is also noted by Stéphane Roy in his discussion of Eco: "Thus, depending on the context and the type of message, an optimal degree of redundancy and information (neither too much nor too little) produces a maximal level of signification" (Roy, 2003: 268, my translation).[185] For Roy, as for Eco, *redundancy* is something which is opposed to information. When something is redundant, i.e. it is reiterated or reinforced in some way, it doesn't represent something new, and consequently it presents no information.

As for the **LI-domain**, which is where the critical information is taken to reside according to this premise, I have already pointed to top-down processes on several levels – the phonetic, syntactic, grammatical and semantic – that provide the basis for making predictions in terms of what is likely to come at each moment so as to reduce potential ambiguities created by noise and degraded utterances (cf. section 3.6.1 and 3.6.2). These processes can also be seen as affording redundancies on these levels, since they restrict the number of likely alternatives for interpretation. However, the other domains will also be important here. In particular, the **AF-domain** can present some information for the maximal voice, but only in as much as it supports or emphasizes the verbal aspects. As for the **ID-**, **SE-**, and **TCM-domains**, they will ideally only present information for a listener at the *onset* when the vocal persona starts to speak in the short initial phase when a listener has to construct a contextual framework for interpretation of the verbal content (cf. section 5.1.4). But after this initial phase, any new information in these domains might be disruptive for verbal processing, and therefore it will have to be kept to a minimum. As for the remaining domains (**VG** and **SQS**), they will ideally have to stay unattended in the background for the maximal voice, and therefore will represent no information. The lack of attention directed

[185] French original: "Ainsi, selon le contexte et le type de message, un degré optimal de redondance et d'information (ni trop, ni trop peu) produit un niveau maximal de signification".

toward these domains will be dependent on their redundant behaviour, however. Lastly, for the maximal voice, the information in the different domains will have to be consistent and coherent between domains so as to create redundancies that ease interpretation, rather than make it more complex. Contradictions, such as when a male voice states that “I am a pretty woman” or when an adult person cries like a baby, would require a somewhat higher level of interpretation, and therefore represent higher information density.[186]

It is important to note that the balance in information density is relative to the listener’s competence and knowledge. The same information can be too difficult to process for one individual, whereas for another it can be processed at such an ease that it becomes trivial and uninteresting. The relationship between one’s cognitive resources and available information is a central issue in the so-called cognitive load theory (CLT), and here it is precisely an *alignment* of these two that will be optimal for complex cognitive tasks (Paas et al., 2004). But it is not only cognitive overload that causes degraded cognitive performance; according to Paas and colleagues, the general view within cognitive load theory is that both cognitive overload and an *excessively low cognitive load* will cause lowered performance (*ibid.*). Therefore, if one feature of a sound is presenting no new information to the listener, for instance in a static, continuous sound, this tends to affect cognitive processing negatively, for instance in that distractions will more likely occur (Lavie, 2005). In certain cases, where information density is below this zone of alignment, an increase in information can actually improve performance. This was the case, for example, in an experiment by Potter and Choi, showing that listeners had improved memory for excerpts from radio broadcasts that were more structurally complex (Potter & Choi, 2006). That humans have a an ideal range of information density, is also in line with Lieberman, who sees the encoding and decoding of speech as balanced through evolution – human beings have simply evolved perceptual systems that can cope with a density of information corresponding to what one is able to produce (Lieberman, 1991: 45, 57-59).[187] We will see below how too low and too high information density is related to each of their modes of the minimal.

[186] In most cases this will create a need for interpreting the verbal content as something else than a matter-of-factly proposition. It will then have to be interpreted so that the words stated by the person who speaks do not refer to the person that is verbally referred to, for instance as in reading or quoting another person’s statements. According to Nass and Brave, inconsistencies between linguistic and emotional content involve more brain regions than when they are consistent (Nass & Brave, 2005: 87-88).

[187] However, it must be noted that speech can be comprehended at rates faster than normal. In comparison with Warren’s average rate of normal speech at 10 phonemes per second (Warren, 1982), Orr and colleagues found that with training, their subjects could comprehend speech at a rate of up to 30 phonemes per second (Orr et al., 1965). Foulke and Sticht reported that subjects could be trained so as to partially comprehend speech up to four times the normal rate (Foulke & Sticht, 1969). As for speech production, auctioneers are examples that speech rate also can be increased with training.

To sum up how information is configured for the maximal voice:

- **TCM-domain**, **SE-domain**, **ID-domain**: Information is dominantly conveyed at the *onset* of the first phase of the vocal persona, and after this, features remain relatively constant, hence implying no new information.
- **LI-domain**: Information density is *balanced* so that it is optimized for perception and cognitive processing.
- **AF-domain**: Information density is low, and any information is experienced as supporting or strengthening the information in the **LI-domain**.
- **SQS-domain** and **VG-domain**: These domains are unattended and therefore do not represent information.
- The information across the different domains is redundant rather than inconsistent.

6.1.2 Two modes of the minimal

Since for this premise the maximal voice represents something between maximal and minimal information density, it is possible to envisage two diametrically opposed poles on each side of the maximal.

On one side, one has the possibility of a situation where information density is at its minimal, that is, where almost no information can be inferred from the sound at all – in short, nothing new happens. This corresponds to what I would like to call *the reduced mode of the minimal voice*. If we see this in relation to the configuration of information for the maximal voice above, there is already little or minimal information in all domains but the **LI-domain**, implying that it is mainly in this domain that information can be reduced. At the very minimal configuration for the reduced mode, one will therefore have a minimal information density, which is equivalent to maximal degree of predictability, in all of the domains, including the **LI-domain**.

On the other side, there is the option of having so much information that it is difficult to process all of it; there are simply too many new things happening at a time. A situation where so much information is presented during a certain period of time that most of it escapes us because of our limited perceptual and cognitive resources, is what I would refer to as the *noise mode of the minimal voice*. As we will see below, however, there is a possibility that such a situation can approach a noisy percept which has lost all connection to voice – i.e. it can turn into non-voice.

Before I go on to discuss how these modes of the minimal will be used in evaluations of musical excerpts, I will discuss two issues that are closely related to information density, namely predictability and complexity.

6.2 Predictability

I have implied so far that predictability and information are negatively correlated – what we know for sure will happen represents little information, and what we regard as having very low probability for happening represents much information. This is in line with some of the basic ideas of information theory, for which information can be measured and expressed as the predictability of an event (Fiske, 1990: 11).[188] Compared to classical and popular musics, in which rhythmic, melodic, harmonic and formal patterns provide a basis for a relatively narrow range of predictions, electroacoustic music seems to offer less in terms of predictability of different cues. In a great number of electroacoustic works, where there are no cues related to melody, harmonic relationships, metric rhythm and large scale form, cues related to sound sources and causes we can link to everyday experiences still provide an important basis for making predictions. The technology involved in the composition of the works can also afford such a basis. Taken together, we can see that many of the aspects that were discussed in chapters 2 and 3 on the experiential domains can be related to the process of making predictions of the range of sounds that can occur, and how the sounds can follow each other. Admittedly, these predictions can be a lot more uncertain than in a classical sonata or a pop-song, however.

There appears to be one paradoxical aspect to the relationship between information and (un)predictability: The more unpredictable an event becomes, the more information it represents. Still, a totally unpredictable sequence of phonemes will no longer appear to represent information at all, it will be experienced as merely random and "uninformative". It turns into what Eco has pertinently described as "undifferentiated chaos" (Eco, 1989: 65). Eco touches briefly upon the possibility that there might be a threshold or breaking point where

[188] Within information theory the relationship between information and predictability has been formalized and given a mathematical expression, rendered verbally by Eco as follows: "*the quantity of information conveyed by a given message is equal to the binary logarithm of the number of possibilities necessary to define the message without ambiguity*" (Eco, 1989: 46). Or, in mathematical terms it can be expressed as: $i = \log_2 (x/y)$, where x is the probability that the outcome is known *after* receiving message (known as the *order* usually 1), and y is the probability that the outcome is known before receiving it.

information increases with decreasing predictability and where it starts to decrease towards "undifferentiated chaos". As I see it, it is more rewarding to discuss this threshold or breaking point in the light of *complexity*, which is related to the concept of information and predictability. Thus, for now I will merely suggest that there exists such a breaking point, and that the conditions of this breaking point will be the subject of a more detailed explication in the following section.

Lastly, I will add one important point made by David Huron, who has investigated the role of expectancy in music, mainly Classical Western, in his book *Sweet Anticipation – Music and the Psychology of Expectation* (Huron, 2006). One of Huron's points is that predictions in music deal not only with *what* will happen, but *when* it will happen (*ibid.*: 175). Predictability thereby has a *temporal* side to it, which is highly important in music. A harmonic cadence consisting of tonic – subdominant – dominant, for instance, creates expectations of a resolution to the tonic on a following downbeat. For the voice, such temporal expectations are also relevant, for example, when an audible inhalation make us expect that a vocalization will follow it right afterwards.

6.3 Complexity

The *complexity* of sounding events is something that can affect the experienced information density. For example, syntactically complex sentences take more time to process, have more errors, and are therefore more difficult to process than simpler ones (Wingfield et al., 2003; Gibson, 1998). In such cases, it seems that higher complexity implies higher information density. We shall see below, however, that this is only the case up to a point.

Since information that is carried by loudspeaker mediated voice tends to involve interaction between several domains, features and elements, I would like here to apply the term *complexity* to mean *relational complexity* in the sense proposed by Halford and colleagues (Halford et al., 1998). According to these writers, relational complexity is primarily dependent on the *dimensionality* of a relation, which is equivalent to the number of variables in the relation, where complexity will increase with the number of dimensions. A relation can be of any sort, from the relation between money, state of hunger and the choice of restaurant to the relation between mother and child, and it can be mathematically as well as linguistically expressed. A binary relation can for example be expressed as "cat is bigger-than mouse" or 5>2. Here, the relation is in both cases "bigger-than" and the arguments, whose slots in the relation are equivalent to dimensions or variables, are (cat, mouse) and (5,2) respectively. A proposition such as “John played cricket at the oval on Sunday”, for instance, is seen as a relation with four dimensions, where the activity “play” is the relation and (player, game, location, day) are the four dimensions which are filled with the four values or attributes (John, cricket, oval, Sunday). Such values or attributes can both be entities and other relations.

However, the number of dimensions that can be processed in parallel is not unlimited. According to Halford, there is an upper “soft limit” of four dimensions or variables that can be processed in parallel, and it is this limit rather than limits related to the amount of information in bits (which depends also on the number of possible alternatives for each dimension) that is seen as relevant for relational complexity (*ibid.*). The authors regard this as a parallel to Miller’s classic proposal that the human limitations of processing parallel “chunks” of information extension lies approximately around seven, even if their number represents an adjustment downwards on the basis of more recent empirical findings (Miller, 1956). The dimensions in the framework of Halford and colleagues correspond to some degree with Miller’s concept of a “chunk”, i.e. a unit of information with arbitrary size, in that both represent varying amounts of information. In line with Miler, the latter authors also see the processes of reducing processing load through segmentation and “chunking” as important in establishing the effective cognitive load, where “chunking” implies a re-coding of information consisting of many chunks into fewer higher-level chunks with more information per chunk. Hence, if one can re-code smaller entities, relations or dimensions into larger entities, relations or dimensions, complexity as well as processing load will decrease. This process can be seen as a form of learning that requires time.[189]

It is worth noting that certain relationships are easier to chunk than others. I would here like to mention some factors which will normally ease chunking operations:

- **Event hierarchies:** Hierarchical structuring of a series of elements tends to ease perception, encoding, organization and remembering of information (Bigand, 1993: 255; McAdams, 1989).
- **Segmentation/grouping:** Temporal discontinuities and internal similarities can form a basis for segmenting a temporal succession of events into groups (*ibid.*:247-49; Godøy, 2006). Each group can then be processed as one chunk.

[189] See also Paas et al., 2004.

- **Regularities:** A high degree of regularity will generally imply low complexity (Tononi et al., 1998). Temporal regularities such as beats and loops can contribute in making events easier to chunk.[190]
- **Learning/familiarity:** As soon as a way of chunking information has been learnt it can be used in other contexts to reduce the processing load. When trying to memorize random sequences of letters, for instance, having learnt how to encode binary numbers into numbers of the decimal system, will help reduce the number of elements to remember (Halford et al., 1998: 810). Familiarity with a certain type of structure might also ease chunking. The sequence GTIRAFFBICAA, for instance, might be easier to remember and recall if it is segmented into four familiar abbreviations/letter combinations GTI (as cars with a sporty appearance are often labelled), RAF (Royal AirForce), FBI (Federal Bureau of Investigation), and CAA (Canadian Automobile Association).

All these factors can contribute to chunking, and the reduction of cognitive load can therefore be seen as negatively defining complexity.

There is one last point about complexity that has to be discussed: Increased complexity can induce an analogous situation to what Godøy refers to as a change of *resolution* in listening: "The notion of resolution enables a *hierarchical understanding* of musical substance by seeing a musical object in most cases both as consisting of possible *sub-objects* and as something that might possibly be included in a *supra object*" (Godøy, 1997: 73). With a complex and dense texture with several layers of voices, one can easily be directed towards what Godøy here refers to as the *supra object*, which in this case can be regarded as a more comprehensive and over-reaching perspective where global relationships are more important than the behaviour of single objects or events. If the complexity on the sub-object or object levels is sufficiently high, one can therefore be directed towards choosing what we might refer to as a "coarser" resolution, where the level of complexity is more in line with what we are able to process.

The "forced" change of resolution when complexity reaches a certain critical level can imply that the relationship between complexity and information density is reversed: At one point, the complexity can get so high that one instead will have to focus on supra objects which have *more* to offer in terms of predictability. Take the sound from a crowd of people

[190] In Western classical and popular music, temporal regularities often constitute highly hierarchized metric structures (Bigand, 1993: 251).

engaging in conversation in a big indoor hall, for instance, which takes on the quality of murmuring noise when heard at some distance. This murmuring noise will have a rather uniform timbre and be quite predictable as a whole, even if there will be constant irregularities on a smaller scale. Much of this uniformity can be explained by the frequency range of the human voice and the reverberant characteristics of the hall. Also, if one mixes a larger number of voice recordings of a single vocal phrase with random onsets, the result will come through as noise. In **sound example 6.1** I have mixed as many as 1000 versions of one or several vocal recordings on top of each other, first with a single vocal phrase (male speaking voice), then with 10 different spoken phrases by the same male voice, and finally with 10 different spoken phrases by a female voice.[191] For all three sounds, the time values which designated playback start time and amount of time skipped in the beginning of the sound file, were randomized. As one can hear, the result is three compact and noisy sounds, where the vocal origin is barely recognizable except from the starting and ending phases of each sound. Still, the differences between the three sounds are clearly noticeable. The first sound is a lot darker than sound number two, which in its turn has brighter and darker components than the third sound. Moreover, one can, despite the noisiness in the sound, hear that sound number three is composed of another voice than number one and two. Anyway, the point is that all these three sounds have a degree of homogeneity in the middle part despite the very high number of vertical elements involved, and that this does only reveal a small amount of information (here, maybe the gender difference between the first two and the last sound can be inferred). In other words, even if certain features of the sound are completely randomly distributed, the randomness in this case creates a uniform sound with less information. This is actually in line with the view of complexity presented by Tononi, Edelman and Gerald, where complexity increases with increasing regularity up to a point and then decreases as regularity turns into randomness (Tononi et al., 1998). Moreover, the uniformity that this kind of randomness can represent can form a link to the reduced mode of the minimal because both represent minimal variation. The consequence of this is that maximum information density has to be defined as located at the *breaking point* where increasing complexity will start leading to less available information for the listener. I would like to define the noise mode of the minimal as the range between the breaking point of complexity and a noisy state which can still somehow be related to vocal production. When any information related to vocal production is lost, however, I will regard it as non-voice.

[191] The example was synthesized using *csound* applying randomized triggering of playback starting time and skip time into the sound files.

6.4 Factors potentially influencing information density

The factors that I see as most important in influencing information density are in addition to *complexity*, which I have just discussed, *rate*, *inter-domain inconsistencies*, and *need for re-listening*.

- **Complexity:** To sum up the most important points from the discussion above, increased complexity tends to increase information density up to a critical limit, where increasing complexity may cause a change in resolution towards supra-objects and thereby decreased experienced information density. Complexity is in its turn affected by:
    - The number of elements or dimensions conveying information. More elements/dimensions => higher complexity
    - Hierarchical structuring => lower complexity
    - Segmentation/grouping => lower complexity
    - Regularities => lower complexity
    - Familiar structures => lower complexity

- **Rate:** The number of information chunks that occur *per time unit* can be seen as a *rate* factor that affects experienced information density. This rate must be seen in relation to the rate with which features can normally be processed. For example, when a verbal phrase is time-stretched, articulated very slowly, or extended by inserting pauses between units, it will usually mean that there are fewer phonemes, words and semantic units per time unit compared to the maximal voice, which can be associated with "normal" speed of articulation, which for English is found to lie in the range from about 3 to about 4.5 syllables per second.[192] Conversely, when there are more information units for the different domains than designated for the maximal voice, one will as a rule have information density in the direction of the noise mode of the

[192] The speakers in an experiment by Apple and Krauss spoke with a mean speech rate of 3.27 syllables per second (Apple & Krauss, 1977). In an experiment by Versfeld and Drechsler the figures were slightly higher, with 3.6 for the female speaker and 4.6 for the male speaker (Versfeld & Dreschler, 2002). Orr and colleagues took 175 words per minute to be "normal" speech rate (Orr et al., 1965). This corresponds to 2.9 words per second, which for English equals about 4.35 syllables per second, calculated from the mean values for syllables per word in English found by Yaruss (Yaruss, 2000).

minimal.[193] Examples of higher and lower rate than the maximal voice will be presented in sections 6.6.1, 6.6.4 and 6.6.5.

- **Inter-domain inconsistencies:** As a rule, inconsistencies between information in different domains imply decreased redundancy and increased information.

- **Need for re-listening:** I will postulate a correlation between the number of times that one needs to listen to a segment and how far the segment should be located towards the noise mode of the minimal voice, at least up to the breaking point of complexity where re-listening will not lead to any new information. We can also see that this is reflected in the learning/familiarity factor mentioned in the section on complexity. It is clear that what is unfamiliar during the first listening, will be familiar after several listenings. This factor is related to what I will refer to as *the dynamics of listening* in section 6.6.

6.5 Evaluation of the information density premise

The evaluation of this premise is more complicated than for the other premises due to the two modes of the minimal, which extend in different directions from the maximal voice. Moreover, what complicates the evaluation further is that information can be inferred from all the different experiential domains, and that the information density for different domains need not be identical. This was already evident from the maximal voice, for which information was dominantly inferred from the **LI-domain** (cf. section 6.1.1). It is therefore necessary to draw up some guiding lines for the evaluation of this premise in the following.

One point should be clarified right away, and that is that I see maximal voice as constituting a special and privileged case in terms of information density: It is maximized in terms of efficiency, specificity and range of significations compared to other kinds of vocal expressions.[194] This is a result of the unique status of language in human communication, where our means of producing, perceiving and comprehending are specialized and finely tuned to speech (Liberman & Whalen, 2000). Therefore, I will not refer to the other domains

[193] From chapter 3, especially tables 3.1, 3.2, 3.3 and 3.4, it is apparent how differences in speech rate can be linked to identification of voices, of identity features like age and personality (**ID-domain**), and to the designation of emotions (**AF-domain**). Changed rate can in other words not only imply that information density is lowered, but also that the information itself can be altered.

[194] This parallels to some degree Barry Truax' view of the continuum among the three acoustic systems of communication, *speech - music - soundscape*. Truax views the continuum as having increasing information density as well as increased specificity of meaning, moving towards the right of the continuum, i.e. towards speech (Truax, 2001: 51-52).

than the **LI-domain** as having a maximal evaluation for the information density premise, even if they are presented in a way that appears "balanced", i.e. it is neither too much nor too little – they will still not reach the same level of signification as the **LI-domain**. For instance, a voice singing a melodic motif of a few bars, which is easy enough to perceive and retain, but which still has some complexity that makes it interesting, might be considered as having balanced information density in terms of melodic processing. Nevertheless, since a melody is a structure which is subsumed in the **SQS-domain**, I regard it as having less information than the maximal voice.

Evaluations are most straightforward when they can be directly related to the **LI-domain**. It is relatively easy to judge, for example, if the information density is higher or lower than the maximal based on the *rate* with which the verbal structures are presented. In this regard, electroacoustic music offers many examples, some of which will be discussed in the following section, of verbal material that is presented with a markedly *slower* rate as well as material that is presented with a markedly *faster* rate. Here, vocal sounds that are "frozen" in time, i.e. they have zero rate and are completely static, constitute a minimum, and I will see them as residing at the reduced mode of minimal voice. In **sound example 6.2** I have provided a sound which pretty much fits this description. It starts out as a spoken vocal phrase, but at one point the [o] of the word "flowing" is "frozen" so as to keep pitch, timbre and loudness constant.[195] A higher rate than for the maximal voice, on the other hand, will imply problems with intelligibility and comprehension, and the evaluation will then reside on the side of the noise mode. With a rate that is artificially made very fast, however, all intelligibility will be lost at one point, hence actually lowering information density. At this point, which can be compared to the complexity breaking point (cf.6.3), any further increase of rate will be experienced as increasingly noisy. And, similarly to what was the case for complexity, I will regard the noise mode of minimal voice as the range between where all intelligibility is lost and where the sound can no longer be related to vocal production. An example of a transition from a voice approaching maximal evaluation throughout the whole range to the noise mode of the minimal is shown in **sound example 6.3**, which is a 4 minute excerpt from a speech by Eldridge Cleaver at the Berkeley campus in 1968, progressively speeded up from the unmanipulated rate to 100 times the unmanipulated rate at the end.[196]

[195] The sound is generated by analysing a vocal recording with the phase-vocoder analyisis utility in *csound* (pvanal), and then resynthesizing it with the pvoc-opcode so that a short portion of the sound (0.01 s.) is time-stretched 2000 times.

[196] The speech was downloaded from URL: http://negroartist.com/black%20panthers/misc1/Eldridge%20Cleaver%20-%20Blacks%20In%20America.mp3,

Here, we can follow the discussed transitions from an intelligible voice to a sound which is no longer recognizable as a voice at the very end.

A somewhat less extreme version of the reduced mode of the minimal than in the "freeze"-example above can be experienced when a short segment of a vocal phrase is repeated in an "endless" loop, thereby constituting a situation of maximum regularity (cf. section 6.3). In such cases, one will relatively soon have drawn most of the information from the verbal content (**LI-domain**), the paralinguistic inflections (**AF-domain**), identity (**ID-domain**), localization and environment (**SE-domain**) and maybe also something related to technology and mediation (**TCM-domain**) from the sound segment.[197] I have made a sound example that can exemplify such a situation (**sound example 6.4**). Here, former US president George W. Bush introduces himself before making a public address during the Iraq war, thus presenting information in the **ID-domain** for the listener. Most people would probably have recognized his voice without this introduction, being perhaps one of the most widely broadcasted voices of the first decade of the 21st century. One might also recognize that the recording is a low-quality mp3 rendering (**TCM-domain**), and infer from the lack of background noise and reverberation that he is speaking within an isolated indoor (studio) environment (**SE-domain**). Since all this information is presented at the beginning of the speech and then becomes common to all the phrases, it will already be redundant information long before the looped segment occurs. At the onset of the loop, information density will therefore already be low. However, there is one important difference with this example and the "freezing" example. In this case, there might still be interesting to follow the fine details of the vocal phrase, both as "pure" sound (**SQS-domain**) and as vocal articulation (**VG-domain**). Details on this level will normally escape a listener during continuous listening sessions, but when a relatively short phrase is repeated continuously like here, such details will suddenly become more available. Schaeffer's strategy for using looped segments to achieve the sound-source uncoupling of reduced listening is an apparent equivalent to this situation (also cf. the unpredictability factor in section 5.4.3). After some time, however, a listener might feel that even the information in the **SQS** and **VG-domain**s is exhausted, depending on the number of times the segment is repeated, the duration of the looped segment, and its complexity. Hence, this example shows how regularities can quickly make

accessed 23/04/2010. It was processed using the PSOLA based time manipulation in Praat (Information and download can be found at URL: http://www.fon.hum.uva.nl/praat, accessed 23/04/2010) .

[197] The context in which the loop appears, however, can affect information density further. If most of the information from the mentioned domains is retrieved from previous vocal phrases that the looped phrase is a continuation of, there will be even less information to draw from the repeated phrase.

some information redundant, and how it may cause the listener to direct his/her attention to features that are not yet exhausted. While extensive looping therefore approaches the reduced mode of the minimal, one can argue that the possibility of accessing the finer details of the **SQS-** and **VG-domains** at one point in the listening process, justifies an evaluation that will have to go on for an even longer time before information density approaches the reduced mode of minimal voice.

The examples with rate and looping are relatively straightforward in terms of evaluation. More complicated issues arise when there are conflicting tendencies concerning information density between the different experiential domains compared to the configuration for maximal voice, for example if the **LI-domain** carries minimal information, i.e. *less* than for maximal voice, and the **TCM-domain** carries a lot, i.e. *more* than for maximal voice. This would then create a situation where one would have to judge these two tendencies against each other.

Such a situation can be encountered in **sound example 6.5**, which is an excerpt from John Cage's *Solos for voice 2* (1960), in a realization by Gordon Mumma and David Tudor (Various artists, 1968, 5:32-6:20).[198] Here, what I experience is a small number of voices, all manipulated with different means of analogue electronic processing. At first listening, I was not sure of how many voices were present at the same time, nor could I decide if all sounds were related to vocal production, but after listening several times I discerned up to three layers simultaneously.[199] It was difficult, however, to attribute the different sounds to recognizable vocal personae mainly due to the processing, which is quite heavily applied at times. In combination with the back and forth panning, this made it difficult to establish a firm image of the whole scene. As for the manipulations, they appeared to change frequently, both in kind, parameter setting and amount. At first listening I thought I could recognize ring modulation, in addition to the use of panning, distortion (due to high gain amplification) and feedback. After additional listening I thought I could also recognize the use of filtering, artificial reverberation and echo played back in reverse. As for the sounds produced by the voices, I will describe them as largely wandering freely back and forth between a kind of unconventional singing and noise making, without recognizable and meaningful words. There is therefore no information that can be attributed to the **LI-domain** here. For me as a listener with more than average experience with the effects of sound processing technology, I would

[198] This is strictly speaking not an acousmatic piece, but in this context I will include it as an example because I am listening to it as one, i.e. without having the live performers to consider.

[199] This is probably due to the use of some kind of delay manipulation or a tape recorder during the performance by one or both performers.

get quite a lot of information from the **TCM-domain**, though. Thus, the relationship between these two domains in terms of information density is reversed compared to maximal voice, something which creates conflicting tendencies for the evaluation. What I find most important in this excerpt is the experience of a lot of things going on which partly escape my attention. I hear a number of different voices and a number of different processing techniques, but to be able to establish more firmly how many voices there are and what processing techniques are used, I have to listen to the excerpt over and over again. Since in this case the **LI-domain** appears to be irrelevant, i.e. the lack of verbal meaning does not seem to be made a point of, I find it appropriate *not* to weigh this so much in the evaluation. Rather, what dominates my experience for this excerpt is that there is a lot of vocal activity which is markedly processed, and that I can only code some of it into information on the fly. This goes particularly for the **TCM-domain,** but also partly for the **ID-domain**, i.e. the number of vocal personae. For these domains, I seem to be able to get more information when listening several times, something which we have seen is associated with the noise mode of minimal voice. Taken together, I therefore evaluate this example as lying towards the noise mode, despite the conflicting tendencies in this case.

6.6 A note on the dynamics of listening

Before turning to more examples of evaluations of this premise, I would like to say a little more about the relationship between information and the number of times one has listened to a piece. The change of the listening experience from the first time one listens to a piece of music to subsequent listenings of the same piece is what I would like to call *dynamics of listening*. This dynamics of listening appears to present a clear link between the number of times one has to listen to a piece or a sub-section of a piece in order to get access to the information it presents. For the maximal voice, information is optimized so that everything can be taken in during the first listening. This is also evident from the link to the radio voice – a voice on radio cannot be replayed; it has to convey all necessary information in just one session. A consequence of this is that the maximal voice has to be regarded as a mode of voice *adapted for the first listening* of something, because it maximizes the possibilities for unambiguous interpretation, and that subsequent listening sessions to sequences of maximal voice will reward the listener with little further information, except for the **VG-** and **SQS-domains**. However, when one moves towards the breaking point of maximal information density, the opportunity for re-listening that one usually has in the case of acousmatic electroacoustic works (at least those recorded on CD), can nevertheless give a listener the possibility to get access to some of the information that he or she would miss due to the increased complexity or speed. By listening over and over to the same passage while focusing on different aspects of the sound, a listener can get a hold on much of the information that was unavailable during the first listening. And, the more information that is present in a segment of music, the more times a listener will probably have to repeat it to get access to that information.

6.7 Evaluation of musical examples

Since the range of evaluations between the reduced mode and the noise mode of the minimal is larger than what is the case for the other premises, I have had to choose another strategy in the evaluation of examples. Here, I have chosen to discuss some examples more briefly and others more extensively so as to cover most steps between the minimal modes without having too lengthy of a discussion. However, the step between the intermediate and the minimal of the noise mode of minimal voice is not discussed with a separate example. Rather, I will regard the example from *Shifting Ground* as covering this category, since I here discuss a transition from the maximal and all the way to the noise mode of minimal. Moreover, I will not present an example of maximal voice, since this is already covered by sound examples in the previous chapters, in particular **sound example 5.1**.

6.7.1 Reduced mode - minimal: Dieter Kaufmann, *La voyage au Paradis*

This excerpt from Dieter Kaufmann's *La voyage au Paradis* (1987, on Various artists, 1988, 6:15-6:32, **sound example 6.6**) is a good example of temporal "freezing" and thereby of the reduced mode of minimal voice as discussed above. Here, three subsequent syllables uttered by what appears as one single woman are superimposed, "sta-", "spra-" and "la-", all retaining the initial consonants and "freezing" part of the vowels. For the first of these, the "frozen" vowel is heard for seven seconds, and during these seconds no new information is presented, thus indicating an evaluation as the reduced mode of minimal voice.

6.7.2 Reduced mode – intermediate-minimal: Steve Reich, *Come out*

The beginning of Steve Reich's piece *Come out* from 1966 (Reich, 1987, **sound example 6.7**, 0:00-1:15) exemplifies a transition towards the intermediate-minimal category of the reduced mode with its excessive iteration of short segments of a spoken phrase. This piece is based on the recording of Daniel Hamm saying "I had to, like, open the bruise up and let some of the bruise blood come out to show them", and this phrase is presented three times in its entirety at the opening.[200] Each of these repetitions contributes to increasing the redundancy for a listener, in other words to reducing the information density. The repetitions give the listener a good opportunity to make sure that the verbal content (**LI-domain**) is comprehended correctly, and also to pick up more features related to the **VG-**, **ID-** and **AF-domains**. For instance, one can maybe note the particular features of Hamm's socio/dialect, as well as his almost childish mispronunciation of "bruise blood" with an [l] instead of an [r] in the first word.

Then, what follows is a continuous loop of the last second or so of the recording containing the words "come out to show them". This continuous reiteration relatively soon makes the information in the **LI-** , **AF-** and the **ID-domains** completely redundant. As in the example with the looped Bush phrase, this invites a change of focus towards the **SQS-** and **VG-domains**, domains which can represent some information.[201] For instance, one can note that the phrase constitutes a quasi-melody with the intervals falling minor 3rd – unison – rising major 2nd – falling major 2nd. One can also note the increased temporal distance between the two versions of the loop when it has been going on for some time.[202] These changes are still very subtle and take place over long stretches of time, implying that they represent very low information density. Therefore, after listening to the loop for some time one can infer very little information from it.[203] As a consequence, an evaluation of the whole excerpt would have

[200] This statement was made in relation to a case where six black youths were accused of committing a murder which only one of them was known to have committed. The statement refers to a situation in which Hamm had punctured one of his bruises to convince the police that he had been beaten (Connor, 2001; Schwarz, 1993).

[201] It is often commented how this piece makes the voice lose its links to person and text and change into sound or music (see e.g. Connor, 2001; Sherburne, 2004; Smalley, 1993) . It may also invite the listener to focus on what I designated in chapter 2 as the body and mind domain, something which is noted by Wim Mertens in his article on minimal music, where he refers to Phillip Glass' emphasis on the immediate physiological effect on the listener (Mertens, 2004).

[202] After 0:30 one can gradually notice a very short delay between the left and right channel, and this delay grows very slowly longer and longer, until at one point the onsets in the left and right channel of the [k]'s in "come" are clearly distinguishable.

[203] It has been commented for the piece as a whole, however, that the details of the ongoing processes of increasing delay between the two versions might have points of less predictable nature which therefore induce interest for the listener, as for example is noted by Connor: "Unexpected overtones and counter-rhythms are derived from the coming apart and regathering of the sound of Hamm's voice" (Connor, 2001: 479).

to start close to the maximal, and then gradually decline towards the intermediate-minimal category during the extensive looping.

6.7.3 Reduced mode – intermediate: Lars-Gunnar Bodin, *Anima*

The excerpt from Lars-Gunnar Bodin's *Anima* from 1984 (Bodin, 1990, 0:00-0:22) in **sound example 6.8** demonstrates how information in different domains can point in different directions relative to the maximal configuration presented in section 6.1.2. Here, one can hear several superimposed artificially sounding soprano voices singing sustained vowels, [a] and [u], with a typical *bel canto*-like timbre and vibrato. One of the voices is clearly in the foreground, whereas two are in the background singing the same notes all the way through the excerpt. The foreground voice presents a melodic motif consisting of three pitches (f4,h4,f5) sung in ascending, descending and finally ascending order. The second time the motif is presented, the last note is omitted, however.

In this example, the information in the **LI-domain** is restricted to two phonemes without semantic reference, clearly indicating evaluation towards the reduced mode. With the melody and harmony as perhaps the most salient features in the example, the **SQS-domain** will probably be the focus of attention for most listeners. I, for my part, could infer some information from these features; the melody was easy to decode and retain in memory in terms of motion (ascending-descending-ascending), structure (full motif, incomplete motif, full motif), and intervals (tritone-tritone). Hence, this information should imply a higher density than in the maximal configuration, where the **SQS-domain**, being unattended, does not represent information. In this case, the information in the **SQS-domain** can be characterized as not too far from being balanced, but still with a relatively redundant form, having only three pitches on seven notes, and a structure both on the single note level and on the motif level that consists of many similar or close-to-similar elements. In my view, this is therefore a pertinent example of the *intermediate* category of the reduced mode.

6.7.4 Reduced mode – maximal-intermediate: Charles Dodge, *In celebration*

In this example from Dodge's *In celebration* (1975, on Various artists, 2006a, 4:13-4:21, **sound example 6.9**), it is the artificially slowed down rate of the ending part of this vocal phrase (cf. rate factor in section 6.4) that led me to place it in the *maximal-intermediate* category of the reduced mode. Having already discussed issues related to rate, I feel that further comment is unnecessary.

6.7.5 Noise mode – maximal-intermediate: Hans Tutschku, *Les invisibles*

If a vocal phrase is articulated at a rate significantly higher than for the maximal, either by being actually articulated very rapidly or artificially time-compressed in any way, the information density will increase (cf. *rate* factor, sect.6.4). This is the case in an excerpt from Hans Tutschku's *Les Invisibles* (1996, on Tutschku, 1999, 8:15-8:27, **sound example 6.10**). Here, one line of the poem *Es wird später* by Karl Lubomirski, "über graue Krume stehn schon Lerchen" ("over gray crumbs larks are already standing"), is presented in only 1.3 seconds (equalling almost 8 syllables per second), hence about two times as fast as in a normally uttered sentence. For me, this phrase is experienced as rapid and dense, clearly reducing the intelligibility of what is said, but when one knows the lines of the poem, it is still easily recognizable.

6.7.6 Noise mode – intermediate: Trevor Wishart, *Globalalia*

Examples of relatively high information densities, especially for the **ID-domain**, can be heard many times throughout Trevor Wishart's *Globalalia* from 2004 (Various artists, 2005). In this piece, a multitude of different voices appear, all which have been recorded from radio broadcasts. In the accompanying DVD-booklet one can read that voices in as many as 26 languages were recorded and then cut into over 8300 spoken syllables.[204] All the different voices appear only in very shorts snippets of sound, comprising only one single syllable, often juxtaposed into longer sequences containing a great number of syllables and speakers. In the sound example (**sound example 6.11**, 0:00-0:10**)** one can hear the first 10 seconds of Wishart's composition, which demonstrates the high density of information for the **ID-domain**.

These features have been most important in the evaluation:

- The rate (cf. *rate* factor) with which the vocal segments are presented varies somewhat, but averages to about 6 segments per second for the part from 2.5 - 9.5

[204] This information is available on the booklet in the DVD-release on which the piece is published. The information regarding the number of syllables was found at Durham Research Online: URL: http://dro.dur.ac.uk/4405/ , retrieved 09/05/2008.

seconds in the excerpt. For information in the **ID-domain,** this implies a relatively high rate of onsets for new voices, meaning that it will be little time to "tune in" to each of the individual voices.

- The number of elements in the whole excerpt is too high (42) to be able to process and memorize all of them individually (cf. number of elements/dimensions, *complexity* factor). The most salient or the most frequently repeated might be remembered. Here, the segments which end (and begin) a sequence will probably be more easily remembered than those which are in between others (Gardiner, 1983; Murdock, 1962). Thus, it will be easier to process and recall the first sequence of three sounds, where the first and last segments make up two thirds of all the sounds, than the last one, containing as many as 30 segments.
- Few global features contribute to linking the segments into larger structures so that they form a unified percept (cf. hierarchical structuring, *complexity* factor). Hence, even if features like short duration, vocal origin and syllable content are common to all the syllable segments, they are too different in terms of vocal gestures (from spoken to shouted), identity features (men/women), and pitch to form any holistic structure.
- The duration of each vocal segment is short; the longest segment is approximately 0.65 seconds, but the majority of segments are shorter than 0.25 seconds.
- Due to the constantly varying duration of the segments, there is no clear rhythmical structure in the excerpt that would have made it easier to mentally structure and process.
- The pitches in the excerpt do not conform to Western tonality, thus making a mental structuring of the sequence into a melody unlikely for a Western listener.
- I do not recognize any individual vocal identities. Any familiarity will have to be achieved through repeated listening.

However, the excerpt from *Globalalia* also possesses features that can counteract some of the above mentioned factors that will increase the information density:

- There is very little overlap between the individual segments, resulting in minimal vertical complexity.
- Some of the voice segments stand out from the others in different respects, thus enabling a mental structuring in a perceptual salience hierarchy (cf. Bigand, 1993: 273

and hierarchical structuring of *complexity* factor). Thus, the listener can choose to reduce the total amount of information by focusing only on the most salient events:

- One vocal segment makes up a recognizable word; the German word "Traum" (dream) spoken by a male speaker.
- One vocal segment consists of an unvoiced sustained uvular fricative ([χ]). The sustained noisy and gritty quality of this segment makes it stand out from the others.
- The first vocal segment consisting of the syllable [dʒ] stands out due to its shouted voice quality and increased perceived loudness. In addition, this vocal segment is the very first sound in the sequence as well as in the piece in its entirety, and stands out for that reason.

The overall impression of this excerpt from *Globalalia* is that it is far from maximal voice, especially through the large number of segments, all with different vocal identities, the high rate in which these are presented, and the short and uneven duration of each of the vocal segments. In light of the presented framework, this excerpt therefore aspires in the direction of the noise mode of the minimal voice, and I find it therefore pertinent to classify it as being in the *intermediate* category.

6.7.7 Noise mode – transition to minimal: John Coulter, *Shifting Ground*

An example that is particularly interesting in relation to the information density premise is found in an acousmatic passage of John Coulter's audiovisual work *Shifting Ground* (Coulter, 2005, **sound example 6.12**, 1:54-4:22).[205] Here, one can actually follow a gradual development from something close to maximal voice to the noise mode of minimal voice. Primarily, the changes along the information density premise are quite interesting to follow. This passage contains segments of voices of several vocal personae in both sexes and many different ages apparently answering questions from an (non-audible) interviewer about what is most important to them in their life. Apart from the first segment, which I guess is in Māori language, most voices speak English with an Oceanic flavour. Since one of the male voices speak about "living in New Zealand" in one of the segments, I infer that the interviewees are

[205] In most of this section except the last few seconds the video part is just a black screen. Thus, even if this work does not strictly belong to the genre of electroacoustic music, the lack of any visual components in this section makes the conditions of reception not too different from an acousmatic work. John Coulter is educated as an electroacoustic composer, and clearly uses techniques from the repertoire also in this audiovisual work.

people living in New Zealand. In the beginning of the passage, each voice speaks uninterruptedly and in isolation presenting one or more full sentences lasting many seconds at a time. Gradually, the speakers are heard for shorter and shorter durations, first speaking incomplete sentences, and then only just a few words or one single word. Simultaneously, the pauses between the voices get shorter and shorter, until at one moment they start to overlap. At the end of this passage, each voice speaks mainly single words while the number of voices that overlap, gradually increases, until a point where single voices no longer can be recognized. This process goes so far as to create a thick texture of noise in which it is hard to notice any vocal characteristics at all. Thus, during the last phase of this passage one will have to start struggling to process all the information that is presented, both because too much is presented at a time and because voices start to mask each other. In the end, all information will be lost to the listener, resulting in a highly chaotic percept.

The major tendency of gradually increasing information density is also accompanied by an opposite tendency that must be commented. As the piece goes on it gets progressively clearer and clearer that the semantic dimension of all vocal segments centres around values or issues that are considered important by the individuals speaking, all of them probably living in New Zealand. At the end this is possibly extended to include life-mottos or guidelines for the speakers (e.g. "live life to the fullest"). Several times the speakers refer to "values" or use the adjective "important" in conjunction with several notions that one can easily understand are important for people, like family, friends, health and religion, or related notions. In so far as such notions are mentioned by speaker after speaker, the whole concept becomes somewhat redundant, hence reducing information rather than increasing it. Nevertheless, among these notions there are also occasionally some of the things mentioned that are less expected, like "winning games during the football season", "hair" and "alcohol". There is even a word which sounds like it is being spoken in an Asiatic language (1:48) which is highly unexpected. Notwithstanding the overall semantic redundancy, the semantic content of each utterance is far from predictable, and the redundancy effect must therefore be characterized as rather weak.

On the basis of these observations, I will conclude that the excerpt all in all displays a continuous process from the maximal to the minimal through a gradual increase of information density, despite the increased redundancy in the semantic domain that was noted. In contrast to the excerpt from *Globalalia* the passage from *Shifting ground* reaches a point at which the density of voices gets so high as to make every identification of any kind of information impossible. At one point, the result is a dense, noisy mass of sound where the only kind of information available is related to the sound qualities (**SQS-domain**) of the noise. Thus, at one point there is a transition from minimal voice into not-voice, i.e. where it is no longer possible to identify any parts of the sound as originating from vocal production. At which point in time this happens is hard to identify exactly , but somewhere around 2:12 in the acousmatic part, I find it very difficult to hear any traces of voice whatsoever. From about 2:06 I cannot hear any individual voices, but I can still hear some instances of sibilants that point to vocal production. Between these points in time, then, one has an extreme example of minimal voice just before it turns into non-voice. And, as for the information density premise, this is a very good example of the noise mode of minimal voice.

6.8 Chapter conclusions

The present chapter has given an explication of the second premise of the maximal-minimal model. This premise has related the concept of information and the way that information can be inferred from music with a given temporal density to information theory, cognitive load theory, relational complexity theory and a theory of musical expectation. The existence of a negative correlation between information and concepts like redundancy and predictability has been central in the discussion. Here, maximal voice was described as a special case in terms of information density, where the possibilities for processing verbal content with high efficiency and certainty were ideal, and where minimal information was presented in the other domains but **LI**. Due to the possibility of having both a higher and a lower information density than for the maximal voice, I designated a *noise mode* and a *reduced mode* of minimal voice for these situations, respectively. Several factors like *complexity*, *rate*, *need for re-listening and inter-domain consistencies* were taken to affect the evaluation of the premise. For the reduced mode of minimal voice, temporal "freezing" of a vocal sound exemplified the most extreme evaluation. For the noise mode of the minimal, the evaluation was complicated, due to what I called the complexity breaking point – a point at which the increasing density of cues would result in a change of focus towards a more global supra level with higher uniformity. The noise mode of the minimal was then defined as the zone between this breaking point and the point at which a sound would cease to be heard as related to vocal production. Several examples were presented that illustrated different evaluations of the premise, where rate, complexity and need for re-listening were seen as affecting the evaluations.

7.0 Naturalness

> **Premise three of the maximal-minimal model:**
> *Naturalness*: The sound has maximal resemblance with one produced by a human being and his/her vocal apparatus.

The main idea of this premise is that maximal voice is heard as produced by a *human* vocal persona, and that a maximal number of its features therefore are attributed to the human physiological organs involved in vocalization (cf. section 3.2). Thereby, this premise is first and foremost related to the most direct and material aspects of the voice, namely those comprised within the **VG-domain**. One can also imagine that it could be possible to assess the naturalness of features related to the other domains, but since such features in any way would have to be conveyed by the **VG-domain,** it would in most cases be sufficient to evaluate this domain. Consequently, I will concentrate on this domain in the following discussion.[206]

In this chapter, as in the preceding two chapters, I will link the premise to existing theory and empirical research from fields of research that are involved with relevant issues, and on that basis, formulate a set of factors that can be helpful in the evaluation process. Furthermore, I will explain and demonstrate how this premise is applied in the evaluation of segments from musical works, with special reference to the **VG-domain**.

7.1 Naturalness in voice research

If we take a look at the body of research directed at improving text-to-speech (TTS) and other synthetic speech interfaces, it presents strong indications that most listeners are highly attentive in detecting vocal features that are somehow non-human or synthetic in character. Studies within this branch of research, which either use the term *naturalness* (e.g. Nusbaum, 1997; Hawkins et al., 2000; Keller, 2002; Moore & Tan, 2003) or, more rarely, *humanness* (e.g. Couper et al., 2004; Huang et al., 2001) show that subjects consistently rate recorded

[206] The **ID-domain** would seem to be the most obvious candidate if one were to assess the naturalness of other domains. For example, at 04:11 to 04:57 in the *Prologue* part of Alain Savouret's *Don Quichotte Corporation* (1981, on Various_artists, 1993) one can hear a voice that is less natural since it comprises vocal features from men, women and children simultaneously. However, since it would be possible to assign the reduced naturalness of this voice to a disproportion between spectral envelope and pitch (see section 7.2), it would give a sufficiently pertinent picture of the naturalness of this segment if one focused on the features relating to the **VG-domain**. One can also imagine that one could evaluate the naturalness of the **AF-** or the **LI-domains**, as for instance when a voice sounds cold, machinic and inhuman or when a verbal expression would seem to lack fluency. Yet, these kinds of lacking naturalness are not so much tied to issues of sound production, since they can also be projected by human actors. These features therefore do not correspond to my conception of naturalness, and I will not consider them in the following discussion.

natural voices as more natural, or human, than synthetic ones.[207] The term *naturalness* is also used in a musical setting in the research on synthetic singing, where the studies have been an integrated part of the development of synthesis systems, rather than being conducted for their own sake (Rodet, 2002). Since electroacoustic music can include synthetic and processed voices in all shades between speech and song, both branches of research seem relevant in this context.

Many of the studies on naturalness in speech converge in their view that naturalness is far from a "simple" quality that can be easily defined and described in relation to a single perceptual attribute of the voice. There seems to be a general agreement in recent research that naturalness is a meta-quality related to a number of different aspects of the voice such as glottal source, intonation, and rhythm/timing (Polkosky & Lewis, 2003; Nusbaum, 1997; Mayo et al., 2005; Keller et al., 2002; Murray et al., 1996; Hawkins et al., 2000). Of these aspects, intonation and voice quality seem to be most frequently mentioned.

Similarly, the premise of naturalness is regarded as a meta-quality where a great many different features contribute to the global evaluation. The most important of these features will be included in the discussion of factors in section 7.2, and although intonation and voice-quality are not included in my list of factors, they are included under other labels: Voice quality/glottal source are included in the discussion of phonatory spectrum, and intonation is included in the discussion of fluctuations, discontinuities, speed, pitch/spectral envelope relationships, precision, and effort and register.

Regarding the literature on electroacoustic music, I have already discussed how Bossis' and his concept of *vocality* is related to the naturalness premise. In seeing vocality as "the similarity of a sound to a voice", his notion resembles this premise to a large degree (Bossis, 2004: 95). His suggestions of a "vocality index", which he envisions as a set of qualities that together indicate "a greater or lesser proximity of the sound phenomenon to the voice" (*loc.cit.*) is interesting in this context, since it resembles my inclusion of a set of factors in the evaluation of the premise. However, whereas I am concerned with experiential issues, Bossis regards the vocality index as being based on an acoustical analysis of the signal rather

[207] The concept of *naturalness* is also applied to aspects related to the fluency of delivery (Noyes, 2001; Schaeffer & Eichhorn, 2001) and the consistency between linguistic and paralinguistic layers of speech (Schröder, 2001; Murray & Arnott, 1996). In these branches of research, however, naturalness is *not* a quality that is opposed to artificiality and not so much related to vocal production, but rather it is seen as the opposition to different types of speech disorders, thus implying that the term denotes a kind of normality for these aspects. Interestingly enough, there is also one neurological imaging study done by Lattner and colleagues showing that both synthesized speech and pitch-manipulated speech elicited a so-called mismatch negativity event-related potential when compared to natural speech, which points to a low-level pre-attentive processing that can separate natural speech stimuli from manipulated and synthetic ones (Lattner et al., 2003).

than dealing with experience. Still, one might regard the factors below as an expansion and partly a concretization of Bossis' suggestions, in so far as they in some cases point to acoustical properties or their perceptual equivalents.

7.2 Factors potentially affecting naturalness evaluation

The factors that I have included are *fluctuations*, *technological artifacts*, *discontinuities*, *speed*, *pitch/spectral envelope relationship*, *precision*, *phonatory spectrum*, *articulatory features*, and *effort and register*. These factors are chosen with a basis in the research discussed in the previous section as well as my own experience with electroacoustic works, but I do not regard it as exhaustive. The factors will be presented with links to research literature where I have been able to find it, and exemplified with excerpts from pieces of electroacoustic music or sound files which I have prepared myself.

- **Fluctuations:** One important aspect of vocal production seems to be that certain parameters, in particular pitch and loudness, incorporate fluctuations with different degrees of deviation and regularity (cf. section 2.4.2.2), which are partly dependent on the type of vocalization:
    - **Irregular micro-level fluctuations:** Irregular micro-fluctuations like *jitter* and *shimmer* can contribute to experienced naturalness (cf. section 3.3.5.1). Jitter and shimmer come about involuntarily, and are an integral part of human vocalization. Therefore, they typically have to be implemented in vocal synthesis systems to make them sound natural (Childers et al., 1987; Aoki & Ifukube, 1996). Moreover, if micro-fluctuations are artificially removed from the voice, they can also sound less natural, as in **sound example 7.1**. Here, four sustained vowels produced by a female speakers are first presented with artificially smoothed pitch and amplitude, and then in an unmanipulated version.[208] Even if the processed vowels still sound human, they are clearly less natural than the unprocessed ones. If micro-fluctuations are manipulated so as to enlarge the amplitude of the deviation, the result may also have reduced naturalness.

[208] The pitch smoothing was done in Praat using PSOLA manipulation, whereas amplitude smoothing was done using serially connected compressors with high ratio setting. The onsets and endings of the vowels were smoothed by simply fading in and out from silence.

  - **Vibrato:** For certain styles of singing, in particular bel canto, *vibrato* is an integral part of the style. The perceptual importance of vibrato for the naturalness of the singing voice appears to be high, and can even be decisive for the attribution of a sound as voice.[209] In the beginning of Viñao's *Chant d'Ailleurs*, also discussed in section 4.5, (1992, on Viñao, 1994, **sound example 7.2**, 0:00-0:30) the transformation from a wind instrument to a singing voice reaches a phase for a few seconds where is has a vowel-like, but synthetic quality (ca 0:19-22). It is not until the pitch fluctuations set in and have reached a certain rate a few seconds later that the sound clearly appears as a human voice.

- **Technological artifacts:** Technological *artifacts*, i.e. sonic "side-effects" that in a sense functions as fingerprints or traces pointing at the processing methods or algorithms applied, tend to reduce experienced naturalness (Wouters & Macon, 2001, Bonada et al., 2002, Arslan & Talkin, 1997 and Cook, 1996). Clicks, extraneous noise and spectral "colouring" are typical examples of such artifacts.[210] In **sound example 7.3**, which is an excerpt from *When I am with you*, the first movement of Charles Dodge's *Speech Songs* (1973, on Dodge, 1994, 0:09-0:15), one can notice loud clicks and a marked "warbling" quality in the last phrase, both which in my opinion qualify as technological artifacts.

- **Discontinuities:** Discontinuities, i.e. abrupt changes in *pitch*, *spectrum* or *loudness* in a vocal sound that surpass or violate the acoustic constraints of the human vocal system tend to be experienced as unnatural (see e.g. Childers et al., 1987 and Klabbers & Veldhuis, 2001).Vocal production is constrained by the gradual way in which we

[209] This was demonstrated by John Chowning, who synthesized a 15 second long sung note in three stages: 1) A 400Hz sinusoid; 2) Harmonics were added one by one, appropriate to the sung vowel; and 3) A mixture of random pitch variation and vibrato was added to the signal. He could report that "not until the random deviation and vibrato are added at stage 3 do the harmonics *fuse*, becoming a unitary percept and identifiable as a voice" (Chowning, 1999; 264).

[210] "Colouring" as an opposition to naturalness has been studied by Moore and Tan (2003). Here, the authors investigated the effects different filters that created spectral distortions had on listeners' judgment of sound quality of speech and music, defined on a scale from "very natural - uncolored" to "very unnatural – colored". They found highly consistent results with correlations between depth of spectral ripples/width of spectral tilts and rated naturalness/increase in colouring (Moore & Tan, 2003). Regarding extraneous noise, a study by Stevens and colleagues evaluated naturalness of voice accompanied by different levels of noise. Here, subjects judged the naturalness of recorded versus synthesized speech stimuli, in a high-quality and low-quality condition, where added white noise was used to create the low-quality condition (Stevens et al., 2005). They found that recorded speech without the added noise was consistently rated more natural than synthetic speech with noise.

have to move the organs involved.[211] While marked discontinuities occur in many types of vocalizations, like stop consonants and changes between modal and falsetto register, we appear to be sensitive to discontinuities which do not match the audible traces of the continuous and coarticulated vocal gestures. In electroacoustic music abrupt *cuts* in a recorded sound, what Chion called "audible montage", are perhaps the most poignant form of discontinuities (cf. sect. 2.5). An example of excessive use of abrupt cuts can be heard in the excerpt from Maja Ratkje's *chipmunk party* (Ratkje, 2002, 0:00-0:16) presented in **sound example 7.4**. Other kinds of less poignant discontinuities will be discussed in section 7.4.

- **Speed:** The organs involved in vocal production are constrained in terms of the *speed* with which they can move, and audible deviations beyond the boundaries of these constraints may lead to decreased experienced naturalness. One important aspect of these constraints are the *speed-accuracy trade-offs*, i.e. the tendency that increased speed will lead to decreased accuracy. For example, one has found that in singing, pitch accuracy will decrease as speed increases (Bella et al., 2007). In speech, natural rapid spoken sequences tend to be characterized by less accurately pronounced vowels (vowel reduction) where pauses and steady state segments tend to be shortened more than transitional aspects (Kent & Read, 2002: 236). Vocalizations with artificial modifications in speed that do not appear to follow such speed-accuracy trade-offs will therefore tend to sound less natural.

- **Pitch/spectral envelope relationship:** Audible disproportions regarding the relationship between pitch and spectral envelope may cause the experience of naturalness of a voice to decrease. In section 3.4.1 I noted how f0 and spectral envelope (often measured as formant frequencies) are correlated with perceived age, gender, and size, but that the differences between voice categories (man, women, child) are proportionally greater for f0 than for spectral envelope. When the relationship between these variables exceed the ranges which occur for men, women and children, it will usually be heard as less natural.[212] Manipulations that affect one or

[211] *Coarticulation*, discussed in 3.6.2, also testifies of the continuous way in which the organs involved in speech production operate; when phonemes are produced in sequences, they tend to "colour" each other.

[212] As Cook notes, if formant values in themselves are outside "reasonable ranges", they will result in a sound that "is no longer speechlike" (Cook, 1999: 144). I still see such a case as a disproportion between the two variables, though.

both acoustic variables may therefore be experienced as having reduced naturalness. The classic case is the manipulation of playback speed, which was a standard technique in *musique concrète* and related kinds of tape music. Works like e.g. Schaeffer and Henry's *Symphonie pour un homme seul* (Schaeffer, 1998), Berio's *Thema – Omaggio á Joyce* (Berio & Maderna, 2006), and Takemitsu's *Vocalism AI* (Takemitsu, 2004) all contain speed manipulated voices with markedly reduced naturalness. Lastly, it has to be mentioned that experience with a particular voice or even with a particular recording may sharpen the sensitivity to any changes in pitch or spectral envelope, that is, *if* the particular voice is recognized.

- **Precision:** Vocalizations with too high of a precision for several features like tempo, pitch (onsets), timing and dynamic variation can be experienced as less natural. Human vocalizations usually involve characteristic irregularities and inaccuracies (as the micro-fluctuations discussed above) when compared to the perfect linearity and exactness that a machine can create. For example, Akagi has showed how small pitch deflections in the form of *preparation* and *overshoot* characterize pitch changes for classically trained singers; *preparations* are deflections to the opposite direction of note change observed just before the note changes, whereas *overshoot* is a deflection exceeding the target note after note changes (Akagi, 2002). He also showed that these features made listeners evaluate synthesized sung phrases as more natural than when they were not present. Another common example of "inhuman" precision is when events are organized in perfectly metronomic time, since human performances involving rhythmic articulations tend to be characterized by more or less systematic *deviations* from metronomic regularity (Waadeland, 2001). Furthermore, exact repetitions of a sound segment, that is, looping, is an example of machinic precision in electroacoustic music. Lastly, precision can also be seen relative to speed, as mentioned above, obeying the basic principle of what is referred to as the *speed-accuracy trade-off*: When something is speeded up, it tends to be less precise (see e.g. Tro, 2000) . For a machine, however, such limitations do not apply.[213]

- **Phonatory spectrum:** Spectral characteristics related to the phonatory component of vocalizations (cf. section 3.3.5) that deviate from the range of natural vocalizations

[213] Sometimes when a feature or property cannot be achieved by a human, but only by a machine, it is referred to as *superhuman* (e.g. in Georgaki, 1999).

may lead to an experience of decreased naturalness. In electroacoustic music the use of manipulations that apply source-filter decomposition of vocal sounds (cf. section 3.3.3) open for replacement or modification of features related to the source/phonatory component. For example, in Wishart's *VOX-5* (1986, on Various artists, 1989) one can find a passage (**sound example 7.5** , 2:39-2:45) where, as I hear it, the voiced phonation part of the sounds has an inharmonic quality – while at the same time the articulation can be recognized as human, pronouncing something that sounds like [d^{h}ʊː], [dɑ̃ː], and [zɔ̃ŋ].

- **Articulatory features:** In the same manner as the properties associated with the voice/phonatory source might entail an experience of reduced naturalness, the properties associated with filter or articulatory component (cf. section 3.3.6) of the voice may also do the same. Such properties are usually associated with what is referred to as the *spectral envelope* of the signal, which among other things is of great importance in giving the vowels their characteristic quality. For example, in Jean-Claude Eloy's *Shânti* (Eloy, 1979) there is a section (**sound example 7.6**, 6:48-7:05 of *Face I*) where a speech-like sound appears, but where the articulation is very blurred and indistinct in a way that cannot be attributed to a human speaker, hence reducing naturalness markedly.

- **Effort and register:**[214] We have seen that vocal effort is related to both loudness and sound spectrum (cf. section 3.3.4.2), and that register is interrelated with pitch and timbre (cf. section 3.3.5.1) for naturally produced vocalizations. Manipulated or synthesized voices in which these relationships are not retained, may be experienced as having lower naturalness. Hence, changes in loudness which are not accompanied by the appropriate changes in sound spectrum, and changes in pitch over a certain range without any spectral changes related to a switch of register, can make vocalizations appear less natural.

[214] This is not a factor that is usually mentioned in studies of experienced naturalness, but I still would like to propose that this can be relevant, since transformed vocalizations in the electroacoustic repertoire demonstrates how violating this coherence can lead to decreased naturalness. See the discussion in section 7.4.1 for details.

7.3 Evaluating the naturalness premise

After having presented and exemplified factors that can contribute in the evaluation of the naturalness premise, it remains to discuss how these can be linked to the actual evaluation along the continuum from maximal to minimal. I will relate this evaluation to the discussion of category boundaries in section 4.7. Here, I argued that the boundary of the maximal-minimal model was dual: (1) On one side, it was based on strict "all-or-nothing" categorization of sound source, and on the other side, based on (2) a fuzzy evaluation of the likeness of the properties of the sound to a human vocal apparatus. These two ways of categorizing can also be applied in the evaluation of this premise. I have redrawn figure 4.6 above, showing the dual category boundaries (1) and (2) with the zone of ambiguity displaced to the right. I have added the five categories from maximal to minimal as a multi-coloured "band" in between the arrows representing the two categorization models, where each of the five categories have been given a separate colour (**figure 7.1**). As I argued in section 4.7, minimal voice is associated with the zone of ambiguity of the first of these boundary models (1), which is localized somewhat above the very lowest evaluation of likeness of properties (2). When it comes to the remaining four categories along the continuum, the evaluation will have to be based on an evaluation of likeness (2), so that the greater the likeness, the further towards the maximal the evaluation will have to be localized. The sound source category attribution (1) will in all these cases be "voice". To further clarify the dual basis of the evaluation, I have made a table (**table 7.1**) in which the five categories from maximal to minimal voice are associated with source attribution and evaluation of property likeness.

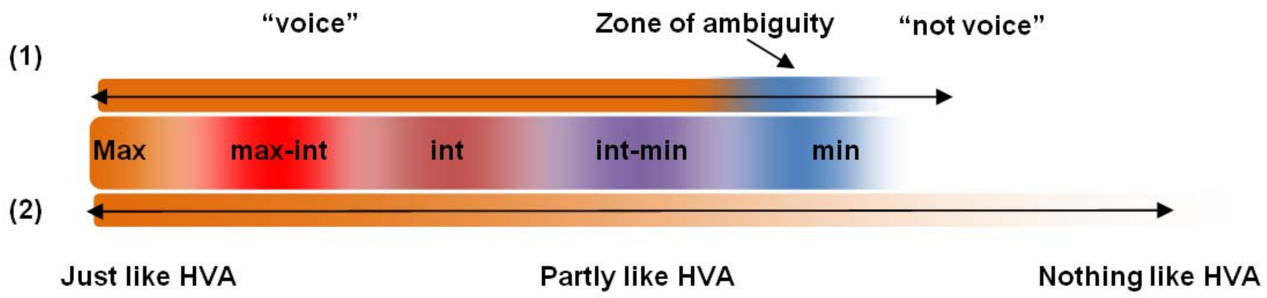


Figure 7.1: Guide for the evaluation of the naturalness premise related to the dual boundaries of the maximal-minimal model as delineated in figure 4.6. (1) represents the strict boundary between the categories "voice" and "not voice" with a zone of ambiguity between them; (2) represents the fuzzy boundaries related to abstracted properties ranging from maximal likeness to human vocal apparatus (just like HVA), via what sounds partly as a human vocal apparatus (partly HVA) to minimal likeness (nothing like HVA).

Evaluation	Source attribution (1)	Evaluation of property likeness (2)
Maximal	Voice	Maximal
Maximal-intermediate		Most
Intermediate		Partly – higher
Intermediate-minimal		Partly – lower
Minimal	Ambiguous	Some

Table 7.1: Guide for the evaluation of the naturalness premise for five categories between maximal and minimal.

If we return to a previously discussed sound example, namely the one with the beginning of Viñao's *Chant d'Ailleurs* (**sound example 7.2**), we can see how a movement along most of the range of the voice-likeness continuum can be realized. The identification of the Eastern wind instrument which can be identified at the beginning, implies that this sound is categorized as "not voice", but the pitched quality and the sustainment of the tone bearing the characteristics of a blown instrument still gives it some resemblance to a voice.[215] Then, around 0:10 and to about 0:12 the timbre changes gradually, losing some of the characteristic nasal and lightly buzzy quality of the beginning along with idiomatic pitch inflections. From this point until about 0:19, the likeness to a vocal sound increases somewhat, but in this phase, the sound has a rather synthetic timbre which is still instrumental in character. Around 0:19, a new transition starts, which introduces resonances in the sound, progressively giving it the quality of a sung vowel, and hence further increasing voice-likeness. At this point, I experience that I start to engage in re-defining the sound in terms of source recognition, but that the highly static character (cf. the *fluctuations* factor) still prevents the sound from fully crossing the boundary into the "voice" category – in other words, the experience has entered the zone of ambiguity of minimal voice. When the slow, but gradually accelerating, pitch fluctuations enter at 0:22 the ambiguity recedes gradually along with increasing voice-likeness – the sound at this point is positively a voice, but is still lacking a bit in naturalness since the vibrato appears artificially slow, stiff and somewhat discontinuous (cf. *discontinuities* factor). Then, the voice-likeness increases rapidly with vibrato rate, and around 0:27, I characterize the voice as fully voice-like, i.e. maximal in terms of the naturalness premise. Consequently, it is apparent how the sound has moved from right to left along the voice-likeness axis, where at one point it crossed the category boundary from "not-

[215] That is, in the beginning of the excerpt there are sections in which two pitches can be heard simultaneously. Nothing is indicated in the liner notes of the CD whether this is due to the particular characteristics of the instrument or if two instruments are playing here.

voice” to “voice” through the zone of ambiguity. While Viñao’s instrument-voice metamorphosis thereby illustrates very well the movement along the continuum of the premise, it still remains to see how evaluations can be made according to the categories along the continuum between the maximal and the minimal, something I will take a look at in the following section.

7.4 Evaluation of musical examples

I will now evaluate several excerpts from an electroacoustic piece in order to exemplify how evaluations of this premise can be done, namely Charles Dodge’s *In celebration* (1975, on Various artists, 2006a). This piece can be seen as a continuation of the composer’s work with computerized manipulation of the speaking voice, beginning with *Speech Songs* (1973). In both of these works, Dodge based his compositions on recordings of Mark Strand’s poems read by the composer himself and subsequently manipulated using computerized techniques of analysis/resynthesis. For that reason, these compositions are often classified as text-sound compositions (cf. section 1.3). In both works, the output from these manipulations, which typically include time-stretching and compression, pitch manipulation, and substitution of voiced excitation with noise, is the only sound material used in the composition. The synthetic

Text / phrase / sound example #	Time	Naturalness evaluation
You want to wave but (**7.7**)	4:13-4:16	Maximal-intermediate (most natural in the piece)
So you wait (**7.8**)	6:58-6:59	Maximal-intermediate
Cannot raise your hand (**7.9**)	4:15-4:22	Intermediate
You have seen it happen before (**7.10**)	4:06-4:13	Intermediate
Feeling your lungs (**7.11**)	6:28-6:37	Intermediate
In a (**7.12**)	4:26-4:29	Intermediate-minimal
[ʃɛɪ] (**7.13**)	0:19-0:26	Intermediate-minimal
Bad (**7.14**)	3:05-3:13	Intermediate-minimal
You (**7.15**)	0:10-0:11	Minimal, among the most unnatural in the piece

Table 7.2: The author’s evaluation of naturalness for selected phrases from Charles Dodge’s *In Celebration* (1975, on Various artists, 2006a) .

quality of the voice in these two pieces, and maybe *Speech Songs* in particular, have been commented on by Dodge himself as well as other commentators, and it corresponds well with my overall impression (e.g. Dodge, 1989; Byrne, 1999). But even if the general impression of *In Celebration* is that of reduced naturalness, there are still considerable differences in naturalness between the most and the least natural sections of the piece. To give an impression of these differences I have chosen a set of phrases that cover the range from the most to the least natural in the piece. The phrases and their evaluations are listed in **table 7.2** along with start and end times, text and sound example number.

In the following I will comment on the factors which I regard as having contributed the most in this evaluation:

- **Discontinuities:** One of the factors that contribute most to the differentiation in naturalness for *In Celebration* is the different degrees of natural continuity in the stream of vocal sounds.[216] Whereas the phrase "you want to wave" (**sound example 7.7**) has no apparent unnatural discontinuities, at least when listening to the phrase as a whole, many of the phrases in the piece have noticeable, and often striking, discontinuities. There are several examples of discontinuities in pitch, from small scale ripples in pitch contour to marked discontinuities between subsequent pitches. The beginning of the phrase "feeling your lungs" (**sound example 7.11**) and the end of the phrase "cannot raise your hand" (**sound example 7.9**) are both good examples of the former kind. In the more extreme examples, the greater distance between pitches makes it so the discontinuities are felt even stronger. E.g. in the phrase "bad" (**sound example 7.14**) there is a rapid (but decelerating) alternation between two pitches a minor ninth apart, which would be impossible for any voice to perform, and these discontinuities are therefore felt as unnatural. There are also discontinuities between un-pitched (noisy) and pitched sounds that sound unnatural in this piece. They are probably not so acute during listening, but still contribute to a more artificially sounding voice. For example, the transition between the [f] and the [i] in "feeling your lungs" (**sound example 7.11**) is marked by a relatively sudden switch between noise and pitched sound that raises its artificiality and reduces its naturalness, albeit subtly.

[216] This shows that the stream integration premise and the naturalness premise are partly interconnected.

- **Effort and register:** Another factor which plays an important role in this context is the degree to which pitch, sound spectrum and loudness correspond to natural changes of effort and register. In many phrases throughout *In Celebration*, high pitches are not accompanied by vocal timbres that characterize the appropriate voice quality, register and effort, and this makes the phrases sound unnatural. A good example of this can be heard in the phrase "you have seen it happen before" (**sound example 7.10**) where the pitch of the upper voice ends up at E6. One thing is that this pitch is outside the range of most male voices.[217] Another thing is that if a man with an extraordinary pitch range would sing a tone this high, it would have the marks of a rather strained falsetto. Clearly, this is not the case here, where the timbre of the highest pitch has lost almost all the timbral characteristics of a voice. This is even more felt in phrases like "bad" (**sound example 7.14**), where the rapid alternation between a G3 and a G#4 is not accompanied by any changes in voice quality, which it would for a naturally produced vocalization.

- **Fluctuations:** Many of the tones that are transformed so as to be heard as "sung" have an artificially static pitch, and sometimes also spectrum and intensity. As opposed to many styles of singing, there is no vibrato, which usually tends to increase the experience of naturalness. In other words, these tones lack both natural and culturally imposed (vibrato) fluctuations, something which affects experienced naturalness negatively. Two examples among the phrases in the table are "In a" (**sound example 7.12**) and [ʃɛɪ] (**sound example 7.13**).[218] For both examples, the individual notes have no pitch fluctuations, something which makes them static and unnatural. For the latter example, the pitch also undergoes a stepwise but overlapping transposition in ascending octaves, which further reduces naturalness.

[217] The range between the highest (about E6) and the lowest (Bb0) pitch in the entire piece surpasses what most male voices can produce. However, there are individuals who have extraordinary vocal abilities who can exceed what one would take as normal physiological limits of the voice. For example, according to Wikipedia, the record for high pitches sung by males is C#8, and the record for lowest sung pitch is B-2, corresponding to a fundamental frequency of 8Hz, which is lower than the human audible range. Moreover, the world record holder with the largest vocal range spans six octaves (Wikipedia contributors, 2007b).

[218] I was not able to decode this phrase into a comprehensible word.

- **Technological artifacts:** Compared to *Speech Songs* (cf. **sound example 7.3**) there are significantly fewer technological artifacts in *In Celebration*, but there are still some present. For instance, the excerpts in the two lower rows in table 7.2 have features that can be regarded as artifacts. In the "bad"-phrase (**sound example 7.14**), between 5 and 7 seconds out in the example, the sound has a muffled quality compared to the preceding and the following part, which I hear as an artifact resulting from low-pass filtering. For the phrase "you" (**sound example 7.15**), it is the pitch sequence rather than the spectral brightness that bears the marks of computer technology; the pitches appear random in a way that I would attribute to some kind of electronically or computerized randomization process, and especially so since the randomization is combined with high speed and equal duration of each of the pitches.[219]

- **Speed:** The most natural parts in the piece are those which are articulated with a speed close to normal speech, and an example of this can be heard in the phrase "you want to wave" (**sound example 7.7**). Here, the speed or rate of delivery lies within the normal ranges of speech.[220] When speech-like phrases are exceptionally fast or slow, however, the result often sounds unnatural. The phrase "feeling your lungs" (**sound example 7.11**), which only contains three words in nine seconds, is the one that most clearly demonstrates this.[221] However, for the phrases that had been transformed into "singing", i.e. in which only the vowel parts of the syllables had been time-stretched and where the pitches were set to steady tones, I did not experience slow rates as less natural *per se*. This is the case, for example, with the phrase, "you have seen it happen before" (**sound example 7.10**), where the time-stretched vowels instead contribute to make the phrases sound more song-like. In a few other passages in the piece, it is not the rate of words or syllables that appear unnatural, but rather the rate with which the defined pitches change, which can exceed the limits of the physiologically possible. This is the case for both the two phrases in the bottom of the table, "bad" (**sound example 7.14**) and "you" (**sound example 7.15**). For the latter of these, the pitch changes vary rapidly and

[219] It is far from evident that this sequence of pitches would have been perceived as computer generated if they had been presented as long notes *ad libitum*.

[220] With 4 one-syllable words in 1.4 seconds one will have a speaking rate of 170 words / syllables per minute, something which also falls within values that has been measured for normal spoken and read passages (Foulke & Sticht, 1969).

[221] There are also some phrases in the piece which appear to have been speeded up, but not to the degree that it radically alters the experienced naturalness of the phrase.

apparently much faster than human capabilities, hence contributing to a lower naturalness evaluation.[222]

- **Precision:** Many of the phrases in *In Celebration* are song-like in that they present sequences of sustained, pitched notes based on Western twelve-tone tonality. For most of these "sung" notes, pitch accuracy is "perfect", i.e. without any deviations such as overshoot or preparation or fluctuations, something which gives them an unnatural level of precision. This is particularly striking for the larger intervals, like the minor ninth in the "bad" phrase (**sound example 7.14**), which would have been very difficult, if not impossible, for a human performer to execute with similar precision. I also experience that the "you" phrase (**sound example 7.15**) has, in addition to perfect precision in pitch, an unnatural precision in timing, with metronomic regularity of each of the notes in the phrase.

As for the factors of *pitch/spectral envelope relationship*, *phonatory spectrum* and *articulatory evaluation* they do not distinguish the evaluation of the phrases included in the table to any large extent, and I will therefore not discuss them any further here.

7.5 Chapter conclusions

In this chapter, we have seen how the premise of naturalness is linked to issues of sound production, i.e. to features belonging to the **VG-domain**. I have shown that the evaluation of naturalness is well established within fields of research involved with different types of synthetic voice, and that these fields see naturalness as something that can be evaluated both with reference to particular features and with reference to the total impression of a vocalization. With basis partly in these fields of research and partly in my own experience of listening to electroacoustic works, I set up a number of factors that I saw as potentially influencing the evaluation of the premise. Then, I linked the criteria for evaluating the premise to the discussion of category boundaries between "voice" and "not voice", discussed in section 4.7, where minimal naturalness was seen as corresponding to the zone of ambiguity between the two. Subsequently, the different evaluation categories above the minimal were

[222] By slowing the phrase down it was possible to count the number of pitches per second, which resulted in a value of 15.

seen as corresponding to different degrees of voice-likeness up to maximal likeness for the maximal voice. Hence, both the strict and the fuzzy categorization models, corresponding to sound source attribution and evaluation of property likeness, respectively, were seen as forming the basis of the evaluation of the premise. Lastly, I showed how these criteria could be used for evaluating a set of phrases from Dodge's *In celebration*, which displayed different degrees of naturalness, from the maximal-intermediate to the minimal. The evaluations were then related to specific factors, showing how they contributed in the process.

8.0 Presence

Premise four of the max-min model:
Presence: The listener experiences a sense of a shared 'here and now' with a vocal persona.

The premise of presence refers to the sense of a vocal persona "being there" for the listener in a shared 'here and now'. With Bazin, "presence [...] is defined in terms of time and space. 'To be in the presence of someone' is to recognize him as existing contemporaneously with us and to note that he comes within the actual range of our senses" (Bazin, 1967, cited in Norman, 1996: 2).[223] For this premise it is the strength and the character of this kind of experience that are central in the evaluation process, as we will see below.

My view of presence is supported by the writings of Katharine Norman and Simon Emmerson, who are, as far as I know, the two writers who have discussed presence most extensively with reference to electroacoustic music. Although Emmerson's discussion of presence in his book *Living Electronic Music* is more general than mine, referring to presence in the very wide sense that *something* is there, be it objects or agents, both in the physical, psychological and social sense, many of the issues he raises have relevance for the current premise (Emmerson, 2007). Katharine Norman has, albeit in free literary style, perhaps less systematic and formalized than that of my own, also discussed many important issues related specifically to voice and human presence in book chapters and articles (Norman, 1996; Norman, 2000; Norman, 2004a). Although I won't discuss her writings directly in this text, she certainly has influenced and supported many of the ideas and notions that I deal with, and consequently I will frequently make references to her texts at appropriate moments.

In addition to drawing on the writings of Emmerson and Norman, I have also based my account of presence on empirical research within psychoacoustics and virtual reality (VR) studies. Even if the latter field offer a much wider definition of the term, often linking presence to the general experience of being present within some environment (often of a multi-sensory character), there are many interesting parallels that open for transference of concepts and ideas to the field of electroacoustic music. As we shall see, parallels can also be drawn between media-theory and electroacoustic music.

The layout of this chapter will follow the now familiar structure of first giving an outline of the theoretical basis of the concept and the premise, then, to explicate a set of

[223] Thus, the notion of presence in my framework is of another kind than what was described by Wesling and Slawek, who saw presence mainly as a correspondence between person, self and voice, albeit with an illusory status modifying its metaphysical implications (cf. section 4.1). Dyson's discussion of the presence of the voice also touches upon many of the same issues as Wesling and Slawek (cf section 4.2).

factors that potentially may influence evaluation, before the criteria for making evaluations are discussed. Lastly, I will apply the evaluation criteria on a set of musical examples.

8.1 Theoretical discussion

8.1.1 Spatial, temporal and social presence

As a preliminary theoretical distinction, one can separate the "here and now" into a *spatial* and a *temporal* dimension of presence. The spatial dimension of presence can be subsumed under Emmerson's notion of *physical presence*, which deals with listeners' reconstruction of physical action and agencies, including all types of (living) sound sources and actions, and their disposition in and interaction with an environment (Emmerson, 2007: 18-23).[224] Thereby, features related to the **SE-domain** will be important for this premise. As for the temporal dimension of presence, there is a conflation of the presence of the voice with the temporal present. As Dyson notes: "For the voice being sonorous, necessarily speaks in the present, is accompanied by the actual bodily presence of the speaker, and is heard (perceived) at the moment of its production" (Dyson, 1994: 176). When we experience voices in acousmatic electroacoustic music, however, this simultaneity is illusory, but might still be experienced to some extent. Both the temporal and the spatial dimension can be thought of in terms of *distances*, even if for both dimensions these distances tend to be difficult to assess accurately; the spatial distance will reach from the proximate "here" into the furthest audible distance, whereas temporal distance usually relates to relative and approximate durations between "now" and some past moment, or between two or more past events.

An additional aspect of presence is the *social* one, implying that there are social implications in the experience of a voice, even if the experience of acousmatic music is one-way and thereby unsocial at its core.[225] Studies within so-called media equation research have indicated that social presence can be created in mediated experiences. In particular, Clifford

[224] To the basic physical dimension of presence, Emmerson adds *psychological* presence, which he links to the process when the listener interprets aspects of the music in terms of *will*, *choice*, and *intention* (*ibid.*: 23-29, cf. section 2.5.3). Emmerson links psychological presence both to composition and performance, i.e. the **TCM-domain** in this framework, but since I restrict my notion of presence to that which is directly linked to the vocal persona, "will, choice and intention" pertaining to anything else will be less relevant here.

[225] In Emmerson's discussion of *social and personal* presence, he argues that important aspects of musical meaning arise through the interaction between listeners and makers, drawing both on personal aspects of experience and aspects that are more social. While I have already made a point of how a listener's personal knowledge and experience is important for the listening experience (cf. for instance section 2.5.3 and 2.5.4.2.3), I do not so much want to enter into a discussion about sociological issues related to reception and production in this context, because my study is primarily based on personal listening.

Nass and his colleagues have shown in numerous experiments that humans react to mediated experiences as they do to real people, poignantly formulated in a *media equation*, expressed as *media equals real-life* (Reeves & Nass, 1996; Nass et al., 1997; Nass & Gong, 2000 Huang et al., 2001; Nass & Brave, 2005). They have found that people in a wide range of settings assign attributes like gender, personality, age and emotions to both recorded and overtly synthetic voices, and that they react in a *socially determined manner* according to these attributes, thus providing *social presence* constituted through these voices. More specifically, they have run empirical studies showing that the evaluation of voices was independent from which machine it emanated from: The same voices produced the same evaluation regardless of which machine played it, and two voices were interpreted as two distinct actors even when the voices were associated with a single box. The general explanation for these media equation phenomena is first and foremost rooted in biology and evolution, according to Nass: The human brain is "wired for speech" through 200,000 years of evolution, so that it equates voices with people and acts quickly and consistently on that identification, and its definition of speech seems to be so broad as to include synthesized speech (Nass & Brave, 2005: 2-3).

Although the element of interaction distinguishes most of the media equation research from listening to acousmatic music, this research still demonstrates that a listener can interpret many socially relevant cues from the voice, verbally as well as non-verbally. In the discussion on factors below, I will particularly consider the way in which the voice can project different social space frames (cf. section 2.5.4.2) and how these can relate to experienced physical distance.

8.1.2 Focus and transparency

In addition to the spatio-temporal aspect of presence, *attention*, or more specifically the aspects towards which attention is directed, is often linked to the experience of presence (Norman, 1996: 2-3; Lee, 2004b; Nash et al., 2000; Witmer & Singer, 1998). More specifically, presence involves directing full attention to the object or being in question, while *turning away* from unrelated aspects, so as to "overlook" them. We can see, thereby, how this resembles the *focus of attention* premise, indicating that these two premises are partly interrelated. The importance of "overlooking" the aspects related to (production and mediation) technology to achieve a sense of presence has been a major point for virtual reality studies. In this field of research, "overlooking" the technology (usually computer driven) of the mediated experience has been regarded as a defining criterion for presence:

> Presence […] is a psychological state or subjective perception in which even though part or all of an individual's current experience is generated and/or filtered through human-made technology, part or all of the individual's perception fails to accurately acknowledge the role of technology in the experience. Except in the most extreme cases, the individual can indicate correctly that s/he is using technology, but at 'some level' and to 'some degree', his/her perception overlook that knowledge and objects, events, entities, and environments are perceived as if the technology was not involved in the experience (Lee, 2004a, 32)

In the case of electroacoustic music, without this "overlooking" one would simply be listening to a loudspeaker rather than a "person" when hearing a voice.

This kind of "overlooking" the technology of mediation to achieve a sense of presence can also be recognized in the notion of *immediacy* in media theory. More specifically, media theorists Bolter and Grusin have described what they call the logic of *transparent immediacy*, which has pervaded parts of Western art and mediated expressions for centuries and still does. According to this logic, the interface or the medium will erase itself "so that the user is no longer aware of confronting a medium, but instead stands in an immediate relationship to the contents of that medium" (Bolter & Grusin, 2000: 23-24). They argue that different artistic practices, from linear perspective drawing to *trompe l'oeil* painting and virtual reality computer graphics, have strived to efface the traces of mediation so as to give the viewers an experience of immediacy with what is represented. The opposite situation, where the signs of mediation are multiplied or where the illusion of realistic representation is stretched or ruptured, so as to make us highly aware of the mediated nature of the experience, Bolter and Grusin calls the logic of *hypermediacy* (Bolter & Grusin, 2000: 34). This logic is in the authors' view most apparent in the "windowed" style of computer desktop interfaces, web pages, multimedia programs and computer games.

Hence, there seems to be a correspondence with the research on virtual realities and media theory in that they both see attending towards technology and mediation as inhibiting the sense of presence. This has also got some empirical support, in that studies have shown that degraded audio or media-specific noise will decrease the sense of presence (Reeves & Nass, 1996, Lessiter et al., 2001). Moreover, it also corresponds with the general tendency for the objective quality of technology as a factor positively affecting presence, as described by Lee (Lee, 2004b).

It is important to note, however, that the experience of audio quality and the degree in which something is experienced as "marked" by recording technology are highly relative to historical and socio-cultural conditions. For example, the implicit claim of "perfect fidelity" for acoustic phonographic recordings made in e.g. the tone tests by the Edison Phonograph company early in the 20th century will evoke a smile for the listener used to digital quality, and demonstrates the importance of the listener's frame of reference when it comes to what is commonly called the "quality" of the reproduction.[226] Thus, if the pattern of media-related noise, i.e. what I have labelled *implicit transformation*, is different from what the listener is used to, it can reduce the sense of presence of the vocal persona to the listener. This reduced sense of presence can also imply increased temporal distance. As Stan Link notes, the hiss and crackling noise of a gramophone record tend to signal a kind of patina and nostalgia of a long gone past, introducing a distance between the virtual "here and now" of the musical performance and the actual listening situation (Link, 2001).

If we relate all this to the current study, we see that it implies a negative correlation between presence and attending to the **TCM-domain**. This will be reflected in one of the factors (implicit/explicit transformation) presented below. What is more, we can see that the distinction between transparent immediacy and hypermediacy resembles in many ways the naturalist/médiatiste distinction discussed in section 2.5.3. For loudspeaker mediated voices it also corresponds by and large to Norman's distinction between the 'clean' and the 'unclean' edit (Norman, 2004a: 114-16).[227] The former designates a situation in which the sensation of mediation is reduced by removing errors, making 'clean' (i.e. inaudible) edits and reducing unwanted noise, while the latter denotes exactly the opposite.

8.2 Factors potentially contributing in the evaluation

For the premise of presence the factors that can potentially affect the evaluation are *implicit/explicit transformation*, *multi-modal associations*, *salience/loudness*, *temporal continuity*, *contextual linkage*, *spatial distance*, *social distance* and *duration*.[228] All of these factors will be discussed below, with links to relevant theory where this was found. Examples are provided where necessary.

[226] See Sterne for a discussion on sound fidelity and its social genesis from the invention of the phonograph in 1877 and through the first couple of decades of the 20th century (Sterne, 2003: 215-286).

[227] In her discussion, Norman considers both technological and voice-related issues – what she calls *edit* and *enunciation*. Whereas the former is most relevant for this chapter, issues of enunciation will be relevant in chapter 9 on the clarity of meaning premise.

[228] It should be noted that other factors related to the listening situation can play a part in creating the feeling of presence for a listener. Several such factors are often mentioned in presence research in addition to features of the VR systems themselves. E.g. do Witmer & Singer mention isolation and interface awareness (Witmer & Singer, 1998). Interestingly, it appears that the performance practice of electroacoustic music has tried to increase isolation, prevent distraction and decrease interface awareness through having only sparse lighting during the diffusion of a piece so that the presence of the loudspeakers in the room does not seem too pressing. Still, what these different factors boil down to is to direct the listener's attention fully to the virtual world "behind" the loudspeakers.

- **Implicit/explicit transformation:** Implicit/explicit transformations, i.e. experienced traces of technology involved in recording/mediation/coding (implicit) and transformative processing (explicit) (cf. section 2.5.3.2), can contribute to directing focus towards the **TCM-domain**, and thereby to reducing the sense of presence. Here, explicit transformation tends to affect presence negatively more than implicit transformation. Moreover, where implicit and explicit transformation display *changes*, this will affect presence negatively more than where it appears static. Thus, if a sudden change in noise level or frequency response occurs, this will likely call the implicit transformation to attention, and the sense of presence will thereby be reduced, at least temporarily.

- **Multi-modal associations:** The degree to which the voice can be linked to other multi-modal associations can affect the experienced presence of the vocal persona.[229] As Norman notes, vision is particularly important here, in that the sensation of presence implies creating an internal visual image on the basis of what we hear (Norman, 1996: 3). Multi-modal associations are in their turn related to the *familiarity* with the voice in question; usually, a person who is *familiar* to us will evoke many more associations than one that is completely unfamiliar (cf. 3.4.2). The degree to which the features of an unknown person can be identified can also have an impact on the degree to which multi-modal associations can be made. For instance, if features such as gender and age are highly ambiguous, it will be far more difficult to evoke multi-modal associations for the voice.

- **Salience/loudness:** The salience of a sound can affect the sense of presence, in that if the sound is too soft in loudness or masked by other sounds, it won't be experienced as present for the listener. One thereby has a close link between this factor and the salience premise, which will be discussed in more detail in chapter 10. For now, it will suffice to focus on loudness, which also Norman relates to presence: "Loud sounds can disorient us through an apparent spatial proximity – the hypothetical object *seems*

[229] Some support for this can be found in research on virtual reality, where there seems to be an agreement that the larger the number of senses involved in an experience, the stronger the sense of presence (Nash et al., 2000; Lee, 2004b; Lombard & Ditton, 1997).

to be here, now" (Norman, 1996: 3, my italics).[230] However, a review of studies of presence in the VR field, suggests that increased loudness in itself can play a role only up to a point, and that sounds that are too loud may indeed lead to decreased presence ratings (Lombard & Ditton, 1997). This is in accord with my own personal experience. When I listened to a particular vocal recording at a comfortable listening level, the sense of presence was maximal. When turning the volume up, however, I found that when the loudness reached a certain level, the perceptual "insistence" of the loud sound actually decreased my feeling of presence. Rather, it was like the sound almost ceased to be referential, and became instead a pure sensation, approaching the painful.[231]

- **Temporal continuity:** The sense of a shared temporal continuity between listener and vocal persona can play an important part for the listener's feeling of presence.[232] When there are signs that temporal continuity is broken, the temporal flow in which the vocal persona "exists" will momentarily lose its track, thus making it more difficult to maintain the feeling of presence. Those signs can be:

    - **Disruptions/cuts:** The temporal unfolding of an event is halted in a manner that implies an artificial cut, like, for instance, in this short phrase from Ratkje's *Intro* (Ratkje, 2002, 1:03-1:06, **sound example 8.1**)
    - **Skips:** The temporal unfolding of vocal phrase is discontinuous in a manner that implies a skip from an earlier to a later point in the temporal unfolding. The section where the voice of what sounds as a young female voice, perhaps a child, enters in Justice Olsson's *Up!* displays a light kind of skips (Olsson, 1991, 10:25-10:36, **sound example 8.2**).
    - **Exact repetitions:** Several versions of the same vocal phrase can be heard after each other and/or superposed over each other. For example, when a segment is repeated in the exact identical manner, as in the *Blue Tulips*

[230] Thus, loudness can also be linked to the factor of spatial distance, since loudness plays a part as a perceptual cue to actual distance: Close sounds will sound louder, distant sounds softer.

[231] This effect of loudness is difficult to demonstrate with a sound example, since the perceived loudness will ultimately be dependent on the settings of the playback equipment. Therefore, the interested reader is urged to repeat this kind of "armchair" experiment of the effect of turning the volume knob on the sense of presence on his/her own.

[232] In VR-research continuity and connectedness are mentioned as general factors that aid in creating realism of an environment and thereby increase presence (see e.g. Witmer & Singer, 1998 or Nash et al., 2000).

example, discussed previously in chapter two (Wishart, 2000b, 0:28-0:45, **sound example 8.3**).

- **Superposition of several versions of the same voice:** When several versions of what is recognized as the identical vocal source is superposed on top of each other, this can highlight the "constructed" and "out of time" relationship between the versions, thus creating the feeling that they are not present in the same continuously unfolding flow of time.

- **Contextual linkage:** When one senses the presence of a vocal persona through voice, he or she can still be felt as present even if the sound of the voice is absent.[233] This can happen through what we can describe as a metonymical linkage to the context within which the vocal persona is situated, i.e. the context being everything that is experienced as intrinsic to a certain recording, like the sounding indices of recording (**TCM-domain**) or the ambience from the recording environment (**SE-domain**). Since such contextual features are a part of the whole recording where the voice also can be heard, the part can in given situations stand for the whole recording, including the voice, when the latter is absent.[234] It has to be noted that without any indications of the vocal persona being present, the listener will, as time passes, gradually become more and more uncertain whether he/she is present or not, until one will have to assume that the vocal persona is no longer there – even if the contextual linkage continues to be present.[235] What is more, the sense of presence will depend on the *temporal continuity* of the contextual sounds. If silence, a completely different environment, or different recording technology than was associated with the vocal recording suddenly replaces contextual sounds, the sense of presence will be cut off – at least if this new environment is present for more than just a short time. The following example, an excerpt from Westerkamp's *Kits Beach Soundwalk* (1989, on Westerkamp, 1996, 1:13-1:47, **sound example 8.4**), demonstrates this. Here, we can hear first an unmodified version, then, this is followed by a version where two sections of vocal pause (during which sound from the environment is heard continuously in the

[233] This is commented by Chion : "lorsque vous avez sur la banda deux phrases dites par une même voix et séparées par plusieurs secondes voire une minute, le personnage continue d'être présent dans l'intervalle pour l'auditeur, invisiblement et inaudiblement" (Chion, 1991: 86).

[234] Hence, one can call this a form of auditory *synecdoche*.

[235] The exception to this might be if a listener can infer from a more or less regular occurrences of indications of the vocal persona, where the time gap is of some length. In that case the listener will base his/her experience of presence on the statistical inference that the vocal persona will reappear within some time span. Hence, both the duration of the absence and the regularity of indications of presence can potentially play a part here.

unmanipulated version) are replaced with silence and sounds indicating another environment/recording technology, respectively.

- **Spatial distance:** That the acoustic correlates of spatial proximity affect the sense of presence, is in my view implied in the term "presence" and its reference to *here*.[236] Thus, to be present means, at least partly, to be close. If loudspeaker-mediated vocal sounds have cues that correspond to a larger physical distance, like being softer, having a larger portion of reflected sound than direct sound, and having less high frequency content, it can therefore imply a decrease in the sense of presence. In the following sound example from Tōru Takemitsu's *Vocalism AI* (Takemitsu, 2004, 0:13-0:21, **sound example 8.5**), we can clearly hear voices located at different distances, where the female vocal persona resides relatively close and the male persona at a further distance. Clearly, the female voice is more present than the male voice.

- **Social distance:** Social distance as it is reflected in vocal effort/voice quality and subject matter, earlier delineated through the four social space frames in section 2.5.4.2, may affect the sense of presence. Social distance may or may not correspond to experienced spatial distance. For instance, one can imagine a vocalization that signals intimate social distance, e.g. whispering, located at a far distance, as well as shouting, which signals public social distance, located very close.

- **Duration:** For the sense of presence to be established, a minimal duration of the sound segment is required.[237] This argument is easy to consent to, since a minimum duration is also required to recognize a sound as a voice at all and to recognize features like familiarity, distance and implicit/explicit transformation. Thus, short fragments of vocal recordings, like those in the intro of Ekeberg's *…des cantiques*, are most likely

[236] E.g. in Merriam-Webster Dictionary & Thesaurus one of the significations is "the part of space within one's immediate vicinity" (In *Merriam-Webster Online Dictionary*, accessed 4/23/2010, from URL: http://www.merriam-webster.com/dictionary/presence), and in the English dictionary on wordreference.com a similar signification reads "the part of space within one's immediate vicinity" (In *Wordreference.com*, accessed 23/04/2010 from URL: http://www.wordreference.com/definition/presence).

[237] Heeter also argues on a general basis that one needs to be located within an environment for a minimal duration for the sense of presence to be established (Heeter, 2003).

not to evoke a sense of presence, or to evoke it only faintly (1997, on Ekeberg, 2001, 2:23-2:52, **sound example 8.6**).[238]

8.3 Evaluation of the premise

In the evaluation of this premise, I have found it difficult to set up specific criteria to be used when dealing with musical examples. Rather, I have had to proceed by comparing experiences of different sound examples so as to gradually build mental reference points for the different categories between maximal and minimal. In this process, I started with the extreme evaluations, which I found easiest to decide upon, and thereafter proceeded towards the two adjacent categories (*maximal-intermediate – intermediate-minimal*), and finally, I settled for what I considered belonging to the *intermediate* category. In the following, I will present excerpts from some of the works I have considered in the process, which I regard to be the best examples of the five categories along the maximal-minimal continuum. As previously, I also present a judgment of which factors that I regard as having influenced the evaluations the most.

8.4 Evaluation of musical examples

8.4.1 Maximal: Hildegard Westerkamp, *Kits Beach Soundwalk*

The intro of Hildegard Westerkamp's *Kits Beach Soundwalk*, from which I have provided an excerpt (1989, on Westerkamp, 1996, 0:20-0:50, **sound example 8.7**), is a very good example of maximal voice as such, as well as of maximal presence. Here, the speaking voice of a woman, which I learn from the CD liner notes is Westerkamp herself, is present in the foreground throughout the whole intro, albeit with some longer pauses in between each vocal phrase, where the environmental sounds that make up the background are heard alone. The factors that are most important in the evaluation are:

- **Temporal continuity/contextual linkage:** The vocal phrases are in themselves without discontinuities of any kind, and the constantly present sounds of water, birds and the city in the distance link the phrases together into one coherent situation.
- **Spatial/social distance:** The voice appears spatially quite close, approximately within one metre, or as it would sound in a personal conversation, in other words within a

[238] The temporal discontinuity for these fragments will probably also contribute in reducing presence.

personal space frame. The vocal effort and subject matter also indicate a personal space frame, hence spatial and social distance correspond well here.

8.4.2 Maximal-intermediate: Christian Zanési, *les mots de Stockhausen*

Christian Zanési's *les mots de Stockhausen* (1996, on Zanési, 1996, 4:24-4:30, **sound example 8.8**) is an interesting piece with regard to presence, since the voices heard in it recurrently play with different features that establish and destabilize presence. In this piece, one can for instance hear long passages containing looped vocal fragments evoking presence only minimally, as well as intimate breath sounds that are mechanically repeated so as to counteract the intimacy. There are also relatively long sequences featuring Karlheinz Stockhausen's unmanipulated speaking voice in this composition which for me approach maximal presence. In the excerpt I have chosen, we can hear two relatively soft, slightly processed and partly superimposed versions of Stockhausen's voice speaking French, where the following factors influence the sense of presence, as I hear it:

- **Implicit/explicit transformation:** Compared to other vocal phrases in the piece which have been close to the maximal voice, these phrases appear to be lightly band-pass filtered. Hence, they are experienced as an explicit transformation, albeit a rather gentle one.
- **Salience/loudness:** Compared to many other vocal phrases and other non-vocal sounds in the piece, these phrases are relatively soft in loudness.
- **Temporal continuity:** Since the vocal phrases are identified as belonging to the same vocal persona, and since the phrases are partly superimposed, there is a sense that this is a scenario which is a product of the composer's organization of the sound material rather than interaction between two vocal personae. In other words, this situation makes me attentive of aspects belonging to the **TCM-domain** and thereby a reduced sense of presence.

For all of these factors, the effect on the experience is relatively gentle, and the result is consequently an evaluation of *maximal-intermediate*.

8.4.3 Intermediate: Tōru Takemitsu, *Vocalism AI*

In the same manner as Zanesi's piece, *Vocalism AI* (Takemitsu, 2004) by Takemitsu can be heard as actively playing with different degrees of presence for the voices in it. The vocal personae here signal everything from intimate to public social distances, and they appear to be localized both close to and distant from the listener. Moreover, many degrees of explicit transformation further contribute to varying the degree of experienced presence for the different vocal phrases. At the end of the piece, from which the excerpt (3:32-3:37, **sound example 8.9**) is taken, we can hear one of the voices where the manipulation is moderate; obviously this voice has increased playback speed and added artificial reverberation. This voice, which monotonously repeats the word "ai" alternately at one low and one high pitch, projects the image of a vocal persona entering the scene from a distance, moving gradually closer, then from side to side, before it once again gradually recedes into the distance. The excerpt in the example is taken from the phase where the vocal persona has recently for the first time approached my virtual point of listening from somewhere further away. A female vocal persona located relatively close is also present here, although it is not her presence that I will evaluate in this case. At the beginning of the excerpt, the voice is moderately low-pass filtered compared to how it appears a little later, indicating, perhaps, that the vocal persona is located behind some wall or barrier. Taken together, the factors that I take to affect the experience of presence for this passage are as follows:

- **Implicit/explicit transformation:** In this case, the moderate speed manipulation clearly affects the identity of the vocal persona, making it ambiguous in terms of gender and age, although I am closest to experiencing this as an adult male, perhaps in a kind of shrunk version. While the voice for me clearly stands out as processed compared to all the vocal phrases in the piece that are not subjected to speed change, the unchanging application of the processing throughout the whole ending of the composition reduces the focus on the processing and thereby the **TCM-domain**. The low-pass filtering that I identify for the voice in this excerpt nevertheless makes me direct my attention somewhat more towards the **TCM-domain** than the section following the excerpt, where there is no filtering.
- **Social/spatial distance:** The situation in the ending part of this piece, with a vocal persona that mechanically repeats the word "ai" over and over again, gives the impression of somebody behaving in an almost machine-like manner, without any

social orientation towards others surrounding him/her. Socially, therefore, the vocal persona appears distant, almost as if in a separate sphere of existence, although the spatial cues indicate that he/she is somewhere not too far away. Spatially, I would say that the vocal persona is localized somewhere between the stage and the arena space frame.

8.4.4 Intermediate-minimal: Bodin, For Jon: Fragments of a Time to Come

In an excerpt from the piece *For Jon: Fragments of a Time to Come* (1977, on Various artists, 1997, 15:54-16:12, **sound example 8.10**) by Lars-Gunnar Bodin, we can hear a female speaking voice followed by a female singing voice which is localized at a relatively far distance. The former voice is included for the sake of comparison, so that it is the latter voice which is subjected to my evaluation here. The factors that I consider relevant for this voice are mainly related to the two types of distance:

- **Social/spatial distance:** The high effort applied by the singing voice in this example signals a public space frame. Spatially, however, it seems that the voice is localized further away than what would be the case during a performance. Rather, I experience it as if I am *outside* the arena space frame within which the voice is situated. Or, alternatively, that the voice is situated in the landscape space frame relative to my virtual position, almost so far that it approaches my acoustic horizon, i.e. the outer limits of listening.
- **Salience/loudness:** Being localized at a far distance, loudness is correspondingly low for the singing voice, and it stands in marked contrast to the much louder speaking voice that precedes it.

8.4.5 Minimal: Paul Lansky, *Notjustmoreidlechatter*

In Paul Lansky's *Notjustmoreidlechatter* (1988, on Lansky, 1994b, 0:00-0:20, **sound example 8.11**) I experience that presence is minimal throughout most of the piece. The short vocal fragments one hears in this excerpt, which is from the very beginning of the piece, together create a dense and complex texture with a harmonic content, where the individual vocal personae are as good as totally absent in the experience. I take the following factors to be most important for the minimal sense of presence:

- **Implicit/explicit transformation:** The vocal fragments in this excerpt are on the whole experienced as explicitly transformed, where the majority of fragments are marked with a "buzzy" and artificial quality, sometimes bearing the marks of having shifted spectral envelope.
- **Duration:** The fragments are quite short – they seem to consist of one, sometimes two syllables.
- **Temporal continuity**: Since each fragment is very short, and since it is very difficult to link several fragments together into one coherent utterance, the temporal continuity is low for this excerpt. Instead, the individual fragments first and foremost function as rhythmic and harmonic elements making up the whole sonic texture.

8.5 Chapter conclusions

In this chapter, I have given an explication of the premise of presence. This premise was related to the experience of sharing the "here and now" with the vocal persona, thus implying both spatial proximity and simultaneity. In the theoretical discussion, I also argued that voices, even if they are acousmatic and artificial to a higher or a lower degree, can project a *social* presence. This social presence can be described using the space frames introduced in section 2.5.4.2. Moreover, I argued that theoretical accounts of presence within research on virtual realities as well as media theory could be related to this premise. Here, attending to the technology involved in creating an expression was seen as something which stood in opposition to the sense of presence. Thus, the premise of presence was seen as partly related to the premise of attention, for which attending to the **TCM-domain** would imply an evaluation towards the minimal. Some types of implicit transformation may also indicate a temporal distance between the "now" of the vocal persona and the present of the listener. The factors that I saw as potentially relevant for experiencing presence were *implicit/explicit transformation*, *multi-modal associations*, *salience/loudness*, *temporal continuity*, *contextual linkage*, *spatial distance*, *social distance* and *duration*. For many of these factors, I referred to empirical research that supported the correlation with presence. For others, however, I relied on my personal experience. When evaluating this premise, I did not put forward any specific criteria, but rather based it on comparing a number of segments from different pieces. Among these segments, I chose five that I regarded as the best examples of the five categories from maximal to minimal presence. Hence, these examples can be regarded as reference points for other evaluations, even if the factors involved in the excerpts discussed necessarily won't be the same.

9.0 Clarity of meaning

> **Premise five of the max-min model:**
> *Clarity in meaning formation:* Meaning can be constructed from the voice with a high degree of clarity – also implying specificity, certainty and coherence.

Meaning, and the clarity with which it is constructed by the listener in the encounter with the voice, was a central issue for all the theoreticians that I discussed in the chapter on the maximal-minimal model (cf. section 4.1 and 4.2): Wesling and Slawek saw indeterminacies and ambiguities related to identification of the speaker, his/her identity, location and socio-cultural linkage as related to the minimal, and both Dyson and Børset emphasized the clarity and the meaningfulness of the maximal/radio voice. The premise of *clarity of meaning* is intended to include such aspects into my framework, where clarity is taken to subsume also *specificit*y, *certainty* and *coherence*, concepts that will be explicated during the course of this chapter.

Regarding the notion of *meaning* in the premise, I have to stress that the meaning I will discuss in this chapter is *not* musical meaning *per se*, or meaning related to a global evaluation of a musical work in its entirety. Rather, it is the meaning that is primarily linked to the ontological level of the vocal persona, its utterances and its localization in and relation to an environment, I will be concerned with. To be more specific, I will mainly deal with the features belonging to the **LI-**, **ID-** and **SE-domains** in this chapter, even if other domains might also in principle be involved. Here, issues of clarity of verbal structures (**LI-domain**) will be central, hence dealing with issues of verbal articulation/intelligibility/ comprehensibility that were introduced in section 3.6.2. I will consider aspects that have a *contextual* function (cf. section 3.6.1), which are influential for the listener's interpretation of meaning for the vocal utterance in relation to the whole situation it appears in. While these in principle can include all the domains, I will primarily focus on the **ID-** and the **SE-domain**, since they tend to be very important for our interpretation of vocal utterance. For these domains, I will consider the specificity with which one as a listener can categorize different features. I will also argue that the *coherence* within and between features related to different domains can be important. The discussion of **LI-domain** clarity, contextual specificity and coherence will then form the basis of a list of factors that can potentially influence the evaluation of the premise and criteria for making the evaluation. Finally, these criteria will be demonstrated in an evaluation of excerpts from five electroacoustic works.

9.1 Theoretical considerations

9.1.1 LI-domain

Issues of intelligibility have been considered important in the appreciation as well in composition of electroacoustic works with verbal material since the 1950s (see e.g. Schaeffer, 1952: 67; Murphy, 1999; Jones, 1987; Lane, 2006). The most famous example from the composer's perspective is possibly Stockhausen's serialization of "Verständlichkeitsgraden" (degrees of intelligibility/comprehensibility) in seven degrees from "nicht verständlich" (not intelligible) to "verständlich" (intelligible) as a part of the composition of *Gesang der Jünglinge* from 1956 (Stockhausen, 1992:60-61).[239] From the listener/analyst perspective, intelligibility has also been subjected to some study (Stacey, 1989a; Broening, 2006; Segnini & Ruviaro, 2005). In Segnini and Ruviaro's paper on "music and language intersections" in electroacoustic music, that I discussed in section 4.4, *intelligibility* was used as one of the axes in a 'music-language sonic space', using it as a basis for a listening based evaluation and comparison of electroacoustic works (Segnini & Ruviaro, 2005). Segnini and Ruviaro also briefly list three processes that might influence intelligibility, namely 1) code manipulation – compliance with linguistic rules, 2) realization – degree of sonic alteration, and 3) context – circumstances of presentation (*ibid.*). These processes will all be considered in the course of this chapter. Besides this, I have not found much in electroacoustic studies that might further enlighten what issues are involved in perception and comprehension of verbal material.

Many of the issues related to the evaluation of this domain have already been discussed earlier in this dissertation. I would like to sum up some points from the discussion of the **LI-domain** (section 3.6) which will have relevance for this premise:

1) The listener's competence or knowledge on multiple levels, including the phonetic, lexical, syntactic, grammatical and semantic, will affect the linguistic processing, something which is usually referred to as *top-down* or *contextual* influence.
2) This competence provides a set of expectations that makes certain configurations more probable than other, thus providing a certain degree of *redundancy*, so that

[239] Stockhausen attests to the difficulties of measuring "Verständlichkeit", and states that it is only through multiple examinations and interrogations of the responses from several listeners that such a series can be established (*ibid.*).

if cues are degraded, ambiguous or missing, plausible interpretations can still be made.

3) Perception of verbal structures are also influenced by low-level or bottom-up processes, which are dependent on the salience of relevant acoustic cues, their clear differentiation and disambiguation and of their integration into a coherent sound stream.
4) The interaction between top-down and bottom-up influence depends on the clarity of the input and the strength of prediction enabled by the different levels mentioned in point 1).
5) Novel voices or new types of degradations and manipulations can affect processing negatively and require a perceptual retuning by the listener that will take time and attention.
6) Familiarity with, or alternatively, training or repeated exposure to, voices, types of degradations and manipulations, can aid processing.
7) Additional information about the sound source and verbal content given in writing may affect recognition and decoding of verbal material.
8) Prosodic parameters like pitch, loudness and duration, as well as rhythm, tempo, stress and boundary cues like pauses, changes in duration or adjustment in pitch can have an effect on linguistic structuring (segmentation), and semantics.
9) The rate with which verbal units such as words, syllables and phonemes are presented to the listener can affect how well the units are perceived and comprehended (cf. table 3.5), particularly when the rate is higher than in normal conversational speech, which in English is found to lie between 3 and 4.5 syllables per second (cf. section 6.4).

With regards to point number 7), I would like to add that such additional written information is in many cases available when dealing with works of electroacoustic music. To include the verbal content of a composition in accompanying texts such as programme notes or CD liner notes, i.e. so-called *paratexts*, is a relatively common practice.[240] I will also assume that it is relatively common that listeners actually read this information, either prior to or after listening. Moreover, many electroacoustic pieces apply verbal material that has been issued or

[240] The term *paratext* has been primarily used within the literary realm to designate those texts that mediates the book to the reader; titles, subtitles, forewords, dedications, epigraphs, prefaces, intertitles, notes, epilogues, etc. (see in particular Genette, 1997). I find the term appropriate also for the texts that accompany different kinds of presentation of music.

circulated in other contexts, and therefore might be familiar to the listener on beforehand. This can be in the form of literary works (Bossis, 2005: 200-210; Dreßen, 1982), radio or television broadcasts (Bosma, 2003: 11), or commonly known material like the alphabet or number counting (Lansky, 2002). My focus in this thesis, however, is not to provide the richest possible context for interpretation using paratexts or other documents, but to centre on what happens in the listening situation. Therefore, I will not assign too much weight to paratexts in the evaluation of this premise.

9.1.2 Contextual issues

Features from domains other than the **LI-domain** can be very important in determining how a verbal statement is interpreted by a listener, something that I have referred to as *context* (cf. section 3.6.1). These features include not only the features that I have referred to as *environmental* setting (**SE-domain**, cf. section 2.6.2.3), but also in particular the identity of the speaker (**ID-domain**) and his/her affective state (**AF-domain**), and in some cases even the **TCM-domain**. At the outset, a verbal statement taken in isolation can be applicable to a large number of different interpretations, but when it is seen together with features from these domains, the statement will take on a much clearer and more specific meaning – the statement will be *particularized*, according to Geoffrey Leech, who lists three ways in which this can take place in spoken communication:

> (A) Context eliminates certain ambiguities or multiple meanings in the message (e.g. lets us know that *page* in a given instance means a boy attendant rather than a piece of paper)
>
> (B) Context indicates the referents of certain types of word we call deictic (*this, that, here, there, now, then,* etc.) and of other expressions of definite meaning such as *John, I, you, he, it, the man*.
>
> (C) Context supplies information which the speaker/writer has omitted through ellipsis (e.g. we are able to appreciate that *Janet! Donkeys!* Means something like 'Janet! Drive those donkeys away!' rather than 'Janet! Bring those donkeys here!', or any other of the indefinitely many theoretical possibilities) (Leech, 1981: 67)

Thus, if it is difficult to determine questions related to the identity of the voice, the environment and the situation that he or she is in, the interpretation of the whole vocal event might result in less clear and more ambiguous or unspecific meaning. Therefore, it is not difficult to see that the premise of *clarity of meaning* also has to embrace the degree to which contextual issues can be specified in an unambiguous manner.

The features of the **ID-domain** are among those contextual features that can play an important role for the interpretation of verbal meaning, and often does in electroacoustic music, in my experience. For instance, when listening to pieces that use voices that are well known from radio or television, the experience will be very different if one recognizes the identity of the voices or not. By recognizing a particular individual, the meaning will be particularized, i.e. be more specific, and hence clearer than if no recognition occurs. In **sound example 9.1**, I have processed a section from Giuseppe Rapisarda's *The day before* (2006, on Various artists, 2006b, 0:08-0:19) featuring a voice that is probably well-known for those who have attended to European mass media in the last decade or so. The original version then follows, revealing the identity of the vocal persona, at least to those who are familiar with this voice, making the interpretation of it much more specific and clear.

But also if a voice seems difficult or ambiguous to define according to features such as gender, age and social/regional background, the meaning of the utterance as a whole can become more elusive and ambiguous.[241] I will give a more detailed discussion of the specificity in categorizing identity from the voice in the following section.

9.1.2.1 Category specificity

The evaluation of the degree of specificity of identity can be linked to categorization, a phenomenon already discussed in sections 4.6 and 4.7. An aspect of prototype theory that weren't discussed at that point is directly related to the question of category specificity, and therefore will be pertinent to address here. This regards what Rosch describes as "the vertical dimension of categories", which refers to the hierarchical levels on which categories include other categories in a *taxonomy* (Rosch, 1978).[242] Taxonomic systems order categories in hierarchical levels, where categories on lower levels are included in categories at a higher level. In Rosch's vertical ordering of categories, she distinguishes between *superordinate categories*, *basic level categories*, and *subordinate categories*. One of her examples of these taxonomies is the superordinate category *furniture*, which can include basic level categories

[241] Nass and Brave also report of a number of studies showing that people strongly prefer clarity in classifying and categorizing people (Nass & Brave, 2005: 16)

[242] Rosch defines categories as "a number of objects that are considered as equivalent" (Rosch, 1978: 30) and *taxonomy* as "a system by which categories are related to one another by means of class inclusion" (*loc.cit.*).

like *chair*, *table* and *lamp*, which in turn can include the subordinate categories (*kitchen chair, living-room chair*), (*kitchen table, dining-room table*) and (*floor lamp, desk lamp*), respectively. From this example, one can see that the degree of specificity for the categories is increasing when descending in Rosch's vertical dimension; from the superordinate, via the basic level, to the subordinate level.[243]

A similar way thinking of categories in levels has been applied in sound source theory. Martin has delineated a sound source recognition category-abstraction space, ranging from very general/abstract to very specific categories, with basic-level categories in between (Martin, 1999: 11-13). In this space, more specific categories represent more information, require less uncertainty about properties and demand more (specific) knowledge. This corresponds well to how I will regard the specificity of vocal features. When it comes to categories that are relevant in specifying the identity of voices, however, they do not comprise concrete objects, but physiologically and culturally defined attributes. For some of the relevant categories, it is not difficult to see that one can operate with hierarchical or taxonomic organization constituting different levels of specificity. For example, a very ambiguous sound might only be possible to identify as being "human voice", whereas in other cases, one might be able to identify a voice as being a "lyric soprano", in other words a lot more specific label. Between these two categories one can imagine several levels of specificity, which, ordered from the more general to the more specific, for example could be: "human voice" -> "female voice" -> "female singing voice" -> "female *bel canto* singer" -> "soprano" -> "lyric soprano".

Category specificity can be used as a qualifier for several of the attributes that are subsumed in the **ID-domain**, with gender as perhaps the most important exception, having only two equally specific categories. For instance, one can define *regional belongingness* with different degrees of specificity. Here, categorization often works as a definition of the geographical area from which the language group, language, accent or dialect originates. Thereby, it is possible to set up a taxonomic hierarchy with different degrees of specificity. An example of this kind of taxonomy is given in **table 9.1**, where Scandinavian (the language

[243] According to Lakoff, basic-level categories have a privileged status in several respects: They render fast identification, they are representable in a single mental image, they are correlated to a general motor program, they have the shortest, most commonly used and contextually neutral words, and they have most attributes of category members stored at its level (Lakoff, 1987: 47). While basic categories thereby appear to be privileged in perception, function, communication and knowledge organization, I still maintain that when it comes to the evaluation of the current premise for ID-related issues, subordinate categories will provide the highest degree of specificity and least ambiguity.

group) is given as low specificity category, and where the specificity of the categories increases with *specific language*, *dialect group* and finally, *dialect*, for which specificity is high. Eriksson notes, similarly to Martin, that specificity, or *resolution* which Eriksson calls it, is dependent on the listeners' knowledge of and exposure to the dialect in question (Eriksson, 2007). For instance, I am able to distinguish dialects from the area where I have grown up at a relatively high degree of specificity, whereas all Slavonic languages sound approximately the same to me, at least so that I am unable to tell languages from each other with high certainty. Other features as age, social belongingness, social role/function/ occupation, and to some degree, personality, can also be defined with different degrees of specificity, and can therefore play a part in the evaluation of clarity of meaning.

Low specificity ↔ *High specificity*			
Language group	Language	Dialect group	*Dialect*
Scandinavian	Norwegian	Trønder (from the area Trøndelag)	*Trondhjemmer (from the city of Trondheim)*
			Orkdaling (from the area Orkdal)
		Vestlandsk (from Western Norway)	*Bergenser (from the city of Bergen)*
			Moldenser (from the city of Molde
	Swedish	Norrlandsk	*Øverkalix dialect*
			Burträsk dialect
		Sørsvenska mål	*Skånska*
			Blekinge dialect
	Danish	Jysk	*Nordjysk*
			Sønderjysk
		Øydansk	*Sjællandsk*
			Fynsk

Table 9.1: Taxonomy of categories for regional belongingness from low to high specificity.

Specificity can also be defined through the *combination* of several identity features, since our definition of identity is usually a result of the combination of a whole set of features. For example, when categorizing a voice as a “female human being” one uses only the feature of gender, thereby rendering an identification which is specified only for one single feature. When combining several features, specificity will increase. For example, in identifying a voice as belonging to a “female, upper crust, outgoing radio host from Western Oslo in her twenties”, features like gender, age, occupation/role, social and regional belongingness and personality/mood are combined so as to constitute a compound category which define the voice with a relatively high degree of specificity.[244] A combination of several features, and in particular those that are highly salient in identification of voices, will therefore increase specificity, and increase the clarity of meaning for the utterance as a whole. Here, gender and age are the two features that are obviously most salient, but other features, especially regional dialect, have proven to be a highly salient feature in the identification of speaking voices (Eriksson, 2007, see also table 3.2). I will therefore regard these three features as particularly important in identification, so that if they are ambiguous, this will have a stronger effect on the evaluation of specificity than any other feature.

9.1.2.2 Familiarity

In the discussion on the identity domain in section 3.4.2, it was made clear that the *familiarity* with a person strongly influenced identification and recognition. It is therefore not surprising that familiarity also can be important in determining the degree of specificity in voice identification. Persons with whom one is highly familiar will usually provide basis for a wide range of features and associations, as well as biographical information and quite detailed inference of personality, mood and attitudes. It would therefore be much easier for a listener highly familiar with someone, to interpret meanings from recorded vocal utterances with a high degree of clarity and specificity, than for a listener with no knowledge of the person whatsoever. However, very few listeners have the privilege of hearing the voices of highly familiar persons in works of electroacoustic music.

Through the omnipresence of media like the Internet, television and radio in our society, most people will probably have a wide range of people with whom they are quite

[244] See Lakoff, 1987: 145-148 for a discussion of compounds or what he refers to as processes of complex categorization.

familiar through their media appearances – I am thinking here of different kinds of celebrities, from music, TV and film “stars”, to sport icons, politicians and radio or TV hosts. Many celebrities give the general public a “piece of themselves” through personal interviews, documentaries etc., so that many people in the audience feel that they know them to some degree. Even if the identities thus projected through media don’t necessarily match the identities that friends and relatives construct through a personal relationship, they can still contribute to providing a basis for a lot more specific identification of these persons than of other people. This can include biographical information, personality, regional and social belongingness, occupation and social role, and so forth. For many politicians and television and radio hosts, however, it is the specific function or role that will dominate the impression on the general public.[245] When one turns to the field of electroacoustic music, one can actually also find several examples where voices of people that are famous through media are used as material in the compositions, as in the Rapisarda example discussed above. In cases where voices of people highly familiar through media are recognized in a composition, therefore, I find it legitimate to regard these as having a high degree of specificity.

In addition to presenting the listener with a familiar voice from which he or she can infer a wide range of identity related features, one can also see that many works use audio clips from media that are associated with a particular historical situation as well as a location.[246] In these cases, the recognition of a particular voice combined with a certain verbal content can indirectly link these voices to a wider context than what is directly referred to in words, thus providing a higher degree of specificity. Again, Rapisarda’s *The day before*, featuring an excerpt of Great Britain’s (now retired) Prime Minister Tony Blair during a political speech about the Iraq war, is a pertinent example. In this case, most Western listeners that have experienced that war through mass media coverage will be familiar with Blair’s visual appearance, the typical setting of political speeches as well as the tense political situation at that specific time in history. Thus, through the recognition of Blair’s voice, the speech style of a political speech and the particular issues he address verbally in that speech, one can evoke a relatively specific historical context.

[245] That this mediated form of familiarity also can be reflected in the ability to recognize voices from as short excerpts as two seconds was shown in a study by Van Lancker and co-workers. Here, 16 out of 45 voices of famous people (male entertainers, politicians, and others well-known in film, radio and television) was recognized with a rate of correct responses above 80% (Van Lancker et al., 1985a).

[246] In a footnote Bosma lists several works where this is the case (Bosma, 2003: 11).

In the corpus of works that I have used as a basis in this research project, the majority of works have presented voices that cannot be linked to any particular situation or context. This is most often because:

- Voices that are not previously known to the general public are used.
- Voices are transformed beyond recognition.
- Voices have been recorded in a studio environment, so that a minimum of room reverberation and accompanying sounds can be heard.
- The virtual setting that the voice is given through accompanying sounds added in the mix, does not provide a recognizable place, time or situation.

One might argue, of course, that the setting that a studio recording implicitly projects – an enclosed indoor space, usually quite small, where only the vocalist is present, but where technician(s) are located nearby, usually with eye contact through a window – is a rather specific setting in itself, comparable to any recognizable natural or cultural setting. Nonetheless, I want to maintain that a setting or situation that is recognized *explicitly* provides a *more specific* basis than the setting or situation that can be only *implicitly* recognized through the absence of natural reverberation and accompanying background sounds. In my view, this is due to the conventions, technology and practices that have converged in rendering the studio recording into a process that projects *transparency*, i.e. it is not meant to be heard or recognized as such.[247] I contend that when reverberation and background sounds are present, artificially created or a bi-product of the recording situation in itself, a listener will be more inclined to use this in constructing a specific context within which the voice is located.[248]

9.1.2.3 Coherence

There is one further issue that can contribute to the clarity of meaning, namely that of *coherence* or *consistency* between different aspects of a vocal phrase, belonging to the same

[247] Cf. the discussion of presence in chapter nine.

[248] This might seem contrary to what I propose as the typical maximal voice, namely the informative radio voice. However, I will claim that voices in radio broadcasts are often explicitly and by verbal means situated in a particular radio studio, in a particular area or city associated with a particular radio station. Hence, the lack of specificity in spatial cues is often counterbalanced by verbal cues specifying the environmental context.

or different experiential domains. When several aspects of a phrase point in different directions in terms of meaning, this can introduce a degree of ambiguity in terms of how the utterance should be appropriately interpreted, and in such cases clarity of meaning can be reduced. For instance, the voice of a child speaking with a typically adult terminology about an adult subject matter would probably be puzzling to a listener, and would likely make him or her doubt the veracity of the verbal content or take it simply as a kind of joke. In that case, features of the **ID-domain** will be incoherent with the **LI/sem**. It is also possible for a listener to detect inconsistencies between the affective (**AF-domain**) and the verbal aspects (**LI-domain**) of an utterance. A typical example would be a person yelling "I am not angry" in an angry voice.[249] Or, if two voices belonging to the same vocal persona are heard simultaneously in two distinct locations in a virtual space, one can experience it as an inconsistency between source recognition (**ID-domain**), which indicates one single source, and the **SE-domain**, indicating two spatially separate voices. One can also have inconsistencies *within* one domain. The lack of compliance to linguistic rules or codes, as mentioned by Segnini and Ruviaro (2005, section 9.1.1), can be regarded as lack of coherence within the **LI-domain** of an utterance. The language used by Maja Ratkje in her composition *Intro* (Ratkje, 2002) is a good example of this, with a lot of quasi-words, a mix of Norwegian and English words, only partial obedience to syntactic and grammatical rules, and mixing standard with non-standard pronunciation.

Inconsistencies between different aspects of an object or an utterance have been investigated extensively within experimental psychology, and for many kinds of inconsistencies, cross-modal as well as intra-modal, one sees that perception and processing get more demanding for inconsistent than for consistent stimuli, particularly in resulting in longer reaction time.[250] Internal inconsistency or uncertainty within non-verbal aspects can also have an effect of interfering with the consistency and clarity of the verbal content.

[249] It is important to note, however, that many such inconsistencies have been conventionalized and are used rhetorically. To use irony and sarcasm effectively, for instance, one needs often use a tone of voice that is inconsistent with the verbal meaning; When saying "this looks easy" when given a task that one really think is difficult, tone of voice is used to communicate that this is ironically meant, especially in cases where it would not be self-evident that the task would be experienced as difficult.

[250] Many such studies have been done in relation to the so-called Stroop-effect, which refers to the increased reaction time when there are inconsistencies in the stimuli. The effect was first studied during naming tasks where the ink colour was inconsistent with the semantic reference to colour, such as when the word "blue" is written in red ink. See MacLeod, 1991 for a review of the Stroop-effect and analogue studies.

9.2 Factors potentially affecting evaluation

With a basis in the points in section 9.1.1 and the theoretical discussion in section 9.1.2, I would like to list a set of factors akin to what I have done in the previous four chapters. Here, however, I consider it unnecessary to include a separate discussion of each factor, since I have already dealt with these issues above. The factors are as follows:

- **Language/code competence**
- **Familiarity**
- **Contextual redundancy**
- **Adaptation (training) to speaker/type of manipulation**
- **Repetition**
- **Rate of presentation**
- **Prosody**
- **Sequential integration**
- **Salience of relevant cues**
- **Selective attention**

As one might notice, the last three factors correspond more or less to other premises of my framework. Once again, therefore, one can note how the premises are partially interdependent of each other. Hence, for these factors the chapters dealing with the premises in question will give a further elaboration of relevant issues.

9.2.1 Evaluation of the premise

9.2.1.1 Methodological issues

Before I turn to the criteria for the evaluation, there are two methodological issues that are particularly related to the clarity of the **LI-domain** that I have to address; 1) the relativity caused by differences in listeners' knowledge and competence, and 2) issues related to the dynamics of listening.

To start with the former, one methodological problem arises due to the fact that verbal comprehension depends on linguistic competence. Being a native Norwegian speaker, I clearly have a handicap in assessing languages other than my own. Thus, it might not be possible for me to comprehend everything that listeners having native fluency in other languages will understand. For the pieces that I deal with in this dissertation, my verbal comprehension might therefore be less specific and rich than native speakers of languages like American/English, French, Spanish and German.

However, this raises some issues regarding whether one should posit an "ideal", i.e. maximally informed listener, when assessing this premise. If so, only the most informed listeners would have enough contextual knowledge to provide the best grounds for an evaluation; a certain voice might only be identifiable by a certain generation of listeners, environments or situations might only be recognizable for locals, and certain semantic issues might only be meaningful (thus comprehensible) for specialists. Setting up criteria for the "ideal" or "maximally informed" listener in this manner might result in narrowing down the range of acceptable evaluators to a few people, possibly including only the composer and his/her co-workers. This might give interesting results, of course, but there might be other problems related to such an approach; issues related to the composition process might be drawn into the evaluation, since a composer usually has knowledge of recordings *prior to* any subsequent processing which can make her or his listening experience quite different from that of others. The "ideal listener" might thereby be restricted to only a few (privileged) people that share the majority of experiences and knowledge with the composer. As a researcher, one could try to approach this level of experience and knowledge by extensive studies of the composition process, by e.g. studying compositional sketches or interviewing the composer.[251] Still, one can question if such a process would be more centred on *creation* and *production* rather than *reception* – hence, if it would be *poietic* rather than *esthesic* (Molino, 1990).

Another possible path that I will suggest as a viable alternative, is to accept that listeners differ greatly in knowledge and experience, and that no particular level or range of experience is privileged. Focus will instead be directed towards the *relationship* between the listener's, in this case my own, knowledge background and aspects of the experience.[252] This

[251] This is a rather typical approach in work-centred musicological studies (see e.g. Ondishko, 1990 or Bergsland, 1999).

[252] In this way, I would approach what has been labeled "situated knowledge" or "strong objectivity" by writers like Donna Haraway and Sandra Harding, meaning that knowledge is always presented and regarded as relative to a particular subject with a particular background and experience, rather than being objectively given (Haraway, 1991: 183-201; Harding, 1992). Their point is that by situating or localizing a subject within a social context, one can demonstrate that knowledge is not about disinterested and impartial subjects "uncovering" or "discovering" facts that are shown to them by nature. Rather, it is always *constructed* by subjects with specific social, cultural and institutional backgrounds, interests and competences: "Critical reflexivity, or strong objectivity, does not dodge the world-making practices of forging knowledges with different chances of life and

would be more in line with the phenomenological approach taken in this dissertation. A consequence of this approach is that if other listeners might attempt the same kind of evaluation, the differences in knowledge and experience might render a different result. But nevertheless, the *same relations* will still be responsible for the differences.

One thing that can be problematic due to the relativity created by relating an experience to differences in knowledge and competence is *comparison*. Comparing a piece containing verbal expressions in a completely unfamiliar language with one in one's native language would probably render great differences, naturally enough. Therefore, to claim some degree of consistency in the evaluations in order to be able to better compare them, I have only chosen pieces with English verbal material. Hence, my language/code competence will be the same for all the pieces, so that the differences in the evaluations will rely more on other factors.

The second difficult methodological issue is related to the effect of *repetition* on clarity, since repeating a verbal phrase may greatly improve a listener's comprehension of the verbal material. Again, therefore, one can see that what I labelled *the dynamics of listening* in section 6.5 can be of great importance also for this premise, since if one listens to a piece for the 10th time, the verbal content might be radically easier to perceive than during the first listening. I pointed out in section 6.6 that the maximal voice is optimized for the first listening, and to be able to evaluate whether a voice is truly at the maximal pole, one should ideally have access to the first listening experience.[253] In most cases, however, access to the first listening is not available. It is usually only in rare cases that we can listen for the first time to a work with the intention of evaluating one particular aspect in mind. Sometimes, one can have memories of one's first listening, but it is not unproblematic to base a judgment upon this. Often it is difficult to tell which memories are first impressions and which have appeared later. What is more, unless one makes a written log, it is difficult to have a clear idea of how many times one has previously listened to a piece and since how long one has done

death built into them. All that critical reflexivity, diffraction, situated knowledge, or strong objectivity 'dodges' is the double-faced, self-identical god of transcendent cultures of no-culture, on the one hand, and of subjects and objects exempt from the permanent finitude of engaged interpretation on the other" (Haraway, 1997: 37).

[253] For one particular work, Paul Lansky's *The Lesson* (1989, on Lansky, 1994b), I could investigate the effect of the dynamics of listening on verbal comprehension directly, since I had never heard this piece prior to the investigation. I found that even though the content of some verbal phrases was picked up at the first listening, and then retained in subsequent listenings, the tendency was that I could perceive more words the more times I had heard a particular vocal phrase.

this.[254] As for me, I do have memories of first listenings for some works, especially some that I have listened to in a concert situation that I have particularly enjoyed or disliked. But, for the majority of electroacoustic works that I am familiar with, my memories are vague and less certain.

My solution to this methodological dilemma has been to combine an experiential strategy with an analytical one. The idea is that I will evaluate the influence of the factors listed above, both those that are mainly concerned with issues related to competence, context and top-down processing (*Language/code competence*, *Familiarity*, *Contextual redundancy*, *Adaptation (training) to speaker / type of manipulation*, *Repetition*), and those that are mainly related to bottom-up issues (*Rate of presentation*, *Prosody*, *Sequential integration*, *Salience of relevant cues*).

The influence of these factors is then evaluated against an experientially based judgment of the degree of ambiguity of the phonetic segments in question. In this judgment I want to use an approach that I have called *reduced phonetic listening*, since it is not too far from Schaeffer's *reduced listening* (cf. section 2.5.1). In this approach, I try to get rid of or 'put in brackets' all the contextual knowledge I possess and to focus on the quality of the phonemes and syllables *as sounds.* I will also use similar techniques as Schaeffer in achieving this, namely isolating shorter segments of the sonic flux, and then repeating these over and over. However, instead of using a typo-morphological framework to qualify these sounds, I will compare them to mental templates and actual acoustical realizations of phonetic segments, so as to decide which phoneme(s) the experienced sound is most similar to, and also the degree of ambiguity with which this judgment can be made.[255] When presenting the transcriptions of this kind of reduced phonetic listening, I will use this nomenclature for the different degrees of ambiguity:

Highly ambiguous with no candidates	**Ambiguous, with several candidates**	**Ambiguous**	**Unambiguous**
[(?)]	[(b/n/v)]	[(b)]	[b]

Table 9.2: Nomenclature for transcriptions of phonemes resulting from a reduced phonetic listening.

[254] I am familiar of no studies of the degree to which aspects of electroacoustic works are forgotten by a listener over time, but one should assume that, as with other kinds of experiences, long-term memory of these pieces fade over time.

[255] Two valuable resources in this process has been Peter Ladefoged's book (with CD) *Vowels and Consonants – An Introduction to the Sounds of Languages* (Ladefoged, 2005) and the interactive IPA chart found at the web site of the Department of Linguistics at the University of Victoria, Canada, available at URL: http://web.uvic.ca/ling/resources/ipa/charts/IPAlab/IPAlab.htm, accessed 23/04/2010.

In the final phase of the evaluation, I will then judge whether any ambiguities that I identified in the reduced phonetic listening could be resolved by referring to contextual, top-down oriented issues. If several contextual aspects *converge* in giving high probability for a particular word or phoneme, it will be evaluated as being *unambiguous*, and if, on the other hand the aspects render conflicting indications or if they indicate very little that can be of aid, it will be evaluated as *ambiguous*. For example, if a phoneme sequence is repeated several times and one certain phoneme is only ambiguous in one of these repetitions, it is more likely that it is the same phoneme in all cases. Thereby, ambiguity will not likely lead to reduced intelligibility. Conversely, if the semantic context provides few cues for the interpretation of an ambiguous phoneme, and this opens up for several possible interpretations, ambiguity can have a negative effect on comprehension. With this approach, I will be able to make an assessment of the ambiguity both for sentences and words as well as for phonemes which will be used when locating a vocal phrase somewhere along the maximal-minimal continuum for this premise.

9.2.1.2 Criteria for the evaluation

In the evaluation of clarity of meaning, I will use three different, but interrelated measures: 1) The clarity of the **LI-domain**, 2) the specificity of the context within which verbal material is situated, in particular the identity categories (**ID-domain**) with which the vocal persona is identified, and 3) the coherence within and between the aspects of the different domains. The evaluation will then be a judgment of the combination of these. How these criteria are related to the five categories between maximal and minimal, is given a schematic representation in **table 9.3**. This table will be explained below.

For the first of these measures, I will use the strategy discussed in the previous section, with a reduced phonetic listening strategy combined with an analytical evaluation of the influence of the factors from section 9.2. The main idea in the evaluation is to assess the *proportion* of (un)ambiguous words or phonemes in a vocal phrase relative to *listening order*, i.e. whether something can likely be comprehended at first listening or at subsequent listenings. To do this, it is necessary also to evaluate the *number* of words, syllables or phonemes that can be heard in the vocal phrase altogether. One can then simply judge how many unambiguous/ambiguous words or phonemes a phrase contains *relative to* the total number. But, to decide upon the total number of words or phonemes in a vocal phrase, is not

Domains **Evaluation**	**LI-domain clarity**	**Contextual specificity**	**Within domain / between domain coherence**
max	**1st listening:** Whole sentences / all words comprehended **Subs. listenings:** as 1st listening	High specificity High certainty in identification	Highly coherent ↕ Highly incoherent
max-int	**1st list.:** Some-most words recognized **Subs. list:** Up to all words recognized	Medium specificity Uncertainty for some of the less salient features	
int	**1st list.:** None-some words recognized, or all phonemes **Subs. list:** Up to int. words	Ambiguity for some of the most salient features	
int-min	**1st list.:** Some-most phonemes recognized **Subs. list:** Up to all phonemes recognized	Ambiguous for most features	
min	**1st list.:** None-some phonemes recognized **Subs. list:** Up to int. phonemes	Highly ambiguous / unspecified for all features	

Table 9.3: Criteria for evaluation of the clarity of meaning premise. Key to abbreviations: max = maximal, int = intermediate, min = minimal, 1st list. = first listening, subs.list. = subsequent listening.

always straightforward, unless the number of ambiguous units is low, or if a written version of the verbal content of a piece exists. If there are many ambiguous units, however, it can be impossible to say exactly how many words or phonemes there are. An alternative, then, can be to base the evaluation on *syllables* instead, since it is often easier to identify the number of syllables in a vocal phrase, especially compared to words. In its turn, an evaluation on the syllable level can be related to the word level by multiplying the number with an average value of syllables per word for the language in question. For English, this number is 1.5, according to Yaruss (Yaruss, 2000). For example, if I have identified that a vocal phrase has 14 syllables, dividing it by 1.5 gives an approximation of the number of words of about 9. Finally, one can then make an assessment of the proportion of ambiguous words or phonemes relative to those that are not. As for the maximal-minimal continuum I will use five categories in this operation, corresponding to the five main evaluative categories of my framework, but referring specifically to *proportion*; *all*, *most*, *intermediate*, *some* and *none*, where *all* and *none* should be interpreted loosely, i.e. as roughly all or none.

In the evaluation, I will start by assessing the proportion of (un)ambiguous words that can be identified, and assess the likeliness that the identification would occur at first listening or subsequent listenings. This is represented in the second column from the left in **table 9.3**, labelled "**LI-domain** clarity". Here, we can see that if any words can be identified at all, the evaluation will be somewhere between intermediate and maximal. If the words are part of a sentence which likely can be comprehended without ambiguity and in its entirety at the first listening, the evaluation is maximal. If no words can be identified, the evaluation will be based on the proportion of (un)ambiguous phonemes, and if so the evaluation will be between the intermediate and the minimal. If between *none* or *some* phonemes are unambiguously identified, the evaluation is minimal.

The second measure is to judge the specificity of contextual features. Here, the features belonging to the **ID-domain** will be particularly important, but the features of the **SE-domain** will also be considered. This measure is represented in the third column of the table, labelled "contextual specificity". As we can see from the table, when these features are highly specific and can be recognized with high certainty, the evaluation is maximal. When the most salient features, such as gender and age, are ambiguous, the evaluation is intermediate, and when all features are highly ambiguous or cannot be specified, the evaluation is minimal.

The third measure is to assess if there is any incoherence within and between features belonging to the different domains, where the highly coherent implies maximal evaluation and highly incoherent implies minimal evaluation, as one can see from the rightmost column in **table 9.3**. However, it might be difficult to assess the degree of coherence when features are ambiguous or low in specificity, since incoherence features that clearly point in opposite or different directions. When the contextual specificity is evaluated as lowered, I will therefore give correspondingly less weight to the coherence measure.

Lastly, these three measures will have to be seen together in a global evaluation, where one will have to assess the importance of each of the measures in the whole picture. Then, one will have to weigh the evaluations of the different measures, so as to reach an overall judgment. In theory, this might seem like a complex operation. In my experience, however, it seems that the three measures are partly correlated, so that the evaluations will not be too problematic. This will become evident in the following section, where I will apply the explicated measures and criteria on excerpts from five different electroacoustic pieces, so as to demonstrate different evaluations in the five categories along the continuum from maximal to minimal.

9.3 Evaluation for musical examples

9.3.1 Maximal: Hildegard Westerkamp, *Kits Beach Soundwalk*

Again, I want to use the beginning of Hildegard Westerkamp's *Kits Beach Soundwalk* as an example of the maximal category, since it highlights many central issues in the evaluation of this premise. In the first one and a half minute or so of this soundscape composition from 1989 (Westerkamp, 1996, 0:00-1:37, **sound example 9.2**), we can hear a woman speaking, who is apparently situated in a natural environment close to water, but where the noise of a big city can be heard in the background. The following is a transcription of the verbal content:

> It's a calm morning. I'm on Kits Beach in Vancouver. It's slightly overcast and very mild for January. It's absolutely wind still. The ocean is flat – just a bit rippled in places. Ducks are quietly floating on the water. I'm standing among some large rocks full of barnacles and seaweed. The water moves calmly through crevices. The tiny clicking sounds that you hear are the meeting of the water and the barnacles. It trickles and clicks and sucks and…The city is roaring around these tiny sounds, but it's not masking them.

LI-domain clarity: The female vocal persona speaks calmly and clearly and with short and most often concise sentences, so that it is very easy to comprehend what she says. In doing a reduced phonetic listening of this example, I found that there were only very few ambiguities in this relatively long example, and that the ones I detected were very easily resolved. For example, the phoneme sequence [ɪtslaɪtlɪ] can be interpreted as both “it’s lightly” and “it’s slightly”, and both interpretations are almost synonymous and fit with the context. Since the [s] is relatively long, I find that “it’s slightly” is a quite probable interpretation. Moreover, the [d] in “seaweed” is masked by the background sound, but contextual and lexical redundancies makes “seaweed” into a highly probable interpretation. The few other ambiguities that I found did not pose any problem for comprehension, and I assume that I would very likely comprehend all of what was said during the first listening, something that implies a *maximal* evaluation.

Contextual specificity: As for the **ID-domain**, there is little verbal information to be gained from the **LI/sem** here, perhaps except that the female speaker appears to be somebody interested in the sounds of nature.[256] Other vocal features, however, tell me that this is a middle-aged woman, probably in her late thirties or early forties, who, judging from her English pronunciation is North-American (with my limited knowledge of North-American dialects I can’t identify her regional belongingness more specifically). The calm and slow way of speaking, with a relatively low tone of voice, suggests a balanced personality with self confidence. Hence, the **ID-domain** can here be specified according to three of the most salient features; gender, age, and regional belongingness, in addition to suggesting a few aspects of personality. Other less salient features, however, cannot be specified. Thus, according to **table 9.3** this implies a *maximal-intermediate* evaluation.

Moving on to environmental setting in which the speaker is situated, we can see from the transcription of the verbal content above that it is described in a specific and rich manner. We get information about the time of the day and of the year, location and geographical setting, weather conditions, and what creatures, objects and substances are present in the surroundings.

[256] When listening to the rest of the piece, one can probably also infer that the person speaking is a person that is familiar with techniques of sound processing, that she has knowledge of experimental music (Xenakis), that she has negative attitudes towards the city (referring to it as a “monster”), and that it is a spiritual person in referring to the healing powers of sounds and sounds in dreams.

The sound-source related aspects of the **SE-domain** also reveal a whole lot of information about the environmental setting:

- There are water sounds that vacillate between immediate foreground and middle ground.
- There are clicking sounds associated with water.
- One can hear several kinds of birds, among which I think I can identify ducks and crows, in the middle ground.
- There is a blanket of noise from a city in the background. One can also hear a car horn and a propeller air plane in the middle ground.

Therefore, much of the information about the environmental setting – like the presence of water, the city in the background, and the ducks on the water – is presented both as source/cause-related sounds and as verbal content (**LI/sem**). Some of the other bird sounds besides the ducks, the car horn and the air plane sounds, however, can only be inferred from the environmental sounds in the middle and the background. Hence, the **LI/sem** and the features of the **SE-domain** converge in some aspect and complement and enrich each other in other aspects, together creating a *highly specific* image of context. Moreover, it is strongly suggested that one should interpret the location of the speaker as identical to where the verbal information tells us that she is, rather than interpreting this as a constructed scene.[257]

Within and between domain coherence: Taken together, we have seen that the verbal and non-verbal cues along with the accompanying familiar environmental sounds create a highly *coherent* context which one can use in the interpretation of the verbal contents conveyed by the voice.

[257] However, whether this is the case or not is more difficult to verify by listening alone. The lack of reverberant characteristics that colour all sound sources present can make a listener infer that the voice has been either recorded in the studio, or alternatively recorded with a separate microphone from the rest of the soundscape having very short range, thus minimizing reverberant sound. The information provided on the CD, stating that this is a piece for "spoken voice and tape", suggests that the spoken voice is recorded in a separate session from that of the soundscape. The reference to time and place by the speaker, pointing to a "here and now", may therefore prove to be fictional. Still, my experience of this scene as projecting an almost documentary setting is quite strong.

Global evaluation: Taken together, the three measures of **table 9.3** give evaluations that are as follows: 1) **LI-domain** clarity: maximal, 2) contextual specificity: maximal for the **SE-domain** and maximal-intermediate for the **LI-domain**, and 3) within and between domain coherence: maximal. While this evaluation thus implies that the clarity of meaning could potentially have been somewhat higher with even more specificity for the **ID-domain**, the global evaluation is still clearly lying within the *maximal* category.

9.3.2 Maximal-intermediate: Katharine Norman, *Losing it*

In the excerpt from Katharine Norman's *Losing it* (Norman, 2004b, 0:07-0:48, **sound example 9.3**) one encounters a voice which resembles what one can sometimes hear in radio or television interviews with people whose voices are scrambled to protect their identities for some reason: The fundamental frequency seems "blurred" and unclear, and the spectral envelope appears to be "warped" downwards so that the voice projects a slightly monstrous quality. A verbal transcription of the example reads as follows:

> Losing it. Losing it. Close your eyes. Close your eyes. Close your eyes. Sleep. Sleep (sleep). You're losing it. You're losing *it*. You're losing it. Trying to smudge that white line of consciousness.

Clarity of the LI-domain: The clarity of the **LI-domain** is high in this excerpt despite the relatively heavy processing that clearly has been applied here. From a reduced phonetic listening, I could only detect a few places with ambiguous phonemes and where my transcription did not correspond to a word:

- At 0:20-0:24, the phoneme sequence [sliː] is repeated twice, hence creating potential ambiguity. However, this potential ambiguity is easily resolved, since the word "sleep" and the phrase "close your eyes" have been heard just a few moments earlier.
- At 0:32-0:34, several phonemes in two syllables of the phoneme sequence [yɔɹluːs(?)(?)t] are masked by accompanying sounds. Already before the ambiguities occur, though, it is evident that this is merely a repetition of the phrase "you're losing it", that has been heard several times before.

Consequently, the few ambiguities of the **LI-domain** are easily resolved as a result of the multiple reiterations of most of the verbal phrases.

Contextual specificity: As for the specificity of the **ID-domain**, the processing of the voice makes the gender of the voice highly ambiguous – the spectral characteristics suggest a male voice, but something in the pronunciation suggests a female voice.[258] Other features are less ambiguous: the voice undeniably sounds adult, and it is clearly British. Moreover, the pronunciation suggests to me a relatively well-educated person, even if I am less certain of this. Together, these issues imply an evaluation as *intermediate*.

The features of the **SE-domain** are even more ambiguous. Here, short snippets of many types of sounds are juxtaposed and superimposed, together creating a relatively chaotic impression, whose features are related more to the **TCM-domain** than the **SE-domain**. There are a few of these snippets that suggest natural environments, projecting very briefly the image of a forest with singing birds in it. Together with the many clearly synthetic and processed sounds subjected to different types of organization (cf.2.5.1), however, the whole situation is what I would characterize as ambiguous for most features, i.e. implying an *intermediate-minimal* evaluation.

Within and between domain coherence: I would not say that features within or between domains are particularly incoherent relative to each other, unless one chooses to see the natural environment suggested by the short snippets of bird song as incoherent with the synthetic sounds that accompany them. Rather, the verbal material appears coherent in presenting a situation which suggests a person's inner speech in trying to fall asleep. If one interprets the accompanying sounds as metaphorical representations of a form of "mental noise" that one can experience when struggling to sleep, the whole picture appears quite coherent, thus implying an evaluation as *maximal-intermediate*.

Global evaluation: The high clarity of the **LI-domain** is here counterweighed by the lack of specificity in the **ID-**and **SE-domains**, which together created a relatively low degree of contextual specificity. As for the measure of coherence, I do not experience that it plays an important role in the experienced clarity of meaning. In sum, therefore, I will evaluate this excerpt as being *maximal-intermediate*.

[258] For some listeners, this ambiguity might be resolved when the female voice enters later in the composition. For others, though, the female voice might be seen as accompanying the ambiguous one.

9.3.3 Intermediate: Trevor Wishart, *Stentor* from *Two Women*

The next example is taken from the third movement of the piece *Two Women* by Trevor Wishart (1998, on Wishart, 2000b, 0:38-0:42, **sound example 9.4**), titled *Stentor*. In this sound example, the short excerpt that I will evaluate is first presented in isolation, thereafter in context, accompanied by the preceding and following few seconds (0:33-0:48). In this short excerpt, we can hear a markedly processed voice with a highly noisy sound spectrum, which forced me to listen many times until I was able to decode the verbal material. This was despite the fact that I was familiar with the written transcription of the text of the whole movement in the CD-booklet:

> Ian Paisley (on Margaret Thatcher): "Oh God, defeat all our enemies … we hand this woman, Margaret Thatcher, over to the devil, that she might learn not to blaspheme. And, Oh God in wrath, take vengeance upon this wicked, treacherous, lying woman … Take vengeance upon her O Lord" (CD liner notes, Wishart, 2000b)

Clarity of the LI-domain: The transcription based on my reduced phonetic listening can be seen in **figure 9.1** below. As devised by **table 9.2**, ambiguous phonemes are put in parentheses, but in this case, where there is more than one likely candidate, the different candidates are stacked on top of each other in the four panes in the figure. As can be seen from this transcription, many of the phonemes, especially the consonants, are ambiguous.

(n)	ɛ	s	w	i	k	(ʌ)	(n)	(t)	ɹ	ɛ	n	(t)	ɹ	ɛ	(n)	s	(x)	ə	n	(ɻ)	ai	(d)	i	ŋ	w	o	m	ə	(n)
(θ)						(ə)	(d)	(k)				(tʃ)			-		-			(v)	ai								
(l)								(f)												(l)									
																				(h)									

Figure 9.1: Transcription based on reduced phonetic listening of an excerpt from Wishart's *Stentor* from *Two Women,* using phonetic symbols (IPA, see Appendix A). The IPA symbols that appear within parentheses are ambiguous. Possible candidates for these phonemes are placed underneath one another.

The ambiguity of the phonemes can indeed affect the interpretation of the verbal message, since it creates many ambiguities on the word level. The most important of these are:

- [(n)/(θ)/(l)ɛs] can be interpreted as the suffix "–ness", the word "less", as the ending of the word "(un)less" and as the word "this", spoken with a dialectal variation.
- [wik(ʌ)/(ə) (d)/(n)] could be interpreted as "we can", "weaken", "weak and", "weekend", "we could", and "wicked". Seen in the light of an earlier reference to a collective "we", both "we can" and "we could" appear as likely candidates here. Moreover, due to the negatively laden context words like "weak and", "weaken", and "wicked" also seem like probable alternatives. Only "weekend" stands out as an unlikely choice in this case.
- The phonemes [ən] have only a few candidates; "an" and "and". Since all the possible candidates for the next phonemes are consonants, "an" is a very unlikely candidate in this case, making "and" the only likely word here.[259]
- [(ɻ)/(v)/(l)/(h)ai(d)iŋ] could potentially be interpreted as "riding", "lying", "vying", and "hiding". Combined with the prior two phonemes, [ən], it could alternatively be interpreted as "inviting". Given the negatively loaded context, the word "lying" appears more likely than "riding", maybe with "vying" and "hiding" as intermediately probable candidates.
- The word "woman" seems to be one single and quite likely candidate at the end of the excerpt, transcribed as [wʊmə(n)]. This appears especially likely on the basis on the other references to a female person that can be heard in the piece; "woman", "her", "she" and "Margaret Thatcher".

Based on these considerations, I will conclude that for the whole excerpt, the words "and" and "woman" are the only candidates that stand out as relatively unambiguous choices for an interpretation, and that the word "lying" is only slightly more likely than its contenders based on contextual information. This is due to factors such as (the lack of) *prosody*, very short

[259] It must be noted that this goes contrary to the written transcription in the CD-booklet.

adaptation time to the kind of processing applied, and low *salience of the acoustic cues*, with weakened difference between vowels and consonants, foreground and background. And, to resolve some of the ambiguities here, one would likely have to listen to this excerpt several times. Hence, only two of the five words “this wicked, treacherous and lying woman” are likely to be unambiguously interpreted after many listening sessions, hence indicating an evaluation as *intermediate*.

Contextual specificity: As for the identity of the speaker in this excerpt, it was given in the CD liner notes as Ian Paisley. Neither Paisley nor his voice were familiar to me prior to listening to this piece, and consequently knowing the name did not affect my experience of the **ID-domain** much in this case. However, this information made me relate all the vocal events, including what I focus on here, in the movement to one single identity, for whom I could specify some identity features; it is a middle-aged man with an Irish accent, who has a speaking style with high vocal effort, using short, quasi-rhythmic phrases separated by relatively long pauses, placing heavy emphasis on each phrase. On those grounds, I assume that the voice is addressing a larger number of people and that he is attempting to communicate clearly and to use persuasive power to convince his audience. In other words, he communicates within a *social space frame*. Together, this implies medium specificity and *maximal-intermediate* evaluation.[260]

The environmental setting of the movement as a whole is quite confusing, with many kinds of conflicting and ambiguous cues, suggesting simultaneously an indoor *and* an outdoor setting. Furthermore, some cues, in particular the banjo, synthetic bass drum and hand clap together with a cheering crowd, suggest a kind of concert or fair, whereas other cues are very difficult to link with any particular setting at all. For the short excerpt in question here, however, it is hard to suggest anything regarding the environmental setting and the **SE-domain**. And, since the excerpt differs quite a bit from the rest of the movement with its overall noisiness, it might be perceived as happening in another place than the remainder of the movement. Hence, I will regard it as being ambiguous for most features, implying an *intermediate-minimal* evaluation.

Within- and between-domain coherence: There are many issues that can be experienced as conflicting and incoherent in this movement as a whole in addition to the mentioned

[260] If one considers the excerpt in isolation, however, the certainty in identifying these features is somewhat lower.

ambiguities regarding environmental setting. For example, that the voice several times appears to be split into several voices with different pitches, but with synchronous articulation, and that the vocal persona at some point appears to morph gradually into a cartoon character, might be regarded as incidences of within-domain incoherence. This excerpt in isolation can be seen as in part coherent with the rest of the movement in that it presents the same vocal persona, but also incoherent since it presents it with a different voice and in a different context. Moreover, the blurring between the vocal persona in the foreground and the environmental context in the background can also be regarded as a form of incoherence or ambiguity. The total picture is therefore one of an intermediate situation between coherence and incoherence.

Global evaluation: Here, the relatively low clarity of the **LI-domain** is accompanied by maximal-intermediate specificity for the **ID-domain** and intermediate-minimal specificity for the **SE-domain**, hence with two measures that counterweigh each other. Adding intermediate coherence to this picture, I find it quite straightforward to assign an *intermediate* global evaluation of this excerpt.

9.3.4 Intermediate-minimal: Trevor Wishart, *The Division of Labour*

In Wishart's *The Division of Labour*, the first movement of *Fabulous Paris: A Virtual Oriatorio* from 2007 (Wishart, 2007), there is a passage from about 3:35 where phonemes or syllables are presented in a sequence of progressively shorter and shorter fragments. The excerpt, which I have included in **sound example 9.5** (3:30-4:10), begins with a sequence of looped syllables, which gradually fade out and slow down while the fragments are introduced from about 0:05 in the example. Many of the syllables and phonemes seem to reappear several times during the sequence, during the course of which the duration of the fragments and the rate with which they are presented increase.

Clarity of LI-domain: In doing a reduced phonetic listening of this excerpt I found that I was able to identify a lot of phonemes in the example as a whole, but none of these could be joined into words. However, the proportion of ambiguous phonemes appeared to change during the excerpt, probably because the fragments containing the phonemes became progressively shorter and overlapped more and more.[261] While I found that in the first seconds after the fragments enter (about 0:07) *most* phonemes were unambiguous, only *some* were identified in the last few seconds of the excerpt. Hence, clarity moves between the upper and the lower boundary of the *intermediate-minimal* evaluation, as it was defined in table **9.3**.

Contextal specificity: As for the identity of the vocal fragments in the excerpt, it does not seem to be assigned to one single vocal persona when I listen to the fragments in isolation, one at a time. When I compare phonemes and syllables, some of them have similar vocal quality, but others are quite different. All clearly belong to a male adult speaker, but I can't specify anything other than the gender from the fragments alone. When I listen to the whole excerpt continuously, however, I experience everything as related to the same vocal persona, and this vocal persona is one that I recognize from the introduction of the whole piece – a male English speaker with what I take to be a Scottish accent reading a text about the industrial production of pins. Thereby, the identification of the features of the vocal persona depends partly on the context of the rest of the piece in this case. One can therefore argue that the features of this excerpt are in themselves more ambiguous and less specific than the initial voice in terms of the **ID-domain**, resulting in an evaluation as *maximal-intermediate*, perhaps approaching the boundary of an *intermediate* evaluation.

When it comes to the **SE-domain,** the individual fragments, especially the ones with the longest duration, bear the mark of being recorded in a studio, with minimal accompanying noise and reverberation, and the microphone placed close to the speaker. On that basis, contextual specificity is not maximal for the fragments in isolation. When heard together, however, the back and forth movement in perceived distance and horizontal location imposes a marked ambiguity, and I would therefore say that specificity is *intermediate-minimal* for the environmental setting.

Within- and between-domain coherence: An obvious incoherent feature related to the **SE-domain** is that several versions of the same voice appear simultaneously and moving very fast from side to side, clearly indicating that the whole situation is constructed and artificial. In my view, the whole situation implies *intermediate-minimal* coherence.

[261] This process culminates as the fragments get progressively blurred so as to almost fuse into one granular stream, which also gets gradually coloured by a chord, apparently created with a bank of band-pass filters. Then, the fragments get gradually more distinct from each other, in what seems as something close to a reversal of the initial process.

Global evaluation: The low evaluation of the clarity of the **LI-domain** is to some degree counterbalanced by contextual specificity for the **ID-domain**. I still feel that these measures are not sufficient to shift the evaluation to the next category on the evaluation continuum, and that this example clearly has a lower evaluation than the excerpt from *Stentor*, discussed above. I therefore find it pertinent to evaluate this as having *minimal-intermediate* clarity.

9.3.5 Minimal: Jean-Claude Eloy, *Shànti*

In Jean-Claude Eloy's monumental work *Shànti* (1973, on Eloy, 1979), a strongly distorted voice appears in the first section of the piece, from which I have chosen an excerpt (6:34-7:00, **sound example 9.6**). Accompanying this voice are several layers of synthetic, pitched sustained sounds, many with a high degree of fluctuation in several parameters.

LI-domain clarity: In doing a reduced phonetic listening of this excerpt, I found that only very few phonemes, i.e. somewhere between *none* and *some*, could be perceived without ambiguity, and these were mostly vowels of long duration, as for example the [ɛ] about 8 seconds into the example. Therefore, the **LI-domain** clarity is evaluated as *minimal* in this case.

Contextual specificity: The distortion and the varying degrees of masking by the accompanying sounds make most of the **ID-domain** features ambiguous. The voice is not straightforward to define in terms of gender and age, but the cues hint at an adult male in my ears. The high ambiguity regarding linguistic cues also make regional and social belongingness ambiguous, even if the intonation contour for me suggests an English speaker, perhaps with a British accent. Lastly, it is not evident what function/social role/occupation the voice projects. Nevertheless, the phrase structure, the accented syllables, the degree of vocal effort, the pauses – in short, many of those features that were subsumed under the label of *prosody* in section 3.6.3 – suggest a kind of vocal rhetoric associated with a speech or public address. Hence, the identity features can only be assigned with a high degree of ambiguity or uncertainty, owing to the heavy processing and the partly masking background accompaniment which lower the salience of the acoustic cues (cf. *salience of relevant* *acoustic cues* factor). I will therefore regard the features related to the **ID-domain** as being ambiguous for most features, i.e. *intermediate-minimal*.[262]

As for the sounds accompanying the voice in this excerpt, they are all highly abstract and very difficult to relate to an environmental setting. The only feature one might infer in this case is related to the setting implied by the vocal qualities hinting at a public address. It is impossible to tell, however, whether this is made in public or via broadcasting, if it takes place indoors or outdoors, etc. Therefore, since these features are relatively ambiguous, they do not represent a marked increase in specificity, and I will therefore regard specificity as being somewhere between *minimal* and *intermediate-minimal*.

Within-domain and between-domain coherence: Since the ambiguity and specificity of the other measures are as high as they are, there are no features that are manifestly incoherent. Consequently, I will not assign any weight to this in the global evaluation.

Global evaluation: Having *minimal* clarity of the **LI-domain** and somewhere between *intermediate-minimal* and *minimal* for contextual specificity, I end with a global evaluation that is *minimal*, but that still is somewhat higher than the theoretical minimum of this premise.

9.4 Chapter conclusion

In this chapter, I have shown that the premise of clarity of meaning is a relatively complex one, relying on an interplay between many factors related to many different experiential domains. For this domain, I had to relate the premise to three different, but partly interrelated *measures*. The first of these was the clarity of the **LI-domain**, which was closely related to issues of linguistic processing discussed in section 3.6. As a result of the effects of the *dynamics of listening*, I had to do an assessment of the available linguistic cues through what I called a *reduced phonetic listening*, which implied a judgment of the degree of ambiguity for short linguistic segments by listening repeatedly and isolatedly to them. This evaluation was then seen in relation to contextual cues and competence so as to estimate the probability that either sentences, words or phonemes could be unambiguously perceived and comprehended. The second measure was related to all the aspects that were seen as contributing to the overall

[262] From the liner notes on the CD one can read that the voice of Eldridge Cleaver, recorded in 1968 when he was leader of the Black Panther Party, can be heard in the first section (*Face I*) of the piece. However, while one can easily associate the much less processed voice that occurs at about 11:00 with Cleaver, the differences in voice quality make the association between this voice and the highly distorted one quite weak. Nevertheless, I suspect that since information about no other persons is included, the distorted voice is Cleaver's.

meaning of a vocal phrase – i.e. including features from all the domains in this framework. These features could to a higher or lower degree define the *contextual specificity*. Here, I argued that features of, in particular, the **ID-domain** and the **SE-domain** were central in defining such contextual specificity. Lastly, the coherence or lack of coherence within or between the features of the experiential domains had to be taken into account, since incoherence between features can lead to ambiguity and lack of clarity. In the evaluation of musical examples, I then demonstrated how these measures could be combined so as to make a global or overall evaluation of the premise.

10.0 Feature salience

> **Premise six of the max-min model:**
> *Feature salience*: Vocal sounds and features "stand out" perceptually – for themselves and relative to other sounds and features.

In research on human perception and cognition the term *salience* is often defined as an attribute of an object that makes it stand out or pop out relative to what surrounds it, thereby making them easier to detect and giving them higher potential for capturing our attention (see e.g. Murphy et al., 2003: 16; Kayser et al., 2005). In the first place, this premise deals with the relationship between a vocal sound and other sounds that might accompany, precede, or follow it.[263] In the maximal end of this premise, however, vocal sounds are unaccompanied by other sounds altogether, or accompanied by sounds that are situated clearly in the background, so that they do not interfere with or mask them. Conversely, in the minimal end of the continuum, vocal sounds are barely heard among accompanying sounds. In the second place, the premise also states that the *features* of vocal phrases have to be salient, and by that I refer to the material features of the voice belonging to the **VG-domain**. If a vocal phrase is processed so that some of its features stand out less or are experienced as missing, I will regard it as having less feature salience, even though there are no other accompanying sounds or features present. This can for instance be the case if a spoken phrase is heavily high-pass filtered so that only the sibilants in the high frequency region can be clearly detected.

The premise of salience has partial overlaps with several other premises. It is closely linked to attentional capture (cf. section 5.1.3), and was therefore regarded as one of several factors that influenced *focus of attention*. Moreover, salience of relevant cues is a factor for the *clarity of meaning* premise, hence making that premise partially dependent on this one. The current premise also deals with many of the same issues as the premise of *stream integration* to be discussed in the next chapter, but in many ways, from an opposite point-of-view.[264] Whereas the *feature salience* premise deals with the degree to which properties of the

[263] The premise can also be seen in relation to Dyson's designation of the radio voice as *singular*, since when heard in the company of many other voices, it will either have to increase the volume or risk that it is rendered inaudible: "For instance, the dissident, interviewed on the street, in front of a screaming crowd, some distance from the microphone, has to shout in order to be heard. Perhaps some words get lost in the din […]" (Dyson, 1994: 181).

[264] Bregman discusses the relationship between stream fusion and masking: "We see that masking and fusion seem to be affected by the same sorts of acoustic relations between a target component of a larger sound and the remainder of this sound. Generally speaking, these are relations that, in a natural listening environment, are useful for deciding whether different spectral components have arisen from the same acoustic event. Despite differences in the measurement operations that define fusion and masking, the two may be affected by common underlying mechanisms" (Bregman, 1990: 325).

vocal sound *stand out* or are separated from other sounds, the premise of *stream integration* deals with the degree to which the parts, segments or components of the vocal sound *hold together*. In other words, the former is primarily oriented at forces of *segregation* while the latter is oriented toward forces of *integration*. Many of these forces are undoubtedly complementary, e.g. are both partly dependent upon relations of similarity and dissimilarity, but where dissimilarities and contrasts increase the experienced salience, the opposite is the case with the stream integration premise; here, it is primarily similarity and contiguity that integrate parts into wholes. [265] What is more, *loudness* is an important factor in determining how salient a sound is to a listener (see section 11.2), but it is less important in integrating a sound stream. Hence, a vocal sound can be well integrated into a separate sound stream, but may not be very salient due to its low loudness.

10.1 Theoretical perspectives

Salience can be linked to two issues in auditory perception that can be seen as related to each other, namely *feature contrasts* and *masking*. I will start out by giving a brief theoretical account of these before I go on to list the separate factors that can affect the evaluation of this premise.

10.1.1 The perceptual importance of contrasts

An important feature of our perceptual system is that it is tuned to detect *differences* or *contrasts* in the environment. In general, objects that are markedly differentiated from their surroundings tend to attract our attention and be more easily detected than objects that are similar to their surroundings (Truax, 2001: 19). The perceptual predisposition for detecting differences can also be observed in the temporal domain, where we tend to perceptually enhance sounds that have properties that distinguish them from what we are already hearing.[266] What is more, if a sound is completely static without any temporal changes in it, we will gradually be *habituated* or *adapted* to it, so that after some time, the sound will be less prominent to us, and even be perceived as being softer than it was at its onset (Jerger,

[265] Ockelford sees similarity and contrast as lying along a continuum from *sameness*, via *similarity* and *difference*, to *contrast* where none of these terms can be defined in absolute terms (Ockelford, 2004: 35).
[266] This enhancement can be observed when noise with a notch in the spectrum is followed by uniform white noise, which results in that the uniform white noise is heard as having a spectral peak in the same location as that of the notch (Moore, 2003: 279-80).

1957).[267] For example, in the office where I am writing this dissertation, there is a ventilation system with an air outlet only one metre from my desk. Most of the time, I don't even think of this sound as being present in my environment because my perceptual system has "tuned it out". If I arrive early at the office, however, I will be there at the moment when the ventilation is turned on, something which makes me immediately notice and be annoyed by it, and be baffled by the sheer loudness of the sound. After some minutes, I have gotten used to it, and my perceptual system and attention can again "tune it out" to focus on other things. Thus, what makes the ventilation sound salient at one point is the *contrast* between what I hear before and after its onset. In other words, this is a kind of contrast that is articulated in time, which we can formalize as *old situation* || (contrast) || *new situation*, where there is *continuity* within the old and new situations, but a *discontinuity* between the two. Moreover, we can also regard this situation as a *dissimilarity* between the old and the new situation.

If we turn to music theory, we can see that several authors have given perceptual discontinuities as well as issues of similarity and dissimilarity a central position in their theories. For Rolf-Inge Godøy, the articulation of discontinuities in time is seen as crucial for our understanding of all kinds of temporal phenomena in the world, and sound and music in particular. (Godøy, 1997). With support from phenomenology, hermeneutics, cognitive psychology and morphodynamical thought, he contends that all understanding happens in a *quantal* manner, i.e. it proceeds in a discontinuous manner from one perceptual "snapshot" to the next (*ibid.*: 28-31). With regards to music, this quantal way of understanding and perceiving tends to make perceptual incisions correlating with discontinuities in the physical world that are within the range of our perceptual sensitivity, as for example in the onset or ending of a sound.[268]

In addition to dealing with temporally articulated discontinuities, music theoreticians have focused on how contrasts or dissimilarities affect *simultaneous* features of perceived (musical) sound. For instance, Deliège defines a *cue* as a *salient* element prominent at the musical surface, and which plays a foreground role in the musical work, against other elements that take on more of a background role. Thus, this appears to be equivalent to the Gestalt notion of *figure-ground* relationships. In many respects this is also dealt with in

[267] Neurologically, habituation has been observed as a decrease in neuronal activity towards an asymptotic level during repeated stimuli. With longer time intervals between stimuli, it has been shown that it takes longer to reach this level (Fruhstorfer et al., 1970; Ritter et al., 1968).

[268] As Godøy emphasizes, detecting discontinuities in the perceptual flux is a process involving a number of factors in addition to physical discontinuities, such as "evolutionary embodied capacities, intentional constitutions, stylistic conventions, or a combination of several factors at the same time" (Godøy, 1997: 60). Categorization based on cultural convention will naturally also play a role here.

Thoresen's *layer analysis* (Thoresen, 1985: 78-90). This type of analysis attempts at describing different aspects of "simultaneous units in a texture", and in the part focusing on "functions and profiles", the link to salience becomes noticeable: "The layers or elements that distinguish themselves as being the more prominent will be said to have a *strong intensity of profile*. When the same layer has strong intensity of profile for a certain time, it is said to have *foreground function*" (*ibid.*: 78, my emphasis). Thoresen's analysis, which considers the "prominence" or "profile intensities" of layers, fits very well with my idea of an evaluation of the salience of simultaneous vocal layers compared to other layers. In my framework, I will see this kind of *vertical analysis* of the total musical environment as complementary to that of the *horizontal analysis* with its focus on temporally articulated discontinuities.

When doing comparisons of features, either vertically or horizontally, one often needs to take into account not only the features in themselves, but also how they change or modulate through time.[269] A sound with varying pitch, for example, will usually stand out from a sound with no pitch variation. A speaking voice usually has continuous modulations both in pitch, spectral type (e.g. between *complex* in unvoiced fricatives and stops and *pitched* in vowels and glides), spectral profile (e.g. different shapes for different vowels) and dynamic profile. Often, it is precisely these characteristic modulations that will make speech stand out from many other kinds of sounds, both those with no or little modulation and those with modulations of other kinds. Since modulations or changes can emerge in a nearly unlimited number of ways, the space of possible differences is very large. One can for instance imagine modulation *shapes* with different overall direction, different number and placement of peaks and valleys, differently shaped irregular or regular fluctuations with all kinds of amplitudes and speeds also applicable to change, different degrees of smoothness of shape, etc.[270] This multitude of possibilities implies that it is very difficult to give a simple assessment of the degree of difference between one sound and another, something that can also affect the evaluation of salience.

10.1.2 Masking

The phenomenon of *masking* is an important issue when it comes to the premise of salience, since when one sound masks another partly or completely, the masked sound will stand out

[269] This is what Godøy in his dimensional approach to the qualification of the sonorous object refers to as "second or n-th order axes" (Godøy, 1997: 192).

[270] This is equivalent to the several of the criteria described in section 2.4.2.2.

less for the listener. Masking is defined in the psychoacoustic literature as: "1. The process by which the threshold of audibility for one sound is raised by the presence of another (masking) sound. 2. The amount by which the threshold of audibility of a sound is raised by the presence of another (masking) sound" (Moore, 2003: 65-66). In this way, masking is defined as a perceptual phenomenon that occurs at the specific point (the threshold of audibility) when the masked sound no longer can be heard in the presence of the masker. The definition therefore describes masking as an all-or-nothing phenomenon.[271] In this framework, however, it might be better to view masking in a relative sense, that is, as a phenomenon which can take place to different degrees – from the situation where no accompanying sounds are present, via situations where the accompanying sound(s) make it difficult to detect all the properties of the (vocal) target sound, to cases where the (vocal) target sound can barely be heard due to the presence of the accompanying sound(s). Alternatively, vocal sounds can mask each other, so that even if one cannot distinguish between any vocal target sound and a masking accompanying sound, one can clearly observe that it is difficult to get access to the properties of the individual voices, precisely because they mask each other. We will see in the discussion below how this is the case in the acousmatic part of John Coulter's *Shifting Ground*.

Owing to the limitations of this dissertation, I will not go into detail about the underlying physiological basis of masking or the relationships between acoustic parameters and masking thresholds. Still, there are some general observations of masking that can be useful to note in this context. First of all, masking is often seen in relation to the so-called *critical band*, which designates the hypothetical auditory pass band filter that is believed to be a part of the process of frequency analysis that one locates to the basilar membrane (Moore, 2003: 66-69). And, it is only when the masking sound contains energy in a spectral band overlapping or adjacent in frequency to a given target that it is effective in masking it (Bregman, 1990: 320). Thus, if one has two sounds that comprise frequency ranges that are well separated from each other, they will not mask each other, no matter how high they are in intensity relative to each other. Since we can see here how a marked difference in spectral range can affect masking negatively, we can thereby also note how *contrasts* in the spectral domain plays a part for masking. We can observe this in **sound example 10.1**, where a band pass filtered recording of a female speaking voice is accompanied by band pass filtered pink

[271] In psychoacoustics, masking is subjected to quite detailed measurements that can give a set of values for the difference in sound pressure level (SPL) between masked and unmasked conditions for different properties of the two sounds.

noise with a bandwidth of two octaves.[272] Since the frequencies of the speaking voice that pass through the filter range only from 600 to 1000Hz, and the frequencies of the pink noise range from 4-12kHz, the auditory filters for the two sounds won't be affected by each other. The result is that no masking occurs even if the noise is gradually turned up in intensity to a quite high level. If, however, there is overlap between the frequency spans that each of the hypothetical auditory filters activate, masking might occur, that is, if the sound levels are not adjusted to prevent it. In **sound example 10.2** the pass band of the filtered pink noise has the same energy and the same two-octave range, but this time the filter is set to 200-600Hz, which is thereby adjacent to the pass band of the vocal sound. As one can hear in the example, the vocal sound will at some point be quite effectively masked by the filtered noise as it gradually increases in level. In this case, the lowest frequencies of the vocal sound will activate the same auditory filters that are activated by the highest frequencies of the filtered noise. Actually, one would also experience that filtered noise in a range further from the pass band of the vocal sound would mask it, albeit less effectively. In the psychoacoustic literature, one can find so-called *masking patterns* that show how the masked frequency range increases with the SPL of the masker (Moore, 2003: 88). For example, a band pass filtered noise with a centre frequency at 410Hz masks the range from 300 to 600Hz at 20dB, and at 80dB it masks the whole range from 100 to about 3500Hz. Thus, one sees that the range extends particularly far on the high-frequency side of the centre frequency, a phenomenon which is referred to as the *upward spread of masking* (*loc.cit.*). So, the louder a sound is, the wider the masking frequency range will be, and especially upwards in frequency. Barry Truax sums up these issues pertinently: "Low frequency and mid-range sounds, and certainly broadband sounds, are more likely to create masking effects than purely high frequency sounds" (Truax, 2001: 82). Consequently, loudness and spectral width will therefore be particularly important factors in determining the masking potential of sounds. This also implies that vocal sounds that are loud and that span a wide frequency range are more "resistant" to masking than softer and more narrow-band voices like the one that could be heard in the examples.

The discussion on masking has so far dealt with what is referred to as *simultaneous* masking, i.e. masking of features present at the same time. Yet, there are also masking effects that are more temporally oriented that need to be considered, since such effects potentially can have large effects on the salience of the properties of the vocal sound. In contrast to simultaneous masking, these temporal masking effects are not restricted to situations where

[272] Pink noise has equal amounts of energy in all frequency bands, having approximately -3dB roll off in energy per octave.

one can distinguish between a masker and a (vocal) target sound. Rather, one will often deal with situations where different segments of the vocal sound mask each other. This might happen simply because of the presence of reverberation or because other factors contribute to "time-smearing" of the segments so they overlap each other to a greater or lesser degree, thus potentially creating masking effects.[273]

One can distinguish between the temporal directions of masking, having either forward or backward direction (uni-directional) or both directions simultaneously (bi-directional). For the cases with uni-directional masking, I have used the terms *tail masking* and *reversed masking*.[274] To demonstrate the effects of these kinds of masking, I have made three sound examples:

- **Uni-directional masking**:
    - **Tail masking** (reverberation, delay etc.): Here, preceding segments mask following ones. Tail masking tends to affect the beginning of phrases to the least degree, particularly if the beginning is preceded by longer periods of silence. In **sound example 10.3**, one can hear a recording of a female speaking voice with added artificial reverberation with long reverb time and a high proportion of reverberated "wet" sound, compared to the direct, or "dry", signal. This makes it very difficult to perceive the properties of the vocal sounds clearly, particularly at the end of the phrase.
    - **Reversed masking** ("gated" or reversed reverberation): Here, following segments are masked by preceding ones. Reversed masking tends to affect the ending of the vocal phrases to the least degree, particularly if the ending is followed by longer periods of silence. In **sound example 10.4**, the same female speaking voice as in the preceding example has been reversed, then artificial reverberation added with approximately the same settings as before, and lastly the sound file has been reversed again. Thus in the final version, the original direction is restored, but the reverberation is reversed. The same

[273] One can argue that that perceptually, this is the same as simultaneous masking, since the reverberated and the dry sound will be heard simultaneously. Since the "source" of the masker, so to speak, is present at another point in time, I will still refer to it as a form of temporal masking. See footnote 274 for two masking effects that are truly temporal.

[274] The terms *tail* and *reversed* masking are chosen above *forward* and *backward* masking because the latter terms refer to masking phenomena inherent in the human auditory system where a preceding masker affects a following target sound (forward masking) and vice versa (backward masking), but where masker and target do not overlap temporally. Since the time range of these effects is rather short (up to 100ms, see Elliott, 1971), I will not take them into consideration here.

difficulties with perceiving the properties of the vocal sounds can be noted, this time particularly at the beginning of the phrase.

- **Bi-directional**: For this type, preceding and following segments mask each other. **Sound example 10.5** is prepared using the phase vocoder-based *pvsblur* opcode in csound, with gradually increasing "blur time", hence gradually making longer and longer segments more and more indistinct. As one can notice, the individual speech sounds are impossible to distinguish towards the end of the example, because of this temporal blurring.

There is one final issue that has to be mentioned when discussing the phenomenon of masking, namely our abilities to restore or compensate for different types of masking. Perhaps the most famous example of this is the so-called *phonemic restoration* effect (Warren, 1970). This effect occurs when a short segment of a speech sequence is removed and replaced with extraneous noise that potentially would have masked the phoneme, such as a cough or another broadband noise.[275] Even if the short segment is not actually present to the listener, most listeners report that they perceive the missing phoneme(s).[276]

An analogous effect, albeit dealing with other types of sound, can be observed in what Bregman refers to as the *illusion of continuity* (Bregman, 1990: 28). This illusion can be experienced when one has an alternately rising and falling sine tone which is periodically interrupted by a loud burst of broad band noise loud enough to potentially mask the sine tone. This situation is illustrated in the upper part of **figure 10.1**. There is actually no masking here, since the sine tone is turned off at the moment the noise burst enters, but listeners still tend to hear the sine tones as one continuous glide. If the noise is removed, however, as in the lower part of **figure 10.1**, most listeners tend only to hear short isolated segments, not a continuous sound.

[275] The requirement that the interrupting sound has to be loud enough to potentially mask the missing speech sound is referred to as the *masking potential rule* (Kashino, 2006).

[276] The replacement of the noise with the appropriate phoneme will have to be done on the basis of redundancies created by different levels of context; lexical, syntactical, grammatical and semantic (cf. section 3.6.1). In other words, one should assume a certain level of top-down influence. Moreover, one should think that effects of coarticulation (cf. section 3.6.2) should also be at work in phonemic restoration, since the articulation of both the preceding and following phonemes would in most cases be affected by the phoneme that is missing.

A third example of perceptual compensation for masking is that listeners appear to compensate for temporal self-masking caused by reverberation. Watkins could report that when given the opportunity and time to get used to the acoustic effects of particular reverberation characteristics on a particular voice, listeners could more easily distinguish between two words that differed slightly than if they did not have this opportunity (Watkins, 2005).

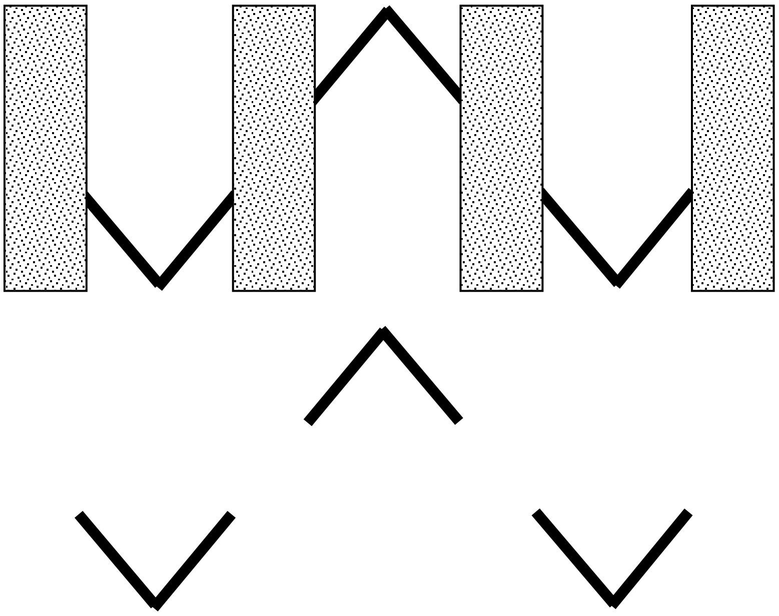


Figure 10.1: The continuity illusion. The solid lines represent sine tones where the horizontal axes designate frequency from low to high, and the vertical axes designate time. The dotted squares represent broad band noise. Adapted from Bregman, 1990: 28.

Taken together, then, these three examples show that the effects of masking can be partly compensated for if the context allows one to collect information that can aid in this compensation, so that the *effective* experienced masking can be lower than the actual masking.

10.2 Factors potentially contributing to feature salience

The factors that I will include for this premise are all highly related to the preceding theoretical discussion. They are *masking* (*simultaneous* and *temporal*), *temporal discontinuities*, *simultaneous contrasts* and *condition of vocal features*.

- **Masking:** Even if, as we have seen, masking is related to many of the feature contrast factors, I want to retain this as a separate factor, since one can have a lack of contrasts in many of the features without masking necessarily occurring. Masking affects salience negatively, and I regard masking as being one of the factors that contributes most in locating a vocal sound towards the minimal end of the premise. In accordance with the discussion above, I want to distinguish between two types of masking:
  - **Simultaneous masking**
  - **Temporal masking:** Includes tail masking, reversed masking and bi-directional masking (cf.10.1.2)

- **Temporal discontinuities:** In the time domain, the *abruptness* with which any changes might take place can affect the experienced salience of the vocal sounds.[277] Here, the *attack* (onset) and *ending* of a sound can play a part, since abrupt attacks and endings will make the discontinuity more salient, whereas very gradual ones will make a sound stand out less (cf. section 2.4.2.2). The degree that one sound is *similar* to the sound preceding it will also affect salience, in that high similarity will render low salience and vice versa. Here, *loudness* is an important feature.[278] A loud sound will in general be more salient than a soft one, at least if no other features distinguish the two, and if none of them are completely masked by other sounds. The greater the positive loudness difference between a target vocal sound and any other sound, the more salient it will be, and vice versa. Furthermore, contrasts in spectral features can also affect the salience of one sound relative to another. This can involve pitch as well as all features included in the explication of sound spectrum in section 2.4.1, such as spectral type, harmonicity, nodality, spectral brightness, width and density.

- **Simultaneous contrasts:** For one sound to stand out relative to another, its features need to be contrasted to the other sound. In general, the more similar two simultaneous

[277] This view is in accordance with attempts of modelling perceptual salience and with research indicating that neurons in the auditory cortex are sensitive to stimulus *edges* in frequency and time. See Coath et al., 2007, Coath & Denham, 2005 and deCharms et al., 1998.

[278] It has to be noted, however, that loudness is not a simple parameter, since it is dependent on several physical parameters such as intensity, spectrum and duration (Roads et al., 1996: 1055). From the so-called equal-loudness contours one can clearly see how perceived loudness changes with different frequencies for sine tones. A sine wave with a frequency of 1000Hz and SPL of 30dB will for example be perceived as equally loud as a sine wave with a frequency of about 80Hz and SPL of 60dB (Moore, 2003: 129).

sounds are to each other, the less one will stand out compared to the other. Just as with temporal discontinuities, loudness and spectral features have a large effect. Other features primarily related to energy articulation (cf. section 2.4.2) might too contribute to experienced salience. Whether sounds are short and impulse like, sustained or iterative, might therefore be influential. Moreover, the degree and manner of variation in a feature can particularly affect the salience of these features. Thus, features such as shape/direction, regularity, abruptness of changes, fluctuations and grain can all play a part here. Lastly, contrasts in spatial location may also contribute positively in making a sound stand out for a listener.

- **Condition of vocal features (VG-domain):** Features belonging to the **VG-domain** might be experienced as degraded or absent altogether due to processing, conditions of the recording (**TCM-domain**), spatial configurations (**SE-domain**) or attributes of the vocal gestures themselves (think of somebody speaking with their mouth half closed). When the features are in prime condition, it will imply high salience, but if they are degraded or absent, it will imply a lower degree of salience.

10.3 Evaluation of the premise

The criteria for evaluating this premise were stated relatively clearly in the premise formulation and the introduction. Hence, it has to do with 1) the relationship between vocal sound and other accompanying, preceding or following sounds, if any are present, and 2) the degree to which the features of the **VG-domain** stand out perceptually so they can be clearly heard. These two points are sometimes independent and sometimes interrelated, and when evaluating the premise, as I will do in the following section, I will assess the combined effect of the two.

10.4 Evaluation of musical examples

In this section, I will evaluate excerpts from four electroacoustic works and one acousmatic passage from an audiovisual work, in order to demonstrate the five different categories along the continuum from maximal to minimal in the evaluation of the feature salience premise.

10.4.1 Maximal: Trevor Wishart, *Red Bird*

The excerpt from Trevor Wishart's *Red Bird* (1977, on Wishart, 1992, 19:10-19:20, **sound example 10.6**) is taken from a section of the piece that follows a long build-up of a virtual sound landscape resembling that of a forest or garden. At the end of this section, the sound of a shattering glass breaks off the environmental sound, and the listener is left with long passages where silent stretches are only rarely broken off by short vocal utterances. The excerpt features such a silent stretch containing one of the occasional short vocal utterances. The factor that is most important in locating this short vocal utterance at the maximal end of the scale is:

- **Temporal discontinuities:** When seen in relief against the long stretches of near silence that precede and follow this vocal phrase, the temporal discontinuity constituted by the single vocal utterance is marked. When presented in an almost silent setting, the discontinuity/contrast in *loudness* becomes particularly striking. With the gradual attunement to the faint noise of the recording and the sound system during the "silent" phase, the listener might even be startled by the sudden occurrence of the relatively loud vocal sound. The abrupt onset of the utterance also contributes here. The sound ends with a fading reverberation, and it might have been experienced as even more salient if the ending had been just as marked as the attack. Even though this reverberation could potentially have created a temporal masking effect, this is not the case here, since the initial vocal sound is very short.

10.4.2 Maximal-intermediate: Paul Lansky, *Things she carried*

In this excerpt from the beginning of the first movement of Paul Lansky's *Things She Carried* (1996, on Lansky, 1997, 0:00-1:01, **sound example 10.7**) carrying the same title, we can hear the speaking voice of Hannah MacKay, Lansky's wife. Her voice is accompanied by two layers of sound; one layer containing short percussive sounds and one containing pitched sustained synthetic sounds which change relatively slowly and vary in loudness throughout the excerpt. The factor that affects the experienced salience here is:

- **Simultaneous masking:** The layer of sustained pitched sound creates a gentle *masking* effect for some of the softest vocal sounds. For instance, the ending of the word “carried” ([-ɹid]) at about 0:10, is partly masked by the relatively loud accompaniment at this point. Moreover, the ultimate [m] of “comb” at 0:23 in the example, can only be heard with difficulty, and it is not possible to identify the phoneme between the two final [s]’s in “pencils” (approx. 0:45).

I have to note that masking here varies with the loudness of the vocal phrases and of the accompaniment, so that the salience vacillates between *maximal* and *maximal-intermediate*. As for the slight reverberation added to the voice in this excerpt, it does not cause any tail masking and reduced salience.

10.4.3 Intermediate: Trevor Wishart, *Two Women - Stentor*

The third example in this section from Wishart’s *Stentor* (1998, on Wishart, 2000b, 0:38-0:42, **sound example 9.4**) has already been discussed in the previous chapter (section 9.3.3), but I found it highly pertinent in this context as well. In the short excerpt that is presented first in isolation, and then in its immediate context, the contrasts between the background sound and the vocal sound are less than what was the case in *Things she carried*. Here, one can hear what appears as a noise-vocoded speaking voice, accompanied by continuous noise that is difficult to give any definite source attribution.[279] The factors that contribute to reduced salience in this excerpt are:

- **Temporal discontinuities:** In the short pauses between the speech sounds in this excerpt, one hears high-level noise that is easily misattributed as phonemes, and this makes the contrast in spectral type between “pauses” and speech sounds minimal. Moreover, the spectral brightness of the noise in between speech sounds lies within the range of the variation that is found in the speech sounds themselves, so that the differences might not stand out as something other than what is found within the speech stream. This contributes to a blurring of the onsets and endings of the phonemes – it is not always easy to detect where the speech sounds end or begin and where the pauses in between end and begin.

[279] In addition, one can also identify a layer of warbling noise in the mid-high frequency area between 0.8 and 1.6kHz in the example. This layer, though, is not easily identified because the vocal layer tends to draw all attention to itself.

- **Simultaneous contrasts:** The difference in *loudness* between speech sounds and background noise is relatively small here.

- **Condition of vocal features:** For this phrase, there are is no pitched phonation, although the phrase does not bear the marks of being whispered. Rather, I experience that the intonation contour is something that is absent from this phrase. Hence, I experience it as degraded. Moreover, I also experience that many of the speech sounds are degraded.

When all these factors are seen together, there are not many factors that contribute to the segregation of the spoken phrases from the other present noise. As I hear it, the modulation in the spectral profile, characteristic of speech, is the only feature that makes the vocal phrases stand out, so that they can be perceived as some form of vocal utterance, albeit with some difficulty. All in all, the excerpt is therefore evaluated as *intermediate*.

10.4.4 Intermediate-minimal: Francis Dhomont, *Á l'orée du conte*

In the fourth example, which is from Francis Dhomont's *Á l'orée du conte* (1994-96, on Dhomont, 1996, 0:27-0:40, **sound example 10.8**), one hears at the beginning a complex, sustained chord-like sound consisting of both pitched and noise-based components. The sound appears to contain components mainly in the low and mid-to-high spectral regions, where the lowest component appears as a drone, and the higher components can be identified as a relatively bright, dense, and harmonically unidentifiable chord. About half way out in the example a low pitched sound slowly emerges, but it is not until towards the last quarter of the example that it is possible to identify this sound as a male voice uttering "una".[280] As the sound continues, a sustained male voice with a granular quality grows out of the complex sound until it ends up in the foreground. Hence, the experienced salience in this case begins in the minimal end of the scale and ends with something close to intermediate feature salience, maybe even higher. Consequently, it is the *middle part* of the vocal sound, in the phrase between 9 and 12.5 seconds, that I refer to when classifying the excerpt to be in the

[280] If listening to the continuation of this passage in Dhomont's piece, one will later recognize this sound will re-appear as the Spanish version of "Once upon a time", namely "una vez".

intermediate- minimal category. The factors that contribute to the low salience for this excerpt are:

- **Simultaneous masking:** This emerging low pitched vocal sound lies in the frequency area of the drone component of the complex chord, something which renders the masking relatively effective in this case.

- **Simultaneous contrasts:** When the vocal sound gradually becomes louder at the same time as it enters into slow modulations in pitch, brightness, and loudness level (at about 10 seconds), it begins to disengage from the masking of the drone. Here, the features of the voice are quite similar to those of the chord, but the modulations make the voice stand out partially. At that point, the granular quality of the vocal sound can be relatively easily recognized, and this possibly contributes to the increased salience of the last second of the example.

- **Temporal discontinuities:** The onset of the vocal sound is very gradual, hence making temporal discontinuities minimal here.

10.4.5 Minimal: John Coulter, *Shifting Ground*

The last example that I want to discuss is from John Coulter's *Shifting Ground* (Coulter, 2005, 1:58-2:12[281], **sound example 10.9**). This sound example demonstrates quite pertinently, in my view, how sounds that are very similar to each other mask each other most effectively. In this case, an unidentifiable number of voices are heard together so as to create a dense texture of sound. In this texture, individual voices manage to stand out from the chattering only for very brief moments at a time in the beginning of the excerpt, where the density of voices is somewhat lower. Since the number of voices is high, the voices all lie within a not too wide range of variations in pitch, loudness and spectrum, and since all have similar kinds of modulations in these parameters, it becomes impossible for single voices to stand out for a listener for more than short moments. If the voices had not been located in different locations in the stereo image, these moments would probably have been even fewer and the effect of the masking even higher. In this case therefore, it is the combination of many of the factors that

[281] Here, I am referring to the acousmatic passage of this audio-visual work. The corresponding time counter on the DVD release shows approximately 2:59-3:03, indicating an error in the time values.

cause a very high degree of masking, which in turn contributes to minimal salience for this excerpt:

- **Simultaneous masking:** Vocal sounds mask each other – partly in the beginning and then increasingly more during the excerpt.
- **Temporal discontinuities:** The attacks and endings of the vocal phrases are difficult to detect due to masking.
- **Simultaneous contrasts:** Similarities in pitch range, modulation and the combination of pitched and noise based spectral types, make the vocal phrases barely stand out from each other in the beginning, and then gradually less towards the end of the example.

10.5 Chapter conclusions

We have seen in this chapter that the premise of feature salience deals with the degree to which vocal sounds and their features stand out perceptually. One aspect of this was the relationship between vocal sounds and accompanying sounds, if any were present. Here, the amount in which the accompanying sounds were *masking* the vocal sounds was a critical issue. With reference to psychoacoustic research and by presenting examples, I showed how the frequency content and loudness of the masking and the masked sound were of importance for the degree of masking. Moreover, I argued that context could lessen the effective masking to a certain extent. The relationship between vocal and accompanying sounds was also generally affected by the articulation of temporal discontinuities by the vocal sound, as well as the degree of dissimilarity/contrast between simultaneous sounds, in that discontinuities and contrasts tended to increase the experienced salience, at least if no masking was involved. The second aspect of the feature salience premise was that the vocal features in themselves could have higher or lower salience, depending on issues such as processing, reverberation and articulation. After having presented theoretical perspectives of salience and a set of factors that potentially could affect the evaluation of the premise, I presented five musical excerpts that exemplified the five categories of evaluation between maximal and minimal.

11.0 Stream integration

> **Premise seven of the max-min model:**
> *Stream integration*: The sound of the voice is integrated into one coherent and continuous sound stream.

The term "sound stream" here refers to a perceptual representation that clusters together related qualities of an auditory event to form distinct mental entities, usually assigned to one and the same sound source. Whereas the maximal voice is integrated into one stream, we have opposite situations in which what we hear is disintegrated into multiple streams that we can no longer hold together as one coherent object. The continuum between these two extremes will be discussed and exemplified in this chapter. In this discussion, Bregman's *auditory scene analysis* theory, which seeks to explain the human ability to parse or group a complex auditory input into coherent sound streams, will be central (Bregman, 1990).[282]

That a sustained and unaccompanied sung note on a single pitch constitutes a "coherent and continuous sound stream" is rather obvious. For speech, with which the maximal voice is associated, it might not be just as obvious that we are dealing with a *continuous* sound stream, since there are many relatively abrupt transitions between speech sounds with quite different sonic qualities. However, Bregman argues convincingly, with a basis in speech research, that speech, despite its range of qualitatively different sounds, changes continuously in several important parameters (Bregman, 1990: 537-556):

1. **Pitch:** In speech, pitch tends to change continuously, thus forming intonation contours, and if interrupted by stops or silences, it usually picks up close to where it left off.
2. **Spectrum:** The transitions between many speech sounds are continuous when they are linked into words and sentences, because the speaker has to make continuous changes in her or his articulatory apparatus when pronouncing them.[283]
3. **Spatial location:** Speech is usually spatially continuous because of the fact that speakers tend to stay in the same place or move relatively slowly through space. For my definition of the maximal voice, the placement of the vocal persona at close distance will naturally imply spatial continuity.

[282] This process is also frequently referred to as *auditory grouping*.
[283] See section 3.6.2 on *coarticulation*.

In addition, I would like to mention that the spectral features of a particular voice, its *timbre*, will be an important cue in linking segments together into a continuous stream (Payri & Bono, 2007). Hence, there seems to be several theoretical points that affirm the intuitive notion that speech is a continuous sound stream.

The integration of sound into streams can potentially influence many other processes, some of which are included in other premises discussed in the course of this dissertation. We have seen that sequential integration, i.e. the integration of subsequent segments of sound into one stream, was a factor for the *clarity of meaning* premise (section 9.2). And, in the chapter on naturalness, I demonstrated that certain types of discontinuities between segments could affect the evaluation of the naturalness premise negatively (section 7.4). Moreover, we are dependent on auditory integration mechanisms to achieve the mentioned effect of "overlooking" the technology of mediation that we saw were crucial for the premise of presence, separating for example the hiss of a bad tape recording from the recorded music itself (cf. section 8.1.2). Lastly, I argued in the chapter on feature salience how that premise and the current one can be seen as complementary, focusing on segregation and integration, respectively.

11.1 Theoretical perspectives

11.1.1 Primitive and schema-based processes

One can make a distinction between grouping mechanisms that are innate and those that are learned or experience based, the former kind labelled *primitive* and the latter kind *schema-based* by Bregman. The *primitive* integration mechanisms, on one hand, are seen as having evolved through interaction with regularities in the environment by a kind of psychophysical complementarity. This means that our perceptual system has been "tuned" through different phases of the evolution over thousands of generations so as to be sensitive to those aspects of our environment that have been crucial for survival.[284] As for the learned, or *schema-based*, mechanisms of auditory grouping, there are a number of studies on voice and speech perception that clearly indicate that schema-based processes are involved in stream integration

[284] Bey and McAdams list several reasons why auditory stream formation is considered to involve pre-attentive and innate processes (Bey & McAdams, 2002: 844). Firstly, it has been observed that stream segregation has occurred against listeners' intentions. Secondly, in studies of brain responses to sound streams, researchers support claims that stream integration also takes place when not attending to sound (see also Sussman, 2005). Finally, the claim of innateness of stream integration is supported by findings of the presence of the skill early in life and in other species.

when we listen to voice and speech (see e.g. Darwin, 1981; Warren & Warren, 1970). The most striking example is perhaps in the perception of sine-wave speech, which is synthesized by having three sine waves follow the formant trajectories of a speech sequence. In a study by Remez and colleagues, listeners had great difficulties making out what they were listening to, until they were informed that the sound was computerized speech (Remez et al., 1981). Hence, it was not until their cognitive schemas of speech were activated that they could integrate the three sine-waves into a coherent speech pattern.

Although the primitive and the schema-based processes are separated in theory, both work together towards the same goal – that of grouping a complex auditory scene into streams that correspond to separate sound sources in the environment. Still, the two processes appear to work in somewhat different ways, according to Bregman (Bregman, 1990: 405-11). Firstly, as opposed to the pre-attentive primitive processes, *effort* or *volition* can be a factor for schema-based grouping. In cases where two different tones alternate with a particular speed, one might choose to hear this as either one or two streams. Moreover, whereas the primitive processes work to *partition* the input into separate streams, the schema-governed processes work to *select* properties that meet certain criteria. Thirdly, the temporal scope of the two processes appears to be different. The schema-based processes appear to be able to span a much greater temporal range than the primitive processes, which depend more on effects of acoustic variables that are more local.[285] In addition to these differences, one might also point to the relationship between the primitive grouping processes and some of the Gestalt principles, something that I would like discuss further in the following section.

11.1.2 Similarity, continuity and proximity

In the preceding chapter, I argued that marked differences or *contrasts* played an important role when it came to the perceptual salience of a sound, both in the form of temporal discontinuities as well as in contrasts between simultaneous layers. In other words, contrasts were considered *horizontally* and *vertically*, respectively. For this premise, I will focus on *similarity*, *proximity* and *continuity* rather than contrasts, but the two-dimensional approach is in many ways the same. These three notions bear clear resemblance to some of the so-called

[285] This is, for instance, evident from research on the *phonetic restoration* effect. Warren and Warren reported that a key word that tended to affect the interpretation of a masked phoneme in a sentence could be separated from the masked phoneme by three words (Warren & Warren, 1970). The exception to this might be the tendency that streaming effects in some cases need to "build up" temporally until stream segregation occurs. According to Bregman, such a build-up might last 4 seconds or longer. (Bregman, 1990: 410).

Gestalt principles, formulated on the basis of studies of vision in the late 19th and early 20th centuries.[286] According to Bregman, these notions also play a crucial role in auditory grouping: "In general, all the Gestalt principles of grouping can be interpreted as rules for [auditory] scene analysis" (Bregman, 1990: 24).[287] Consequently, *similarity*, *proximity* and *continuity* in the temporal and/or spectral dimensions will be seen as underlying principles that will act as integrating forces for sound streams. I will specify in what ways these three very general notions have an effect on stream integration in section 11.2, dealing with the factors potentially affecting stream integration.

In many ways, these three notions have meanings that are related. A *continuous* sound, for example, will be characterized by not having any abrupt changes, thus implying that at each moment of the sound, the *similarity* with the preceding moment is relatively high. And, if a sound is broken off by silences, the more *proximate* in time one segment is to the following one, the more *continuous* it will be felt. Moreover, if one considers features as constituting axes in the same manner as I have regarded the premises of this framework (cf. section 4.6), *similarity* can be interpreted as *proximity* – the more similar the features are, the closer they will be located on the axis. For those properties that can be represented by a single dimension, however, like pitch, spatial distance and azimuth, we might prefer to talk about degrees of *proximity* instead of degrees of similarity, at least for those cases when two properties are not perceived as exactly the same.[288] For instance, it makes more sense to speak of the notes C3 and D3 as being *proximate* in the pitch space rather than similar to each other.[289]

11.1.3 Level of source coherence

One aspect that becomes problematic when one deals with auditory grouping mechanisms in musical expressions is that one often has several levels on which source attributions can be

[286] See the section on Gestalt principles in the article on "Primary tendencies in perceptual organization" in Encyclopaedia Britannica Online, URL: http://search.eb.com/eb/article-46708, accessed 11/09/2008.

[287] In my view, a principle like "common fate", which Bregman relates to his discussion of amplitude and frequency modulation, can simply be expressed as a form of *similarity* in the way that a property changes over time.

[288] This is in line with Bregman, who states that "perhaps we might reserve the use of the word similarity for cases in which we were unable to physically describe the dimension of similarity, but when we could describe the physical basis of the dimension, we might say that the sounds were near [or proximate] on that dimension" (Bregman, 1990: 198).

[289] There are many aspects of pitch that cannot be represented in just one dimension, for example when it comes to the harmonic relationships between pitches, octave equivalences etc. These aspects, however, are less important for the type of auditory grouping discussed here. Issues of harmonicity will be discussed further below.

made. In music, as in many other situations in daily life, one can often describe a sound source on several levels, depending on whether one focuses on the smaller details of something, or on larger constellations of elements taken as wholes. Take, for example, a situation where you sit in your living room and listen to a recording of a pop/rock band. On one level, you would have to segregate the sound emanating from the loudspeakers from any other disturbing sounds nearby, for example the television set in the room next door. If there are several instruments playing relatively similar material, for instance, if a horn section plays block harmonies, this can be perceived as a coherent unit on a lower level. On an even lower level of organization, you can maybe separate some of the other instruments like the guitar and the bass from each other. And, for some instruments, like the drum set, you can probably fairly easily distinguish different drums and cymbals from each other. In other words, we can in many situations relate what we hear to different levels on which sound sources can be organized. Bregman refers to this as *hierarchies in auditory organization*, but I would like to call it levels of sound *source coherence* (Bregman, 1990: 203-206). As he notes, our "scene analysis process is not parsing the parts in a piece of music as wholly distinct auditory objects. Instead it is creating a hierarchical description, marking out the parts as distinct at one level of the hierarchy and the total composition as a unit at a higher level" (*ibid.*: 461). One will therefore often have situations where sounds are integrated on one level and segregated on another, and that different features can be attributed to one or the other level.

In this context, I will refer to the *local* source coherence level as the level of the vocal utterance attributed to a single vocal persona. Conversely, I will refer to a *global* source coherence level when several streams or objects are attributed to a larger constellation of streams that share a certain number of features. This is the case, for example, if several different voices articulate something in perfect synchrony, so that they more or less "melt" together. By sharing articulatory features, they will constitute a coherent source on the *global* level, which we then might refer to as a group. And, as we will see in section 11.3, this will imply that the evaluation of this premise will move in the direction of the minimal.

This situation is illustrated in **sound example 11.1**, which is a sequence of seven transposed and superimposed versions of the same sequence of sustained vowels articulated by the same speaker.[290] Consequently, the synchronous articulation refers to the *global* source coherence level, while the different pitches of the single voice streams refer to the *local* level. In this case, since the global level is attributable to one and the same vocal persona and since

[290] The voice belongs to actor and professor at Concordia University in Montréal, Nancy Helms.

there is only one feature that distinguishes the streams at the local level from each other, source coherence is not too far removed from the maximal. If the stream at the global level has many properties that aren't shared with the local level, however, I will regard it as being located further in the direction of the minimal.

11.2 Factors potentially affecting stream integration

There are a number of factors that can be linked to stream integration. Those I will consider are *feature similarity/proximity/continuity*, *harmonicity*, *rate of change for features*, *temporal proximity between events*, *similarity in spatial location*, *synchronicity of onsets and endings*, *looping*, *modulation coherence*, *familiarity with patterns/linguistic structures*, and *volition/effort*.

- **Feature similarity/proximity/continuity:** In general, a high degree of similarity, proximity and/or continuity in features like pitch, sound spectrum (e.g. spectral type, spectral brightness, nodality), vocal resonances (spectral envelope), loudness and spatial location between subsequent events will make sequential integration more likely, whereas similarity/proximity between simultaneous layers will make simultaneous integration more likely (Darwin, 1997; Cusack & Roberts, 2000; Bregman, 1990; Bregman et al., 2001; Deutsch, 1999).[291]

- **Harmonicity:** Harmonic relationships among components are considered to be one of the most important cues for simultaneous integration (Darwin, 1997; Bregman, 1990: 232-234; Deutsch, 1999: 302-304). The auditory system appears to exploit the regularity that when something vibrates with a steady period, it will usually form complex sounds where the partials are harmonically related (Bregman, 1993: 27). This means that components that do not have a harmonic relationship to a fundamental tend to be segregated into separate streams. To demonstrate this, I have prepared **sound example 11.2**[292], in which a vowel sung by a female voice is first played with all

[291] This is also related to the so-called "old-plus-new" heuristic, which can be noticed in the tendency to interpret the entrance of a set of "new" components along a set of old continuing ones, as the onset of a new event, rather than a change in the old one. Hence, even if the "new" situation as a total is different from the "old" one, the continuation of the "old" components will be taken as indication that the first event persists along with a new one (see Bregman, 1990: 14).

[292] This example, as well as all consequent examples presenting two successive vowels, is made by extracting the 12 lowest partials of a sustained vowel, and then synthesizing these partials with additive synthesis.

components harmonically related, and subsequently with one partial detuned. The result is that the detuned partial which is no longer harmonically related to the other partials is not integrated with the rest of the components, but forms a separate stream by itself. Another consequence of this factor is that sounds with an inharmonic spectrum type will be more difficult to integrate into a single sound stream, often giving rise to an interpretation of multiple sources.[293]

- **Rate of change for features:** Features that change gradually rather than abruptly tend to signal that a current event is a continuation of the previous situation (Summerfield et al., 1992).[294] This is related to what Bregman refers to as the "sudden change rule", which states that the auditory system will treat a sudden change of properties, in other words changes with a high speed or rate, as the onset of a new event (Bregman, 1993: 19). This is especially the case with sudden changes in loudness.

- **Temporal proximity between events:** The temporal *proximity* of the interval between events can play a part for sequential grouping, in that the more proximate two events are, the higher the chances are that they are integrated into one stream (Bregman et al., 2000; Deutsch, 1999: 319-320).[295]

- **Similarity in spatial location** is a cue for the integration of simultaneous components, even if it is not considered a dominant one (Bregman, 1990: 294-300; Grossberg et al., 2004). This can be heard in **sound example 11.3**, in which a vowel sung by a female voice is first played with all components in both channels, then with one partial in one channel and the rest in the other channel. While it might not be easy to single out the single partial while listening to loudspeakers, it can easily be done if one listens to this example with headphones.

[293] We have already heard in **sound examples 7.5** (cf. section 7.2) from Wishart's *Vox V* how the vocal sounds in that excerpt had an inhamonic spectral type which almost gave the sensation of more than asingle voice.

[294] Bregman notes, however, that the difference between "sudden" and "gradual" cannot be clearly defined, but might be regarded as a continuum from the most abrupt and discontinuous changes to almost unnoticeable transitions. See also section 2.4.2.2 on the qualifications of dynamic profile.

[295] Bregman (1990: 65-67) reported that it was less clear if this was primarily dependent on the inter-stimulus interval (ISI) or onset-to-onset interval, and if it was the intervals *within* each single stream or *across* different streams that were important. A later study that he conducted with a group of colleagues has, however, indicated that it is the inter-stimulus interval *within* streams that is the important one in determining sequential integration (Bregman et al., 2000).

- **Synchronicity of onsets and endings:** Synchronous onsets and endings, the former more than the latter, tend to strengthen simultaneous integration whereas asynchronicity tends to cause components to be perceived individually (Deutsch, 1999; Grossberg et al., 2004; Cooke & Ellis, 2001; Bregman, 1993:17). To illustrate this, one can listen to **sound example 11.4**, which presents two versions of a sustained sung vowel. In this example, the partials of the first vowel all start at different times within a period of 0.65 seconds, then continue together for 0.6 seconds, and then stop at different times within a period of 0.65 seconds. Here, the lack of synchronous onsets and endings of the partials makes them more difficult to integrate into a coherent sound than when all the partials enter together, as is the case with the second sound in the example.[296]

- **Looping:** Depending on factors like *feature similarity/proximity* and *temporal proximity between events*, as mentioned above, different portions of a looped vocal sound can constitute separate streams, thus constituting a factor that has a negative effect on integration (Cooke & Ellis, 2001; Scott & Cole, 1972).[297] Hence, if there are any portions of the loop that is markedly different from the others, and if the temporal distance between the iterations is sufficiently short, i.e. the speed is sufficiently high, these portions can form a separate stream. Often, fricatives with a large portion of high frequency energy can segregate into separate streams under the right conditions. One can hear this happening in **sound example 11.5**, which consists of a single word, “stå” ([stɔː]), which is repeated in a continuous loop with progressively decreasing loop time, as well as a correspondingly increasing temporal compression of the looped segment.[298] At the beginning of the loop, one can hear the word as one continuous articulation, but as the loop speeds up and then stabilizes at a higher speed, one starts instead to hear the repeated syllable [dɔː] as a separate stream, along with a second stream consisting of the [s]’s, but which sounds like a sequence of high frequency “hits” with little resemblance to vocal sound. One other thing that can be observed when listening to the example is that it appears to take some time for stream

[296] This effect would have been even more pronounced if there had been no synchronous microfluctuations in frequency of the partials, something which contributes to a strengthening of integration.
[297] This is often referred to as the *streaming effect*, see e.g. Bregman, 1990: 47-67.
[298] The compression factor varies from 1 to 1.65, whereas the duration of the looped segment varies from 0.43 to 0.26 seconds.

segregation to occur. This is often observed in the research literature on the streaming effect, where it is referred to as *cumulation* or *buildup effect* (Bregman, 1990: 128-133). Thus, for a loop to be effective in provoking stream segregation, it has to last for some time, so that the streaming effect is allowed to build up.

- **Modulation coherence:** Modulations in pitch, loudness and spectrum might, under some circumstances, provide cues for making the modulated components integrate into a stream, or strengthen the integration of the modulated components. Chowning and McNabb, for instance, have reported that when synthesizing vocal timbres from sine waves, a common frequency modulation of the sine tones enhanced integration (Chowning, 1999; Deutsch, 1999). And, it has been shown that, when accompanied by unmodulated sounds, frequency modulated components tend to group together better than when there is no modulation (Marin & McAdams, 1991; Summerfield et al., 1992). This can be observed, for example, in complex sounds where single components that are frequency or amplitude modulated tend to be segregated from the remaining components of the sound, as can be heard in **sound example 11.6**. Here, all partials are frequency modulated with the same rate and amplitude in the first vowel, and in the second vowel, where one partial is given a different modulation rate than the other, the difference is enough to prevent it from integrating into the stream. Coherence in amplitude modulation also appears to strengthen integration under certain conditions (Hall & Grose, 1990). In instrumental music, coherent changes in pitch and/or loudness can cause different instruments to fuse (Bregman, 1990: 489,495,499, 518-520), and it is likely that it will play a part in electroacoustic music as well. Parallel movement in pitch of different parts in octaves or fifths, for instance, will easily be integrated so as to be heard as one single sound (Deutsch, 1999: 336-337; Karydis et al., 2007). When pitch, loudness and spectral cues in a voice vary in parallel, this can act as a strong force for integration, something which is clearly demonstrated by **sound example 11.7**.[299]

- **Familiarity with patterns/linguistic structures/sound source:** Familiarity with sound patterns (e.g. melodies) and linguistic structures can affect stream integration positively in many conditions (Bregman, 1990: 407; Cooke & Ellis, 2001; Slaney,

[299] This example is made by superimposing seven transposed versions of the same speaking voice on top of each other.

2004).[300] Sine-wave speech (see section 11.1.1) demonstrates this pertinently. An example of sine-wave speech can be heard in **sound example 11.8**, followed by the vocal phrase from which the sine-wave version was generated.[301] Usually, it is difficult for listeners who don't know what type of sound this is, nor the linguistic content, to integrate the sine-tones into a coherent percept. And often, one has to listen to the original recording several times to actually hear the sine waves as somehow resembling speech. In addition, I will assume that knowing the range of sounds that can be produced by a source such as the voice can affect integration, although I have not found any empirical research supporting this claim. I will argue, however, that this is the case for many of the musical examples that will be discussed in section 11.4.

- **Volition/effort:** Volition and effort can play a part in some cases where the primitive mechanisms offer more ambiguous cues. On the basis of reviewing Van Noorden's research on the streaming effect, Bregman concludes that the "rates and frequency separations at which one gets segregation depend on the intention of the listener" (Bregman, 1990: 406). The boundaries that Van Noorden could draw on the basis of his experiments, beyond which it was not possible to intentionally choose whether to integrate the sequence into one stream or segregate it into two streams, therefore had a zone in between them where a conscious effort could make it possible to focus either on holding the sequence together, or to separate it into two streams.

In table **11.1**, I have listed all the discussed factors and what for each factor will strengthen and weaken sound stream integration, respectively. For this premise, as for the preceding ones, the ways in which these factors "work together" in combination are in most cases, except those that are highly controlled laboratory created scenes, difficult to determine exactly. Bregman sees the process as a kind of "election" where the mechanisms that contribute in solving scene analysis problems act by "voting":

[300] For example, Bey and McAdams report that listeners that were given the chance to listen to a melody before it was played together with distractor tones that interfered with the pattern of the melody, more listeners were still able to recognize it than if the order of presentation was vice-versa (Bey & McAdams, 2002)

[301] The example was synthesized using a Praat script written by Chris Darwin, downloaded from http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/, accessed 07/10/2008.

	Factor	**Strengthens integration** ◄	► **Weakens integration**
Primitive grouping	Feature similarity/proximity	Similarity/proximity/continuity	*Dissimilarity/distance/continuity*
	Harmonicity	Harmonic relationship	*Inharmonic relationship*
	Rate of change for features	Slow	*Fast*
	Synchronicity of onsets and endings	Synchronous	*Asynchronous*
	Temporal proximity between events	Short	*Long*
	Looping-feature similarity/ proximity	Similarity/proximity	*Dissimilarity/distance*
	Looping - speed of repetition	Slow*	*Fast**
	Modulation coherence	Coherent	*Incoherent*
Schema-based grouping	Familiarity with pattern/linguistic structure/ sound source	Familiar	*Unfamiliar*
	Volition/effort	*Volition/effort to integrate*	*Volition/effort to segregate*

Table 11.1: Factors potentially affecting integration. Factors that are taken to be based on primitive grouping processes are placed in the upper part of the table, and those taken to be based on schema-based processes are placed in the lower part, as indicated by the first column. The properties involved with each factor are listed in the second column. The configuration pairs in the rightmost column indicate what configuration will strengthen or weaken stream integration, respectively. The double pointed arrow in the top row indicates that there is a continuum between each of the configurations in each pair.
*** High speed of repetition tend to cause stream segregation, i.e. integration of similar elements that are not contiguous to each other when being played more slowly.**

> These criteria are allowed to combine their effects in a process very much likevoting. No one factor will necessarily vote correctly, but if there are many of them, competing with or reinforcing one another, the right description of the input should generally emerge. If they all vote in the same way, the resulting percept is stable and unambiguous. When they are faced with artificial signals, set up in the laboratory, in which one heuristic is made to vote for integration and another for segregation, the resulting experiences can be unstable and ambiguous. (Bregman, 1990: 33)

The metaphor of the election is enlightening in that it depicts a process where many forces work together in making a decision one way or the other, but might also be misleading since in most elections each vote has a similar value as another one – and this is not necessary the case in auditory scene analysis, where Bregman himself talks about the strength of different cues. For example, in relation to a reviewed experiment on timbre in streaming, where there was competition between different grouping tendencies, he suggests that "frequency proximity was the strongest cue, followed by timbre […], followed by spatial origin, with trajectory the weakest" (*ibid.*: 170). Moreover, in cases where schema-based processes are likely to be involved, the differences in strength might be relative to the degree of exposition that a listener has had to a particular type of sound, something which is both highly variable and also subject to change.

11.3 The evaluation of the premise

For the evaluation of this premise, I will use three criteria:

1. **The level of source coherence.**
    a. When a sound stream is attributed to the local level of source coherence, i.e. attributed to one single coherent vocal utterance, it indicates an evaluation towards the maximal.
    b. When a sound stream is attributed to a global level of source coherence, i.e. it consists of several sub-streams, either as superimposed streams, as discontinuous juxtaposed objects, or a combination of these, it indicates an evaluation further towards the minimal.
    c. When a sound stream is ambiguous between being attributed to a local level and a global level of source coherence, it indicates a situation towards the minimal.

   d. The higher number of features shared on each level that can be attributed to a vocal utterance and a vocal persona, the closer to a maximal evaluation it will be located.

2. **Sequential integration**
   a. The stronger continuity in a phrase or section, the closer to the maximal it will be located.
   b. The more discontinuities in a vocal phrase or section, the closer it will be placed towards the minimal.
   c. The shorter the within-stream durations between discontinuities, the closer it will be located towards the minimal.
   d. Ambiguities as to whether something is a continuation of something else or not indicate an intermediate situation between maximal continuity and maximal discontinuity.

3. **Strength of simultaneous integration**
   a. When the components and elements in a vocal utterance are strongly and unambiguously integrated into one stream, it indicates an evaluation towards the maximal.
   b. When the components of a vocal utterance are segregated into several streams, it indicates an evaluation towards the minimal.

Based on these criteria, continuous speech or continuous sung phrases are both good examples of maximal evaluations for all criteria. At the minimal end, a hypothetical example regarding the second criterion of sequential stream integration would be a situation with a loosely organized sequence of vocal events that are so short that they would not be experienced as streams in themselves, and where there are minimal continuity between the events for properties like pitch, vocal timbre, type of vocalization, voice quality, spatial localization, etc. In such a situation, the only thing that would link the events together is the shared property of being short vocal events. For the third criterion, a good example of weak integration is sine-wave speech (cf. **sound example 11.8**).

After having presented the criteria for the evaluation, it is time to move on to the assessment of musical examples in the following section of this chapter.

11.4 Evaluation of stream integration in musical examples

In this section, I will present evaluations of excerpts from five electroacoustic pieces according to the current premise, where I will also present links to the factors that were introduced above.

11.4.1 Maximal: Philippe Manoury, *En écho*

In this excerpt from Manoury's *En Écho* (1994, on Manoury, 1998, section I, event no.75, **sound example 11.9**) we hear a phrase sung by a female soprano voice accompanied by synthetic sounds.[302] This phrase is easily integrated in one stream and presents no ambiguities in relation to other streams. One can point to some factors that contribute to the strong integration in this case:

- **Temporal proximity:** The seven notes are sung with legato phrasing, hence having minimal temporal distance between the notes.
- **Feature similarity/proximity/continuity:** Although there are two stop consonants ([t]) at note boundaries that create discontinuities in the streams, the pauses in between are so short that these discontinuities are barely noticeable. The similarities in spectral features (timbre) throughout the phrase, along with pitch intervals that aren't too big, (the largest is a major 6th) are also important here.
- **Similarity in spatial location:** There is no change in spatial location for the voice in the excerpt.
- **Familiarity with sound source:** My familiarity with *bel canto* singing might also be of some assist in integrating this phrase, although I believe that it can be integrated without this familiarity.

Taken together, the assignment of the vocal stream to the local source coherence level, the minimal discontinuities in the stream, and the lack of interfering or competing streams (the electronic sounds in the background are clearly segregated from the voice), indicates *maximal* stream integration.

[302] Strictly speaking, this is not an acousmatic work, but a work for soprano and live electronics.

11.4.2 Maximal-intermediate: Lars-Gunnar Bodin, *Wonder-Void*

In one section of Lars-Gunnar Bodin's text-sound suite, *Wonder-Void*, from 1990 (Bodin, 2006, 9:46-10:01, **sound example 11.10**), one can hear a highly processed voice which utters a set of vocal phrases with intelligible verbal content, albeit with relatively elusive meaning. In this section, the processing of the voice gives the phrases a sort of layered quality, as in a *stratified* sound (cf. section 2.4.1.2) but where it is difficult to pin-point the exact pitch and character of each layer because they appear to behave in parallel. However, the impression I get is not one of multiple voices that engage in synchronous articulation, but of a single voice that is divided into partly distinct streams, resulting in a less integrated voice than in ordinary speech. The factors that contribute to some degree of *weakening* of integration in this case are thereby:

- **Feature similarity/proximity/continuity:** Dissimilarity in pitch between streams, where I can at several instances identify one relatively high pitch and one relatively low. It might also be that a third stream can be identified at times.
- **Similarity in spatial location:** When listening to this excerpt through headphones, one can discern that there is some spatial separation between the high pitched and the low pitched stream.

This division into several streams creates only a slight ambiguity regarding the number of vocal sources present, in that it might project both a male and a female identity simultaneously. Still, my main overall impression is that this is one single stream. When it comes to the factors that contribute to *strengthening* integration, the most important are:

- **Synchrony of onsets and endings:** The different streams behave in perfect synchrony, with all onsets and endings happening simultaneously.
- **Modulation coherence:** All pitch and spectral modulations appear to be synchronized, i.e. they are coherent.

Thus, the overall impression is that the vocal phrases are in some respects segregated into several streams, but that the streams are integrated into one global stream in a relatively strong manner. This also creates some ambiguity regarding the source coherence level. In sum, I will therefore evaluate it as being *maximal-intermediate*.

11.4.3 Intermediate: Trevor Wishart, Fabulous Paris – The Division of Labour

This sound example from Wishart's *The Division of Labour* from 2007 (Wishart, 2007, 7:28-7:36, **sound example 11.11**) exemplifies how a phrase can be cut into fragments so that it loses much of the continuity that ties it together into a coherent stream. In this example, each fragment is very short, often containing only a small portion of a single phoneme. Moreover, the pauses between each fragment appear to vary, but at some instances are relatively long compared to the fragments, so that the impression is that of a severely disrupted stream. We can link this to some of the discussed factors:

- **Temporal proximity between events:** The fragments are temporally close in the beginning of the example, but shortly the distances between each fragment get longer, so that they are separated by a noticeable temporal gap.
- **Feature similarity/proximity/continuity:** The fragments contain different types of phonetic material, pitched as well as unpitched (noise based), and these are organized so that the similarity from one fragment to the next varies from being very different in the first half of the example, to relatively similar in the latter half.

Combined, these factors make the experienced continuity in the phrase considerably reduced. Some factors also contribute to integration, however:

- **Similarity in spatial location:** The fragments all appear to come from a single location.
- **Familiarity with patterns/linguistic structure/sound source:** If knowing the rest of the piece, one will recognize elements from the verbal phrases that are presented in the beginning of the piece and the identity with which they are associated.

We see that this example has factors that both strengthen and weaken integration, and this corresponds to the experience that it is highly fragmented, while the familiarity with the speaker and the vocal phrases presented in the beginning contributes to holding it together to some degree. Consequently, I have evaluated it as being *intermediate* in terms of stream integration.

11.4.4 Intermediate-minimal: Hans Tutschku, *Sieben Stufen*

For this category, I want to consider a very short excerpt from Hans Tutschku's *Sieben Stufen* (Tutschku, 1999, **sound example 11.12**). In this excerpt, we can hear a layered vocal phrase with an iterative energy articulation, bearing the marks of granular processing. One might also recognize the German word "Verfall" (decay).[303]

As for the degree of integration in this excerpt, it is marked by the layered quality and the granularity of the phrase. As I experience it, these factors contribute to weakening integration:

- **Feature continuity:** The grains have a slow velocity, they are quite coarse and relatively sparsely distributed, so that I experience the sound as containing discontinuities, almost as it is perforated.[304]
- **Familiarity with linguistic structure:** The vocal phrase appears to have several simultaneous streams, i.e. it is a *stratified* sound, especially the first syllable, "Ver-". During the last phoneme of this syllable, the uvular [ʀ], one can also hear a very bright stream that is difficult to attribute to a vocal production altogether. The different streams are not too easy to single out in the mix, however.
- **Level of source coherence:** Global, since multiple source identities (**ID-domain**) can be identified throughout the excerpt. While the dominating identity features for the whole phrase are that of an adult male speaker, in parts of the phrase, particularly in the beginning, it sounds like a female or pre-adolescent male identity is superimposed on the male one. Moreover, the male identity does not sound consistent to me, in that I hear different identities in different parts of the excerpt.[305]

There are also several factors that appear to be attributable to the whole phrase, i.e. on the global source coherence level, that contribute in holding it together and strengthening integration:

[303] Whether this would be clear to the listener without having this information is not evident, though.

[304] Bregman discusses granularity as one possible candidate to be added to the list of factors that can contribute to auditory grouping/streaming. See Bregman, 1990, p.117-122.

[305] To be more specific, there is a change in spectral brightness and intensity during the [a] at 0:02 that implies for me a change in speaker identity.

- The granular quality.
- The linguistic content in the form of a comprehensible word: “Verfall” (**LI-domain**).

All in all, I experience that the phrase does not “hold together” as well as an unmanipulated voice does, in that it both contains numerous discontinuities and several vaguely defined streams. The general impression of this excerpt, therefore, is that of a high level of ambiguity, both for sequential and vertical integration. I have therefore chosen to judge this vocal phrase as *intermediate-minimal*.

11.4.5 Minimal: Trevor Wishart, *Two Women - Facets*

The last piece that I want to discuss here is Trevor Wishart’s *Facets* (Wishart, 2000b), one of four parts from *Two Women*, composed in 1998. This example is chosen because it shows how parts of a vocal sound through manipulation can form several streams with only weak integration among them.[306] For this piece, I have chosen to include two excerpts, one from the beginning of the movement (**sound example 11.13**, 0:00-0:06), and one from about halfway through the movement (**sound example 11.14**, 1:22-1:30). This is done to give a sense of the overall development in the piece; from a situation in which streams/events associated with vocal utterances are relatively well integrated as individual streams and relatively clearly separated from each other, to a situation in which streams and events are blurred relative to each other. Sometimes this results in new streams which have minimal association with the initial utterances with which they are linked.

In this movement, grouping processes play an important role in the development of the piece. From the situation in the beginning, where the different versions of the same vocal phrase are easily integrated into individually separable streams, one reaches a situation in which several sound streams, often hard to define in terms of numbers as well as starting and

[306] The material that is used in the piece is taken from a BBC interview with the late Princess Diana, where she among other things, comments on her relationship with the press: “There’s a… There was a relationship which worked before, but now I can’t tolerate it because it’s become abusive and it’s harassment. But I don’t want to be seen to be indulging in self-pity. I’m not”.[306] In the beginning of this piece one can hear a voice which is clearly manipulated, but might still be recognized as Princess Diana, for those who are familiar with her voice. For my part, however, I could only assign the voice in the piece to Princess Diana only *after* reading the notes supplied with the CD, something which can be attributed to a relatively modest familiarity with Diana’s voice along with the effects of the manipulation.

ending point, are juxtaposed and superimposed so that it is difficult to link streams together for more than short moments. The increased spreading out across the left-right continuum also contributes to difficulties in integration. What is perhaps the most striking feature of the second excerpt is that some of the consonants which have noise components with relatively high spectral brightness, namely [s], [f] , [ʃ], [dʒ], [tʃ] and [t], are at this point transformed into sounds with a chirping quality, almost resembling bird sounds. These chirps tend to form streams that are segregated from the remaining parts of the verbal phrases with which they have been associated at earlier points in the piece. Hence, this creates an ambiguous situation in which streams segregate from the source they have been integrated with earlier. This can be related to several of the factors discussed in section 11.2:

- **Similarity/proximity/continuity of features:** The durations of continuous streams are considerably shorter in the second excerpt (0.4-1.5 s.) than in the first (2.5-4.5 s.). It is also difficult to link the individual streams together into longer streams in the second excerpt, hence giving it a relatively fragmented character. The discontinuities between the individual streams can be linked to:
    - Spatial location appears to vary from stream to stream across the whole left-right continuum.
    - The high similarity and short temporal proximity between each chirping sound make the chirps more easily form separate streams.
    - The lack of continuity between consonants/chirps and the remaining part of the phrases prevents integration of these two into a single stream.
- **Familiarity with patterns/linguistic structures:** The resemblance between the consonant chirps and bird sounds may make it easier to accept these sounds as a separate stream. It is increasingly difficult to interpret these sounds as consonants due to the lack of similarity with naturally produced ones. Furthermore, the decreased similarity with consonants decreases the chances for making schema-based integration of the chirps into the streams which still bear some resemblance to verbal phrases.

All in all, the experience of multiple relatively short fragments that are difficult to integrate in coherent streams that can be attributed to the voice heard in the beginning, along with the formation of new streams with a difference source attribution, only having a weak link to the consonants from the beginning of the movement, makes this a good example of an evaluation of the stream integration premise as *minimal*.

11.5 Chapter conclusions

In this chapter, I have presented the premise of *stream integration*, which dealt with issues of auditory grouping, both in the time domain and in the spectral domain, or with the terms I have applied in this context, *sequential* and *simultaneous integration*. The premise delineated a continuum from *maximal* integration – where one experiences a vocal phrase as continuous in time and integrated into one single coherent stream which is attributed to one single vocal persona and one single utterance – to minimal integration, where the vocal material contains multiple streams/objects that can only with difficulty be linked to each other. Research on so-called auditory scene analysis problems provided a theoretical framework for the discussion, which allowed me to propose a set of factors that would potentially affect the evaluation. In accordance with this framework, the factors were divided into *primitive*, i.e. innate and automatic, and *schema-based*, that is, based on learning and experience. The criteria for making the evaluation of this premise were directly related to the degree of temporal discontinuity, the strength of simultaneous integration, and the degree to which streams could be attributed to a local or global *source coherence level*. Here, the local level designated situations where one can attribute a stream to one single vocal utterance and a single vocal persona, whereas the global level designates a situation in which one assigns a stream to several utterances, vocal personae or perhaps even other sources. Combined, these criteria allowed me to evaluate five excerpts from electroacoustic pieces that exemplified the five categories between a maximal and a minimal evaluation of this premise.

Part III

Lansky’s *Six Fantasies*

12.0 An evaluation of Six Fantasies by Paul Lansky

12.1 Introduction

In this chapter, I will apply the framework that I have developed during the course of this dissertation on one single piece of music, namely Paul Lansky's classic computer music piece from 1979, *Six Fantasies on a Poem by Thomas Campion*, referred to simply as *Six Fantasies* from now on (Lansky, 1994a). The reason why this piece was chosen is that it is almost entirely based on recorded vocal sound, and that it incorporates sounds that intuitively appear to span the whole range between maximal and minimal voice. That the six relatively short movements the piece consists of all have a relatively consistent individual character, was also an attractive feature. This allowed me to use an analysis of shorter excerpts from each of the movements as a basis for more general considerations of the piece as a whole.

I will start by giving an introduction to the piece and the poem upon which it was based. Then, I will discuss issues related to the evaluation process and how it should be interpreted by the reader, before giving a brief presentation of a software instrument, which is submitted with this thesis. This tool has been used to create several sound examples which resemble more or less the vocal phrases in the piece. The "hypothetical" sound examples synthesized with this instrument will constitute a relief against which the evaluations of the excerpts from the piece hopefully will seem clearer. The main part of the chapter will be devoted to the presentation of the evaluations themselves, both in graphical and written form, with reference to factors that have been important in the evaluation process. Lastly, I will compare the evaluations from all the six movements of *Six Fantasies*, and hold this against what others have written about the piece.

12.1.1 Text and recitation

As the title suggests, the piece is based on an untitled poem by the English Renaissance poet Thomas Campion (1567-1620), published in his treatise *Observations in the Art of English Poesie* in 1602 (Campion, 1966). The poem was written as an exemplification of a particular lyrical form, namely lyrical verse in *dimeter*, a classical quantitative meter in which the structure is based on the duration and number of syllables rather than accents and rhymes. Lansky describes the poem as being "about Petrarch's beloved Laura, whose beauty expresses an implicit and heavenly music, in contrast to the imperfect, all too explicit earthly music we must resign ourselves to make" (Lansky, 1994a):[307]

Rose-cheekt Lawra, come,
Sing thou smoothly with thy beawties
Silent musick, either other
Sweetely gracing.

Lovely formes do flowe
From concent devinely framed;
Heav'n is musick, and thy beawties
Birth is heavenly.

These dull notes we sing
Discords neede for helps to grace them;
Only beawty purely loving
Knowes no discord;

But still mooves delight,
Like cleare springs renu'd by flowing,
Ever perfect, ever in them-
selves eternall

The poem starts with an invocation of the mentioned Laura/Lawra by the lyrical subject, i.e. the character or persona we experience as presenting the poem *qua* text. We do not learn much about this lyrical subject throughout the text, apart from the feelings that he (or she) directs at Lawra. However, I would guess that most readers will associate the lyrical subject with the pre-modern stereotype of a courting male, expressing love and adoration for his beloved. In addition, readers might associate the lyrical subject with either Petrarch or Campion for obvious reasons. After the invocation at the beginning of the poem, the text revolves around the lyrical subject's experience of Lawra's features, especially her beauty, and the relationship between the two lovers. This topic is dealt with through several metaphors, especially musical ones, where the activity of singing is central, and where harmony and "concent" (an archaic word for the concord of sounds) are contrasted with "discord" and "dull notes", both being metaphors for interpersonal relationships. These musical metaphors are also associated with Lawra's beauty, with "pure" love and with heaven, the divine and the heavenly. Thus, it is not the sensuous quality of the music that is

[307] Ondishko points out that late Renaissance metaphysical poetry tends to invite an earthly interpretation in addition to a spiritual one, and that this also applies to this poem (Ondishko, 1990: 24). Such an interpretation would see "harmony", "beauty" , "heaven" and the "heavenly" as expressions of erotic feelings and see terms like "loving" and "delight" as references to the sensuous and bodily rather than the spiritual.

associated with Lawra's beauty, but rather the abstract idea of music – music that is paradoxically silent. Owing to this metaphor, one can at times get the impression that Lawra is more an idealized image of divine perfection than an actual woman of flesh and blood.[308] There are also other metaphors in the poem, some of them which express a flowing movement, either involving more abstract, immaterial shapes or water in motion. These metaphors are associated with the mentioned harmonious relationship and with the moving feeling of delight.

The recitation of the poem was done by Lansky's wife, Hannah MacKay, who has lent her voice to several of Lansky's compositions.[309] The recording was done in one single take, and in addition to the synthetic accompaniment in the first movement, it constitutes the sole sound source material for the whole piece (Ondishko, 1990: 23). In the last of the fantasies, entitled *her self*, the recording of the reading is presented without any manipulation, other than light reverberation.[310] The fact that it is a female voice that recites the poem, rather than a man, which would make a better fit with the stereotypes of Renaissance courtship, can potentially affect the meaning of the poem for the listener. One interpretation of this feature is that the praise of Lawra's beauty is not associated with heterosexual love, but love and adoration between women, with or without any erotic undertones. Another, and in my view more likely interpretation, is that the vocal source is *not* to be associated with the lyrical subject of the poetic text, thus emphasizing that the presentation is "only" a performance of somebody else's text. In that way, the reading thereby sets up the lyrical subject and the vocal subject(s) as two separated sites of intentionality. As a consequence, what in literary theory is sometimes referred to as "the speaking" and "the spoken" will be distanced or dissociated from each other (see e.g. Wesling & Slawek, 1995: 13).[311] As I see it, this interpretation is supported by the relatively neutral tone and temper of the recitation, without any strong affect laden emphasis or creation of dramatic moments. The partial dissociation between the vocal persona/reciting subject and the lyrical subject makes the central themes in the poem less poignant and perhaps the relationship between the lyrical subject and Lawra less interesting to engage in. This could potentially make the listener try to disregard the reader and the reading

[308] In many ways, Lawra thereby resembles Dante's Beatrice in his *Divine Comedy*.

[309] A few examples are *Idle Chatter*, *Word Color*, *Memory Pages* (all from Lansky, 1994b), *Things She Carried* (Lansky, 1997) and *Alphabet Book* (Lansky, 2002).

[310] The background accompaniment in this movement is an exception, since it is made of time-stretched and formant shifted vowels from the recording. The vocal identity of the accompaniment is only barely recognizable, however.

[311] Instead, the titles of the fantasies, all beginning with the pronoun "her", open up for an association between Lawra and the reciting vocal persona, since "her" can refer to both the reciting voice and Lawra.

in order to focus as much as possible on "the poem itself". Or, on the contrary, the listener could instead direct more attention toward the recitation and the voice.

In Lansky's piece, there are several reasons why I think the latter is much more likely than the former. One reason is that the topic of the poem, especially the general idea of a woman as an idealized image of the pure and heavenly, probably has little resonance in today's largely secularized Western culture, where the dominant object of idealization is instead women's physical attributes. Another reason that I suspect is even more acute for most contemporary listeners, i.e. at least those without higher-level degrees in early English literature, is that the poetic language and form dating several hundred years back make it more difficult to seize the meaning of the poem. For example, one here encounters words that are archaic, not in common use, or are used in a way that is not common nowadays ("thy", "thou", "concent", "either other" for "each other", "discords" and "helps" are used as plural nouns). Moreover, the structure of the poem with lines and stanzas, largely expressed by inserting pauses between the phrases in the recited version, makes it difficult to comprehend the overall syntax of the sentences. This is often owing to the fact that the syntactic function of each word or group of words is not clear until the completion of each stanza. Lastly, I expect that the metaphors earlier discussed increase the overall complexity of the poem for the contemporary reader, so that it is not straightforward to grasp its meaning.[312] It is interesting to note, therefore, that what attracted Lansky to this poem was its sonic properties more than its meanings, namely its play with vowels (CD liner notes, Lansky, 1994a; Ondishko, 1990: 22). In other words, it seems that an inclination to attend more to the *sound* of the voice than the semantic content can already be seen as the result of the choice of poem, the reciting voice and the character of the recitation.

One other important feature of Lansky's piece that further strengthens the focus towards the sound of the reading is that by presenting the text six times over, the linguistic-semantic aspects will be increasingly redundant. The effect of this reiteration could potentially have been weakened by adding or modifying the meanings through non-verbal or musical means, since if sound and words had interacted in ways that had given rise to new meanings, this redundancy would have been much weaker. I do not, however experience that this happens in *Six Fantasies*, at least not more than very mildly. Phrases are rarely given special emphasis or associated with non-verbal or musical meanings, at least not in a way that

[312] For me, it was not until I had sat down and studied the poem in its written form, using time and interpretative effort, that I was able to grasp many of aspects of the poem that I found opaque when listening to it.

gives a clear sense of change of direction.[313] This might also be why I find it hard to hear Lansky's piece as a particular interpretation of the poem, at least not as one where semantic aspects play an important part. Rather, as I experience it, each of the movements encompass a certain "configuration" of the voice or voices, a relatively stable "sound world" where there might be groups or layers of sounds which are similar to each other and where individual sounds change only within a limited range. Consequently, the sense of internal consistency for each fantasy in itself is relatively strong, largely making the differences *between* the fantasies more salient than the differences *within* each of them. For instance, one movement can be characterized by having speech-like intonation, whereas another movement has intonation contours quite similar to singing. Taken together, this gives the impression of a musical work in which the movements present different *facets* of the poem, not primarily as meanings, but as sounds related to the many shadings and configurations of vocal sound.

12.1.2 Overview of the piece

To get an overview of the *Six Fantasies* I here present a brief description of the six individual movements:

her voice (2'52'')

- Here, three adult women's voices articulate the phrases of the poem tightly synchronized and with speech-like intonation and articulation. The features of each individual voice are difficult to tell apart.
- The intonation contours of each of the voices appear to be parallel, but the intervals between the voices vary from phrase to phrase.
- The voices are accompanied by sustained notes in the treble registers and short notes in the bass register, both having synthetic timbre, playing notes and chords within the Western tonal system. These notes largely take on a background function when the vocal phrases are present.

her presence (3'45'')

- The second fantasy presents two female voices which articulate the poem either tightly synchronized or with one voice entering a short while after the other.
- The articulated text is accompanied by sustained static vowel sounds, which appear derived from the voices presenting the text, since they begin at the same time as corresponding vowels in the vocal phrases having the same vowel quality and pitch, often occurring at the end of the phrase. Sometimes this gives

[313] The instances of text painting noted by Ondishko are in my view very vague, and do not stand out as such in listening (Ondishko, 1990: 25-26). Moreover, the *fugato* treatment of **LCSRBF** in *her reflection* which is supposed to be a word painting of the flowing motion ("fugere"/"fugare" from which the word "fugato" is derived, means "to chase"/ "to flee"), is not unique to this phrase: It is also applied to the phrase "these dull notes we sing". Hence, it is more difficult to claim that the fugato and the "flowing" motion of the text.

the impression that the sustained vowels constitute a seamless continuation of the vocal phrase, but most of the time the transition from one to the other is noticeable.

- The intonation of the two voices varies from the speech-like to the quasi-melodic. The intonation contours appear related to each other in most cases, either by being close to parallel, or being inversions of each other.

- An "instrumental" section consisting of slowly changing chordal textures made up of sustained, vowel-like notes follows after the articulation of the last phrase of the poem (2:24-3:47).

her reflection (4'04'')

- The variability of the sounds in this fantasy is much greater than in the previous two fantasies. Here, one can find what appears as unmanipulated vocal phrases, as well as those that have only faint resemblance to a voice due to heavy processing.

- Many sounds appear to consist of excitation plus resonance, with the latter often having a gradually decaying envelope. For some sounds, only the resonances can be heard.

- Words or combination of words from the poem are often repeated several times, in contrast to the previous two fantasies, where each phrase occurs only once.

her song (3'10'')

- This fantasy presents a number of female voices in a choir like setting of the poem, with vocal phrases that appear to be "sung", so as to make up chords and melodies all mapped to the Western tonal system, often resembling those found in jazz.

- The phrases are "sung" in a dominantly syllabic manner and the duration of each syllable is markedly longer than in the first and second fantasies.

- The phrases of the poem are synchronously articulated by the voices, even if the synchronization does not appear as tight as in *her voice*.

her ritual (6'02'')

- In this fantasy the text is articulated by whisperlike vocalizations that often appear to excite resonant components, so as to make up tonal chords or more complex inharmonic textures.

- Words or combination of words from the poem are often repeated several times, even more than in *her reflection*.

- The vocal phrases are accompanied by iterative sequences of short, percussive sounds that together make up audible pitches and chords to different degrees. Sometimes these sounds take on the quality of words or syllables, but mostly they do not sound vocal at all. These sequences are frequently subjected to changes in rate and spectral qualities.

her self (2'31'')

- This movement presents the voice of a woman reading Campion's poem from beginning to end, in a calm and relaxed style, and with no other apparent processing than light reverberation.

- The female voice is accompanied by sustained, vowel like notes with partly overlapping stable pitches, sometimes constituting chords. The vowel qualities and the pitches appear partly to follow the vowels and the intonation curve of the female voice.

12.2 Background and guide to the evaluation

12.2.1 Organization of the material

Because this piece lasts over 22 minutes, I have had to restrict myself to evaluating shorter excerpts from each of the six movements. Consequently, I won't be able to say much about the overall structure in each of the fantasies. Since the *consistency* within each movement is relatively high, however, in that each of them are exploring a limited set of musical possibilities, one will probably still be able to see some tendencies for the piece as a whole by comparing these excerpts.[314] So, even if I am only evaluating parts of the six fantasies, they will nevertheless form a basis for making inductive generalizations of the piece as a whole, and this will be the focus of section 12.4.

Each of the excerpts I have evaluated consists of what I refer to as *vocal phrases* or just *phrases*, which consist of vocal or voice-like articulations of one to six consecutive words from the poem. I have not used the same phrases for all the fantasies, even if one can see that some of the excerpts contain the same vocal phrases. The vocal phrases of each of the excerpts are presented in **table 12.1**, along with reference to provided sound examples, the start and end points of the selection, and the abbreviation of the words contained in each

Movement/excerpt/sound example	**Phrase text**	**Abbreviations**
her voice (1:16-1:42) **Sound example 12.1**	Lovely formes do flowe From concent divinely framed Heav'n is musick And thy beawties birth Is heavenly	LFDF FCDF HIM ATBB IH
her presence (0:36-1:12) **Sound example 12.2**	Lovely formes do flowe From concent (voice 1) From concent (voice 2) Divinely framed Heav'n is musick And thy beawties birth Is heavenly	LFDF FC1 FC2 DF HIM ATBB IH
her reflection (2:34-2:58) **Sound example 12.3**	Purely loving (1st occurrence) Purely loving (2nd occurrence) Only beauty…knows no (fragments) Knows no discord (1st occurrence)	PL1 PL2 OBKN KND1

[314] Lansky has stated that the choice of shorter movements rather than a longer piece gave him more freedom in the composition process: "My thinking then was that doing shorter things gives you a lot more compositional freedom because you don't have to live with the implications of a decision for a long period of time. You can try experimental things" (Lansky's personal communication with Denise Ondishko, cited in Ondishko, 1990: 8).

	Knows no discord (2nd occurrence)	KND2
	But still (1st occurrence)	BS1
	But still (2nd occurrence)	BS2
	But still (3rd occurrence)	BS3
her song (2:25-3:05)	Like clear springs renu'd by flowing	LCSRBF
Sound example 12.4	Ever perfect	EP
	Ever in themselves	EIT
	Eternall	E
her ritual (2:02-2:43)	Lovely formes do flowe	
Sound example 12.5	- phrase type a, 1st sequence	LFDFa1
	- phrase type a, 2nd sequence	LFDFa2
	- phrase type b, 1st occurrence	LFDFb1
	- phrase type b, 2nd occurrence	LFDFb2
	- phrase type c, fragments 1-5	LFDFc1-5
	From concent (divinely framed)	
	- phrase type c, fragments 1-10	FCDFc1-10
	- phrase type b	FCDFb
	- phrase type a	FCDFa
	Heav'n is musick	
	- phrase type b	HIMb
	- phrase type c, fragments 1-4	HIMc1-4
	- phrase type a	HIMa
	And thy beawties birth	
	- phrase type b, 1st occurrence	ATBBb1
	- phrase type c, fragments 1-5	ATBBc1-5
	- phrase type b, 2nd occurrence	ATBBb2
her self (0:37-1:05)	Lovely formes do flowe	LFDF
Sound example 12.6	From concent divinely framed	FCDF
	Heav'n is musick	HIM
	And thy beawties birth	ATBB
	Is heavenly	IH

Table 12.1: Overview of the excerpts taken from the six movements of *Six Fantasies* with reference to sound examples, and of the different phrases contained in each excerpt with abbreviations for each phrase.

phrase. For the phrases from *her reflection* and *her ritual*, where the same text is repeated in several versions, I have also used numbers to indicate the order in which each phrase or fragment of a phrase occurs. For *her ritual*, three different phrase types are indicated by the letters a, b, and c.[315]

12.2.2 Form of presentation

The evaluation according to the seven premises of my framework will be presented graphically in different formats according to focus and level of detail. All graphical

[315] For a description of the three sound types, see section 12.3.5.1 below.

representations are derived from a representation created with the acousmographe software.[316] With this tool, I have been able to place and manipulate different types of graphical objects on top of a spectrogram and a time-domain representation of the signal. Additionally, this program has made it possible to play back shorter or longer portions of the sound file, loop it, change the playback speed and direction as well as to choose particular parts of the frequency spectrum that one wishes to listen to.[317] This has therefore enabled me to work in an interactive way during the evaluation process, going back and forth between listening, evaluation and re-evaluation.

The primary representation of the evaluations is contained in the provided score (.aks) files on the accompanying CD-ROM, which can be opened with the acousmographe software. These score files can then be viewed and listened to in an interactive way, so that it is possible to play back any desired section while watching the evaluations, choosing which graphical symbols to display and which to hide, zooming in or out, etc. For somewhat more immediate access, I have included the graphical representations rendered as Flash/Small Web Format movies (.swf) on the CD-ROM, and these movies can easily be viewed with most Web browsers. These movies contain the graphical representations accompanied by sound, but allow for no interactivity or selectivity of view, only playback. Lastly, I have provided images of the acousmographe representations as figures in the present text, hence providing easy access. To get the most out of this chapter, however, the reader is urged to accompany the reading of this text with interactive use of the score (.aks) files.

12.2.3 Key to the graphical representation

I have chosen to represent the evaluations in two different ways, each emphasizing different aspects of the evaluations:

1) **Time-varying representation:** This representation shows the temporal variations of the evaluations of each of the premises in the form of a horizontally oriented thick line or curve, where each premise has its unique colour. In addition, the noise mode and the reduced mode of the minimal for the information density premise are distinguished from

[316] This software can be freely downloaded from the Ina-GRM site, URL: http://www.ina.fr/entreprise/activites/recherches-musicales/acousmographe.html (accessed 02/02/2009. In the moment of writing it is available in version 3.4 for Windows Vista/XP and Mac OS X. Instructions for utilizing the provided score files are provided in Appendix B.

[317] I have mainly used the information gained from the spectrogram in guiding the placement of the graphical objects on the time-axis.

each other by letting the former have a hatched diagonal pattern. Taken together, this gives the following nomenclature:

i. *Focus of attention*: Bright green 🟩

	Reduced mode	Noise mode
ii. *Information density*: Blue-grey	▦	▨
iii. *Naturalness*: Light orange	🟧	
iv. *Presence*: Red	🟥	
v. *Clarity of meaning*: Turquoise	🟦	
vi. *Feature salience*: Violet	🟪	
vii. *Stream integration*: Yellow	🟨	

The curves/lines are placed in a system consisting of an upper and a lower vertical limit representing the maximal and the minimal evaluations. These lines should perhaps have been made diffuse so as to avoid the impression that they represent a definite, absolute and discrete value, but due to limitations in the acousmographe software, such a representation was not possible. The reader should therefore keep this in mind when viewing the representation. To ease comparison among evaluations, I have drawn the intermediate position between the two extremes as a dotted line. Moreover, I want to make it clear that the placement of these lines within the same "system", so to speak, is done purely for the sake of having the possibility of viewing the evaluations together: The evaluations are based upon different grounds, so that a similar placement along the continuum does not imply the same thing for different evaluations. Thus, the only thing that these evaluations have in common is that they can all be placed on a continuum between the minimal and the maximal, where the maximal end represents a convergence of the premises as discussed in chapter four. This does not mean, however, that it is impossible to compare the temporal evaluation of different premises; whether two or more premises have a common tendency or not for a certain phrase, whether they "peak" at the same time, or whether they reach their bottom levels simultaneously, can still be interesting to note.

2) **Axial representations:** The second representation closely resembles the axial representation in **figure 4.3**. Here, the circular light blue area in the centre represents the maximal voice, and the peripheral light blue circular zone represents the minimal voice. These areas are drawn rather fuzzy or diffuse to once again underline the lack of any

precisely defined limit for these evaluations. The evaluation of each of the premises at one particular time, indicated by a vertical grey line, is then represented by points on each of the seven axes. By defining a line between each of the points and an area enclosed within this line, one will have a representation which more directly than the coloured lines shows the evaluation for all the different premises together, especially since the *shape* and the *size* of the area give a lot of information just at one glance. For example, in cases where the area approaches a circular shape with the crossing of the axes at the centre, one will have evaluations which are relatively similar for all of the premises. Conversely, if the area constitutes a shape with many edges and irregularities or smoother shapes which do not have the meeting point of the axes at the centre, this indicates that the evaluations are more dissimilar, thus covering a greater span between the minimal and the maximal. As for the *size* of the area, it generally gives some indication of the overall tendency of the evaluation. A maximal voice, for example, will be represented as only a small shape in the centre of the figure, whereas a voice which is minimal for many of the premises will have a much larger area covered. However, since the size of the area bounded by the line between the points depends also on the arrangement of the axes, one cannot use size very precisely as an indicator of overall tendency. The area will for example be much larger if there are evaluations towards the minimal for several axes next to each other than if every other evaluation is minimal and then maximal, the latter giving a more edgy shape with a much smaller area. Therefore, size can be reliable only when the shapes are relatively similar, and this is most straightforward with shapes approaching a circular shape having the meeting point of the axes at the centre. Nevertheless, shape and size together can still be useful in discovering similarities in the evaluation of different phrases, since graphical shapes are very easy and fast to compare visually.

In addition to the graphical representations of the evaluations, I have provided text boxes for displaying the text of the vocal phrases. For phrases which occur in several versions within the same movement, I have also included numbers in parentheses, so that each phrase can be uniquely identified at a glance. For some movements I have also given the text boxes different colouring to distinguish them in terms of what category the phrase belongs to (cf. table 12.1). If the reader wants to avoid being affected by the content of these text boxes during listening, for instance if he or she wants to check out for himself or herself if my evaluation of the clarity of linguistic meaning corresponds to theirs, they are advised to hide this layer.

12.2.4 Use of analysis/synthesis model in the discussion

In the preceding chapters, I have used many sound examples to illustrate theoretical points in the discussion of the different premises. These examples have mainly been from existing works of music, but where needed, I have also synthesized many sound examples myself to prove points for which I could find no appropriate examples in the acousmatic repertoire. As a part of the *analysis by synthesis* approach described in section 1.4.3, I will draw on examples that I have synthesized in parts of the discussion, especially in section 12.4. This approach has been motivated by the possibility of seeing the evaluations against what they would potentially have been if the phrases were composed in a different way – what I referred to in section 1.4.3 as the *epistemology of simulations*. To do this, I have created a computer instrument which applies methods of sound processing that are highly similar to those in Lansky's piece.

The main technique that Lansky applied in this piece was Linear Predictive Coding (LPC), developed during the 1960s and 70s, mainly to compress speech signals (Atal & Hanauer, 1971; Atal, 2006). Greatly simplified, the LPC *analysis* of the speech signal makes an estimation of the time-varying filter component, the fundamental frequency of the phonation component (cf. section 3.3.3 on the source-filter model), the intensity of the signal, and whether the signal is voiced or un-voiced, i.e. noise (Lansky, 1989). In the *re-synthesis* process, the analysis parameters are used to decide whether to synthesize a buzz signal (pulse-train) for the voiced parts, or white noise for the un-voiced part – the appropriate signal being controlled by the analysis parameters for intensity (buzz and noise) and fundamental frequency (buzz only). The resulting source signal is then fed through a time-varying filter controlled by the analysis parameters, so as to create a synthesized approximation of the speech signal. By changing the analysis parameters, however, it is possible to manipulate the source parameters, in particular f0 and intensity, independently of the filter component. Since the analysis is made on a frame-by-frame basis, it is also possible to vary the frame rate in the re-synthesis process, thereby changing the playback speed without affecting pitch or spectrum. This enabled Lansky to transform the reading of Hannah MacKay in different ways to different degrees; transposing, inverting, flattening, exaggerating or fully "sculpting" the intonation contour, time-stretching the signal and replacing the buzz with noise.[318] The reading/speaking voice could be thereby be transformed into what sounded like a kind of singing (as in *her song*), a vocal style between speech and song (as in *her presence*), and

[318] The possibility of shifting the spectral envelope (i.e. the filter component) that LPC offered, however, was only applied in the accompaniment of *her self*.

whispering (as in *her ritual*). Lansky also implemented a "chorus" effect by using several pulse generators together with small random variations in fundamental frequency, hence creating a richer sound.[319]

In addition to the LPC technique, Lansky also applied comb filters extensively in *her reflection* and *her ritual*. Each comb filter, which is commonly implemented as a delay with feedback added to the original signal, adds a resonance at a particular frequency depending on the delay time and amount of feedback. By using banks of many double comb filters, Lansky could produce rich resonances, often with a chord like flavour. By setting the delay time higher, however, Lansky could create more typical delay effects, as can be heard in *her reflection*. A small amount of reverberation can also be heard in several of the fantasies.

Even if I have used a contemporary processing tool in creating my own instrument, I have tried to confine myself to the same technological paradigm as Lansky, using LPC analysis/synthesis at 14kHz sampling rate combined with comb filtering, chorus and reverberation. The computer instrument is called *Six Fantasies Machine* (SFM), and was made using the software synthesis and processing tool *csound*.[320] SFM is included on the CD which is submitted with the dissertation, and the reader is encouraged to install and play with the instrument, so that one can get a deeper understanding of how the different parameters affect the evaluations of this framework. This will also provide a better understanding of how the sounds examples in section 12.4 are made, and it may also give an extended experience of the possibilities and limitations of LPC synthesis combined with effects such as comb filtering. A closer description of SFM as well as an installation guide can be found in the manual in pdf format on the accompanying CD.

The SFM instrument produces vocal phrases that are similar to those that can be heard in Lansky's piece, but they are *not* based on the same vocal material.[321] Having no access to Lansky's original LPC files nor to his sound files, I instead used an actress to imitate Hannah MacKay's original reading as it is presented in *her self*. The recording of her reading was then analyzed with the LPC analysis utility in *csound* and used as a basis for the LPC-resynthesis

[319] Personal communication with Lansky, October 2006.

[320] *Csound* was developed on the basis of the so-called *Music N* family of software synthesis languages (Boulanger, 2000; Roads et al., 1996: 787-802). The current version at the time of writing is 5.11 and it can be downloaded for free from http://sourceforge.net/projects/csound/files/ (accessed 4/23/2010). The routines for making the LPC analysis and synthesis in *csound* are modified versions of Lansky's own LPC programs.

[321] The reader is urged to compare the imitations provided in presets 1-10 of the SFM instrument with the original phrases in the piece.

in the instrument. Clearly, the use of another voice is one of several reasons why the sounds created with SFM are not quite the same as the sounds in the original piece.[322]

With SFM it is possible to control several parameters in the synthesis process, among them pitch contour, intensity, time-stretch of segments for each phrase, spectral envelope shift and blend of pitched sound and noise in the excitation signal. One can also apply effects to the synthesized signal, mainly reverberation and different kinds of filtering. SFM can be controlled either from a script or by using a graphical user interface (GUI), but the latter is by far the easiest method since it requires no previous knowledge of the *csound* score syntax.

All in all, making the SFM instrument has both given me an in-depth understanding of the challenges and benefits of the LPC and comb filtering techniques, and enabled me to familiarize myself with the sounding results of the different techniques. Furthermore, I have been able to interactively investigate the results that different settings of the instrument have on the listening experience and the evaluation of the developed framework, something which will be discussed in section 12.4.

12.3 Evaluation of excerpts from the six movements

When presenting and discussing the evaluations of the six excerpts, I have been forced to delimit the level of detail in the descriptions. This has been done both to restrain the length of the dissertation and to maintain the reader's interest. Since many of the same factors and issues have been involved in the evaluations of three of the excerpts, namely *her voice*, *her presence*, and *her song*, I have made a condensed presentation of the latter two in the text, focusing mainly on what distinguishes each movement from the other two. Moreover, since *her self* was much less complex than the other movements, I have also given only a condensed presentation of this fantasy. Therefore, the discussion of *her presence*, *her song* and *her self* goes less into detail for the single premises and phrases than for the remaining three fantasies. For them, I will discuss each of the premises separately, linking the evaluation to the factors that I see as most important in each case, either explicitly in the text or with a reference in parentheses. In all cases, factors will be written in *italics*.

As for the categories along the continuum, I have largely used the five categories as I have practiced in the preceding chapters. For some premises and phrases, however, I have

[322] I also suspect that I used recording levels that were too low, so that the general impression of the voice in my instrument is somewhat thinner than in *Six Fantasies*.

wanted to make further distinctions, so that I have in a sense applied a finer resolution in the evaluation. I will refer to these evaluations as either being "close to", "just above", "just below" or "in between" the five established categories *maximal*, *maximal-intermediate*, *intermediate*, *intermediate-minimal*, and *minimal*.

12.3.1 Evaluation of *her voice*

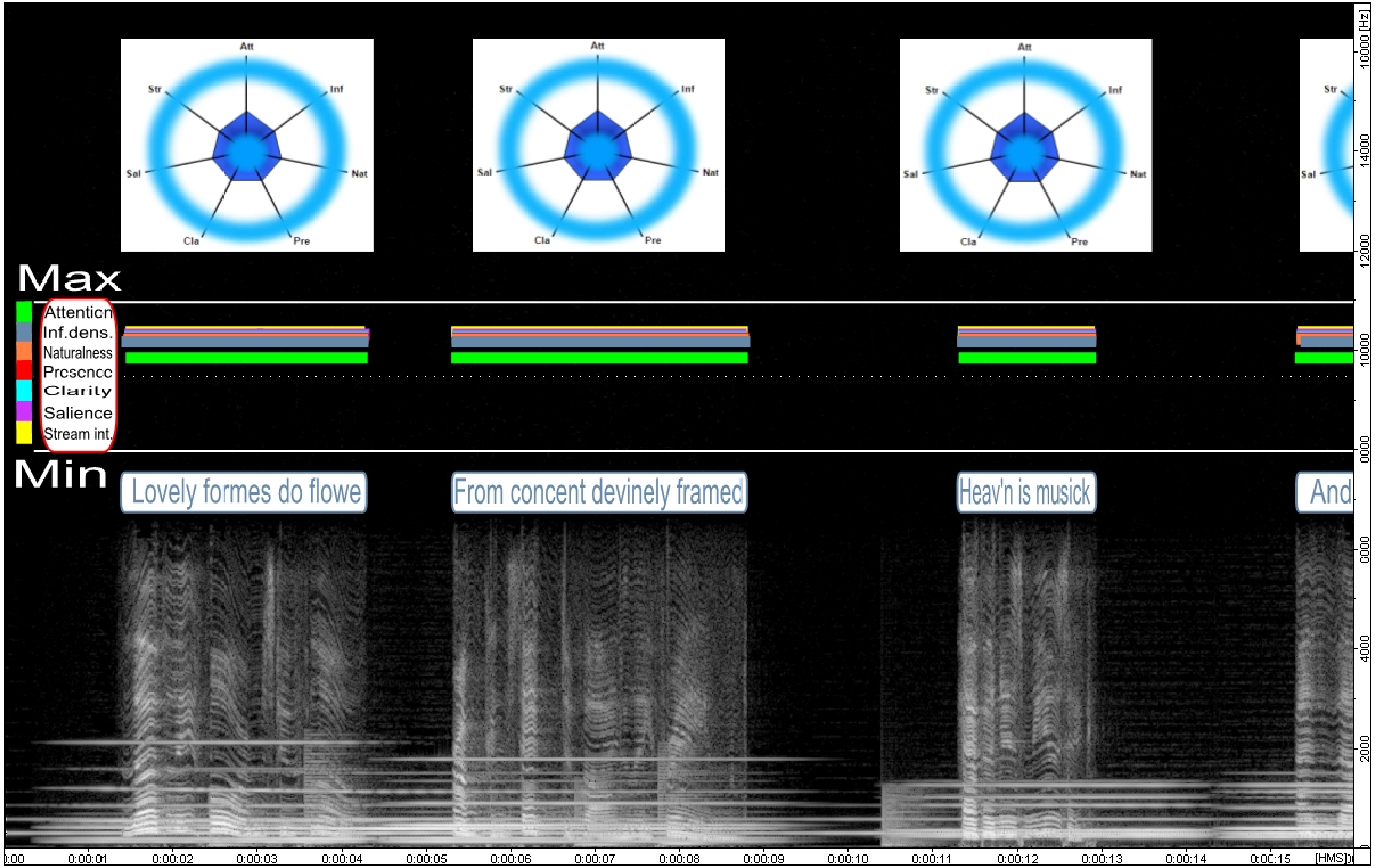
12.3.1.1 Over-all description

The first fantasy, *her voice*, comes in many ways the closest to the original reading as it is presented in *her self*. In particular, this is because the seemingly parallel intonation contours are clearly speech-like. The relatively slow and clear articulation also makes the phrases of the poem easily perceivable. And as in *her self*, the lines of the poem are largely presented as they are originally ordered.[323] However, the slightly buzzy quality of the pitched portions of the phrases gives them a mildly synthetic quality, which sets them apart from the phrases in *her self*, which lack these artifacts of the LPC technique.[324] The features that I experience give *her voice* its momentum are mostly the intonation contours of the vocal phrases, along with the "instrumental" accompaniment, which appear to follow the variations in effort and intensity in the vocal phrases.

The excerpt I have chosen from *her voice* is presented in **sound example 12.1** (Lansky, 1994a: *her voice*, 1:16-1:42). It contains the phrases **LFDF**, **FCDF**, **HIM**, **ATBB** and **IH**, and my evaluation of these can be seen in the graphical representations in **figure 12.1**. An even better view of the evaluations, however, can be attained by opening the acousmographe files at this point. One will then also be able to hear the music while viewing the evaluations one at a time. From the similarity in size and shape of the axial representations and the straightness of the time-varying representations, one can clearly see that the evaluations have no variation. This is not too surprising since the mentioned features of the fantasy, along with those listed in section 12.1.2, do not vary much during the movement. *Her voice* therefore appears highly consistent.

[323] The exception is in the introduction where the name of the female subject of the poem is inserted once before the beginning of the first phrase, as the underlined name shows here: "Lawra, Rose cheek't Lawra".

[324] It is also possible to identify an artificial noise component at the end of some phrases if one listens closely.


Att
Str
Inf
Sal
Nat
Cla
Pre
Max
Min
Attention
Inf.dens.
Naturalness
Presence
Clarity
Salience
Stream int.
Lovely formes do flowe
From concent devinely framed
Heav'n is musick
And
16000 [Hz]
14000
12000
10000
8000
6000
4000
2000
0
0:00
0:00:01
0:00:02
0:00:03
0:00:04
0:00:05
0:00:06
0:00:07
0:00:08
0:00:09
0:00:10
0:00:11
0:00:12
0:00:13
0:00:14
0:00:15
[HMS]

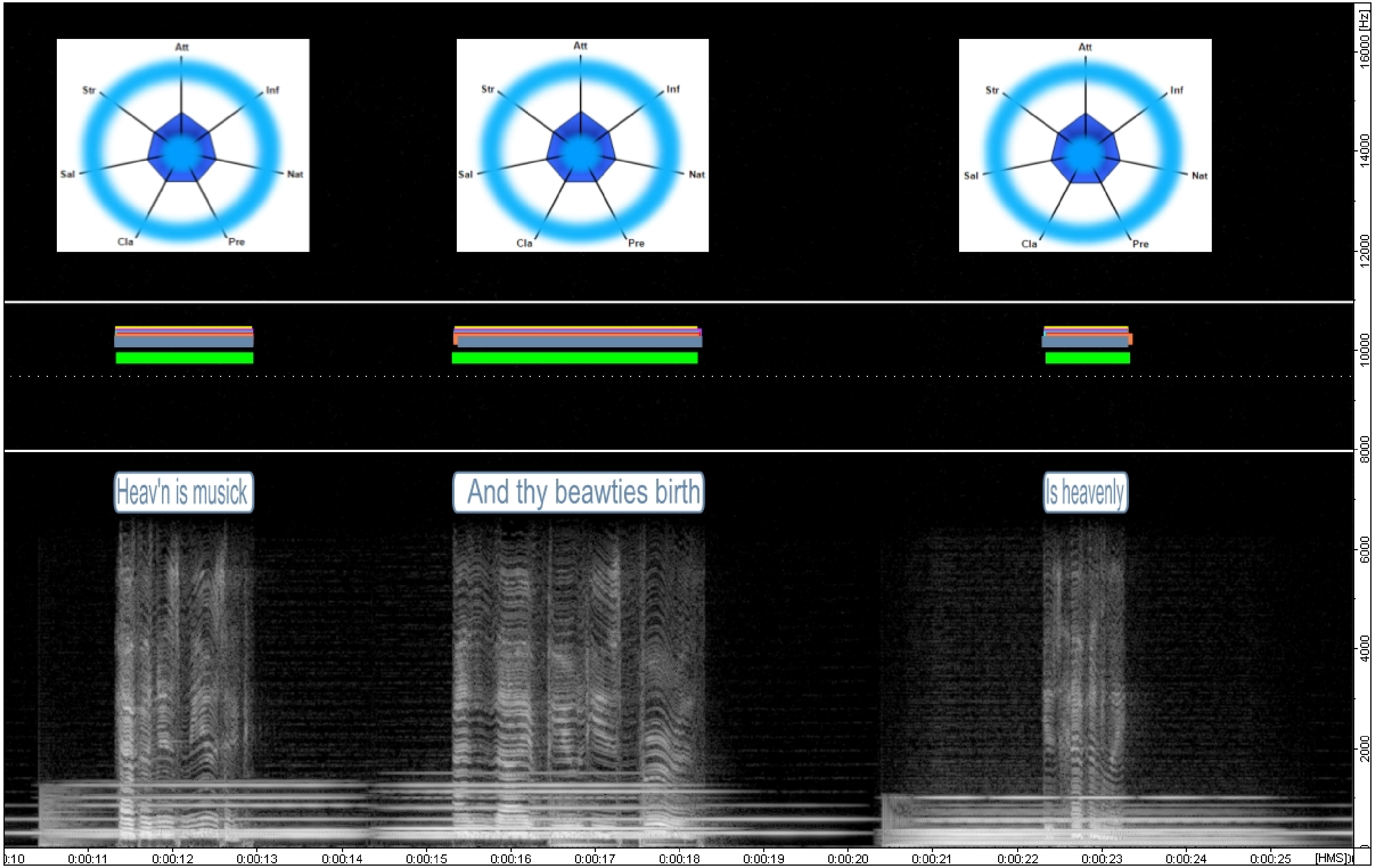


Figure 12.1: Graphical representation of evaluations of the seven premises in Lansky's *her voice*. Note that some of the colored lines overlap so that they are not fully visible in the figure. For representation of the evaluation of single premises, the reader should use the *acousmographe* software and the files provided on the accompanying CD-ROM.

12.3.1.2 Attention evaluation

The mentioned constancy and consistency in this movement are also reflected in the evaluation of this premise, which shows no variation. The constancy also implies that factors such as *novelty/change* and *unpredictability* are therefore not very important for the evaluation of this premise. Rather, I experienced that *relevance/interest* had more to say here. For me, the **LI/sem** aspects did not dominate during listening, not so much because the verbal meaning of the verbal content of the poem is difficult, as was discussed in section 12.1.1, but more because I did not feel that the content of the poem had much relevance for me personally, dealing with issues that I experience as archaic, both on the interpersonal and metaphysical level. Still, I was not completely negligent to **LI/sem** aspects, and at times my attention focused on the surface meaning of the words. However, I found that several other aspects related to both vocal and non-vocal domains caught my interest *along with* the surface meaning of the phrases. These aspects were mainly the articulatory movements (**VG-domain**), the shapes of the intonation contours (**SQS-domain**) and the general level of bodily activity/arousal that they indicate (**VG**/**AF-domain**s). Due to the relative low prominence of these features, I have interpreted this situation as lying between two categories in table 5.1: 1) divided attention between all three loci (*intermediate* evaluation), and 2) low strength distraction from vocal/non-vocal domains (*maximal-intermediate*). Taken together, this has therefore resulted in an evaluation just above *intermediate* for all phrases.

12.3.1.3 Information density evaluation

The evaluation of this premise is partly a result of the relatively constant configuration of features in the fantasy. The fact that many features do not change much, makes them relatively predictable and thereby also redundant. A listener will probably soon be able to predict with relatively high certainty that since these features have remained unchanged up until a point, they will continue in the same way in what is to come.

One can also link this to several of the factors discussed for this premise. First of all, the *need for re-listening* to be able to grasp and take in the features presented is not too large here, since one can get a grip on most of the features, perhaps apart from the semantic meaning of the poem, at first listening. Secondly, the consistency in features can be seen as constituting *regularities* and reinforcing *familiarity* with the way features are configured, something that will reduce the factor of *complexity*. The use of speech-like intonation contours will also be familiar to listeners, hence also contributing to reduced complexity. In addition, one can see that the low number of voices (cf. *number of elements/dimensions* factor) and the clear discontinuities at the beginning and end of the phrases (cf. *segmentation/grouping* factor) play a part in making the phrases in this excerpt relatively low in *complexity*. When it comes to the *rate* factor, I clearly experience that information is presented more slowly than in e.g. a public address on radio, especially if one takes into account the relatively long pauses between the phrases. This factor also contributes, although subtly, in decreasing information density.

Even if the excerpt presents phrases that are relatively redundant and low in complexity in several ways, this is partly counterbalanced by the comparatively high complexity in the **LI-domain**, especially for the semantic level. For me, as I would guess for many contemporary listeners, the semantic couplings between the main character (Lawra), her beauty, heaven/the divine and music represent relatively complex information that requires time and effort to process.[325] There are also other concepts here that may be difficult to get a grip on immediately:

- What are the “lovely forms” that are mentioned? Are they material or immaterial?
- How can they flow from “concent” (harmony, concord of sounds)?
- How can a property such as “beauty” be given birth to and in what respect is this divine?

Taken together, I experience that information density is in the direction of the *reduced mode* of the minimal for this excerpt, but that the semantic complexity still contributes so that information density is not too far from being balanced. Consequently, I have evaluated it as *maximal-intermediate*.

12.3.1.4 Naturalness evaluation

For this excerpt, I experience that *naturalness* is clearly not *maximal*, given the mildly synthetic quality of the voices, but it is not too far from the *maximal* either – the three voices do not move outside what a natural human voice can produce. If we relate the evaluation to the factors, there are two in particular that seem to affect the evaluation here:

[325] I will discuss this further in the section on the evaluation of clarity of meaning.

- **Technological artifacts/phonatory spectrum:** The phonation component (the voiced parts of the phrases) generally has a slightly noisy and buzzy quality. Some fricatives also sound somewhat artificial, like the [f] in "Lovely formes do flow".[326]
- **Precision:** The high synchronicity of the articulation can appear unnatural if one starts to reflect on it during listening. However, it can also be relatively easily neglected.

Based on this, I evaluate the excerpt to be *maximal-intermediate*.

12.3.1.5 Presence evaluation

For the vocal phrases in the excerpt, presence is clearly not maximal, but still not too far from the maximal. And, as for the other premises, there is little change in the evaluation during the excerpt. I regard the following factors as contributing most in the evaluation:

- **Explicit/implicit transformation:** The voices appear to be somewhere between the reproduced, the synthetic and the manipulated. Consequently, I do not experience that the voices have been subjected to marked explicit transformation. The overt constructedness of having three precisely synchronized voices, however, adds to the mildly synthetic character in reducing presence modestly.
- **Social distance:** The voices appear to be located at a personal distance, indicated by moderate vocal effort and modal voice quality.
- **Temporal continuity:** There are no repetitions or temporal disruptions. There is a sense of continuation through all the phrases in the movement, and the pauses between each phrase in the excerpt are relatively short.
- **Multi-modal associations:** For me, multi-modal associations are modest here. I find it difficult to associate the multiple synchronized voices with correlates in other modalities.
- **Contextual linkage:** There are no ambient or environmental sounds accompanying the voices, only an accompaniment with a quasi-instrumental character (one "instrument" playing chords and single notes of long duration, the other playing shorter bass notes). Since there are no obvious links between the voice and the accompaniment in terms of sharing space, the contextual linkages are very weak.

[326] In addition, there are some incidences of relatively subtle technological artifacts in the form of noise pops in parts of the movement that are not presented in the sound example, e.g. in the phrase "like clear springs ren[click]ued [click]by flowing" (2:20-2:24).

Taken together, I experience the vocal phrases in this excerpt as having *maximal-intermediate* presence, something which corresponds well with the contribution of the mentioned factors.[327]

12.3.1.6 Evaluation of clarity of meaning

The evaluation of this premise is largely based on these issues:

1) **LI-domain clarity:** The majority of the words (between some and most, cf. table 9.3) in the excerpt have a relatively high bottom-up clarity and can easily be recognized, owing to phrases and words that are articulated clearly and slowly. The few words with ambiguous phonemes can with a few exceptions be interpreted on the basis of lexical or semantic redundancies.
2) **Contextual specificity:** The environmental setting in this excerpt is unspecified and therefore uncertain. As for the identity of the three speakers, it can be specified to some degree, but their similarity and synchronicity creates some ambiguity as to whether it is one or several vocal personas presenting the text.
3) **Within/between domain coherence:** The ambiguity discussed in section 12.1.1 regarding the relationship between the lyrical subject and semantic content of the presented text, represents an incoherence between the **ID-domain** and the **LI/sem** which contributes to reducing the clarity of meaning somewhat. The metaphors used in the poem, at least if one is not able to interpret them meaningfully, can also represent some degree of reduced coherence.

Based on these criteria, the overall evaluation for all of the phrases in the excerpt is *maximal-intermediate*.

12.3.1.7 Evaluation of feature salience

For the feature salience premise, I have chosen to evaluate only the salience of the three synchronous voices relative to the background accompaniment, and not each of the individual voices relative to each other. The reason for this is primarily that the voices in most respects behave as one sound source, that the relationship in salience between the three voices is

[327] This is also the case with the remaining phrases in the movement.

relatively constant for this movement, and that I would not like that the evaluation should be more complex than necessary.

The salience of the vocal layer against the background accompaniment is, as I experience it, dominantly affected by the following factors:

- **Masking/simultaneous contrast:** The background accompaniment is considerably lower in loudness than the vocal phrases, hence making the contrast to the vocal phrases relatively high. This contrast also largely prevents masking. Here, even fairly soft portions of the vocal phrases are not masked, maybe with the first of the two [f]s and both the [m]s in **FCDF** as exceptions.
- **Temporal discontinuities:** The attack and ending phases of the vocal phrases are relatively marked, due to discontinuities in both loudness and spectrum, including pitch and spectral brightness, hence making them stand out clearly from the background accompaniment.
- **Condition of vocal features:** Most vocal features stand out relatively clearly in this excerpt. Since the phrases contain superimposed parallel intonation contours, the pitches of the individual voices stand out less. Voice quality features are also somewhat opaque in consequence of the artificial quality of the pitched component. Furthermore, the phonemes mentioned in the first point above indicate a degrading of these sounds.

My experience of the vocal phrases in this excerpt is that they are *maximal-intermediate* for the feature salience premise.

12.3.1.8 Evaluation of stream integration

That three quite similar voices articulate the phrases in this movement in what sounds like a perfect parallel, make them appear partly as three different voices and partly as one single voice.[328] Hence, there are forces integrating streams on several levels of source coherence:

- **Feature similarity/proximity/continuity:** The continuity in pitch for each individual voice, and the simultaneous differences in pitch between the three voices, act as

[328] Ondishko's interpretation of this movement as "a chorus of women's voices, behaving as one voice" captures this ambiguity in a pertinent way (Ondishko, 1990: 31).

forces segregating the voices from each other and integrating each voice into a single stream.

- **Similarity in spatial location**: To a modest degree, differences in spatial location act as forces segregating the three voices from each other, and integrating them into individual streams.
- **Modulation coherence:** The synchrony in articulation (synchronous modulation of intensity and spectral qualities), the parallel pitch contours (synchronous pitch modulation) and the relatively similar timbral qualities, integrate the voices into one single stream. This imposes a certain degree of ambiguity relating to the level of source coherence, and thereby whether one can hear several voices (indicated by the cues for segregation) or one single voice with a "chord like" quality (indicated by the cues for integration).

These ambiguities regarding source coherence has made me evaluate this excerpt from *her voice* as *maximal-intermediate*.

12.3.2 Evaluation of *her presence*

12.3.2.1 Over-all description and evaluation

The second movement, *her presence*, stands out compared to the rest of the fantasies with its two part structure, having an "instrumental" part following the part containing the vocal phrases. Even if this second "instrumental" part is based on vocal recordings, the relationship to the vocal origin is so remote as to come through as non-vocal, at least in my ears, and I will therefore *not* deal with this part in the following discussion.

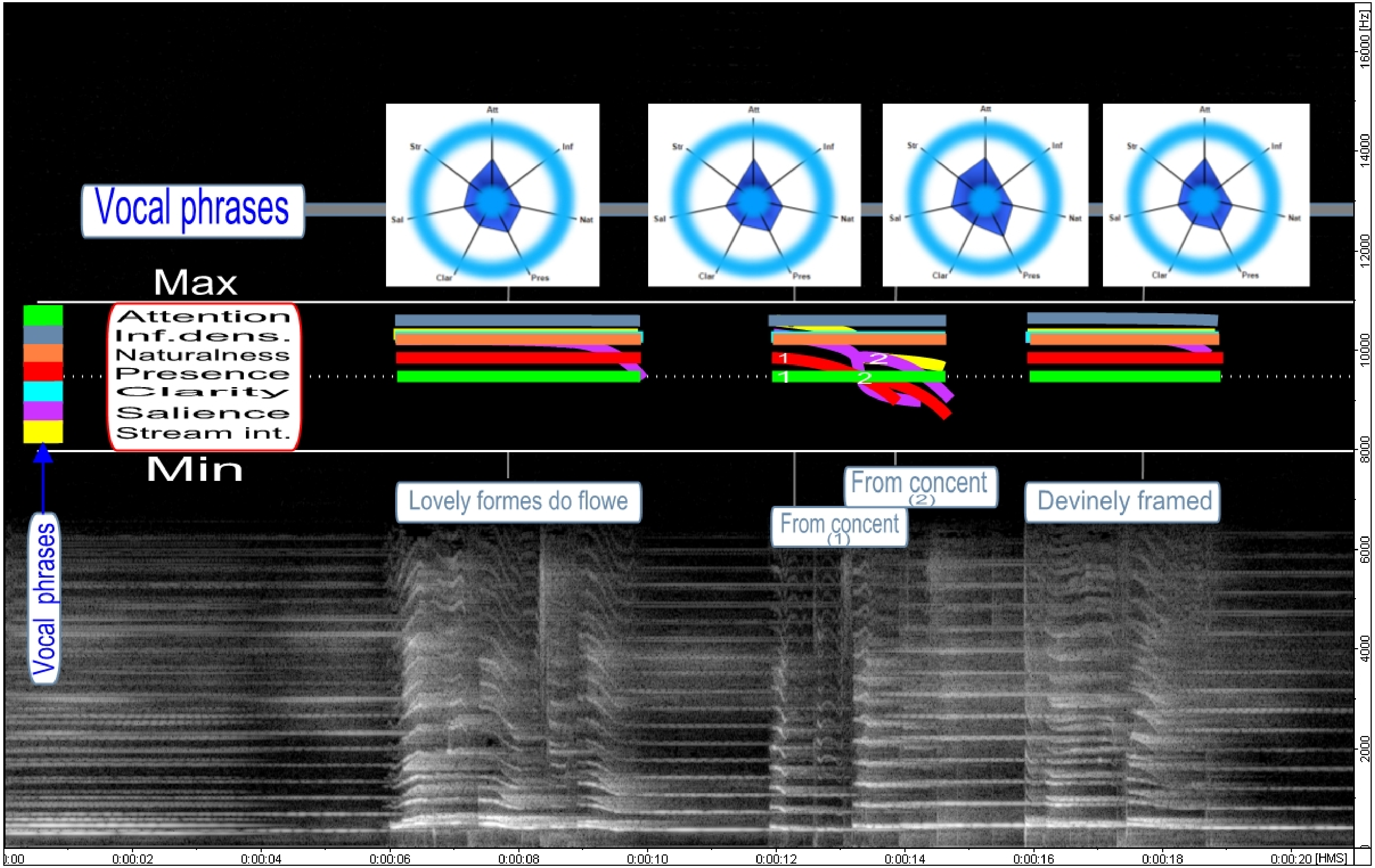
As for the vocal phrases in the first part, the two voices that one can hear are much easier to identify and distinguish from each other than was the case with the tightly synchronized voices in the first fantasy. In this movement, the voices instead seem to move in and out of synchronicity: Sometimes they articulate the phrases simultaneously, other times the second voice is delayed compared to the first so that they are only partly superimposed. In addition, the phrases in this fantasy are accompanied by more reverberation. For many phrases, the two voices differ both in loudness and amount of reverberation added, often so that one voice takes on a foreground function, whereas the other resides more in the background. There is also variation in the way that the intonation contours of the two voices are related to each other: They move in parallel, they move away from (diverge) or towards each other (converge), they "mirror" each other's pitch inflections, or they imitate each other.

The intonation contours of these phrases appear somewhat more "tuned" and sculpted compared to *her voice*, often starting or ending on the same notes, often constituting harmonic intervals, and often landing on a particular scale step. But, since it appears that the rhythm of articulation is maintained and since the pitches often move in a gliding fashion, most of the phrases do not appear like they are "sung", but maybe something in between speech and song. In that respect, it resembles the *Sprechgesang* applied by composers like Schoenberg and Berg, where the performer is to strike the note, but then leave it immediately by rising or falling in pitch (Griffiths, 2009; Anhalt, 1984: 7-9). Moreover, the upper voice reaches considerably higher in pitch in this movement than the previous one, and at the highest pitches, I experience that the phonation quality and the vocal effort are not quite corresponding.

This fantasy introduces an accompaniment that stands in a particular relationship to the vocal phrases, in that it consists of vowel or vowel-like timbres, which sound highly artificial because pitch and spectrum are rather static, lacking the natural fluctuations of a human voice (cf. *fluctuations* factor), and the timbre has a kind of ringing quality to it, not attributable to vocal behaviour.[329] Many of these vowels even loose the vocal quality altogether and sound like wholly synthesized sounds. Still, the fact that the onset of many of these sustained vowel sounds coincides with the articulation of a similar vowel in the vocal part creates a kind of causal connection between the two. This connection differs in strength from phrase to phrase depending on particularly two things: 1) The similarity between the vowel in the vocal phrase and the sustained vowel, and 2) the degree to which the onset of the sustained vowel is synchronous with the articulation of the corresponding vowel in the vocal phrase. This link, or perhaps one should call it *partial* link, can contribute in making the vocal origin of the sustained vowels more apparent to a listener. Moreover, the number of sustained voices sounding simultaneously also affects the degree to which the vowels appear voice like – in general, the higher the number, the more artificial they sound.

The excerpt from *her presence* is featured in **sound example 12.2** (Lansky, 1994a, *her presence*, 0:42-1:12) and the phrases included here are **LFDF**, **FC1, FC2**, **DF**, **HIM**, **ATBB**, and **IH**. The evaluations are shown in the graphical representations provided on the acousmographe files on the accompanying CD-ROM and in **figure 12.2** below. Due to the

[329] The evaluation of the sustained vowels can optionally be viewed in the acousmographe version of the graphical representations.



Vocal phrases
Att
Inf
Nat
Pres
Clar
Sal
Str
Max
Attention
Inf.dens.
Naturalness
Presence
Clarity
Salience
Stream int.
Min
1
2
1
2
Lovely formes do flowe
From concent (1)
From concent (2)
Devinely framed
Vocal phrases
16000 [Hz]
14000
12000
10000
8000
6000
4000
2000
0
0:00
0:00:02
0:00:04
0:00:06
0:00:08
0:00:10
0:00:12
0:00:14
0:00:16
0:00:18
0:00:20 [HMS]

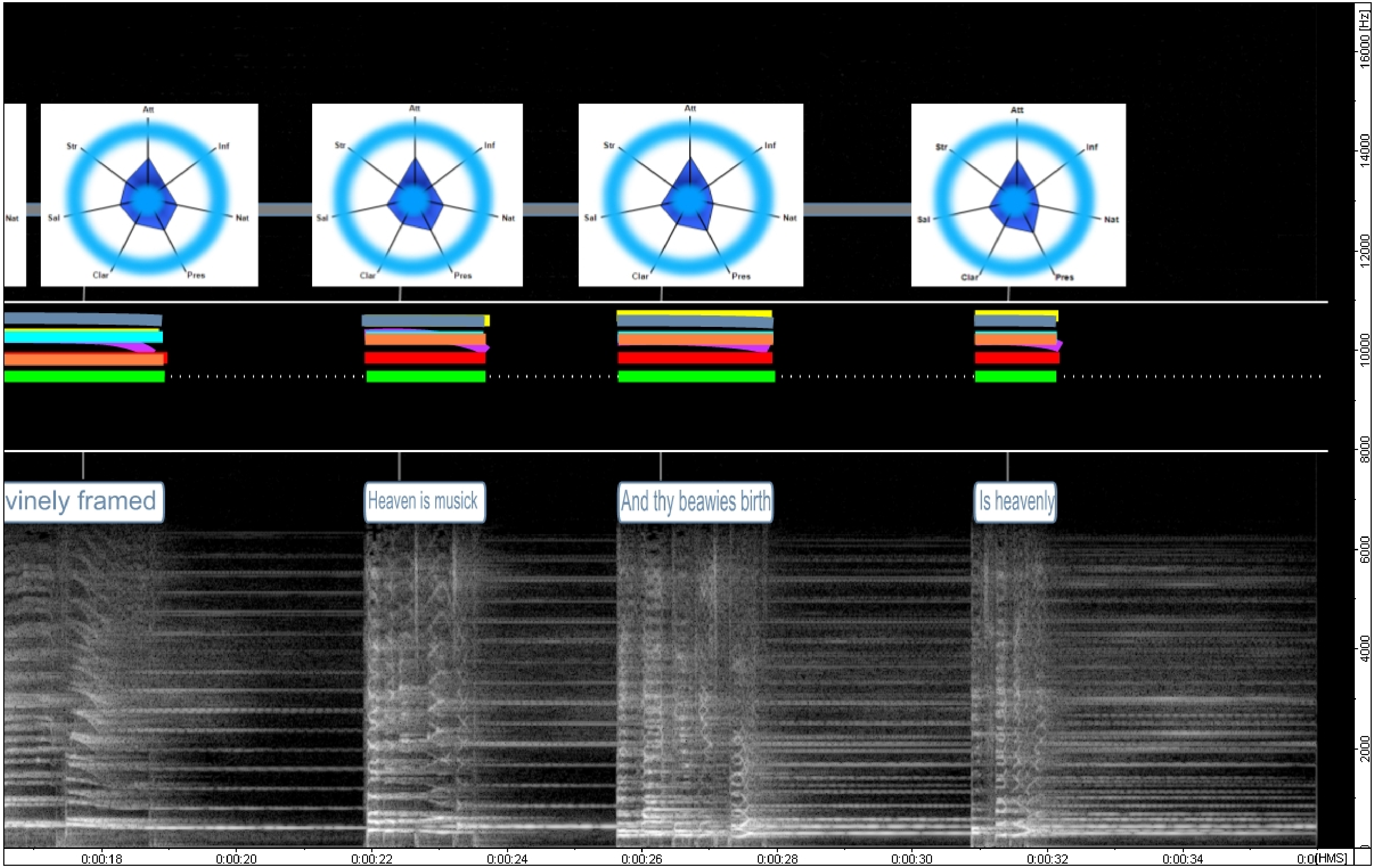


Figure 12.2: Graphical representation of evaluations of the seven premises in Lansky's *her presence*. Note that some of the colored lines overlap so that they are not fully visible in the figure. For representation of the evaluation of single premises the reader should use the *acousmographe* software and the files provided on the accompanying CD-ROM.

fact that many vocal features in this movement resemble the previous one, and that the evaluations thereby are relatively similar, I will not go into so much detail on the evaluations and the contributing factors here. Rather, I will give a brief outline of what distinguishes the movement and the evaluation of this fantasy from *her voice*.

First of all, the intonation contours and the interrelationships between them, with the drift back and forth from arbitrary to recognizable harmonic intervals, are all features that are more prominent and easier to follow than the speech like contours in *her voice*. While the division of attention between **LI/sem**, the non-vocal and the vocal domains are not far from that of *her voice*, these features contribute so that I directed somewhat more attention towards the vocal and non-vocal domains, especially the **SQS-domain**. This has resulted in a slightly lower evaluation of the *focus of attention* premise. Since this variation is not very predictable, it also represents somewhat higher *information density* compared to the first fantasy, as can be seen from the evaluations.

The *naturalness* evaluations are by and large similar to the previous fantasy, but some of the phrases reaching high pitches appear to have a vocal effort that does not quite match what would be expected at these pitches, hence giving them a little lower evaluation than the rest (cf. *pitch/spectral envelope relationship* factor). Moreover, I experience that these phrases, and especially **FC2** and **DF**, appear more distant than in the previous fantasy, thereby giving them lower evaluations for *presence*. For the *clarity of meaning* premise, the differences between the vocal phrases in the first two fantasies are negligible.

One of the marked differences between this fantasy and the previous one is the sustained vowel accompaniment partly linked to the vocal phrases. Owing to this link, they carry with them, so to speak, some of the properties of the vocal phrases, despite that their static quality gives them almost a non-vocal character. For most of these vowels, then, I experience that when I attend to them, attention is gradually shifted away from vocal features at the onset of the vowel towards attending to their spectral qualities as they continue, in other words a non-vocal domain. Consequently, the evaluation of the *focus of attention* premise for these sounds tends to decline gradually towards the minimal until the onset of the following vocal phrase. The static character of most of these vowels also affects the other premises. A static sound is per definition highly regular, and once we recognize its regularity, it will become highly predictable, and thereby low in information. Therefore, the evaluation of *information density* also has the same tendency for declining gradually from the onset of the vowels to the beginning of the following vocal phrase.[330]

The static or close to static character and the highly synthetic quality of the vowels also affect other premises. Both experienced *naturalness* and *presence* are evaluated as *minimal*.[331]

Since the sustained vowels have a partial link with the vocal phrases through the common vowel, I also experience them as linked with the phrases in terms of meaning. Therefore, I do not find the phrases devoid of meaning, despite that they only carry one single sound which is *per se* without any specific meaning. However, the static articulation affects bottom-up clarity negatively compared to the vocal phrases, thus making the evaluations for the vowels lower than the vocal phrases. The high degree of artificiality, which makes the identity of the vocal persona more ambiguous, and the somewhat ambiguous link with the vocal phrases also reduce contextual specificity for the sustained vowels. For the two cases where different vowels are layered (following **ATBB** and **IH**), it is more difficult to make out the qualities of the vowels, and these cases thereby receive the lowest evaluations. All in all, the sustained vowels are evaluated as lying in the range between *intermediate* and *intermediate-minimal*.

For the *feature salience* premise, I experience that the vocal phrases largely stand out markedly against the sustained vowels that accompany them because they are considerably louder, not too far from what was the case in *her voice*. However, the endings of the vocal phrases are somewhat less salient due to the partial continuation by the sustained vowels. Consequently, there is a small decline in the salience evaluation at the end of the phrases. The sustained static vowels are in general evaluated as having lower feature salience than the vocal phrases, since their vowel qualities are less salient, especially when they are layered.

Compared to the phrases in *her voice,* those in *her presence* differ somewhat more for *stream integration*, due to the variation in the configuration of voices. Generally, the lower number of voices and the clearer spatial separation make the two voices fuse less on the global source coherence level in this movement, hence making the evaluations somewhat higher. The two phrases with approximately parallel pitch contours and synchronous articulation, **LFDF** and **DF**, are not too far from the phrases in the previous movement. The

[330] For two of these vowels, namely those following **LFDF** and **DF**, there is a slight gradual change in vowel quality along the way, reducing the decline towards the minimal for both these phrases.

[331] There are slight changes in naturalness for the different vowels, but I still consider them all to be *minimal* in naturalness. For instance, the vowels following **ATBB** and **IH**, both consisting of more than one layer of vowels, sound most artificial, whereas those following **LFDF** and **DF**, are somewhat more natural, due to the small change in the vowel qualities for these vowels.

phrases with non-parallel pitch contours and synchronous articulation, **HIM**, **ATBB** and **IH**, are even more clearly segregated, and therefore have a higher evaluation. For **FC2**, however, the partial integration between the sustained vowel and vocal phrase towards the end of the latter creates a source coherence ambiguity that made me evaluate it as lower than the remaining phrases. For the static sustained vowels, those that are layered are evaluated as lower than the remaining ones, due to the source coherence ambiguity introduced by the layering.

All in all, the phrases in this movement show more variation in evaluation than in the previous one, although there is evidently a lot that is similar. Especially if one takes the sustained vowels into consideration, the variations are marked. All in all, therefore, this fantasy appears a bit "looser" and less consistent than the first one.

12.3.3 Evaluation of *her reflection*

12.3.3.1 Over-all description

Compared to the first two fantasies, the individual phrases of *her reflection* are sonically much more diverse, something which is evident from the excerpt in **sound example 12.3** (Lansky, 1994a, *her reflection*, 2:31-2:58). In this excerpt, there are vocal sounds that appear unprocessed and there are sounds that are hardly recognizable as voice at all due to heavy processing. The heavily processed sounds all appear to incorporate *resonances* of some kind, excited by vocal sound, like for example the strings of a guitar would start to resonate if shouting close to it. For some of the sounds, one can hear the excitatory voice together with the resonances, whereas for other sounds only the resonances are audible. For all of these sounds, however, I experience that the resonances are *external* to the voice, emanating from an undefined object or material structure.

Another thing that distinguishes this movement from the previous two is that there is text repetition here; the verbal content of whole phrases as well as shorter portions of them is repeated – most often with relatively large sonic differences between repetitions. When one considers these sonic differences across all the phrases in the movement, however, one can make a rough classification into four main phrase types, which are all represented in the featured excerpt by one or more phrases:[332]

[332] This corresponds to the four types of sound that Lansky himself operated with in the composition process, which he labelled "whisper", "drone", "sitar" and "original" (Ondishko, 1990: 37). I would guess that these

Type a **PL1** and **KND1**: Phrases which make up chords in a relatively traditional tonal idiom with a relatively rich and dense timbre, a resonating ending phase with a slow decay. At the beginning of these phrases one can hear a female voice that appears to excite the resonances present in the chord. For most of these sound events, the voice can only be recognized in the beginning of the sound, since after a short while it becomes "drowned" in the resonances it seems to excite.

Type b **PL2** and **KND2**: Phrases consisting of inharmonically related components in the mid-high register with asynchronous onset, which are sustained, slowly decaying, and which appear to be excited by a vocal sound. These phrases vaguely suggest the articulatory features (spectral profile) of speech when taken together.

Type c **BS2**: Phrases with a very dense and complex spectrum with no pitched components, where one can vaguely recognize traces of vocal articulation. In some cases these sounds resemble the resonant part of the sound of struck cymbals.

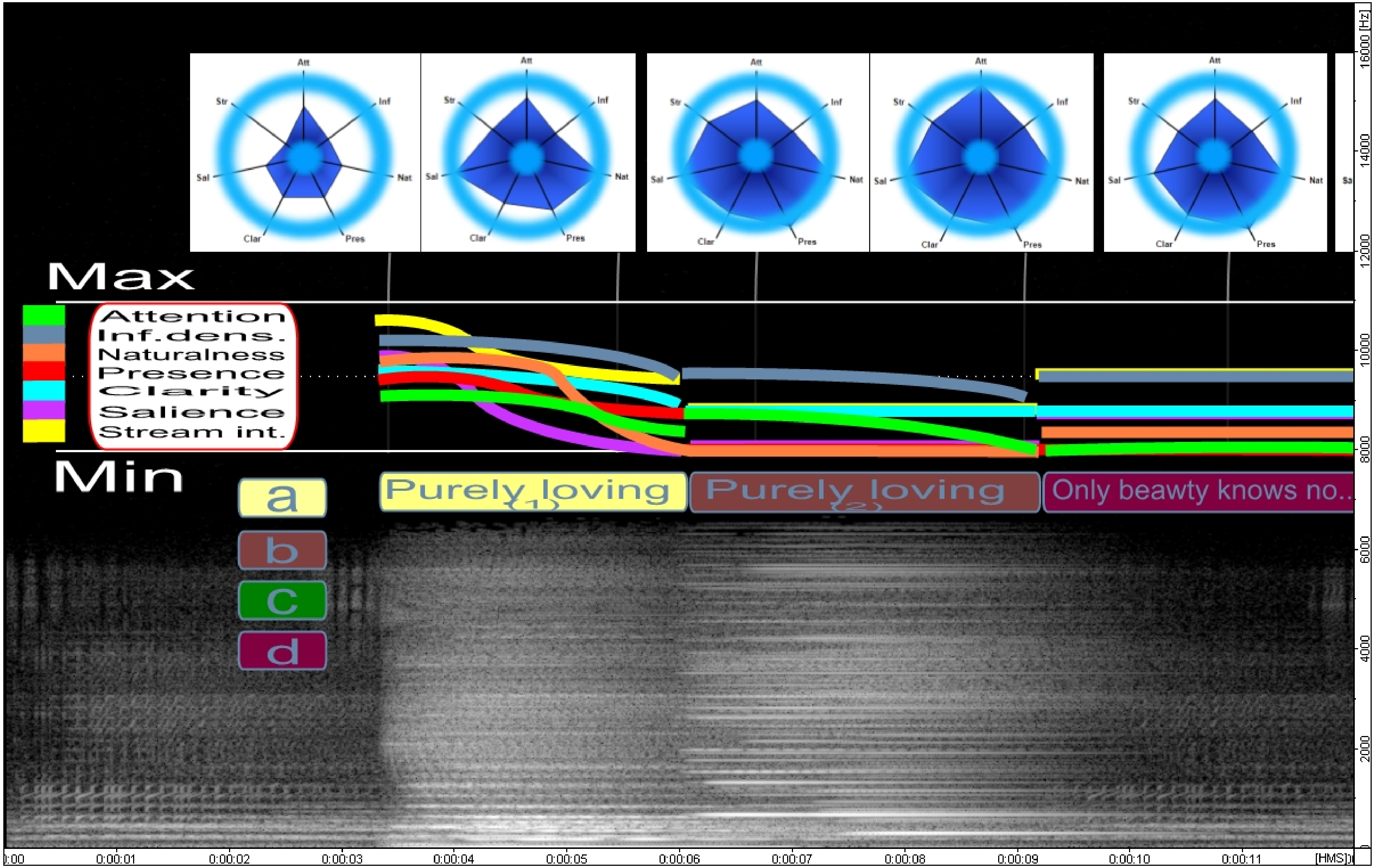
Type d Shorter or longer portions of vocal phrases with gradually decaying echoes with different degree of overlap, different delay time and different decay time. **BS3** as well as the chains of echoed decaying sound fragments that follows **KND2**[333] fall into this category. **BS1,** strictly speaking, has no echoes, but since it is similar to **BS3** except for the echoes, I will regard it as belonging to this type.

The evaluations are presented in **figure 12.3** and the corresponding acousmographe file on the accompanying CD-ROM. Here, I indicate the phrase type with colouring of the text boxes.[334] The noted diversity in the evaluations is also very evident from the large differences in shape and size of the axial representations. It must be noted, however, that by including **BS1** and **BS3** in the excerpt, which stand out in the movement as a whole as being the least processed, the impression of diversity is more prominent than when considering the fantasy from beginning to end.

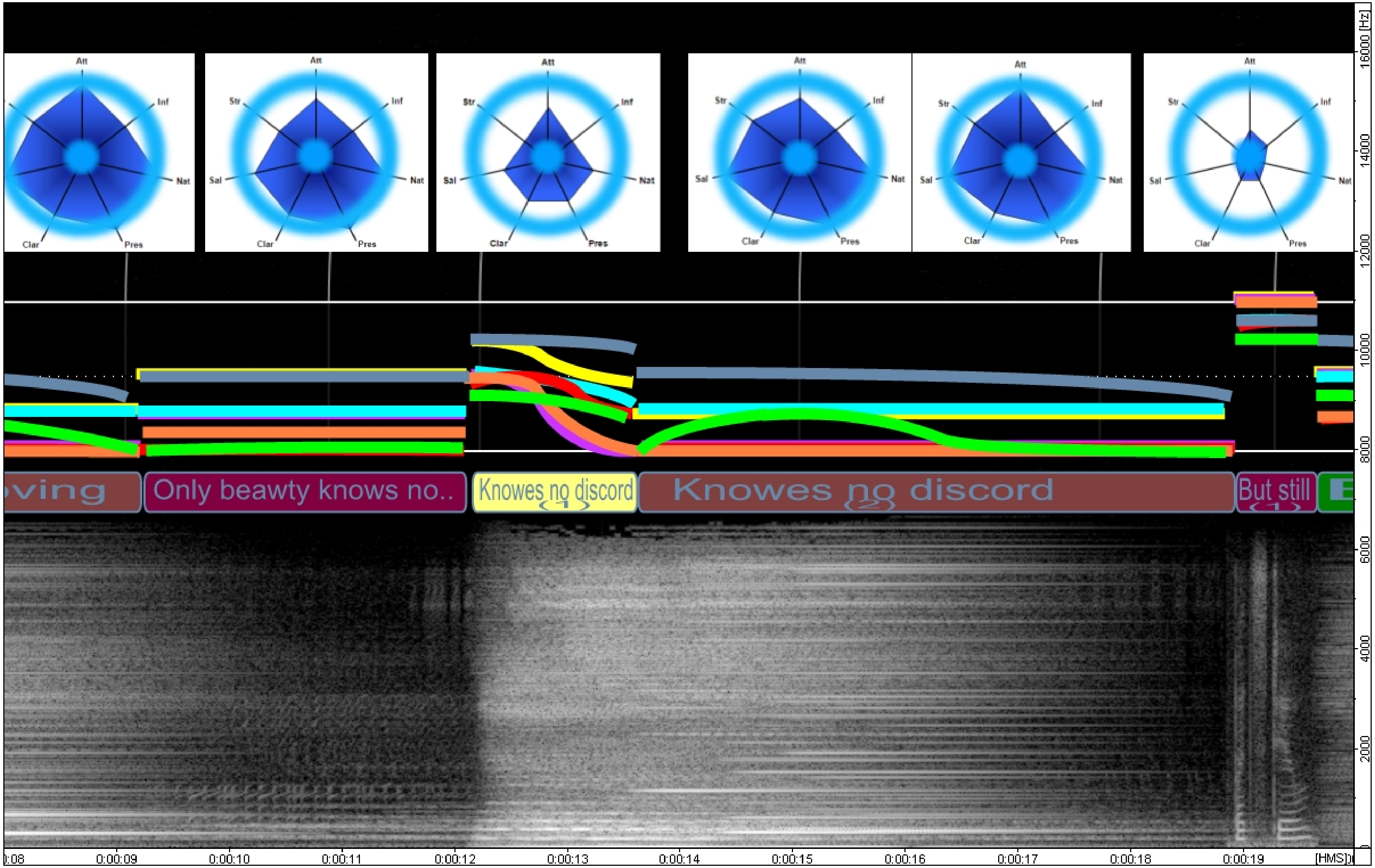
labels fit with phrase type c, a, b and d, respectively. There are also some sounds in the movement as a whole that don't seem to fit this scheme. For example, there is one phrase, "these dull" ("notes we sing"), at 1:54 that begins without definite pitch, but with a course granular quality. Gradually, though, this sound changes into one that is quite similar to type b. There is also another sound, at 2:12 which has some of the grittiness or coarseness of the previously mentioned phrase, only that there is also a chord-like structure that resonates and that all the partials don't seem to be excited at the onset. In that respect it resembles both sound types a and b.

[333] It is also possible to hear this kind of sound at the end of **KND1**, but since it is very low in loudness, I choose not to include it in the discussion.

[334] The placement of the rounded text boxes in the graphical representation is only approximate, however, since many of the phrases are wholly or partly superimposed upon each other.



Att
Inf
Nat
Pres
Clar
Sal
Str
Max
Min
Attention
Inf.dens.
Naturalness
Presence
Clarity
Salience
Stream int.
a
b
c
d
Purely loving (1)
Purely loving (2)
Only beawty knows no..
16000 [Hz]
14000
12000
10000
8000
6000
4000
2000
0
0:00
0:00:01
0:00:02
0:00:03
0:00:04
0:00:05
0:00:06
0:00:07
0:00:08
0:00:09
0:00:10
0:00:11
[HMS]


Att
Inf
Nat
Pres
Clar
Sal
Str
ving
Only beawty knows no..
Knowes no discord
(1)
Knowes no discord
(2)
But still
(1)
E
16000 [Hz]
14000
12000
10000
8000
6000
4000
2000
0
:08
0:00:09
0:00:10
0:00:11
0:00:12
0:00:13
0:00:14
0:00:15
0:00:16
0:00:17
0:00:18
0:00:19
[HMS]

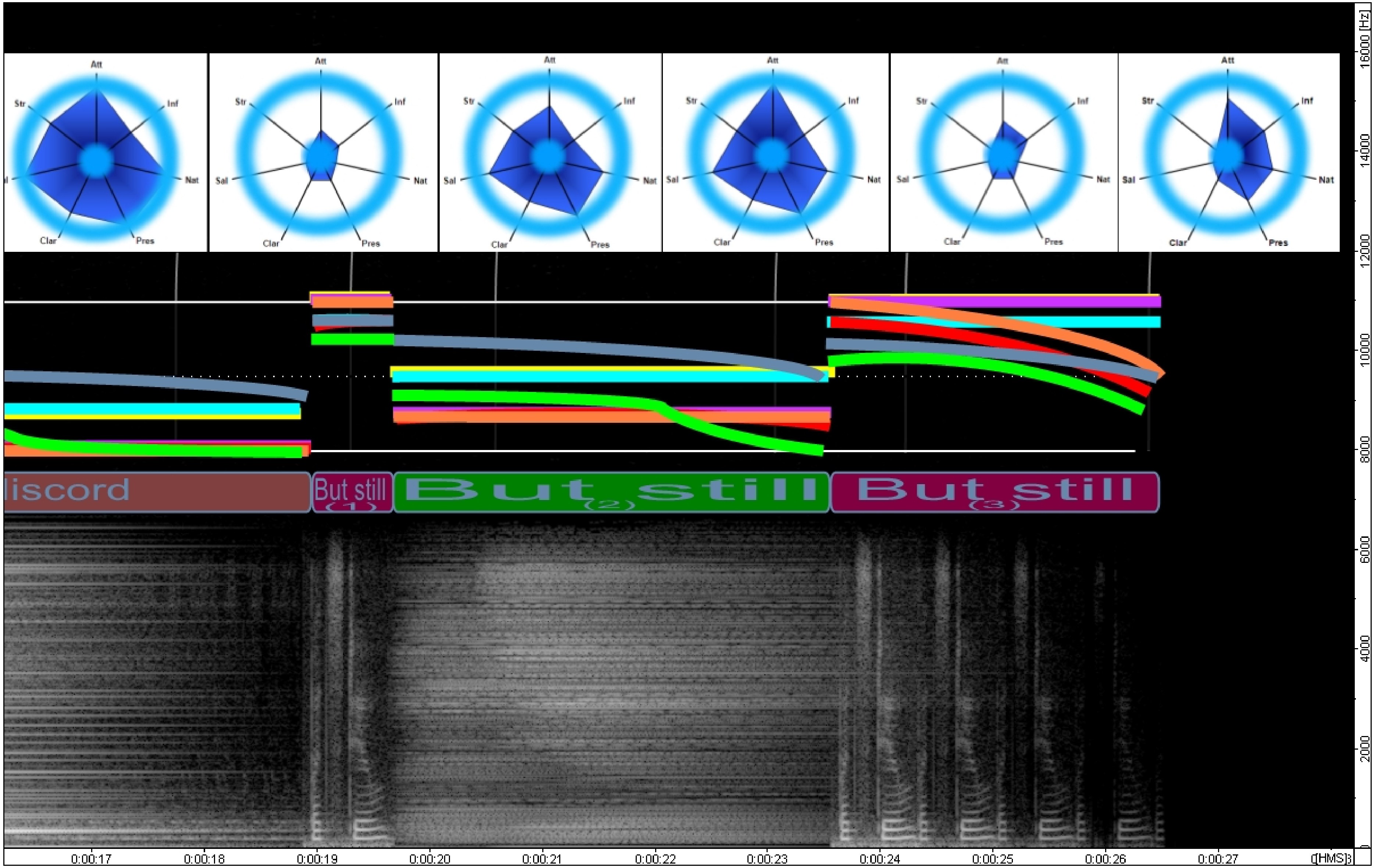


Figure 12.3: Graphical representation of evaluations of the seven premises in Lansky's *her reflection*. Note that some of the colored lines overlap so that they are not fully visible in the figure. For representation of the evaluation of single premises the reader should use the *acousmographe* software and the files provided on the accompanying CD-ROM.

12.3.3.2 Attention evaluation

In this fantasy, the degree to which verbal content can be recognized varies greatly, with many phrases having low salience (cf. *salience* factor) in parts or throughout the whole phrase. Somewhat surprisingly, I found that after having familiarized myself with the poem through repeated listenings of the whole piece, the difficulties that low salience created for the recognition of the lines of the poem in some of these phrases could be overcome. Actually, this made me try even harder to identify and follow them, almost as a kind of challenge. In other words, I directed my attention towards recognizing the often faint components of vocal articulation in the phrases, while the (non-vocal) resonant features represented a form of distraction in this task. My interpretation of this situation is that my focus here was directed towards the vocal domains, with varying degrees of distraction (what one, in a sense, tries to listen *through*) from the non-vocal domains, i.e. primarily the **SQS-domain**, hence indicating an evaluation between *intermediate* and *intermediate-minimal* according to table 5.1. Towards the ending of many of the phrases, however, the resonances in many cases cause masking or blurring so that there are no vocal features left to focus on, resulting in a decaying tendency for the evaluations, often ending at *minimal*.

In contrast, **BS1**, being unmanipulated, certainly stands out compared to the other phrases in the movement. In this very short phrase, the identity of the vocal persona and the details of vocal articulation that has only been hinted at earlier in this movement are here presented with unprecedented clarity. What catches my attention the most here is the identity features of the voice (**ID-domain**), since these features now are presented in a much less ambiguous manner than earlier, the vocal gestures (**VG-domain**) as well as the **LI/sem** level, which gets mildly accentuated by the sudden clarity and lack of manipulation. For me, the relatively large contrast between this phrase and the previous heavily manipulated phrases creates an emphasis that is also transferred to the semantic level. Thus, there is a situation in which attention is divided between **LI/sem** and vocal domains (cf. table 5.1), something which indicates an evaluation as *maximal-intermediate*. **BS3**, which is identical to **BS1,** only that it is followed by a chain of echoes, is evaluated as starting out a little lower and decaying towards the end of the phrase. This is because the words “but still” have now been heard two times already (cf. *novelty/change* factor), and that the exact repetition will also make it less interesting when it comes to vocal features. The obvious artificiality of the chain of echoes (**TCM-domain**) and the sound qualities therefore become more likely candidates for attracting attention as the echoes fade out.

12.3.3.3 Evaluation of information density

The mentioned resonances that characterize many of the phrases in this fantasy also have an effect on this premise, since the ending resonant phase of the sound is highly predictable – nothing new will happen except the fading out of the components that are excited. The result is that for most phrases evaluations tend to have a gradually decaying shape, as can be seen in the graphical representations.

In this fantasy, there are aspects that contribute to increasing as well as decreasing information density. As for the former, this fantasy clearly displays a much greater *complexity* than the preceding fantasies (cf. section 6.3). First of all, the similarity from phrase to phrase is low, with some phrases even being markedly contrasted. Even if there are some patterns here, in that all the phrase types are somehow characterized by resonances or reflections and that the phrases within the four phrase types are relatively similar, it is generally difficult to predict the features of the next phrase. This is also related to the lack of regularity regarding the ordering of the different phrase types. In other words, in terms of variation between phrases there appears to be more *elements/dimensions* and less *regularity* (cf. section 6.4) than in the previous movements.

There are also issues that contribute to reducing the *complexity* factor and thereby information density in *her reflection*. Firstly, there are issues of repetition. In addition to having the same verbal material as in the previous two fantasies (cf. *complexity* factor – familiar structures), many of the lines of the poem are repeated several times, although not always in a condition that is easily perceived.

If we consider the features of the phrases more locally, there are relatively large variations in complexity, because some of the more heavily processed phrases convey only relatively small amounts of information, at least related to vocal features (cf. *complexity* factor – number of elements/dimensions). Especially for the phrases of type b and c, only a minimum of vocal features can be inferred from the phrase. For **PL1**, for instance, only a few of the phonemes can be recognized, and only marginal information from the other vocal domains can be gathered from the phrase; most of the articulatory as well as all of the phonatory qualities are not present. Moreover, in the type a phrases, much information is missing due to a considerable amount of masking in the last portion of the phrase. In contrast, other phrases, in particular **BS1** and **BS3**, are closer to regular speech and thereby carry much more information.

On the whole, I experience that the tendency in this excerpt is that the information density is lower than in the previous two excerpts, mainly due to the lowered complexity caused by the reduced number of elements/dimensions conveyed by the phrases locally. Considering the higher complexity on the structural level, I would evaluate the phrases having the lowest information density as being *intermediate* or slightly lower. Clearly, **BS1** and **BS3** defy this tendency, and their evaluations therefore approach the *maximal*.

12.3.3.4 Naturalness evaluation

As in *her presence*, the experienced naturalness for the phrases in this movement differs greatly between the four phrase types, ranging from the *maximal* to the *minimal*. For most of these phrases, the voice is accompanied or replaced by what I experience as resonances that are excited by the voice but external to it, perhaps residing in the environment/room or some objects within it. Accordingly, one could perhaps interpret this as reduced naturalness on the part of the **SE-domain**. Still, since for the phrases of types b and c the voice can only be heard "through" these resonances, I find it more appropriate to include *all* the components of the phrases in the evaluation, even if this is difficult to relate to the factors discussed in section 7.2:

- **Type a:** Both **PL1** and **KND1** are gradually "drowned" by resonances, which I partly hear as being some kind of heavy room reverberation and partly a low pitched resonance, hence resulting in gradually decaying naturalness for these phrases.[335]
- **Type b:** For these phrases, there is a very low degree of resemblance to the voice. Here, the only features that faintly show some resemblance are the articulatory features, which are only barely recognizable. I presume that it is only when one knows beforehand that the linguistic content is "purely loving" that the verbal content is recognizable at all. Without the link to linguistic content known beforehand, it would probably be difficult to link it to vocal production at all. Therefore I have judged it to be *minimal*.

[335] Since the beginning of the sound invites an evaluation of the vocal sound independent from any resonances, whereas the middle part and the ending make it difficult to disregard the resonances as something fully external to the voice, the evaluation process faces an ambiguity regarding whether one should consider only the voice independently from the resonances or as part of them. A representation of such an ambiguity would have to display two different evaluations. To avoid packing the graphical representation with too much information, I have chosen only to display what I feel is the most plausible interpretation of this situation.

- **Type c:** The resemblance to human vocal production for **BS2** is higher than in the previously discussed phrases, almost approaching whispering. The high degree of temporal blurring, however, has made me evaluate this phrase as *minimal-intermediate*.
- **Type d:** The phrase **BS1** appears as an unmanipulated female voice, and therefore, I have evaluated it as being *maximal*. For **BS3**, on the other hand, a chain of repeated versions of the phrase decaying in loudness soon makes it evident that this is a manipulated sequence, thus giving rise to an evaluation that starts at the *maximal*, but decays towards the *intermediate*. In contrast, for **OBKN** the high degree of overlap and long decay phase blur the articulatory evaluation of the fragments so that they become indistinct and therefore appear much more processed. These sounds are therefore evaluated as close to the *minimal*.

The lack of vocal features in several of these phrase types also restricts the range of factors that may have affected the evaluations. For phrase types b and c, it is only the *articulatory features* that can be associated with vocal sound and that therefore directly can be linked to the evaluation. And, clearly the articulatory features for these phrases are highly blurred and indistinct, especially for the type b phrases. For the type a phrases, both the *articulatory* and the *phonatory features* get progressively blurred and masked during the course of the sounds. As for the experience of reduced naturalness for the chain of decaying echoes for **OBKN** and **BS3**, I hear them as a type of *technological artifacts* associated with artificial delay.

12.3.3.5 Presence evaluation

In the excerpt chosen from *her reflection*, one can experience the great contrasts in presence – from the *minimal* and almost all the way to the *maximal* for the four phrase types:

- **Type a**: As in the naturalness evaluation, experienced presence for **PL1** and **KND1** is around *intermediate* for the first part of the sound. Since presence has a tendency to persist despite that the voice is not actually present, it does not drop as much as the naturalness evaluation; only to the *intermediate-minimal*.[336]

[336] See the discussion of contextual linkage in section 9.2.

- **Type b**: The faint resemblance to spoken articulation creates a situation where the sound stands as an almost ghostly and illusory projection, with only a vague imprint left of what was once a voice. Hence, it is evaluated as *minimal*.
- **Type c**: For **BS2** the presence is markedly higher than in the type b sounds. There is still a situation where the sound one hears is primarily a resonance created by the voice, but in this case the resonance is richer and denser, so that it to some degree resembles the quality of a whispered voice. For that reason I have evaluated this phrase as being *intermediate-minimal* in presence.
- **Type d:** The unmanipulated phrase, **BS1**, creates for a very short moment the sensation of something with close to *maximal* presence. The phrase is very short, so the sensation is almost over as soon as it is sensed, and that is the primary reason that it doesn't reach a maximal level.[337] Some seconds after, however, the phrase is repeated, this time with a tail of decaying delayed versions of itself – artificial echoes. That **BS3** is an exact repetition of the first, does not affect the sense of presence to a large extent in my experience, since the intervening phrase makes it is hard to tell whether this is a mechanical repetition or a man-made one. The decaying echoes of **OBKN**, however, are clearly artificial, in addition to being very low in loudness, so that they appear to be at a distance. Therefore, they are evaluated as *minimal*.

The factor contributing most in the evaluation of the phrases from *her reflection* is probably *explicit transformation*, with many phrases appearing highly processed. The type of processing, where one experiences many of the phrases as faint imprints of vocal production, thereby also imposes a sense of *spatial distance*. For the **PL1** and **BS3** phrases, the exact repetitions of the artificial echoes indicate that *temporal disruptions* can affect the evaluation.

12.3.3.6 Evaluation of clarity of meaning

The phrases in *her reflection* differ greatly for many of the factors that are relevant for the evaluation of this premise. The following issues have had most to say for my judgment of the chosen phrases from *her reflection*:

1) **LI-domain clarity**: For **PL2** and **KND2** the bottom-up clarity for phonemes is very low, with only some phonemes being unambiguously recognizable. For **BS1** and **BS3**

[337] Cf. the factor of duration. See section 9.2.

clarity is very high with all phonemes recognizable, whereas for **PL1**, **KND1** and **BS2** it is something in between. These differences in bottom-up clarity can be clearly seen in the graphical representation. The reduced bottom-up clarity is partially counteracted by repetition, however: All of the phrases in this fantasy have been heard in the previous two movements in versions that were largely comprehensible. Moreover, most of the phrases in this movement are repeated once or several times, often using several phrase types so that the least intelligible might benefit from being repeated in more intelligible versions. This is one reason for not giving the phrases with the lowest bottom-up clarity, **PL2** and **KND2**, minimal evaluation.

2) **Contextual specificity:** The phrases in this movement also differ greatly as to the degree to which identity features can be attributed to them. It is clearly most difficult to attribute identity features to the type b phrases, since the likeness to a voice here is minimal. For the **BS2** phrase, features related to language might be recognized, but gender and age are still ambiguous. For the type b phrases, it is possible to recognize that it is an adult woman, but since bottom-up clarity is lower for the ending of these phrases, features like socio-cultural background might be more ambiguous. The most important identity features for **BS1** and **BS3**, however, are both unambiguous and relatively specific.

For both of these evaluation criteria, the *salience of relevant cues* factor is probably the one which has affected the evaluation the most in the negative direction, making both bottom-up clarity of the **LI-domain** and specificity of the **ID-domain** low for many of the phrases. *Repetition* affected the former of the two in a positive direction.

12.3.3.7 Evaluation of feature salience

The nature of many of the sounds in *her reflection* poses some challenges to the evaluation of salience, since this premise was defined according to the relationship between the vocal phrases and “other sounds”. In this movement, this distinction is far from clear because in many cases, what we hear does not so much appear to be vocal sound in itself, but rather sound taking the form of some kind of reverberation, reflection or resonance. Thereby, this implies some form of causal relationship to the voice which is less direct. This is particularly the case with the phrases of types b and c. Therefore, I have in most cases had to focus instead on how salient the vocal *features* are, i.e. the features that come through as being somehow vocal, compared to those features which point in another direction:

- **Type a:** At the onset of **PL1** and **KND1**, partial masking gives an evaluation as *intermediate* (**KND1**) or a little above (**PL1**), but as the masking increases towards the ending of the phrases, the evaluation drops down to the *minimal*.[338]
- **Type b:** For **PL2** and **KND2**, the minimum of vocal features that can be recognized (only a few vowels plus perhaps the [s] in "discord" in **KND2**) makes both these phrases being evaluated as *minimal*.
- **Type c:** For **BS2**, I have evaluated salience as *intermediate-minimal* due to the relatively large amount of bi-directional self masking, and the cymbal-like resonances which dominate the sound and clearly have no relation to vocal production.
- **Type d: BS1** and **BS3** are both evaluated as *maximal* in salience. **OBKN** is both low in loudness and have a relatively high degree of self masking due to the overlapping fragments, and is therefore evaluated as *intermediate-minimal*.

From these points we can see how, in particular, the *condition of the vocal features* and the *temporal* and *simultaneous masking* factors have played a part in the evaluations.

12.3.3.8 Evaluation of stream integration

Similarly to the other premises, the evaluation displays great variety for the stream integration premise for the different phrase types:

- **Type a:** At the onset of these phrases, there is a relatively clear separation between a vocal stream and a non-vocal stream (the component with a low and stable pitch), which results in relatively little ambiguity regarding source coherence level, especially for **PL1**, which is therefore judged as somewhat higher than **KND1**. During the course of these phrases, however, the increasing blending of the vocal sound with its reverberation or resonances fuses them into one single stream. Since this stream is ambiguous in terms of source coherence level (the stream contains something which is external to the vocal sound), it is evaluated as descending to *intermediate*.

[338] This can be seen as an extreme form of self-masking (cf. *masking* factor, section 11.1.2).

- **Type b:** These phrases are only weakly integrated, with strong cues for segregating the individual spectral components of the phrases. Due to asynchronous onsets and inharmonic relationships between the components, the pitched components are in many cases separately discernible as individual streams. Nevertheless, the modulation of the spectral envelope of all the components together resembling speech articulation, matches familiar speech patterns. And, the similarity between the individual components still provides some cues for integrating them at the source coherence level of the vocal utterance. Therefore, the phrases are evaluated as *intermediate-minimal*.
- **Type c:** Due to the dense distribution of components in **BS2**, it is more difficult to single out and perceive components as separate streams. **BS2** is therefore more integrated than the type b phrases. Consequently, I have evaluated it as *intermediate*.
- **Type d:** Being unaccompanied, unmanipulated and with very little reverberation, **BS1** and **BS3** are also well integrated with an unambiguous source, and are therefore judged to be *maximal*. The decaying echoes of **OBKN** are less integrated in that they appear to "wobble" from side to side between the channels.

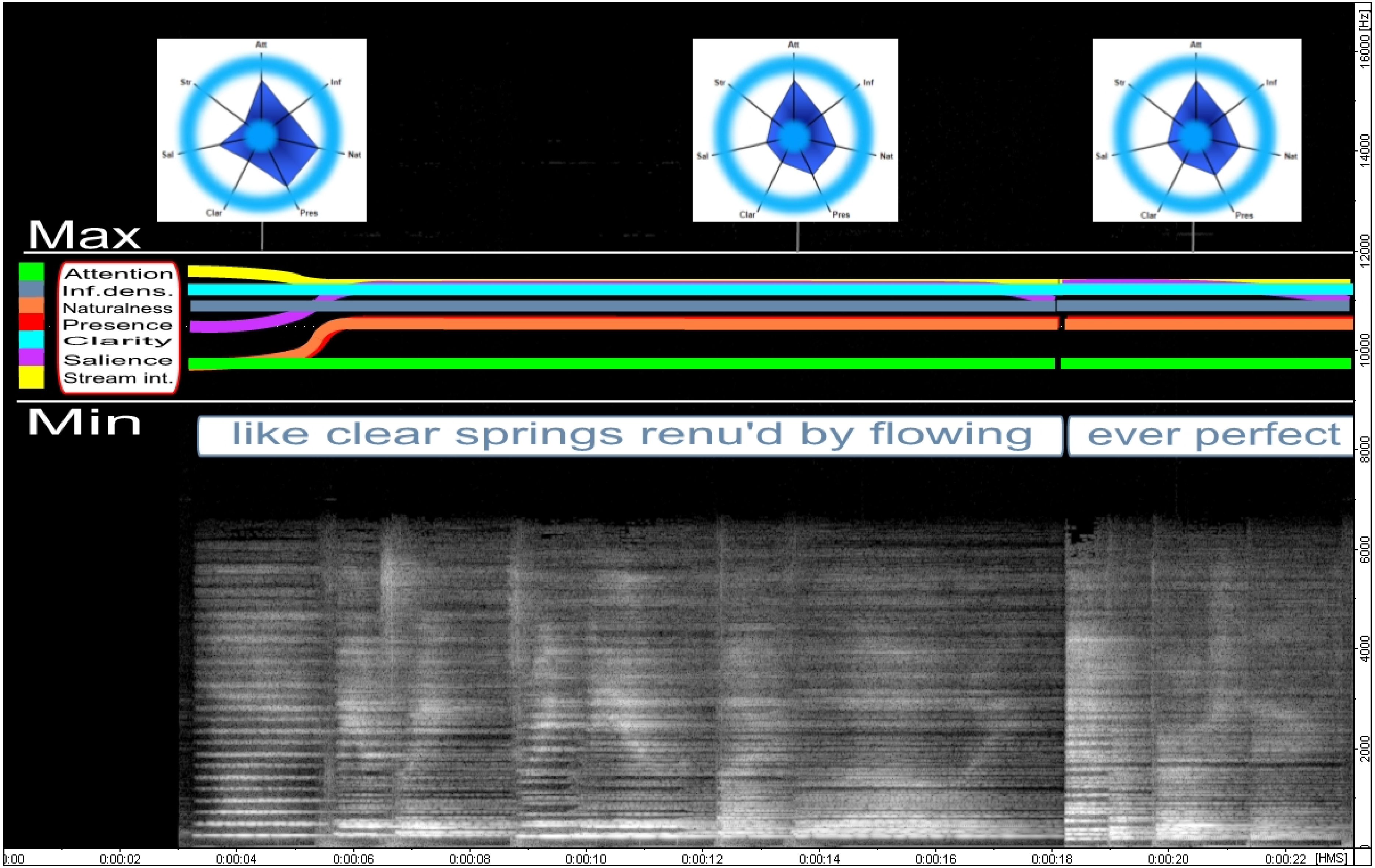
From this, we can infer that the factors *harmonicity*, *synchronicity of onsets and endings*, *similarity of spatial location*, *feature similarity/proximity/continuity*, *modulation coherence* and *familiarity with linguistic structures* have affected the evaluation of the phrases in the excerpt.

12.3.4 Evaluation of *her song*

12.3.4.1 Over-all description and evaluation

Giving the impression of a choir like setting, this fantasy is perhaps the one of all six that resembles a traditional conception of music most closely.[339] Nevertheless, it has several similarities with the two first fantasies. Firstly, intelligibility is comparable in all three fantasies, with largely well comprehensible phrases. Secondly, both the first and fourth fantasies, as well as many of the phrases in the second fantasy, all appear to involve more than one voice in synchronous articulation. Thirdly, *her presence* and *her song* share the

[339] Denise Ondishko has described this fantasy in the following manner: "eight women's voices sing the poem in a popular music style reminiscent of the Andrew Sisters" (Ondishko, 1990: 17). An alternative interpretation is that there is only one single voice articulating harmonies instead of monophonic notes.



Att
Inf
Nat
Pres
Clar
Sal
Str
Att
Inf
Nat
Pres
Clar
Sal
Str
Att
Inf
Nat
Pres
Clar
Sal
Str
Max
Attention
Inf.dens.
Naturalness
Presence
Clarity
Salience
Stream int.
Min
like clear springs renu'd by flowing
ever perfect
16000 [Hz]
14000
12000
10000
8000
6000
4000
2000
0
:00
0:00:02
0:00:04
0:00:06
0:00:08
0:00:10
0:00:12
0:00:14
0:00:16
0:00:18
0:00:20
0:00:22
[HMS]

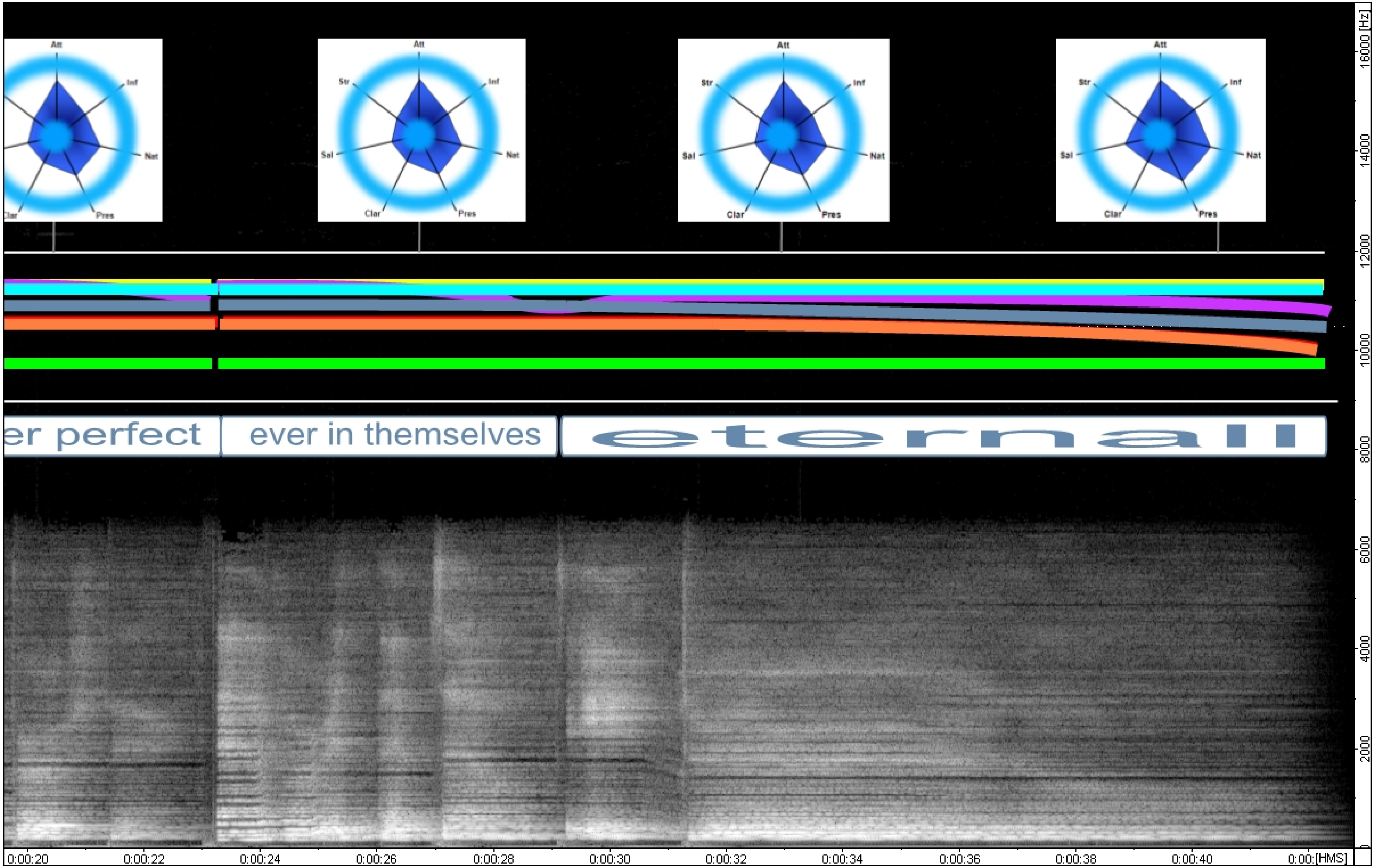


Figure 12.4: Graphical representation of evaluations of the seven premises in Lansky's *her song*. Note that some of the colored lines overlap so that they are not fully visible in the figure. For representation of the evaluation of single premises the reader should use the *acousmographe* software and the files provided on the accompanying CD-ROM.

rooting in Western tonality, even if this is much more pervasive and consistent in this movement.

For this fantasy, I have selected the four phrases at the end of the fantasy as a sound example (**sound example 12.4**, Lansky, 1994a, *her voice*, 2:25-3:05), for which the evaluation of the premises is presented in **figure 12.4**. Owing to this similarity, however, I will not discuss all the features of this excerpt in detail, but focus mostly on the features that are particular for the movement in an overall evaluation.

The redundancies caused by consistency of the choir-like configuration of the voices along with having the same text presented three times over in the previous movements, affect the evaluation of several of the premises. For the *focus of attention* premise, text repetition represents little *novelty/change* for **LI/sem**. Along with the variations in harmonies and the articulation of the phrases in varying manners, the tendency is therefore that it is the **SQS**- and the **VG-domain**s that receive the most attention, with **LI/sem** mostly being in the background.[340] The evaluation of this premise is therefore (cf. table 5.1) *intermediate-minimal*. The mentioned redundancy and consistency also affect the evaluation of *information density*, making the phrases relatively predictable for many aspects. However, the factor that probably has the strongest effect on the evaluation of information density is *speed*, since one of the things that are characteristic for this fantasy is the slow articulation of the verbal phrases – and one can actually find the phrases with the slowest articulation in this movement: The phrase **E**, with only three syllables and one word ("Eternall"), is 13 seconds long, and because the syllables get progressively longer, I have given it a gradually declining evaluation. Taken together, I have evaluated the phrases just above *intermediate***.** Finally, the redundancy of the text also affects the *clarity of meaning premise*. The differences between this movement and *her voice* regarding **LI-domain clarity** and **contextual specificity** are relatively small: Whereas bottom-up clarity perhaps is somewhat lower, the top-down redundancy caused by text repetition compensates so that the evaluation is similar; *maximal-intermediate*.

The evaluation of *naturalness* also resembles that of *her voice* and *her presence*, but the phrases from this movement are somewhat less natural with an evaluation as *intermediate*. This is because the "singing" voices lack the natural pitch inflections that one can observe for

[340] I would like to note is that even if **LI/sem** had little potential for attracting attention, there are certain phrases where this potential seems to increase, as there are some phrases in which there is some *interplay* between the semantic content of the phrases and the musical properties of the phrase. For example, the word "discord" in the phrase "knows no discord", which is not included in the excerpt, is presented as a dense cluster of dissonant tones.

real singers – preparation, overshoot, etc. (cf. *precision* factor).[341] Moreover, the phonatory component (the voiced parts) has a relatively marked "ringing" synthetic quality with a relatively rich timbre, with marks of a type of chorus effect, similar to what was heard in *her presence*, but a lot more noticeable in this case.[342] The word "like" in the phrase **LCSRBF** has a particularly artificial flavour, due to the octave distance between the voices, and it is therefore evaluated as being *intermediate-minimal*.

The evaluation of presence in *her song* is mostly affected by the factor *implicit/explicit transformation*, and since the phrases in this movement appears somewhat more transformed than *her presence*, they are evaluated as *intermediate*.

The evaluation of feature salience is also comparable to *her voice*. Here, one can notice a slight masking of timbral features of the sustained vowels due to the relatively strong "chorus" effect that prevents the vocal features from being maximally salient. In addition, the word "like" (**LCSRBF**) stands out as less salient due to the blurring of articulatory features caused by the octave interval between voices. Lastly, many phrases have relatively sudden onsets but gradual and slow endings, thereby causing a slight descent in the evaluations towards the ending of the phrases.

Regarding the stream integration premise, *her song* resembles *her voice* in that there are a number of voices articulating the phrases of the poem on different pitches. Hence, one can find a similar kind of ambiguity regarding the source coherence level; on one side one can hear several voices articulating something in synchrony, on the other side one can hear the phrases as produced by one single voice articulating the verbal phrases with chords instead of a single pitch. The integrating and segregating forces are largely the same, only with minor differences, and the resulting experience is that the voices fuse a little less in this movement. Taken together, stream integration is evaluated as *maximal-intermediate*, i.e. a little above what was the case for the phrases in the excerpt from *her voice*.

[341] For many of the phrases which are not included in the sound example, the precision of the glissandi is also unnatural. E.g. on the glissando on the "-ly" syllable in the phrase" sing thy smoothly" and the syllable "for" in the phrase "for helps to grace them", the voices move in perfect parallel in a way that is markedly unnatural.
[342] Moreover, when the harmonies have very many voices and get particularly complex, the timbre starts to lose some of the vocal qualities that are associated with pitched phonation. This is the case with phrases in parts of the fantasy not included in the excerpt, e.g. in the syllable "-cord" in the phrase "knows no discord".

12.3.5 Evaluation of *her ritual*

12.3.5.1 Over-all description

This movement contains a lot of sound that one might experience as very far from vocal production, as was the case with *her reflection*. Similarly to *her reflection*, one can find phrases where complex timbres resonate, and where one lacks the sense of pitch for the phonatory component. As a whole, it is the sequences of short percussive sounds that dominate the movement.[343] The variations in spectrum, rate and duration of the elements during the course of these sequences also appear to be correlated, so that the shifts in spectral envelope follow changes in duration and presentation rate, resembling the effects of gradual playback speed manipulations. This is especially evident when the rate of presentation is low, something which make the spectral changes very easily noticeable. It is mostly at these moments that it is possible to recognize the vocal quality of the single elements, and sometimes also to identify the verbal content – usually one single syllable containing at least one consonant. Therefore, these sequences appear as something in between what Smalley calls *unitary* and *continuous transformations*, which vacillates back and forth between a source-bonded base identity and consequent identity with no clear source-cause relationship (Smalley, 1993: 286). The percussive sequences also vary as to the degree in which they excite chordlike resonances.

The other vocal phrases in this fantasy are superimposed on top of these percussive sequences, and mostly they appear as quasi-whispered, noise excited articulations. Many of these phrases also appear to excite timbrally rich resonances in the form of chords or harmonies. As with the resonances in *her reflection*, these also appear to be *external* to the voice.

Taken together, one can distinguish between three main phrase types in this movement:

Type a:	Percussive sequences of short sounds with varying speed and spectral envelope, sometimes recognized as syllables or words
Type b:	Noise excited vocal phrases exciting resonant chords
Type c:	Noise excited vocal phrases

[343] Ondishko notes that the title in this fantasy is therefore well chosen: “This movement is highly repetitious at a microscopic level. Almost every word is echoed again and again before fading away. In a ritual, long standing, traditional – even perfunctory – movements are carried out. Here routine reoccurrences of each sound is carried out” (Ondishko, 1990: 40).

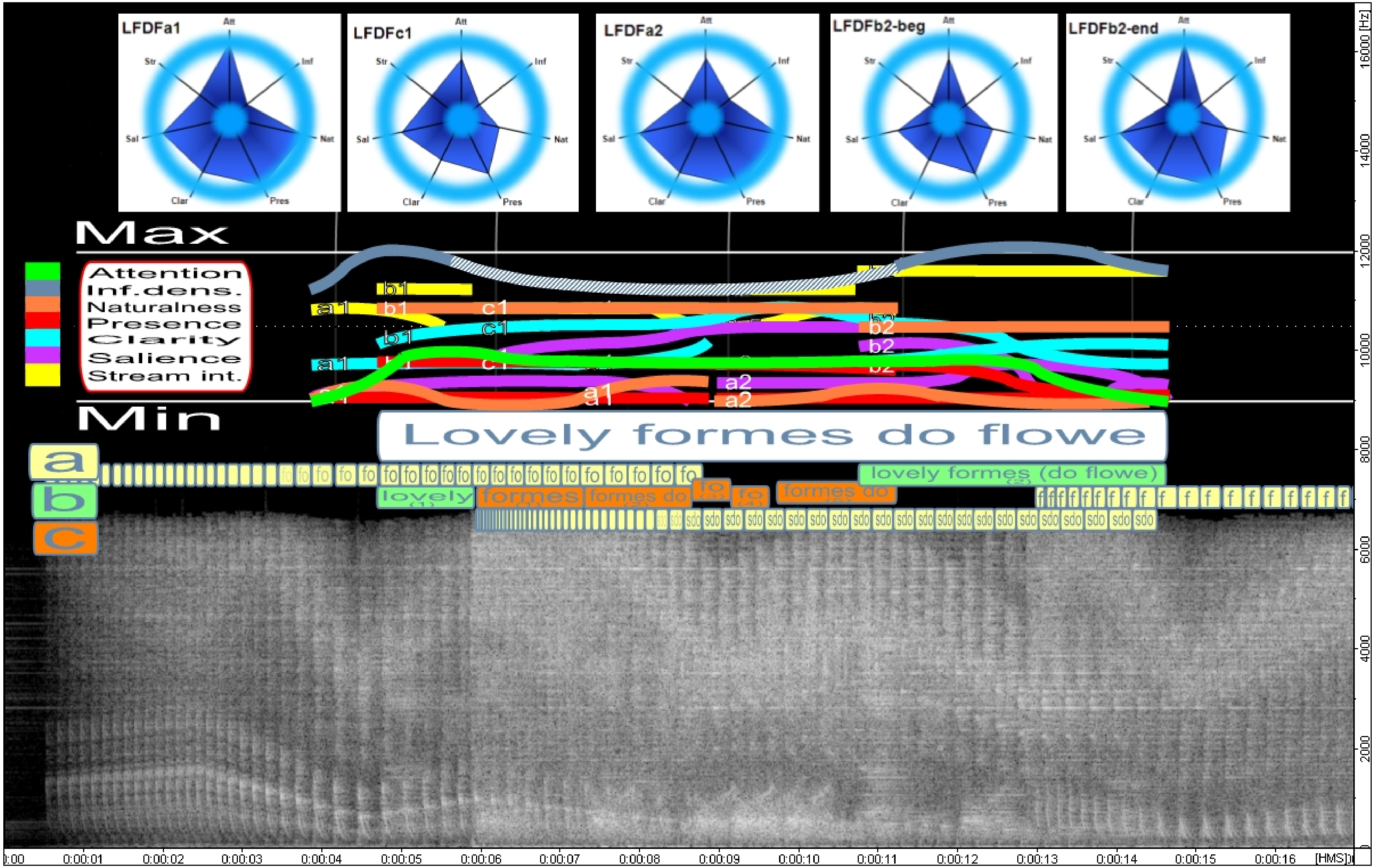
These types of sound are all included in the excerpt I have chosen from this fantasy, which can be heard in **sound example 12.5** (Lansky, 1994a, *her ritual*, 2:00-2:43). The evaluations of this excerpt are shown in **figure 12.5** and the accompanying acousmographe files on the CD-ROM. One can see here that the complexity in this fantasy is high, with many overlapping phrases, in contrast to the other movements, where there was little or no overlap. Owing to the fact that the type a phrases vacillate between voice and non-voice, and that my framework does not transcend the boundary into non-voice, only the percussive sequences *within* the voice category are evaluated.

12.3.5.2 Attention evaluation

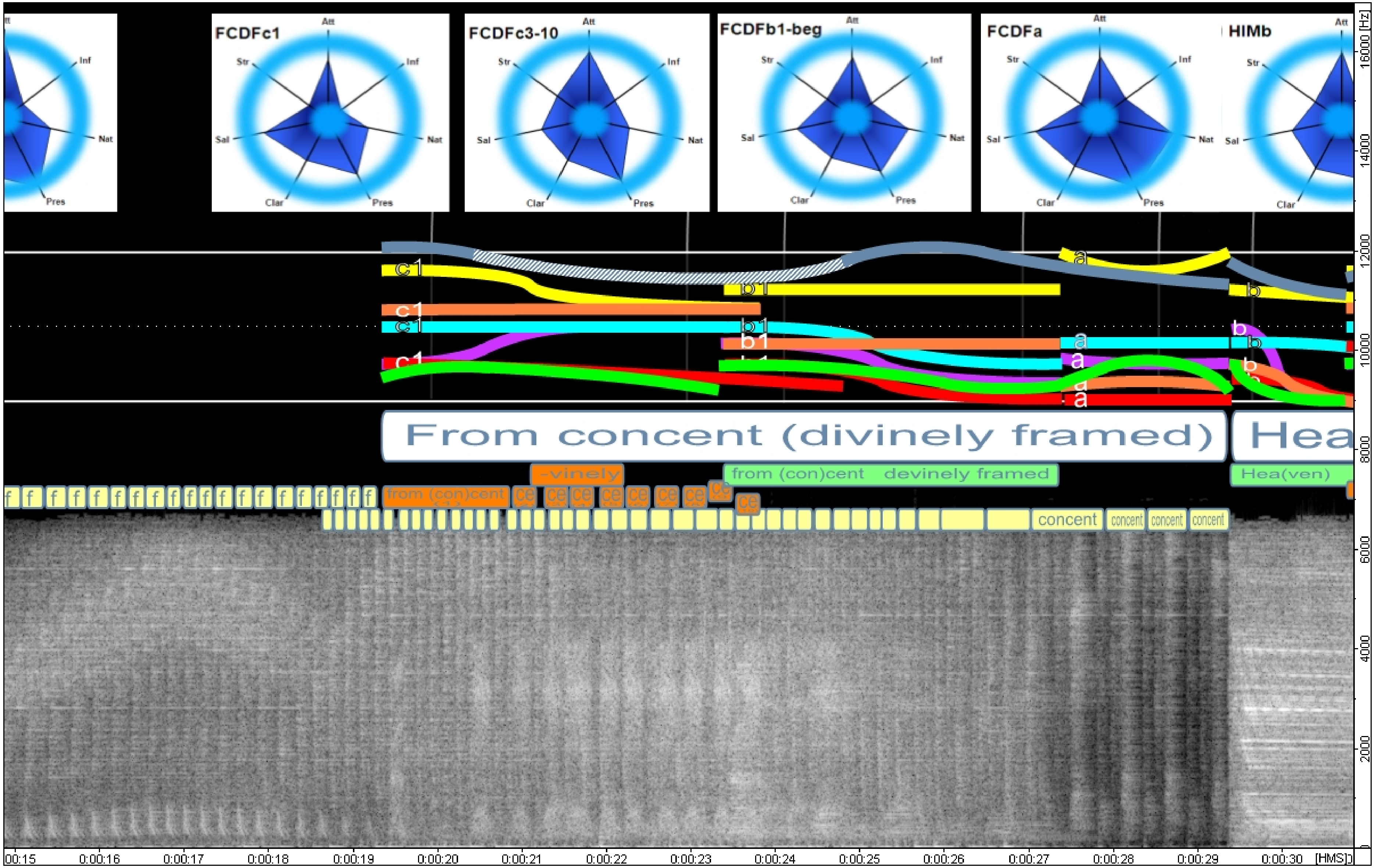
As in most parts of the other fantasies, the repetition of the text from the earlier movements prevents **LI/sem** from attracting too much attention. Instead, the many onsets of short words or phrases of the type b and c phrases, often shifting back and forth between the channels, have a tendency of drawing attention to themselves, in particular the *articulation* of the words (**VG-domain**). But, for the type c phrases I occasionally also attended to the different harmonic flavours (**SQS-domain**) of these sounds.

Due to the temporal overlapping of many of the phrases, many of the phrase onsets are simultaneously *distractions* relative to the previous phrase. Thereby, I had a situation where vocal domains in one phrase represented a distraction of the vocal domains in other phrases, something which gave an evaluation of *intermediate-minimal*, according to table 5.1. For instance, this is the case in **LFDF** (0:08-0:12 of the excerpt), where at many points there are several vocal phrases present at a time in addition to the percussive sequences. Moreover, the difference in spatial localisation from each phrase (i.e. in this case word) to the next makes the onsets increasingly salient (cf. *salience* factor).

Another recurrent situation was when the percussive sequences (phrase type a) approached intelligible words. At these points, attention was shared between vocal articulation (**VG-domain**) and both the **SQS-** and the **TCM-domains**, also giving an evaluation as *intermediate-minimal*. This is the case for **FCDFa** at about 0:27-0:29 in the example, where the percussive sequence for a short moment approaches vocal articulation with partly intelligible content. When the percussive sequences speed up, however, the remnants of vocal articulation are soon not recognizable, and all that is left to attend to is



LFDFa1
LFDFc1
LFDFa2
LFDFb2-beg
LFDFb2-end
Att
Inf
Nat
Pres
Clar
Sal
Str
Max
Min
Attention
Inf.dens.
Naturalness
Presence
Clarity
Salience
Stream int.
a1
b1
c1
a2
b2
Lovely formes do flowe
a
b
c
fo
lovely formes (do flowe)
lovely
formes
formes do
sdo
f
16000 [Hz]
14000
12000
10000
8000
6000
4000
2000
0
0:00
0:00:01
0:00:02
0:00:03
0:00:04
0:00:05
0:00:06
0:00:07
0:00:08
0:00:09
0:00:10
0:00:11
0:00:12
0:00:13
0:00:14
0:00:15
0:00:16
[HMS]


FCDFc1
FCDFc3-10
FCDFb1-beg
FCDFa
HIMb
Att
Inf
Nat
Pres
Clar
Sal
Str
c1
b1
a
b
From concent (divinely framed)
Hea
-vinely
from (con)cent devinely framed
Hea(ven)
from (con)cent
ce
concent
16000 [Hz]
14000
12000
10000
8000
6000
4000
2000
00:15
0:00:16
0:00:17
0:00:18
0:00:19
0:00:20
0:00:21
0:00:22
0:00:23
0:00:24
0:00:25
0:00:26
0:00:27
0:00:28
0:00:29
0:00:30
[HMS]

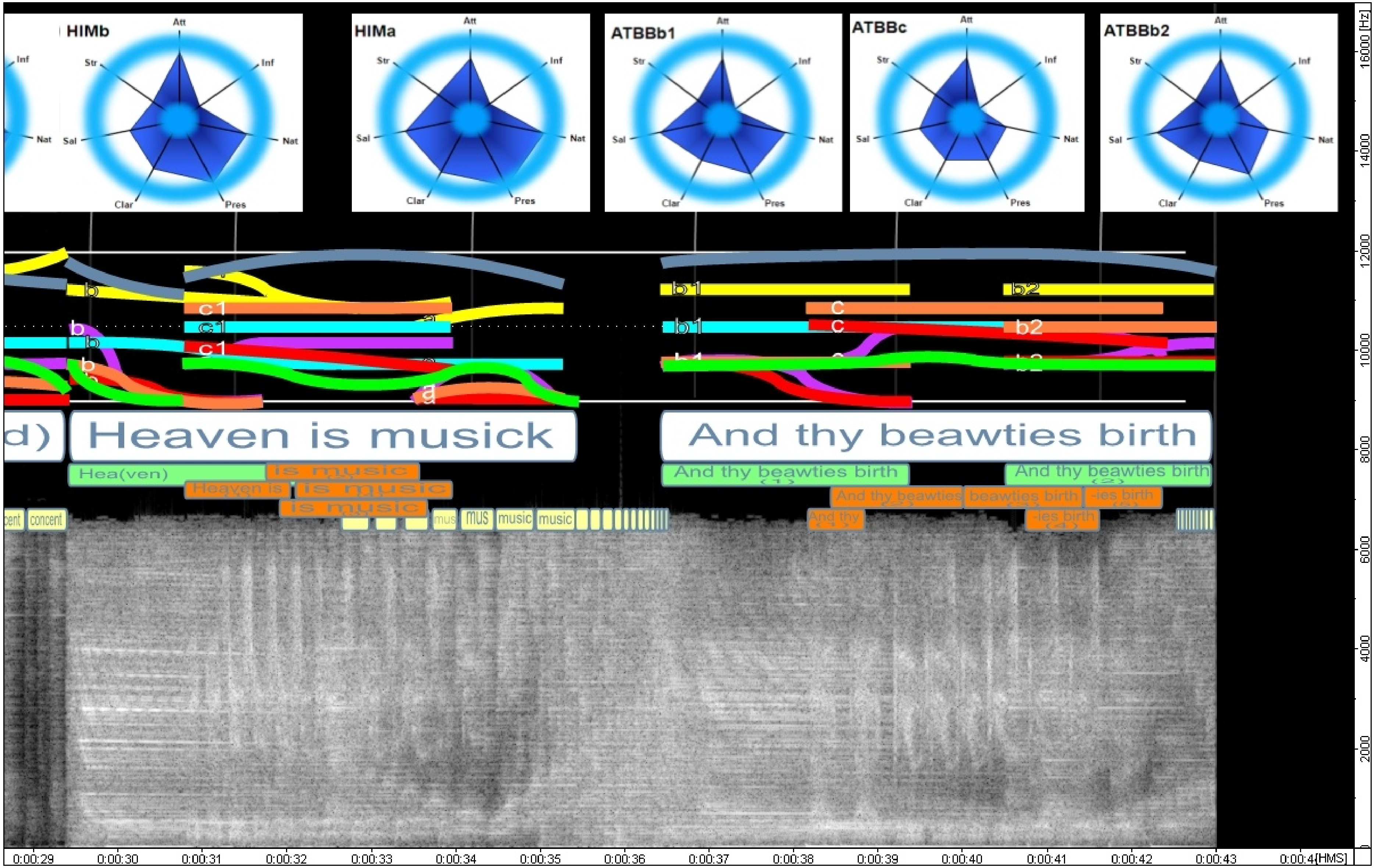


Figure 12.5: Graphical representation of evaluations of the seven premises in Lansky's *her ritual*. Note that some of the colored lines overlap so that they are not fully visible in the figure. For representation of the evaluation of single premises the reader should use the *acousmographe* software and the files provided on the accompanying CD-ROM.

the non-vocal domains, hence implying *minimal* evaluation. Where the sequences lose their vocal quality, it is beyond representation within this framework, and consequently no evaluation is given.

In addition to the mentioned issues of distraction, the factor of *salience* affects the evaluation for this excerpt in that many phrases are partly masked. This is the case for **LFDFb2** and **FCDFb**, which are both relatively low in loudness compared to the percussive sequences that accompany them. Additionally, for **HIMb** the ending of the phrase is masked in a manner similar to many of the type a phrases in *her reflection*, hence causing the evaluation to decline towards the end of the phrase.

12.3.5.3 Evaluation of information density

The superimposition and juxtaposition of longer and shorter phrases with different sound types contributes to making *her reflection* into a fantasy where a lot of things appear to happen at once. This results in evaluations of information density which are a bit higher than what is the case in the other fantasies, resulting in evaluations in the direction of the noise mode of the minimal. This is indicated by letting the curves have a hatched diagonal pattern. For the axial representations, however, the difference in mode is not visible in the representation.

As earlier, it has been necessary to weigh conflicting tendencies against each other to reach a judgment. The two most important factors that have contributed to the conflicting tendency are 1) *repetition* of words and syllables, which contributes to reduced information density, and 2) increased *complexity*, which contributes to increased information density. Here, the latter is a result of several co-existing layers presenting verbal information in different sound types, and little regularity in the way that the sound types and the different levels of verbal segmentation vary.

All in all, therefore, though the linguistic material is relatively predictable, it is presented in a form that makes it more complex, and therefore it contains more information. The result is that one thing counterweights another, hence making it relatively balanced. Said in another way; because we already know the words, we have resources to cope with the increased complexity. Consequently, the evaluations vary around the *maximal*, reaching from *maximal-intermediate* on both the side of the reduced and the noise modes of the minimal, depending on the degree of repetition and complexity. For instance, in **LFDF**, between ca 0:06 and 0:12, and in **FCDF**, between ca 0:21 and 0:25, the complexity is relatively high, with three simultaneous layers with high activity being present at the same time. These sections are therefore judged as being *maximal-intermediate* towards the noise mode. Conversely, the endings of many of the phrases are both relatively redundant when it comes to verbal material and have fewer layers and less activity, making me evaluate it as *maximal-intermediate* towards the reduced mode.

12.3.5.4 Naturalness evaluation

The three categories of sounds that one can distinguish in this fantasy are relatively distinct in terms of naturalness, and it therefore makes sense to consider the sound types one by one:

Sound type a: These sounds appear as a type of synthetic percussion instrument with no or minimal similarity to vocal production for significant portions of the time. For the parts where these sequences approach vocal quality, they are either evaluated as *minimal* or a little above it, due to several factors: Firstly, there are unnatural discontinuities between the syllables in the sequence (cf. *discontinuities* factor). Secondly, the varying degrees of processing, often resembling playback speed manipulation (cf. *technological artifacts*, *speed* and *pitch/spectral envelope* factors) clearly affect naturalness negatively.

Sound type b and c: These phrase types appear reduced in naturalness mostly because of the artificial quality of the noise “phonation” (cf. *phonatory spectrum* factor). The synthetic resonances that accompany the type b phrases make them lower in naturalness than most of the type c phrases, albeit different in degree from phrase to phrase. For example, **HIMb**, which is the phrase where the resonant chords are the strongest, is evaluated as *intermediate-minimal*, whereas **ATBBb** is evaluated as *intermediate*. The type b phrases, on their part, are evaluated a little above this.

12.3.5.5 Presence evaluation

In general, I experienced the sense of presence in this fantasy as low: Mostly, the phrases are evaluated as being *intermediate-minimal* or less, depending on the phrase type:

Type a: For the type a phrases, I only experience *minimal* presence, even in those instances where the sounds approach the voice. This is primarily due to the close to exact repetition of the syllables in the sequence (cf. *temporal continuity* factor), their very short duration (cf. *duration* factor), and their playback-speed-like processing (cf. e*xplicit/implicit transformation* factor).

Type b: For these phrases, the sense of presence lies between *intermediate-minimal* and *minimal*, and for two of the phrases, the evaluation declines towards the ending. The evaluation hinges for the most part on the *explicit transformation* and the *salience/loudness* factors. As for the former, two of the phrases, **HIMb** and **ATBBb1**, stand out. The resonances in these phrases are stronger than the others, and particularly in **HIMb**, the resonances are so dominant there is very little left that resembles a human voice, something which is reflected in the lower evaluation compared to the onset of the other phrases. As for the latter factor, the evaluation is mostly affected by the soft loudness and part masking by other sounds.

Type c: These phrases are the ones where I experience presence to be the highest. These phrases are largely affected by the same factors as the type b phrases, varying in degree from phrase to phrase, but where all phrases largely have a smaller degree of *explicit/implicit transformation*.

12.3.5.6 Evaluation of meaning clarity

What distinguish the phrases types and the single phrases for this premise are mostly the *salience of relevant cues* and *repetition* factors, both related to the criteria of **LI-domain clarity**, and **contextual clarity** for the features of the **ID-domain**:

Type a: Most of the percussive sequences carry no semantic meaning in themselves, nor do they point to features of identity and context. Still, by being part of a continuous sequence of sound that at some points is heard as partly intelligible words spoken by a voice, some of the meaning at these points can "spill over" into the preceding or following parts of the sequence – one can say that they get their meaning metonomically.[344] For example, at 0:27-29 (**FCDFa**), one can, at least with good will, recognize the syllable "cent", which then can be linked to the previous articulations of "from concent". Working in the other temporal direction, some of the sequences are linked to words that can be identified at a later point. For

[344] One could call this a kind of *synechdoche*, since a reference to the whole word is provided through a smaller *part* of it.

instance, the syllable [fɔ] can be identified at 0:04, but only linked to the word "formes" at about 0:08-09. The same "spilling over" can also happen for the features in **ID-domain**, so that the sequences of percussive syllables are indirectly linked to the vocal persona pronouncing the more intelligible words, even if the surface similarity is quite low. Since the percussive sequences are thereby linked to both a meaningful statement and the identity of a speaker, they are not devoid of any meaning. Therefore, none of the evaluations are *minimal*, even if the **LI-domain clarity** is very low.

Type b and c: For these phrase types, **LI-domain clarity** varies quite a lot between the phrases, and is what largely distinguishes the phrases from each other. The factors of *repetition* and *familiarity*, on the other hand, are roughly the same for all phrases. As for **contextual specificity**, it is also relatively similar for all phrases, with unspecified context and **ID-domain** features pointing in the direction of a female adult vocal persona. One single phrase, **HIMb**, is more ambiguous than the other ones when it comes to gender and age, and since only the first two phonemes can be identified with certainty, regional belongingness also becomes somewhat more ambiguous for this phrase.[345] The similarity with the other type b phrases, however, creates a context that reduces this ambiguity. For many of these phrases, there is an additional ambiguity regarding the number of vocal personae; the similarity in vocal identity among the phrases indicates one single speaker, whereas the partial superimposition of the phrases suggests several speakers. The result might be interpreted as a kind of surreal scene or simply that it is a result of a process of *organization* done by the composer (cf. section 2.5.1).

12.3.5.7 Evaluation of feature salience

In this fantasy, the factors *simultaneous masking*, *simultaneous contrasts*, and the *condition of the vocal features* are important in the evaluation. As for the latter, all phrases have in common that there are no features related to intonation present here, hence making the articulatory features more important in the evaluation, in contrast to *her voice*, *her presence* and *her song*.

[345] Again, it is important to note that the phrase can take on identity features through a metonymical link.

Type a: The percussive sequences are evaluated between the *intermediate-minimal* and *minimal*, mostly due to the degraded condition of vocal features caused by the heavy processing, only occasionally making some articulatory feature recognizable.

Type b: For these phrases, *masking* is a much more important factor, since portions of the phrases are often masked by other sounds or the resonating chords. The varying degree of masking is the main reason for the variations in the evaluations, which roughly vary from *minimal* to *intermediate*. The effect of masking and low simultaneous contrasts in loudness/spectrum can for instance be seen in the evaluation of **LFDFb2**. This phrase has an overall decaying loudness envelope which ends up in an almost complete masking of the penultimate syllable, something which explains the shape of the evaluation curve. **H(IM)b**, on its part, differs from the rest of the phrases of this type with its progressive *temporal masking*, in which the heavy resonances make the sound rapidly loose most of its vocal qualities. The beginning of this phrase is thereby not affected as much as the ending, and since the onset of this phrase is relatively loud compared to the preceding phrases, it starts out with at an *intermediate* level.

Type c: Of the three sound types represented in this excerpt, this type is the one in which vocal features are most salient, making in most cases especially the articulation component stand out. Consequently, the phrases of this type have the highest evaluations, reaching the *intermediate* at their most salient, with some variations due to *simultaneous masking* by accompanying sounds and a lowered degree of *simultaneous contrasts*.

12.3.5.8 Evaluation of stream integration

What separates this fantasy from the others is among other things that there are several sound streams, up to 3-4, superimposed upon each other with very little synchronicity among the streams, in contrast to *her voice*, *her song* and partly *her presence*. Here, the activity in the simultaneously sounding phrases is more independent, often being related only occasionally through having some common phonemes, syllables or words. However, in the parts where several superimposed phrases are heard together, their common noisy spectrum type increases the possibilities for confounding several streams, something which creates some ambiguity and therefore reduces the level of source coherence at these points. There are also differences between the phrase types regarding stream integration that are worth commenting on:

Type a: For these phrases, the lack of temporal distance between the individual syllables in the sequences (cf. *temporal proximity between events* factor), along with continuous changes from syllable to syllable (cf. *feature similarity/proximity/continuity* factor), ensures that sequential integration is high, so that one hears the sound as one continuous stream. Still, most of the phrases in this type can be heard as to some degree comprising more than one stream, meaning that there are also forces present that weaken the integration, partly creating separate streams. This is probably a result of the *looping* factor, with the repetitions having high speed and no temporal distance between the iterations. For example, in **LFDFa1** (repeating the syllable "-fo"), one can hear one layer consisting of an almost continuous stream of noise and one quasi-pitched (nodal) and more discontinuous layer; **HIMa**, on the other hand, appears to have at least two pitched components in addition to one noise component (albeit rather merged with the pitched ones). Depending on the strength with which this layering is experienced, I have evaluated these phrases as being between *maximal* and *intermediate*. There are also some cases in which the beginnings or endings of the percussive sequences fuse more or less with the type c phrases. These cases will be discussed below.

Type b: For these phrases, the pitched resonant components are layered in chord or interval-like structures, often including noised components. Consequently, these sounds consist of several streams in the form of clearly distinguishable pitches, albeit still integrated through similar timbre, synchronous onset and common development of their spectral envelopes. Such layering is maybe clearest in **H(IM)b** and **ATBBb1**, where one might identify a major seventh chord (approximately D^{maj7}) and a major triad (C# major), respectively. For **ATBBb2**, however, it is difficult to hear any chord-like harmony, but one can still discern that there is a set of pitched components in the sound. In all cases, the different streams must be attributed to properties of the resonance rather than multiple sound sources, implying therefore a relatively weak form of ambiguity. I have therefore evaluated most phrases as being around *maximal-intermediate*, with minor differences due to the clarity of the streams and the degree to which components blend in with accompanying sounds.

Type c: The phrases in this category generally have fairly well simultaneous and sequential integration. One thing that creates a little ambiguity, though, is the spectral similarity (noise based spectrum) with accompanying sounds, especially the type a phrases. When these sounds are heard together, they sometimes partially fuse, so that it becomes ambiguous which components should be integrated with what sound.[346] There are also some ambiguities when it comes to sequential integration. In some cases, the type a phrases coincide with the type c phrases, making it less clear what sound event is integrated with what stream. This is the case with **LFDFc3**, which can be heard both as a continuation of **LFDFc2** and of the sequence of repeated "fo" syllables in **LFDFa1**. Another factor that creates some ambiguity for the horizontal integrations is the back and forth movement in spatial location for some of the phrases, in particular **LFDFc2-5**, **FCc7-10**, **HIMc1-4** and **ATBBc1-5**.

12.3.6 Evaluation of *her self*

12.3.6.1 Over-all description and evaluation

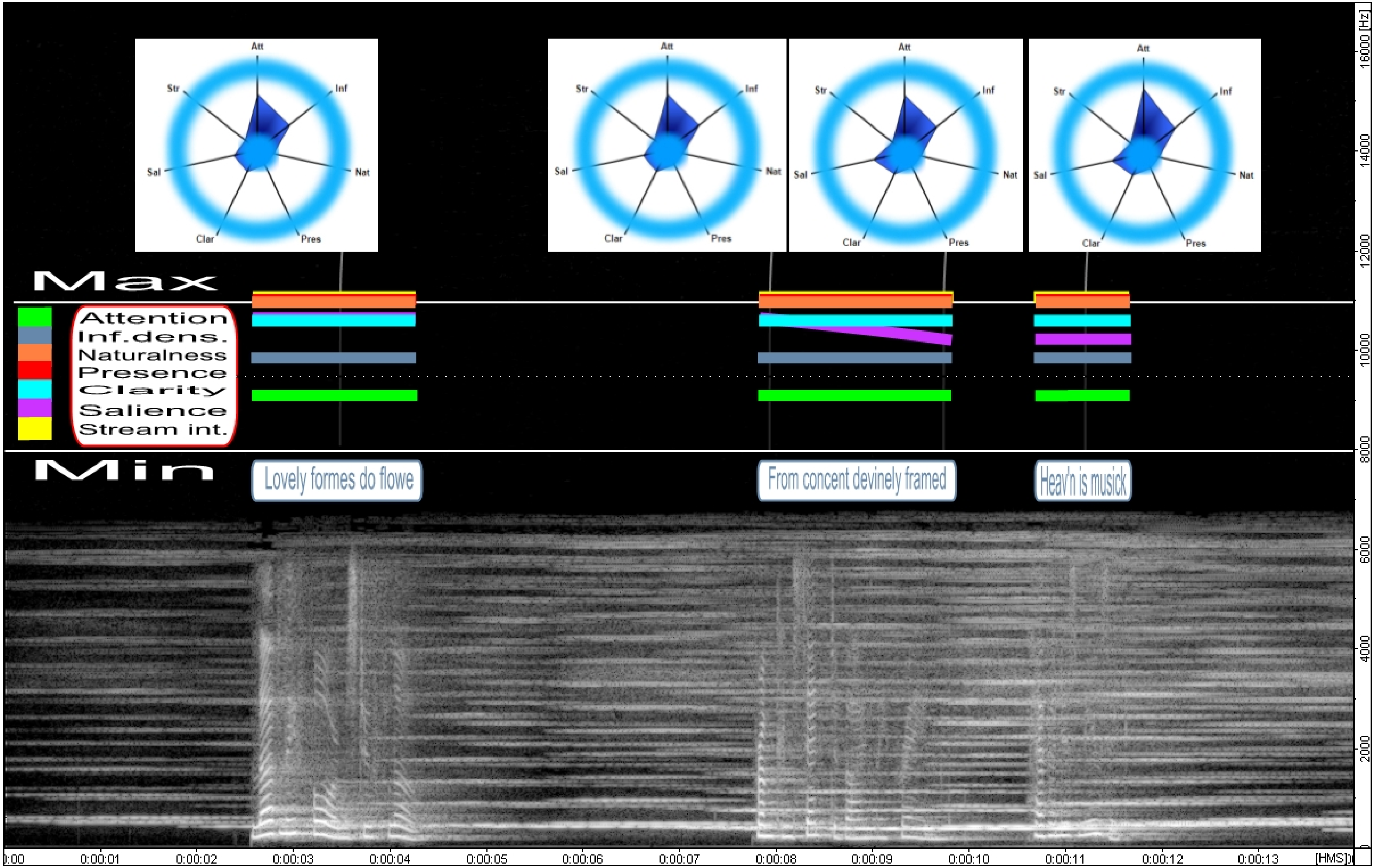
The close to unmanipulated female voice reading Campion's poem in a calm and clearly articulated manner in *her self* has clear similarities with the voices heard in the other fantasies, especially with *her voice*, which has intonation contours and articulatory features that appear to lie not far from this movement. This movement also appears to share the **ID-** and **TCM-domain** features with **BS1** and **BS3** from *her reflection*.[347]

In the pauses between the phrases, which are of varying length, one can hear sustained notes at stable pitches, partly overlapping each other and thereby forming melodies as well as harmonies. These notes are quite similar to the static sustained vowels that formed the accompaniment in *her presence*, but owing to the more marked contrast to the unmanipulated vocal phrases in this movement, the accompaniment appears more synthetic in this context, often losing all resemblance to vocal production. Another reason for the increased artificiality might be the layering of the vowels into harmonies and chords during major parts of the excerpt, which partly blur the individual vowels.[348] When the accompaniment rarely takes on a vocal quality, it is mainly because the sustained notes occasionally sound similar to vowels that have occurred or are about to occur in nearby phrases. In addition, the accompaniment sometimes also appears to imitate pitches of salient portions of the vocal phrases, but most often the short duration of the intonation contours of the reciting voice makes it difficult to compare it with the sustained notes of the accompaniment.

[346] Ondishko also notes that there is "blending [of] the foreground with the background" in parts of this fantasy (Ondishko, 1990: 41).

[347] The voice in **BS1** and **BS3** from *her reflection* have less reverberation, however.

[348] This was also noted for the sustained vowels in *her presence* (cf. section 12.3.2).



Att
Inf
Nat
Pres
Clar
Sal
Str
Max
Min
Attention
Inf.dens.
Naturalness
Presence
Clarity
Salience
Stream int.
Lovely formes do flowe
From concent devinely framed
Heav'n is musick
16000 [Hz]
14000
12000
10000
8000
6000
4000
2000
0
0:00
0:00:01
0:00:02
0:00:03
0:00:04
0:00:05
0:00:06
0:00:07
0:00:08
0:00:09
0:00:10
0:00:11
0:00:12
0:00:13
[HMS]

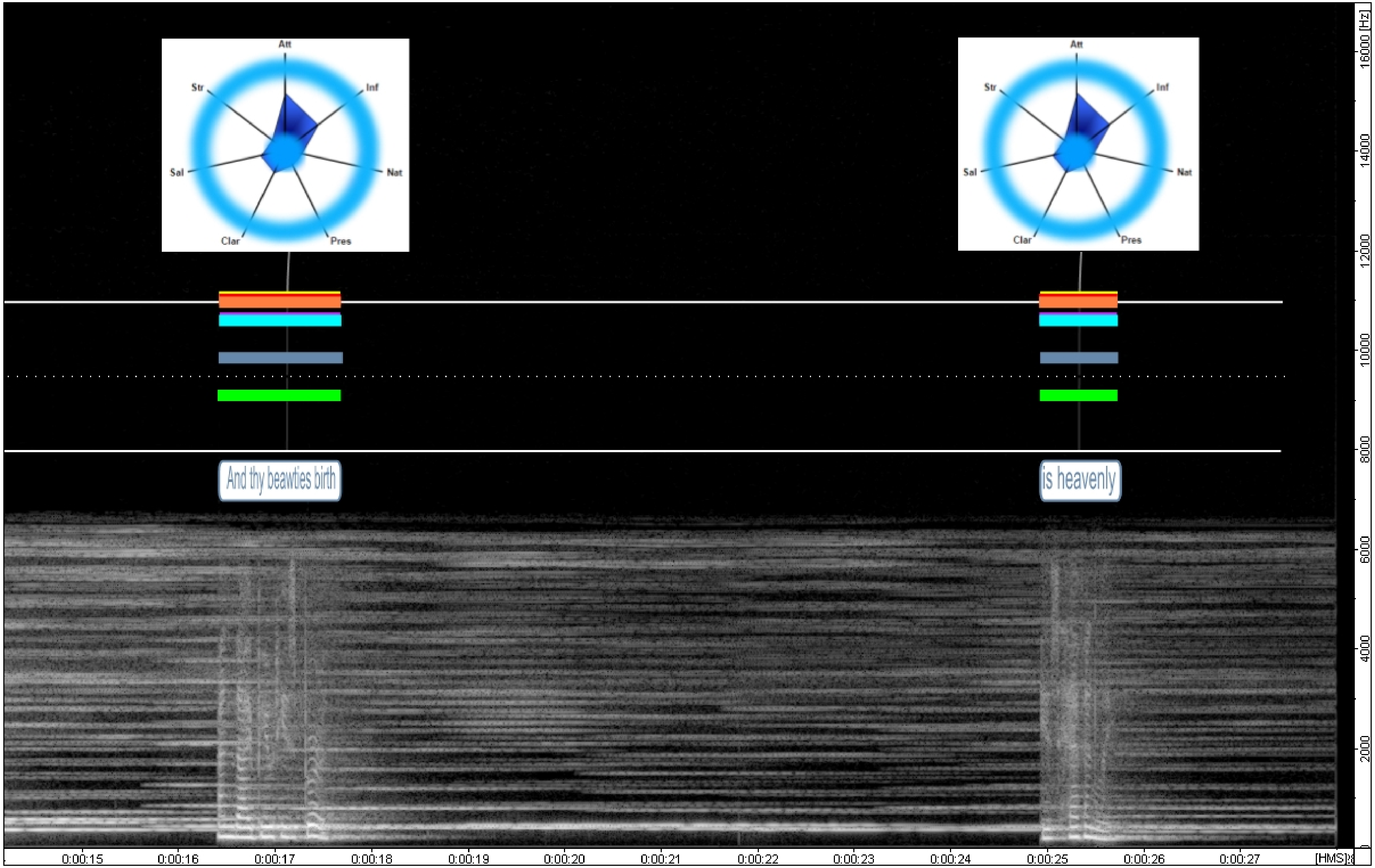


Figure 12.6: Graphical representation of evaluations of the seven premises in Lansky's *her self*. Note that some of the colored lines overlap so that they are not fully visible in the figure. For representation of the evaluation of single premises the reader should use the *acousmographe* software and the files provided on the accompanying CD-ROM.

The excerpt from *her self* is presented in **sound example 12.6** (Lansky, 1994a, *her self*, 0:37-1:05). In the evaluations presented in the acousmographe files and in **figure 12.6** I have chosen not to show the evaluation for the sustained vowels, since they only occasionally sound vocal.[349]

The unmanipulated voice is the obvious focus of attention in this fantasy, even if the accompaniment challenges it a few times during the movement. Having heard the phrases of the poem five times over the course of the five previous fantasies, the redundancy of the repetition should now make it far less interesting to attend to the semantic aspects of the poem (**LI/sem**). In my experience, however, the semantic aspects were not completely exhausted in the listening process; several times during listening to this movement I have found myself partly and fleetingly reengaged with issues of verbal meaning. The explanation to this might have something to do with the fact that this is the first time in the piece that the text is presented as a straightforward recitation, and that this mode of presentation triggers a way of listening and interpreting that is closer to that of speech than in the earlier movements; the speech rate is within normal range, and it is much more evident in this movement that the voice one hears is a human being speaking rather than a product of a complex compositional and technological process.

In addition to the repetition of the now familiar text of the poem, the overall trend in this movement with attention being directed elsewhere than towards **LI/sem** can be related to the mentioned similarity relations or semi-imitations between the vowel quality and the pitches of the sustained notes in the accompaniment and the phrases of the reciting voice. When these relations are noticeable, I have experienced that this is something that attracts attention, so that one focuses more on pitch (**SQS-domain**) and vowel qualities (**VG-domain**) than would have been the case without this similarity relationship. This can be heard most clearly, in my opinion, in the phrase **HIM** (0:10-0:16 in the excerpt). Here, the reciting voice has a rising intonation contour on the first four syllables, whereupon it drops below the initial pitch at the last syllable, "-sic". Following this phrase is a sequence of notes which, even if the notes overlap extensively, follows the same overall intonation pattern, even if it is considerably stretched out in time. One can also recognize some of the vowels from the phrase, which are shown here as underlined: [hε̲vn ɪ̲s mjuzɪ̲k].[350]

[349] For the parts that are most voicelike, e.g. at times between the **ATBB** and **IH** phrases, the evaluations are not far from the lowest evaluations of the sustained static vowels in *her presence*.

[350] The [u] is difficult to perceive as more than a slightly different quality compared to the surrounding [ɪ] sounds.

This linkage appears a lot weaker for many of the other phrases. Nevertheless, what I noticed was that even if there were no easily recognizable similarities between the accompaniment and the vocal phrase in other cases, the similarity relations that had been established for some phrases made me search for comparable relationships elsewhere. Hence, I often paid some attention to pitch and vowel qualities for the other phrases as well, albeit somewhat less perhaps than when I could confirm similarities.

In sum, the dominant situation in this excerpt is one in which attention is divided between vocal (**VG**) and non-vocal (**SQS**) domains, but where there are occasional weak distractions from **LI/sem**. According to figure 5.1 this indicates an evaluation around *intermediate-minimal*, but since the weak distraction from **LI/sem** pulls it somewhat closer to the maximal, I have evaluated all the phrases in the excerpt as a little below *intermediate* for the *focus of attention* premise.

As with *her song*, the evaluation of the *information density* premise has been affected by feature repetition and the constancy of features related to the **ID**, **SE**, and **TCM-domains**, hence linking it to the factor of *complexity*, and to regularities and learning/familiarity (cf. section 6.3). When seen in relation to my evaluations for the other fantasies, I have found it appropriate to give it an evaluation slightly above the *intermediate*.

With vocal phrases consisting of (almost) unmanipulated speech, several of the premises are evaluated as *maximal* or close to *maximal*. This goes for *naturalness*, *presence* and *stream integration*, for which the evaluations are relatively self evident.[351] As for the *clarity of meaning* premise, it is only due to the discussed difficulties with the semantic, syntactical and lexical issues in the poetic text (cf. section 12.1.1) that I choose not to evaluate the phrases in this song as *maximal*, but rather as close to the *maximal*. Lastly, most of the phrases have high feature salience, but since parts of some phrases are moderately masked by the accompaniment, the evaluations vary between close to the *maximal* and *maximal-intermediate*.

12.4 Overall evaluation and interpretation

After having evaluated excerpts from all fantasies during the preceding section, it can be interesting to make some more systematic comparisons of them, so as to see whether this can

[351] Considering the span between the minimal and maximal in the piece as a whole, I have chosen to disregard the slight artificial reverberation added here, which could have been taken as an indication at a departure from the *maximal* for both *naturalness* and *presence*.

say something about Lansky's piece and the process of listening to it, and possibly if this can be related to what others have written about the composition. In the subsequent sections, I will therefore compare the evaluations of each of the premises in turn, so as to be able to discuss similarities and differences between the excerpts, and then consider if this in turn can be used to infer something about the relationships between the six movements.

During my discussion, I will also introduce sound examples that I have made with the SFM instrument to illustrate how changes in certain parameters or compositional strategies can affect the evaluation of the premise in question. By comparing these evaluations with the ones of Lansky's piece, one can hopefully see some of the aspects of the piece in relief, and get a better understanding of the relationship between the listening process, the piece, and some of its prominent features.

12.4.1 Individual premises

12.4.1.1 Focus of attention

Throughout the previous sections, we have seen that the evaluation of the focus of attention premise did not comprise the whole range from maximal to minimal, with the highest evaluation being *maximal-intermediate* (**BS1** from *her* reflection). The dominating tendency for this premise therefore seems to be much closer to what one could expect for vocal music than for music based on a reading of a poem. When looking at the evaluations of all excerpts together, as in the time compressed representation in **figure 12.7**, one sees that a majority of the evaluations are made in the range between the *intermediate* and the *minimal*. In the preceding discussion, several reasons have been pointed out for attention not being directed towards the meanings of the poem (**LI/sem**), despite that the piece at one point presents the poem almost as a straightforward recitation. Instead, there were a lot of interesting and aesthetically appealing aspects that were played out in the other experiential domains, with perhaps the **SQS-** and the **VG-domains** being most frequently attended to. This was the result of several issues: 1) the meaningful changes in these domains; 2) that the same text was presented in all the movements, hence making it increasingly redundant; 3) there were very few situations in which text and music engaged in overt interplay, something that potentially

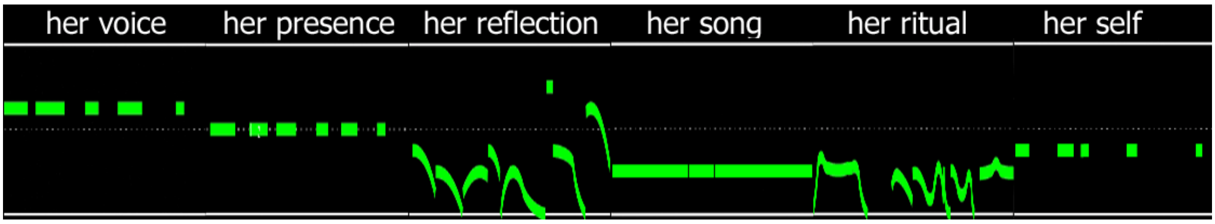


Figure 12.7: Time-compressed representations of all evaluations of the attention premise juxtaposed. Maximal and minimal evaluations are at the upper and lower lines, respectively.

could have modified or altered the semantic meanings of the poem; 4) the poetic text had several features that were perhaps not too relevant for a contemporary reader/listener and that were relatively complex. These issues were linked to several of the listed factors for this premise: *Novelty/change*, *unpredictability*, *relevance/interest* and *salience*. In addition, one can also link these issues to the compositional structure of the piece. A different compositional structure, e.g. with the presentation of the unmanipulated reading in the beginning of the piece instead, either unaccompanied or with very discrete accompaniment, could potentially have made attention turn more towards the verbal aspects than was the case here. One could also imagine other compositional strategies which may have worked in the same direction. For example, if the manipulation of segments in the piece had included a re-ordering of words taken from the original phrases so as to construct new sentences, this could under ideal conditions have altered the focus so that it would potentially be focused more towards the semantic aspects. A hypothetical example of this approach is presented in **sound example 12.7**, generated with the SFM instrument[352], in which words from different phrases are combined so as to create a new and relatively meaningful sentence; “Lawra, thy dull notes neede for helps”. If this sentence was to be presented for the first time in the piece, unaccompanied by any other sounds that may distract the listener, attention could perhaps have been directed more towards the semantic aspects than the vocal or non-vocal features, especially since this sentence presents a totally different stand towards Lawra compared to what is presented in the poem.

The overall tendency in the evaluations from the first to the last excerpt, as seen in **figure 12.7**, can be seen partly corresponding to Lansky’s own ideas about the piece, and especially his views about the function of the last movement in the piece: Lansky has expressed that he wants to “explicate the implicit music within [Hannah MacKay’s reading of Campions poem]”, and that when listening to the final fantasy, even if it consists of an unmanipulated reading, the presented reading will sound like music to the listener (CD liner

[352] This example can also be played with the SFM instrument by retrieving preset 11.

notes, Lansky, 1994a). What I noted was that for most phrases, attention was divided between the **LI/sem** and vocal and/or non-vocal domains, with the **SQS** and **VG-domain**s being the most frequently mentioned and the **TCM** and **AF-domain**s attracting attention more occasionally or weakly. For the **AF-domain**, I see this as a result of the neutrality of MacKay's reading, and of the low emotional charge of Lansky's manipulations. With a more emotionally charged reading, or with manipulations that would successfully sound emotional, the factor of *emotional salience* would perhaps been evoked. Furthermore, the **ID-domain** very rarely attracted attention.

Taken together, even if other domains also attracted some attention, the abstract "musical" qualities clearly got a share of my attention for many of the phrases in the provided excerpts, indicating that Lansky's intentions were, at least partially, fulfilled in my own listening process. This is also evident from the evaluation of the final fantasy. If this fantasy had been given a maximal evaluation for the attention premise, Lansky's intentions would have failed. However, as we saw, several features and strategies contributed in guiding my attention towards (mostly) the **SQS-** and **VG-domains**, something which resulted in an evaluation a little below the *intermediate*.

Considering the fact that the piece is made of extensive manipulations, the amount of attention that was directed to the **TCM-domain** appears rather modest. Hence, my evaluation corresponds with Ondishko's experience of the effect the piece has had on listeners: "Most listeners are not aware of any manipulation going on in the piece and instead are struck by its sonic beauty, supple harmonics and gentle phrases (Ondishko, 1990: 17). Also, Lansky himself has stated that he was interested in downplaying the role of technology in the piece, thus making the technologies involved in the creation process as transparent as possible.[353] As I see it, three reasons for this can be found among the factors of the *focus of attention* and the *naturalness* premises: Firstly, the *technological artifacts*, especially those related to the LPC analysis/synthesis process, are relatively constant throughout large portions of the composition, perhaps with the "buzziness" of the voiced timbre as the most prominent trait. Thus, even if the listener perhaps reacts to the artificiality when hearing the beginning of the piece for the first time, he or she will soon get used to it and thereby focus less on it as the piece unfolds. Secondly, the technological artifacts in the form of clicks and discontinuities that one can sometimes hear are usually very subtle and they therefore also attract minimal attention. Thirdly, a great deal of the manipulations, with those exciting external resonances

[353] Personal communication with Paul Lansky, October 2006.

as exceptions, behaves in a manner that is largely within the range of human possibilities. Thus, the contrast to pieces discussed earlier, like Charles Dodge's *Speech Songs* and *In Celebration*, is clear; while Dodge often explicitly manipulated the voices in these pieces so that technological artifacts and discontinuities were prominent and their behaviour was clearly beyond human limits, the transformations in *Six Fantasies* usually appear more implicit than explicit. However, there are also some clear exceptions from this tendency, especially in *her reflection* and *her ritual*. For the voice excited resonances that could be heard in *her reflection*, however, I did not primarily hear the resonances as products of technology, but more as a result of the interaction between the voice and some virtual object or construction in the surroundings of the vocal persona. The percussive sequences in *her ritual* were probably the sounds in which the largest share of attention was directed towards the **TCM-domain**, especially for the sections were the sequences slowed down and became more voicelike, paradoxically enough.

The fact that the **ID-domain** and the **AF-domain** received little attention, I suspect can be ascribed to the modest variation in the attribution of affective and identity related features. One can imagine ways that these domains could have had a lot more variation, and thereby also potentially attracted more attention. I have synthesized two examples in which I have tried to vary cues related to affective and identity related aspects, respectively. In **sound example 12.8**, there are three subsequent phrases/utterances, all with the same verbal content, but with differences in pitch contour, amount of time stretching or compression and shifting of the spectral envelope.[354] Even if it is not straightforward to attribute the affective states from the different voices, I think that most listeners will sense that there is a difference in the emotional content of these phrases.[355] Due to the similarity in the verbal content and the identity of the voice in these phrases, I suspect that the **AF-domain** will attract more attention here than what was the case in Lansky's piece. In **sound example 12.9**, I have put together five different phrases in which pitch transposition level and amount of spectral envelope shift differ from phrase to phrase, thus affecting the experienced identity.[356] Here, phrases one and three are hopefully experienced as the voice of a child, two and four are experienced as a man's voice, and the fifth is experienced as a woman's voice.[357] I believe that these cues

[354] Cf. SFM instrument, preset 12.

[355] When doing an informal test with two of my colleagues, the first phrase was experienced as "straightforward" or "declarative", the second was experienced as "sad" or "comforting" and the last one was experienced as "worried" or "worked up".

[356] Cf. SFM instrument, preset 13.

[357] Furthermore, each of these "identities" is panned differently from the others.

would give most listeners the impression that several voices with different identities were involved here.[358]

12.4.1.2 Information density

For all of the excerpts evaluated, information density has largely occupied the range between *maximal* and *intermediate* of the reduced mode of the minimal, meaning that the range between maximal and the noise mode of the minimal has rarely been applied.[359] This can be observed in the time compressed representation in **figure 12.8**, where one can see that only some sections of *her ritual* have the hatched light blue and white pattern that designates the noise mode. In other words, from the chosen excerpts, it does not seem that *Six Fantasies* is dominated by excess of information. Throughout my discussion of this premise, this has been partly a result of repetition of the text in the fantasies. Despite that at certain points, especially in the fifth fantasy, *complexity* was significantly increased, repetition of syllables, words or phrases contributed in hindering the evaluation from moving very far towards the noise mode of the minimal. What is more, Lansky's fantasies consistently used time-stretching of the vocal phrases and avoided time-compression, so that the factors of *complexity* or *rate* were never excessively high while listening to these movements.

Based on these observations, it is possible to imagine compositional strategies that would have caused the evaluations to be played out in the direction of the noise mode of the minimal. For example, in **sound example 12.10**, I have synthesized several simultaneous layers of time-compressed phrases, where only a few of them are ordered according to the original sequence.[360] Thus, both *rate* and *complexity* are factors playing a part here, and the loss of regularity in the ordering of the phrases that was so important in Lansky's piece makes it much more difficult to grasp the verbal content of the phrases without extensive re-listening. The phrases are given pitch trajectories and panning so as to prevent the phrases from masking each other too much, something that could have made the phrases merge into one single stream with properties that potentially could approach some regularity and pass the complexity breaking point (cf. section 6.3).[361] Especially if this was to be presented early in a composition, the complexity and the speed of presentation could therefore lead to an

[358] Presenting this example to two colleagues, they confirmed this suspicion.

[359] In preliminary phrases of the evaluation process, before I had chosen to focus exclusively on shorter excerpts from the six movements, I also evaluated some phrases in the beginning of *her voice* as being close to the maximal, but slightly on the noise mode side.

[360] This example was made by mixing several sound files created with the SFM tool.

[361] Also compare the discussion in section 7.3.2 on regularities in noise.

evaluation in the direction of the noise mode of the minimal. If the piece had been dominated by sequences close to the noise mode of the minimal, however, it would probably have been harder to get to know the phrases of the poem during the piece, and the effect of bringing out the music in MacKay's unmanipulated reading in a final fantasy would have been much more difficult to achieve.

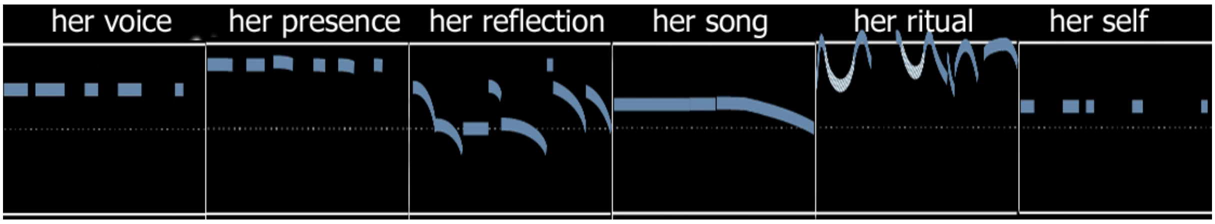


Figure 12.8: Time-compressed representations of all evaluations of the information density premise juxtaposed. Maximal and minimal evaluations are at the upper and lower lines, respectively.

Conversely, one could also imagine that other compositional strategies would have resulted in evaluations that dominantly resided in the range between the intermediate and the reduced mode of the minimal. For instance, this could have been achieved by sustaining the static vowels in *her presence* even more and letting the vocal phrases enter more rarely, by looping words or phrases for long periods, or by time-stretching them extensively. The latter option is exemplified in **sound example 12.11**, which contains the word "Lawra" time-stretched in multiple voices over 41 seconds, hence giving an extremely low *rate* of presentation.[362] Using such phrases, however, would have made it a lot more difficult to use the poem in its entirety without making the duration of the piece very long. It would probably also imply that it would be very difficult to follow any verbal content, because one would have to retain single phonemes, syllables and words in memory for a long time in order to be able to put together whole phrases or sentences. Hence, Lansky's compositional strategy can be seen as a means to retain the structure of the poem for the listener, while at the same time gradually exhausting its semantic meaning so as to allow for a more musically oriented listening when the unmanipulated reading is presented at the end.

12.4.1.3 Naturalness

As one can see from the **figure 12.9**, the evaluations of naturalness move within the whole range from maximal to minimal**.** Considering that the first and the last fantasies have the

[362] Cf. SFM instrument, preset 14.

overall highest evaluations compared to the other movements, one can therefore see that there is a movement from the maximal towards the minimal through the first fantasies, and then back again towards the minimal throughout the last fantasies.

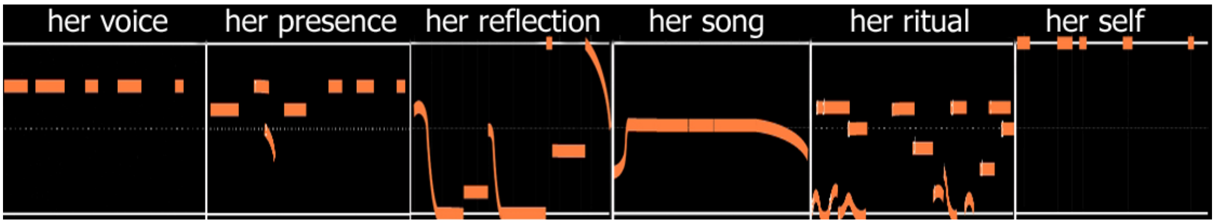


Figure 12.9: Time-compressed representations of all evaluations of the naturalness premise juxtaposed. Maximal and minimal evaluations are at the upper and lower lines, respectively.

What one can also notice about this premise is that for the majority of phrases, the evaluations are constant, and for the ones with temporal changes, these are only relatively moderate. There were some phrase types in *her reflection* for which naturalness decreased towards the end of the phrase due to the dominance of the resonances, and some less marked changes in *her song* due to time-stretching, but largely there were few instances with gradual transitions in naturalness over a greater span of the continuum. The most notable transitions were in *her ritual*, where the percussive sequences appeared to move back and forth between voice and non-voice in a gradual manner, hence constituting a continuous but at the same time unitary *transformation* (cf. section 12.3.5.1).

One can imagine that such *transformations* could have been accentuated much more in the piece by letting them move all the way from non-voice towards the maximal. This is what I have attempted in **sound example 12.12**, where I have synthesized a sound that resembles the percussive sequences in *her ritual*, only that the sequence goes all the way from being non-vocal to what is comparable to *her voice* in naturalness.[363] In contrast to what is the case in Lansky's piece, such a transformation would thereby establish a much clearer connection between the female vocal persona heard throughout all fantasies and the percussive sequences. Such a connection might also have given the impression that the identity of this vocal persona is more "bendable" than what is the case in the composition, where I experience the identity features as quite consistent. Hence, Lansky's choice of avoiding transitions that are too extreme might have resulted in constituting a more coherent and less plastic identity in line with the titles of the fantasies, which all centre on the same "her".

[363] In this case I had to use a scripted input to the SFM tool because of the limitations in the GUI version regarding the number of possible repetitions (which equals the number of voices) of a sound.

Another thing that one can note about the naturalness evaluations is that they do not differ much for the vocal phrases in the three fantasies where the pitched LPC-manipulated voices dominate, namely in *her voice*, *her presence* and *her song*.[364] For all these fantasies, a "buzzy" quality characterized the pitched parts of the phrase, which was associated with the factor of *phonatory spectrum*. This had a relatively mild effect on the naturalness evaluation, however, and one can see from figure **12.9** that it is comparable to the artificiality of the phrases with noise excitation that had the highest evaluations in *her ritual* (most type c phrases and some type b phrases). For these phrases, *phonatory spectrum* was also mentioned as a factor, but here it was the artificiality of the noise rather than "buzziness" that was in focus.

Considering that the LPC technique can potentially produce striking artifacts, the features that I experienced as related to the factor of *technological artifacts* in Lansky's piece were not very prominent, as I have already noted in section 12.4.1.1. These could potentially have been made more prominent with a more overtly "médiatiste" approach. In **sound example 12.13**, I have synthesized several phrases which all relatively explicitly bring out some form of *technological artifact*.[365] Here, two phrases extend a feedback inherent in the filter part of the LPC analysis file, hence causing significantly reduced naturalness while making the feedback more prominent. The last of the phrases in the example is an example of applying time-stretch on the transition from consonant to vowel so that it becomes particularly abrupt and unnatural. Such effects are relatively far from what can be heard in the piece, where it is dominantly the vowels that are stretched. The reason why Lansky chose to not to stretch the consonants in this piece very much might therefore be that he wanted to avoid such an effect. Sudden dips in naturalness may have stood out as technological artifacts thus directing attention towards the **TCM-domain**, which was precisely something that Lansky stated that he didn't want. On the other hand, by stretching mainly the vowel parts of the phrases as in *her song*, the result could more easily be perceived as precisely song-like and therefore more natural.

12.4.1.4 Presence

From **figure 12.9**, one can see that the evaluations of presence resemble the naturalness evaluations relatively much. This was for a large part is a result of the *implicit/explicit*

[364] For the sustained vowels in *her presence*, however, the minimal evaluations stood out markedly in that respect.

[365] Cf. SFM instrument, preset 15.

transformation factor that links presence to experienced naturalness, which thereby attests to the relatively high impact of this factor on presence compared to the other factors. As a result, we can recognize some of the same tendencies here as we did for naturalness; we can see that the evaluations comprise the whole range between maximal and minimal and we can see the irregular arch shape from the first to the last excerpt. There are some differences between the two premises, however, especially in *her presence* and *her ritual*, where presence is lower for some phrases, mostly due to lower loudness (cf. *salience/loudness* factor) and longer experienced virtual distance (**FC2** in *her presence*, cf. *spatial distance* factor) and to masking (**LFDFb2**, **FCDFb** and **ATBBb2**, *her ritual*, *salience/loudness* factor). One other difference is the descending curve for **BS3** at the very end of the excerpt from *her reflection*, which the reader might recall as a phrase where a close to unmanipulated voice was followed by a chain of decaying "echoes". The difference in the evaluations suggests that this type of echo/repetition in this case has greater effect on the sense of presence than it has on experienced naturalness, since it explicitly presents the phrase as recorded and processed, but where the similarity between the individual instances and a natural voice is relatively high.

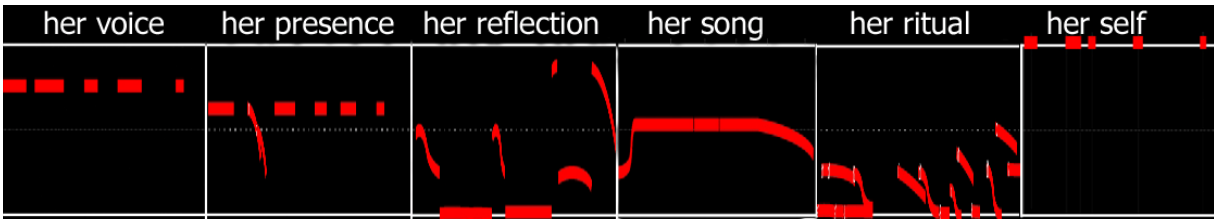


Figure 12.9: Time-compressed representations of all evaluations of the presence premise juxtaposed. Maximal and minimal evaluations are at the upper and lower lines, respectively.

The differences in the naturalness and presence evaluations would probably have been greater if Lansky had engaged more in moving the virtual persona around in virtual space, hence playing more with the *spatial distance* factor, as I have done in **sound example 12.14**. Here, one can hear a noise excited voice whose naturalness is pretty much unchanged, while the apparent vocal persona is moving around, first approaching until it is quite close, and then retreating into the distance at the end of the example. Experienced presence will thereby vary from being distant at the beginning, then relatively close/present, and finally more distant again at the end of the sound, all the time while naturalness does not change noticeably. Having this kind of spatial motion in the piece could potentially have attracted attention to itself, i.e the **SE-domain**, at the expense of both the **SQS-** and the **VG-domains**. Hence, one can thereby see how this had run counter to Lansky's intentions of highlighting "some aspect of her speech—contour, vowels, resonance, articulation, consonants, etc.—in order to explicate the implicit music within" (CD liner notes, Lansky, 1994a).

12.4.1.5 Clarity of meaning

In comparison to the previously discussed premises, the evaluations of clarity of meaning comprise a more limited range of the max-min continuum, as we can see from **figure 12.10** below. The reason why none of the phrases were evaluated as maximal, even with high bottom-up clarity (cf. *salience of relevant cues* factor) was mainly that the semantic opacity of the poetic text made comprehension difficult, at least more difficult than a spoken message about daily matters would be. Alternatively, the evaluations can be seen as a result of my lacking knowledge of the relatively archaic poetic language and semantic issues related to courtship, the divine and beauty (cf. *language/code competence* factor). Anyway, the use of Campion's poem instead of a text of which I had a better knowledge of language and codes appears to be an important precondition for the outcome of the evaluation process.

At the other end of the scale, I have also avoided minimal evaluations altogether, despite that for several of the phrases in *her reflection* and *her ritual*, the *salience of relevant cues* has been very low. This can be seen as a result of the consistent use of a single voice and a relatively narrow range of manipulation types (cf. *adaptation to speaker/type of manipulation* factor), and perhaps also a *rate of presentation* factor that didn't deviate excessively from a normal rate. However, I believe that this is mostly owing to the very clear and redundant structuring of the poetic text throughout the piece as a whole, including excessive repetition (cf. *repetition* and *familiarity* factors). Consequently, one hypothetic possibility to achieve lower evaluations would be to loosen up the structure of the

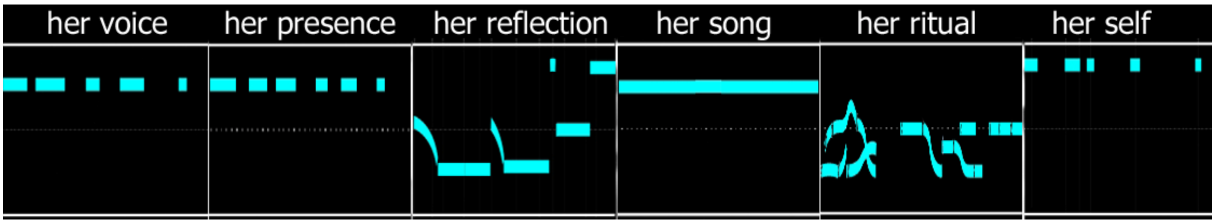


Figure 12.10: Time-compressed representations of all evaluations of the clarity of meaning premise juxtaposed. Maximal and minimal evaluations are at the upper and lower lines, respectively.

presentation of the text, so that it would not be that easy to predict what would come next. One even more radical approach would be to introduce text from other sources or even quasi linguistic utterances. Combined with low bottom-up clarity/*salience of relevant cues*, these approaches would have made it much more difficult to construct clear, specific and coherent meaning from the phrases.

Another thing that could have made the phrases of the poem rated lower for this premise is related to the **ID-domain**. While there are certainly phrases in which the voice is heavily transformed in the evaluated excerpts, the overall consistency of the identity of the vocal persona in the piece still link these phrases to adult female reader with the soft and gentle voice. True, the identity is gently stretched and bent at some instances, e.g. with the high pitched voice in some of the phrases in *her presence*, but one never really gets the experience of a radical change, e.g. affecting the central categories of age, gender and size. This can naturally also be linked to the titles of the movements, each beginning with “her”, which clearly relates the vocal phrases to some female “protagonist” binding all the six fantasies together. If Lansky had played with these identity categories, however, as I attempted in **sound example 12.9**, the phrases with no clear bottom-up identity cues would have been even more ambiguous. Using several vocal identities like this would also probably lead to somewhat lower evaluations of clarity of meaning, since it would be more difficult to relate the different aspects of the text to several vocal personas with different identity features, thus potentially making the link with the verbal phrases more ambiguous.

12.4.1.6 Feature salience

From **figure 12.11** below, one can observe that the evaluation of feature salience spans the whole range between the minimal and the maximal, with *her reflection* containing both the highest and the lowest evaluation. Thereby, it seems that as for *naturalness* and *presence*, the experience varies within the whole range of the premise. For this premise, it was perhaps the *masking* and the *condition of vocal features* factors that generally contributed to the lowest evaluations, which were found in the excerpts from *her reflection* and *her ritual*.

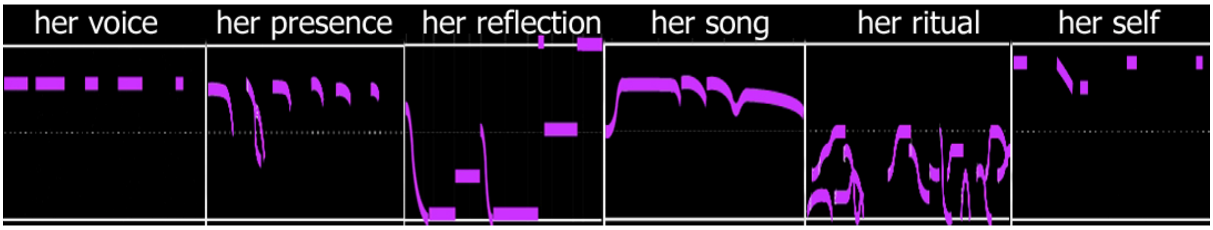


Figure 12.11: Time-compressed representations of all evaluations of the *feature salience* premise juxtaposed. Maximal and minimal evaluations are at the upper and lower lines, respectively.

One can also note that there seems to be somewhat more variation in the evaluations compared to many of the other premises, even if the variation within each phrase is still relatively modest – one can see that there are only a few places where the evaluations change markedly over time. Thus, even if the piece as a whole appears to contain relatively large contrasts when it comes to feature salience, the largest contrasts are not accentuated or played out so much in each individual phrase, at least for the phrases included here. One can imagine that this could have been done by making either abrupt or gradual transitions that greatly affected feature salience in the course of a single phrase. For instance, by introducing a masking sound abruptly upon the vocal phrase and then removing it just afterwards, like I have done in **sound example 12.15**, it would create an abrupt transition from maximum to minimum salience and back again during the phrase, probably creating a relatively dramatic effect.[366]

As for gradual transitions for this premise, one sees that *her reflection* contains many, going from around the *intermediate* to the *minimal*, or the other way around. If such transitions had instead gone from one extreme to the other, they would accentuate the contrast even more. I have made two **sound examples**, **12.16** and **12.17**, which I feel fit this description pretty well. Both start out with a high degree of masking and then gradually they become less and less masked, until the end of the vocal phrase is heard in isolation. **Sound example 12.12** would in a similar manner present a transition from a situation where no vocal features were audible to a high degree of feature salience.

Together, all examples show that the contrasts in feature salience could have been played out *within* individual phrases, and perhaps more extremely than was the case in *Six Fantasies*. That Lansky chose to mainly play out contrasts *between* phrases instead of *within* them can be a result of a structural idea or a compositional method that retains the phrase as a relatively consistent unit in itself.

[366] Cf. SFM instrument, preset 16.

12.4.1.7 Stream integration

What we can see from **figure 12.12** is that *stream integration* appears to be the premise with the highest overall evaluations for the excerpts in question. One can also see that the relationship between excerpts from different fantasies are quite different compared to the other premises. For instance, even if it has higher evaluations for most of the other premises, *her voice* is evaluated as lower than most of the phrases in both *her presence* and *her ritual*, due to the level of source coherence ambiguity created by the synchronous articulation and parallel intonation contour. Similarly, level of source coherence ambiguity created by having several voices superimposed upon each other, also affected the evaluations in *her presence* and *her song* to varying degrees, suggesting one single voice on one side, and several voices on the other.

The lowest evaluations were found in the excerpt from *her reflection*, where the inharmonic interrelationship and the asynchronous onset and endings of the components (cf. *harmonicity* and *synchronicity of onsets and endings* factors) weakened integration, while the factors of *modulation coherence* and *familiarity with linguistic structures* prevented it from disintegrating altogether. And, some of the phrases in *her ritual*, which was evaluated as below *intermediate* for most of the other premises, were even evaluated as maximal for stream integration. One can also see from the figure that the overall tendency with a large spread in the evaluations for *her reflection*, and little variation in *her voice*, *her song* and *her self*, can be seen for this premise.

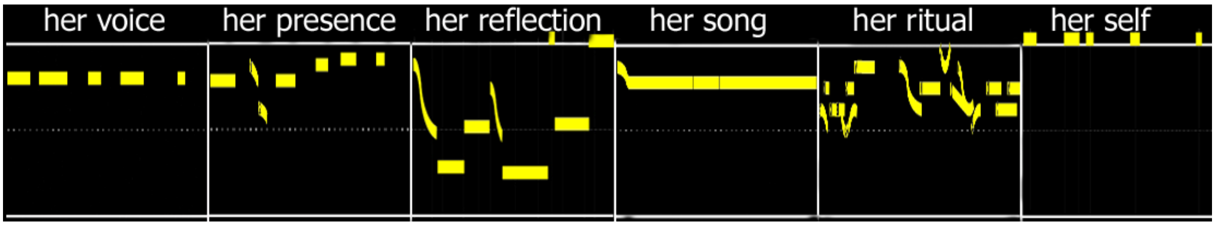


Figure 12.12: Time-compressed representations of all evaluations of the *stream integration* premise juxtaposed. Maximal and minimal evaluations are at the upper and lower lines, respectively.

While there are clearly contrasts in the evaluations for this premise, especially in *her reflection*, these might have been accentuated more by having phrases in which different levels of stream integration were contrasted against each other, or if there had been more marked gradual transitions from lower to higher stream integration or vice versa within the phrase. A hypothetical example of a transition made with the SFM instrument, going from a quite low degree of stream integration (mostly a result of top-down recognition of verbal content) to a well integrated ending, is presented in **sound example 12.18**.[367] In contrast to what was the case for the majority of the evaluated excerpts, this example uses primarily temporal dissimilarity in several parameters to weaken integration in the beginning of the phrase. Such a way of weakening integration will, however, also affect the coherence of the experienced vocal source, almost so that the phrase appears to consist of speech sounds glued together from different speakers. This would have been in conflict with the mentioned idea of attributing the same coherent vocal persona to all fantasies.

12.4.2 Overall comparisons

12.4.2.1 Overall tendencies

On the basis of the preceding sections, one can infer some overall tendencies in the evaluations for the six excerpts. The excerpts which are on the whole closest to the maximal voice are *her voice* and *her self*, where all premises except one or two for all the phrases are evaluated as *maximal-intermediate* or higher. It is difficult to judge which of these two is closest to the maximal, however, since they differ quite a lot in what premises depart from the maximal and how large the deviation is. This difference is very evident from the axial representations, where one can see that while *her voice* is characterized by a relatively even and symmetrical shape around the centre, the representations in *her self* are asymmetrical with a marked "peak" for *focus of attention* and *information density*, but where the remaining parts of the shape is almost entirely within the zone of the maximal voice. The excerpt from *her presence* is not too far from these two, at least if one only considers the vocal phrases. Still, one can see that the axial representations of the phrases have relatively similar shapes which together also set them relatively well apart from the two fantasies mentioned earlier. Here, the shapes are less symmetrical relative to the centre, with a peak for the *focus of the attention* premise, with the *information density* and *stream integration* premises giving most representations "slimmer" shapes down the sides. **LFDF** and **DF2** deviate somewhat from this pattern, with the former of the two approaching a more even shape like in *her voice*. If one

[367] Cf. SFM instrument, preset 17. Here, I have created a phrase in which eight short segments taken from one single phrase are given contrasting properties, including in pitch, amplitude, spectral shift, amount of reverberation and localization in the stereo image, so as to prevent the segments from integrating well into one stream, at least in the beginning. In addition, some of the segments overlap temporally, thus disobeying the regularity of synchronous onsets and endings.

also takes the sustained vowels into consideration, the overall picture of *her presence* is one of regular variations between this configuration and one with several premises evaluated as *minimal* for the sustained vowels. In the excerpt from *her song*, the tendency is that the evaluations lie between *maximal-intermediate* and *intermediate* for all premises but *focus of attention*, which stands out from the rest, having *intermediate-minimal* evaluation. Again, one can see that all the shapes of the axial representations are relatively similar for all phrases, with the beginning of **LCSRBF** as a notable exception. The shapes are clearly more irregular than *her voice*, with evaluations on the left side (*stream integration*, *salience*, *clarity of meaning*) being closer to the maximal than the rest, and the "peak" of the *focus of attention* premise standing out. To some degree, it resembles *her presence*, but the size is markedly larger, thus indicating evaluations further towards the minimal.

As for the excerpts from *her reflection* and *her ritual*, they contain the phrases which are on the whole closest to the minimal, even if both have a fair amount of variation between phrases as well as premises. This can be seen from the axial representations, which clearly have a larger size than the phrases in the other movements, albeit with **BS1** and **BS3** from *her reflection* as marked exceptions. In particular, a few of the phrases from *her reflection* appear to cover almost the whole space between the maximal and minimal zones, and more so towards the ending of the phrases, indicating that these phrases have the evaluations that are furthest towards the minimal in the piece. While the variation between phrases *within* these two excerpts are evidently larger than in the four other movements, one can still see that the shapes of the axial representations (except **BS1** and **BS3**) from *her reflection* are generally less pointed than in *her ritual*, where evaluations of *focus of attention* are closer to the minimal and *information density* and *stream integration* are closer to the maximal, hence together creating a "peak" for the *focus of attention premise*. One can note a similarity in shape *between* phrases in the two excerpts, however, with **PL1** and **KND1** in *her reflection* being not too far from **ATBBc** in *her ritual*.

Taken together, based on my listening experience, *Six Fantasies* appears to encompass a relatively large span of the continuum between the minimal and the maximal voice, having phrases which are not far from these extreme poles as well as phrases which cover the continuum in between, but where there is relatively large consistency within each excerpt. Therefore, it seems like my intuition that this piece contained phrases close to the maximal as well as the minimal voice corresponded fairly well with what the evaluation process has shown in the end. I will expand on the ranges of evaluation in section 12.4.2.1.

One could perhaps imagine compositional strategies that would have extended this span even further, e.g. by choosing a text more easily and directly comprehensible. Given Campion's text as a precondition, however, presenting the poem in an unmanipulated and clearly articulated version with minimal accompaniment early in the piece, maybe with a male reciting voice to avoid any ambiguities regarding the lyrical subject and its relationship to the female subject, would probably have given an evaluation even closer to the maximal. In the opposite direction, one can imagine that by weakening the link to the structure and meaning of the poem, e.g. by recombining and restructuring words and lines to create new meanings and structures, many phrases of the poem could probably be evaluated even further towards the *minimal*, especially for the *clarity of meaning* and *focus of attention* premises.

12.4.2.2 Range of evaluations

Within the wide range of evaluations for the chosen excerpts, there are differences regarding the individual premises and the individual fantasies. For *information density*, *naturalness*, *presence* and *feature salience* the whole continuum was covered.[368] For the remaining premises, the evaluations did not cover the whole range, but at least more than half of it.

When looking at all the phrases I have considered, it seems pretty clear that **BS1** from *her reflection* stands out as the phrase which is closest to the maximal, with most premises being evaluated as maximal or not very far from it. This is especially evident from the axial representation, where one can see that the area covered is just a little bigger than the maximal zone in the centre. Compared to the phrases in *her self*, which also had many premises proximate to the maximal voice, the relatively low evaluations for attention and information density still distinguished all these phrases from **BS1**, notwithstanding that the phrases in *her self* had higher judgments for presence. Anyhow, since **BS1** so clearly stands out in *her reflection*, this phrase might be interpreted as a kind of early premonition of what is to come in the final movement.

On the minimal side, the phrases that appear to come closest are the type b phrases **PL2** and **KND2** from *her reflection*, for which three of the premises are evaluated as *intermediate-minimal* and the rest are evaluated as minimal. This is also very evident from the large size and close to circular shape of the axial representation. One therefore sees that *her reflection* is a movement that contains phrases that lie pretty close to each of the extremes.

[368] If we don't include that the noise mode of the information density premise was only very sparingly represented.

Consequently, *her reflection* might be assigned a function of “opening up”, so to speak, the space between the maximal and the minimal for the listener in this piece by presenting phrases close to the extremes within a relatively short time span.

The range of evaluations for each of the six movements is represented in **figure 12.13**, with the length and placement of the vertical bars indicating the range of the evaluations. Here, one sees that in some movements the evaluations span a narrow range, while in others they cover a wide range. Clearly, *her reflection* stands out as the fantasy with the largest range of evaluation, with the majority of premises covering all or almost all of the whole range from maximal to minimal. For *her ritual*, the evaluations cover the whole range from maximal to minimal when seen together, but the majority vary in the range between the *intermediate* (or just above) and the *minimal*, whereas only two premises cover the range up to the *maximal*.

For the remaining fantasies¸ the evaluations vary markedly less. In both *her presence* and *her song*, about half of the premises don’t vary or vary only slightly, whereas the other half of the premises vary across at least one category (of the five) on the continuum. For *her presence*, three of the premises even cover a larger span than this. Lastly, *her voice* and *her self* have negligible variation or no variation at all. Hence, the excerpts appear to differ greatly in terms of the range of evaluations, from very wide to very narrow; where *her ritual* probably falls somewhere in between these two extremes, and *her song* and *her presence* are clearly more narrow than wide.

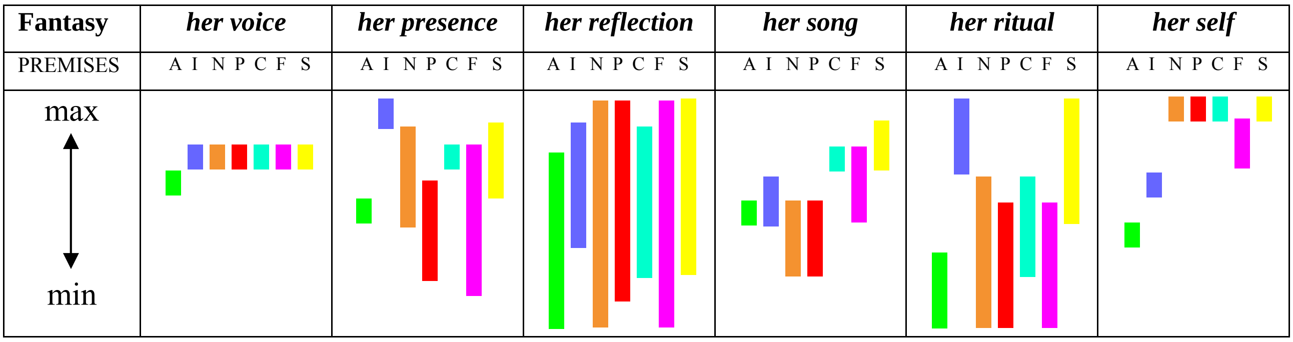


Figure 12.13: Range of the evaluations of all premises for the six excerpts from *Six Fantasies*. For *her presence* and *her self*, the sustained vowels are not included. The premises are shown as coloured bars where the height of the bar corresponds to the range within which the premise varies. The key to the premises is presented in abbreviated form in the second row: A = *focus of attention*, I = *Information density*, N = *Naturalness*, P = *Presence*, C = *Clarity of meaning*, F = *Feature salience*, and S = *Stream integration*. The colours of the vertical bars correspond to the nomenclature presented in 12.2.3.

12.4.2.3 Temporal tendencies

One thing that is relatively obvious from the graphical representations of the evaluations and the accompanying discussion is that the excerpts from the different fantasies differ to a relatively large degree when it comes to the amount of variation/constancy for the evaluations. As for the variation *between* phrases, one can see, for example, that the excerpts from *her voice*, *her song* and *her self* have relatively little variation from phrase to phrase, something which was also clearly seen in **figure 12.13**. Conversely, the excerpts from *her reflection* and *her ritual* are characterized by much larger differences between the phrases. The variations that appear most notable are found in *her reflection*, in which the evaluation changes from *minimal* for most premises to *maximal* for most premises from one phrase to the next (**KND2** – **BS1**). In *her ritual*, the variation in the evaluation is less than in *her reflection*, but markedly higher than in the remaining fantasies. *Her presence* appears to situate itself between these two groupings, however, if one considers only the vocal phrases, the variation seems to be comparable to that in *her song*, with **FC2** being the phrase that departs the most from the overall picture.[369]

If one looks at the temporal evolution of the evaluations *within* the individual phrases, one sees that the excerpts can be grouped almost in the same manner. For *her voice*, *her presence*, *her song* and *her self*, the dominant tendency is that the evaluations are constant, with mostly vertical lines in the time varying representation and with high similarity in shape and size for the axial representations.[370] In *her reflection* and *her ritual*, on the other hand, a significant portion of the evaluations change during the course of the phrases, with the type a phrases in *her reflection* probably have the largest change in the evaluations. This can be seen from the diagonal and curved lines in the time varying representations and large differences in size and shape for the axial representations.

From the time varying representations one can also observe that the direction and the regularity of these changes are quite different in these movements. In *her reflection*, the majority of the phrases with variable evaluation has a descending overall tendency, largely due to the decaying resonances. In contrast, in *her ritual* they have both ascending and descending curves, as well as convex and concave curves.

[369] However, if one includes the static sustained vowels in the picture, the variation becomes formidable, since so many of the premises are evaluated as approaching *minimal* for these sounds and since the vocal phrases are relatively similar to that of *her voice*.

[370] Again, when taking the sustained static vowels into account for *her presence*, the picture is somewhat different, since these are characterized by having evaluations for information density and attention which are consistently having a descending slope.

Taken together, the excerpts from *her voice*, *her presence*, *her song* and *her self* contain little variation, while *her reflection* and *her ritual* contain a lot of variation, both when it comes to variation *between* phrases and *within* them. It is interesting to relate this to Lansky's own statements on the "journey" that listening to the piece from beginning to end can represent:

> It's a jagged journey. You seem to almost have a direction with the first two. Then the third one takes you in another direction and then the fourth one takes you in the opposite direction and then the fifth one takes you into outer space. The sixth one brings you back home (Lansky, cited in Ondishko, 1990)

Clearly, my listening experience can also in a sense be seen as a jagged journey, which to some degree parallels this description, with movements back and forth along the *maximal-minimal* continuum from the first to the last fantasy. Here, the tendency for the evaluations to move slightly in the minimal direction from *her voice* to *her presence* can be regarded as a parallel to following a "direction with the first two". Moreover, most of the evaluations in *her song* are higher than in *her reflection*, hence giving the fourth fantasy "an opposite direction". Finally, one sees also that from the fifth to the sixth, *her self*, one moves towards evaluations close to where one started out, that is, "back home". However, since *her reflection* includes phrases that are localized along the whole max-min continuum, it is difficult to see that *her reflection* can represent "another direction" in my case. Moreover, *her song* only represents the "opposite direction" from the previous movement as long as one disregards the large variation in the evaluations in *her reflection*. All in all, despite that Lansky had other aspects in mind than me when making his statement, and that I have only taken shorter excerpts from the fantasies into consideration, there still seems to be a relatively good correspondence with his statement and my evaluations.

12.4.2.4 Grouping of the fantasies

Another question that can be worth addressing is whether the evaluations suggest that the fantasies are structured in pairs. This is suggested by David Loberg Code, who sees that *Six Fantasies* can be perceived as three pairs of movements; 1 and 6, 2 and 4, and 3 and 5 on the basis of timbre, intelligibility and the closeness to natural speech (Code, 1990: 162). Considering the evaluations of the premises that would correspond best to those criteria, namely *naturalness* and *clarity of meaning*, one sees that my evaluation gives only limited support to Code's suggested pairing: The evaluations of *her voice* and *her self* for these premises are not more similar than *her voice* is similar to *her presence*. As for *her reflection* and *her ritual*, they can be grouped on the basis of being the two fantasies which have the most temporal variation in the evaluations, and having the majority of the phrases with evaluations between the *intermediate* and the *minimal*, something which is evident from the markedly larger size of the axial representations. However, a more interesting claim by Code is that "within this larger framework [of the three pairs of movements], the inner pairs function as two divergent paths along multiple dimensions, both evolving from and returning to the outer pair of movements [constituted by *her voice* and *her self*)" (Code, 1990: 166). For a majority of the premises, i.e. *naturalness*, *presence*, *feature salience*, *clarity of meaning* and *stream integration*, this description seems indeed pertinent. On the whole, however, I regard such attempts of grouping interesting mostly because they can highlight similarities or differences between how one experiences the fantasies – not because they reveal some hidden intrinsic connection between them. And, while regarding such similarities and differences allows one to construct groups or categories, there are also certainly ways of regarding the fantasies so that their differences appear more continuous. It all depends on perspective and resolution, or, in my case on the premises applied in the evaluation and on the number of phrases one regard together.

12.5 Chapter conclusions

In this chapter, I have hopefully been able to demonstrate how the framework developed during the course of this thesis can be used to evaluate the experience of voice in a particular piece of electroacoustic music. Paul Lansky's *Six Fantasies on a Poem by Thomas Campion* has proven to be an interesting work to assess with this framework, even if I was forced to make a selection of excerpts from each of the movements, thereby delimiting the scope of the evaluation.

Part of the process of using this framework has also been to present the evaluations in a graphical form, with the .aks files made with the acousmographe software as the main representation. The two types of representations that I have been presenting have given me the opportunity to get an overview of the time-varying changes in each of the premises, whereas the axial representations have given a good overview of the evaluations of single premises and their interrelationships at particular times. Since the time-varying representations were more difficult to get an overview of when all the premises were presented together, these representations have been complementary, and together they have given me the possibility to compare evaluations within and between phrases as well as within and between excerpts. Moreover, by presenting these evaluations within an interactive environment such as the acousmographe, where graphical layers can be turned on and off according to one's needs and where one can apply the tools for selective playback both in the time-domain and the frequency domain, one has a high degree of flexibility in viewing/listening.

With a few notable exceptions, the differences in the evaluations *between* the six excerpts were particularly prominent, something that was evident from both types of graphical representations. Firstly, we saw that the shapes and sizes of the axial representations were quite similar within the excerpts, with some having smaller variations around one consistent configuration, while others showed variations among several configurations. Secondly, the excerpts differed greatly in the *ranges* that the evaluations spanned, from those covering a minimal range to those covering almost the whole range from maximal to minimal for the majority of premises. Thirdly, the excerpts varied in the degree that the evaluations *varied* over time, where some had dominantly constant evaluations while especially the excerpts from *her reflection* and *her ritual* had more variation, either between the phrases or within the phrases. Here, the majority of phrases in *her reflection* had descending evaluations mainly due to the slowly decaying resonances that caused the endings of the sound to bear only faint similarity to vocal articulation, while *her ritual* had less predictable variation, especially for the percussive sequences that dominated the excerpt. The variations were only rarely presented as *transformations*, though, except the percussive sequences in *her ritual*, which vacillated between non-voice and minimal voice. On the whole, the vocal phrases were experienced as relatively consistent and without drastic temporal transitions in features that might have introduced sudden shifts in the experience.

There were also similarities across excerpts, something which provided a basis for grouping the fantasies. Here, the most prominent difference was between the excerpts from *her reflection*/*her ritual* and the remaining excerpts – the former two being overtly marked by the use of banks of comb filters creating resonances that appeared to be external to the voice. As for the remaining fantasies, the similarities in the evaluations seen together did not present any clear cues for grouping, perhaps surprisingly, since *her self* was not evaluated as maximal for all premises. But, the recognition of LPC processing in *her voice*, *her presence* and *her song*, can make these excerpts stand out in contrast to the close to unprocessed voice in *her self*, something that was apparent from the differences in *naturalness* and *presence* evaluations.

In evaluating *focus of attention* for the six excerpts, we have seen that despite that the piece had its basis in a poetic text, **LI/sem** did not receive attention exclusively. Rather, the **SQS-** and **VG-domains** in particular came increasingly into focus during the course of the composition, while the **LI/sem** on the whole received less and less attention. This downplaying of the text and highlighting of its presentation was caused by several issues, among which the meaningful variations in the **SQS-** and **VG-domains** and the redundancy created by text repetition were perhaps the most important. This gradual re-direction of focus was also in line with Lansky's intentions of making explicit the implicit music within Hannah MacKay's reading. Playing with different vocal styles/utterance modes and voice qualities in the different movements was also a part of this picture; from speaking/reciting via speak-singing to singing, using modal or whispered voice quality. In that respect Lansky's piece falls into the tradition of extending the use of voice in music, advocated particularly by composers like Schoenberg, Berio and Ligeti (cf. Anhalt, 1984). Moreover, one can see that my experience of *Six Fantasies* was not too far from what was indicated by Segnini and Ruviaro (Segnini & Ruviaro, 2005, cf. section 4.4), moving between "speechness" and "musicness". However, whereas I surely focused on the musical qualities of many phrases, I also heard the musical qualities of speech in the last fantasy, something that would seem contradictory to these authors' "music-language sonic space". Moreover, I did also hear musical qualities in phrases with intelligible text, especially in *her song*, hence suggesting that I depart from Segnini and Ruviaro's analysis, which suggested that the piece displayed "musicness" only in the portions with unintelligible text.

Moreover, we saw that the **ID-domain** received little attention mainly because the identity of the vocal persona was not "stretched" or "bent" very much, thus constituting a relatively stable identity base in the piece. This contributed in binding together all the six fantasies and provided the link to the "her" common to titles of the fantasies. That the vocal phrases did not cross the boundaries of human vocalization to a significant degree might also have contributed to this stability. Nevertheless, the multiplications of this identity by juxtapositions and superimpositions in *her voice*, *her presence*, *her reflection*, *her song*, and *her ritual*, created some ambiguity regarding the *number* of voices.

The **TCM-domain** also received little attention, somewhat surprisingly perhaps, since many phrases appear heavily manipulated and have correspondingly low evaluations of *naturalness*. The relatively consistent use of the LPC and comb filter techniques in the fantasies, along with few and relatively faint technological artifacts may be the reason for this.

Throughout the evaluation process, I have made repeated references to the factors presented in the premise chapters. In that way, I have hopefully shown how the evaluations have been related to aspects of the vocal phrases, to perceptual and cognitive issues, and to my previous experience, competence and knowledge.

The application of the SFM instrument in the evaluation process and discussion, which was able to emulate the sounds that occurred in Lansky's piece *along with* creating numerous variants or even heavily deviating sounds, was a part of an *analysis by synthesis* approach adopted in parts of the thesis. This gave me the opportunity to demonstrate how different factors could affect the evaluations. For example, I made sounds with this instrument that demonstrated that one could get sharper *contrasts* and temporal transformations with different tendencies for the evaluation of some of the premises. Furthermore, I synthesized examples that resulted in evaluations beyond the range found in the excerpts for other premises. In that way I was also able to demonstrate how certain control parameters such as e.g. time-stretch factor affected the evaluation of the premises, although this was not done on a systematic basis.

What is almost equally important as the presented evaluations, however, is the knowledge and experience I have gained while engaging in these evaluations. While consciously trying to monitor aspects of the listening experience, I have gained a new awareness regarding the many facets of the listening process, something that I will hopefully gain from when listening to other pieces, both those that are new to me, as well as those that I have heard a number of times before. Hopefully, I have been able to communicate some of this to the reader, so that he or she might appreciate the many nuances of the different kinds of voices in electroacoustic music, and that the current framework may also be a tool for describing and understanding their experiences as well as mine.

13.0 Epilogue

13.1 Summary and conclusion

During the course of this dissertation, I have established a framework for describing and understanding the experience of voices in acousmatic electroacoustic music and related genres. The framework has been developed with a basis in a phenomenological approach, where I have used my own listening experience as the main object of study. Still, by having an interdisciplinary outlook, I have been able to relate aspects of the framework and my own application of it on excerpts from electroacoustic works to issues within perception, cognition, acoustics, composition, physiology and several other fields.

The framework has consisted of two partly interconnected components. The first of these were what I called *experiential domains*. These domains designated a number of aspects or properties of an experience that can be grouped together based on some common feature, relationship or function, and they were the focus of chapter two and three. The second component was the *maximal-minimal model*, which was the main focus of chapters five through eleven.

In chapter two, I began with giving an account of the experiential domains that were applicable to acousmatic electroacoustic works in general, namely the domain of sound qualities and structures (**SQS-domain**), the domain of technology, composition and mediation (**TCM-domain**) and the domain of space and environment (**SE-domain**). These three non-vocal domains were all considered to have an *intrinsic* and *outward* orientation, meaning that they embraced aspects that were experienced as directly related to the musical work and that attention was directed outwards towards the incoming sensory information from the outside world rather than towards one’s own bodily and mental response to this information. After substantiating the distinction of these domains with reference to relevant theories, I then set forth several terms with which each of the domains could be qualified, thus providing means of description of pertinent aspects of the experience.

Subsequently, the three experiential domains were seen in relation to each other and to the *vocal* experiential domains. One important aspect of this relationship was the distinction between reference-oriented and quality/structure-oriented domains, the former comprising the **TCM-**, **SE-** and vocal domains, the latter comprising the **SQS-domain**. I argued that this distinction was not without a grey zone, so that the boundary between the two was difficult to define clearly. Similarly, there were also many ways in which the experiential domains were interconnected, and the boundaries between them were not always easy to define clearly. The relationship between the reference-oriented domains also had certain hierarchical traits, in that a recorded voice necessarily will have to be situated in a space or environment (**SE-domain**), which in turn will have to be the mediated by technological means (**TCM-domain**). In the last part of chapter two, however, I argued that that this basic configuration was in many cases made more complex, in that several layers of space and technology could be nested into each other, hence creating what I called different *ontological levels*, i.e. frames of existence within which one experiences objects and beings to be situated. Lastly, I proposed that the reference-oriented domains could also be qualified by the concepts of *virtuality*, *documentary* and *realism*, which dealt with different aspects of the relationship with the real world, and with the experience of degrees of constructedness.

In chapter three, I introduced the term *vocal persona*, which designates the virtual person or character that one perceives as the source of the voice, and this concept remained central throughout the remaining parts of the dissertation. The main focus in the chapter, however, was the four *vocal* experiential domains, namely 1) the vocal gestures (**VG**) domain, 2) the identity (**ID**) domain, 3) the affective (**AF**) domain, and 4) the linguistic (**LI**) domain. These domains were all presented so as to provide a set of terms that could be helpful in qualifying and describing aspects of the experience of vocal sound. Moreover, I also wanted to provide a link between these aspects and other closely related features – mainly physiological, acoustic and socio-cultural – so as to provide a basis of understanding the interplay of factors contributing in constituting the multi-layered meanings conveyed by voice. By referring to studies that demonstrated how different types of electronic or computer processing of vocal sound affected perceptual and cognitive processes related to each of the vocal domains, I could also postulate how similar types of processing within electroacoustic music could affect the experience of different aspects belonging to the four domains.

In the fourth chapter, I introduced the maximal-minimal model, the second component of the framework I have developed in this thesis. The main idea with this model was that the wide range of vocal and vocal-like sounds in electroacoustic music could be compared and judged against two reference *zones*, constituted by *maximal* and *minimal* voice, which were seen as the extreme poles on a continuum which comprised this whole range. Here, the maximal voice paralleled the informative and clearly articulated speaking voice dominating the radio medium, but was described in this framework by a set of *premises*, i.e. partly interconnected conditions related to one particular aspect or feature of the experience of voice. When these conditions were all fulfilled, they jointly defined the maximal voice. I argued that maximal voice in its fullest sense was rarely found in electroacoustic works, but that vocal expressions *close to* it were quite common. At the other end of the continuum, *minimal* voice was defined as a boundary zone between what is experienced as voice and what is experienced as not voice, which was related to the negative fulfilment of the mentioned premises. The minimal voice was more difficult to define clearly than the maximal voice, however, since there was potentially a multitude of ways in which voice could cross the boundary into not voice, and since only some of the premises appeared to allow for transgressions into non-voice. I then argued that maximal and minimal voice could be thought of as centre and periphery, respectively, and that the premises could be represented as axes running from the centre towards the periphery, thus having similar structure to prototypical categories and cluster models as described within prototype theory.

The following seven chapters were all dedicated to describing and exemplifying the seven premises of the maximal-minimal model. All chapters aimed at 1) giving a more in-depth description of the premises, 2) of relating the premise to relevant research so as to substantiate the inclusion of the premise in the model, 3) setting up a list of factors that could potentially affect the evaluation of the premise, 4) presenting the criteria for making an evaluation of a vocal phrase from an electroacoustic work, 5) presenting examples of evaluations in the range between maximal and minimal of excerpts from electroacoustic music and the factors that contributed in the evaluations, and 6) integrate the framework of the experiential domains in the presentation of the premises and in the description and evaluation of the sound examples. The premises presented in these chapters were 1) *Focus of attention*: The semantic level of the linguistic domain receives sustained and maximal attention; 2) *Information density*: The information density of the experiential domains is optimal for the processing/decoding of the **LI-domain**; 3) *Naturalness*: The sound has maximal resemblance with one produced by a human being and his/her vocal apparatus; 4) *Presence:* The listener experiences a sense of a shared 'here and now' with a vocal persona; 5) *Clarity of meaning:* Meaning can be constructed from the voice with a high degree of clarity – implying also specificity, certainty and coherence; 6) *Feature salience*: Vocal sounds and features "stand out" perceptually – for themselves and relative to other sounds and features; and 7) *Stream integration*: The sound of the voice is integrated into one coherent and continuous sound stream.

Finally, in chapter twelve, the two integrated components of the framework were applied in an evaluation of my own listening experience of Paul Lansky's *Six Fantasies on a*

Poem by Thomas Campion from 1979. In this chapter, the maximal-minimal model contributed with the central criteria in the evaluation, while the experiential domains and the related qualifying and descriptive terms provided more underlying and implicit grounds in this process. Excerpts from the six movements, or *fantasies*, were each given an overall description and an evaluation according to the seven premises of the model. In addition, the evaluation of the premises was linked to several factors linked to properties of the vocal phrases, perceptual and cognitive issues, and my own experience, knowledge and competence as a listener. These evaluations were presented in a graphical form using the acousmographe software, which allowed for a selective viewing of the evaluations while listening interactively to the excerpts. The evaluations were presented in two graphical forms: The first displayed the evaluations as horizontally oriented lines or curves that were drawn between an upper line representing the maximal voice, and a lower line, representing the minimal. The second was an axial representation where the evaluations were marked along the seven axes each representing one of the premises of the model, running from maximal evaluation in the zone enclosing the crossing point of the axes in the centre, towards the minimal zone in the periphery of the representation.

All in all, the framework proposed in this dissertation has provided a systematic way of assessing aspects of the listening experience *beyond* the qualities of the sound itself, hence expanding the more established schaefferian approaches (cf. chapter two), while retaining the emphasis on close listening central to these approaches. The framework has also provided a terminology covering a wide range of aspects that can be inferred from the voice in listening, something which has facilitated description and qualification of vocal phrases. In discussing and applying the framework, it has also become clear that what I called the *dynamics of listening*, i.e. the changes in the listening experience caused by subsequent listenings, and the knowledge and background of the listener are particularly important for several premises and experiential domains in the framework. In comparison to schaefferian approaches, the application of the framework is therefore more *relative* to the individual listener and the listening context/situation.

Similar to the schaefferian approaches, my approach has been dominantly focused on a shorter temporal scale, i.e. on the level of the *vocal phrase*. This has been a deliberate choice to delimit the scope of the project so that what was taken into consideration could be treated with sufficient depth and thoroughness in the dissertation. One result of this choice is that I have mainly worked with *excerpts* from pieces, rather than whole pieces, something which has delimited the context within which I have dealt with the listening experience. For the same reasons, I have also mainly focused on a "domestic" kind of listening experience, mostly disregarding the effects that a public presentation may have on the experience. Hence, these issues might give directions for some future developments of the framework, which I will discuss in more detail in section 13.3.

13.2 Relevance

As I see it, the current study is first and foremost relevant within the field of electroacoustic studies. The framework proposed here can in principle be applied on listening experiences of any electroacoustic work with vocal elements, both for qualifying and describing the aspects belonging to the different experiential domains and for evaluating it according to the maximal-minimal model. Thereby, it can have great relevance for researchers interested in studying the multiplicity of meanings that can arise in listening to voice in electroacoustic music. I also believe that the framework can be applied in conjunction with other approaches taking the reception (esthesic) point of view which are more general in scope, like for example those of Lasse Thoresen (Thoresen, 1985; Thoresen, 2007b; Thoresen, 2007a) and Stéphane Roy (Roy, 1996; Roy, 1998; Roy, 2003). Also, I think that my framework may well be combined with less systematic and more hermeneutically oriented approaches, such as can be found in the writings of Katharine Norman (Norman, 2000; Norman, 2004a) and Lawrence Kramer (Kramer, 1984; Kramer, 1990). In many ways, using the framework developed in thesis makes the listener consider a great many aspects of the experience, and this might subsequently be taken as a basis for a more comprehensive and less systematic interpretation of a certain piece.

I also believe that the two frameworks that I have developed and many of the insights that have come up during the course of the thesis can have relevance for composers and live performers in the electroacoustic domain working with voice. For instance, a composer might be made more aware of which voice-related features in a composition that may potentially draw the listener's attention, and he or she may become more conscious of the effect of the listener's background on their experiences. The link between acoustical parameters and experiential features that were presented as *factors* for the premises in the framework can also be of relevance to composers, in that it may guide an exploration of techniques of synthesis, processing and organization of material. Performers working with electronics and voice may also become more conscious about the potential range of meanings and effects the electronically mediated and processed voice can convey.[371]

The presented frameworks can furthermore prove to be useful as pedagogical methods or tools in dealing with music experience, especially those related to voice. This is something that I have personally already experienced in teaching graduate courses dealing with aesthetic analytical perspectives on music and technology.

Even if the framework developed in this dissertation deals specifically with voice in electroacoustic music, it may also be applied – albeit perhaps with some modifications – to other expressions involving voice mediated by loudspeakers. It may provide researchers with a tool for describing and understanding the experience of works within sound art or radiophonics that explore the expressive ranges of voice(s) and technology. Furthermore, my framework may also be relevant for researchers within popular music studies, in dealing with explicitly 'artificial' vocal effects and sampling of vocal sound, which has expanded greatly in pop music in the last decades (see e.g. Middleton, 2006; Lacasse, 2000).

Lastly, the chapter on Lansky's *Six Fantasies* may have relevance to scholars or students studying Lansky's music and its reception. Here, the SFM instruments can be an interesting tool for demonstrating the possibilities and limitations related to LPC as a technique for musical composition. And, the SFM instrument can itself be applied as a compositional instrument for composers and sound designers interested in exploring the sound world of a technique which in today's musical world might seem obsolete and outdated, but which nevertheless provides a sonically rich palette of expressions.

13.3 Future research

There are several ways in which the work presented in this dissertation can be further developed in the future. Studies of electroacoustic music taking a more systematic empirical approach to listening have hitherto been relatively few in number.[372] One of the most interesting ways to follow up this study, in my view, would be to follow such a path and involve groups of listeners, both to assess the pertinence of the framework and to make evaluations *according to* the framework. Such studies could involve more open qualitative

[371] Naturally, live performers will also have to take their presence on stage into account (cf. Emmerson, 2007: chapter 4).

[372] Here, Landy and Weale's *Intention and Reception* project probably represents the most developed approach (Weale, 2005; Weale, 2006; Landy, 2006). The efforts of Delalande (Delalande, 1998) and Payri should also be mentioned (Payri & Bono, 2007).

methods, similar to those of Weale and Delalande (Weale, 2005; Weale, 2006; Delalande, 1998), as well as more controlled quantitative ones, using different types of musical material from pre-existing compositions, to compositions made for the purpose and material where parameters are controlled and varied in a more restricted manner. In that way, one can produce results that may guide the further development and/or refinement of the framework. Moreover, one can investigate how different parameters/aspects pertaining to voice and the context within which it is presented may affect the evaluation of the premises of the max-min framework. Here, the many tables in chapter 3 presenting reviews of research on different aspects of the voice and the listing of the many *factors* in the premise chapters can potentially function as a guide for possible parameters/aspects to investigate. The effects of the dynamics of listening and previous listener experiences can also be considered included in this type of investigation. This implies getting a firmer intersubjective basis for the frameworks developed here, as well as extending the scope of systematic empirically oriented studies of listening in electroacoustic music.

Another interesting way of extending the research presented in this thesis would simply be to apply the framework to more pieces of music. This can produce a basis for making several kinds of interesting comparisons – between different listeners with different backgrounds, between different works with different types of vocal material and techniques of processing, as well as between sections within the same works. This can potentially open up for discussions that can be interesting both from a theoretical and from a composer's perspective.

The framework in itself can also be further developed theoretically in the future. For instance, the social aspects involved when several voices are heard interacting were not taken into consideration in my framework. And several of the non-vocal domains could have been expanded with further aspects for description and qualification. Furthermore, the graphical representation of the evaluations could be improved. The problematic issues of not being able to show the distinction between the *reduced* mode and the *noise* mode of the minimal for the axial representation of the *information density* premise could possibly be solved by investigating other representations, for instance using three dimensions instead of two. Three dimensions would also enable showing the development of the axial representation *in time*, something which might potentially carry a lot of information in one single representation.

Lastly, I see that my investigation of Lansky's *Six Fantasies* can be further developed in the future, especially related to the SFM instrument, which was originally thought to play a more prominent part in this thesis. By proceeding with the analysis-synthesis approach that I took in parts of this thesis, and further extending the possibilities for user interaction, one could end up in something similar to Michael Clarke's presentation of Harvey's *Mortuos Plango Vivos Voco* (Clarke, 2005b). Here, Clarke allowed the user to play back parts of the composition while presenting spectral and structural features in conjunction to the parts, and to play with parameters of synthesized bell sounds, similar to the ones in the piece. Taking this direction, the SFM instrument could similarly be included in a more comprehensive interactive multi-media presentation of the piece, including material related to the composition process such as interviews, notes and programming code. This would to allow for an exploration of the piece that could complement written material and probably make the research more accessible and entertaining.

I began this dissertation by telling about my first experience with voice in electroacoustic music – an experience that was in many ways *enigmatic*. While I feel that I have greatly increased my own understanding of the voice and how it is presented and experienced in this genre of music, the enigma surrounding these phenomena and my urge to understand them better has nevertheless persisted, and I therefore hope that this dissertation represents not a closed chapter, but the beginning of new and intriguing explorations.

References

Literature

Abramson, A. S. and L. I. Lisker (1967). "Discrimination along the voicing continuum: Cross-language tests". *Proceedings of 6th International Congress of Phonetic Sciences*, Prague.

Abramson, A. S. and L. I. Lisker (1968). "Voice Timing: Cross-Language Experiments in Identification and Discrimination." *The Journal of the Acoustical Society of America* **44**(1): 377.

Adorno, T. W. (2002). "The aging of the new music." In *Essays on Music*. R. Leppert (ed.). Berkeley, University of California Press**:** 181-202.

Akagi, M. (2002). "Perception of fundamental frequency fluctuation". *Forum Acusticum 2002*, Sevilla, Spain.

Allen, J. B. (1994). "How Do Humans Process and Recognize Speech?" *IEEE Transactions on Speech and Audio Processing,* **2**(4): 567-577.

Allport, D. A., B. Antonis, et al. (1972). "On the division of attention: a disproof of the single channel hypothesis." *The Quarterly journal of experimental psychology* **24**(2): 225-35.

Amedi, A., G. Jacobson, et al. (2002). "Convergence of Visual and Tactile Shape Processing in the Human Lateral Occipital Complex." *Cereb. Cortex* **12**(11): 1202-1212.

Anderson, E. L. (2007). "Materials, Meaning and Metaphor : Enveiling Spatio-Temporal Pertinences in Acousmatic Music". *Web proceedings of the EMS2007 : Electroacoustic Music Studies Network*, De Montfort, Leicester.

Andersson, A., A. Eriksson, et al. (1996). "Cries and whispers: acoustic effects of variations of vocal effort." *TMH-QPSR* **37**(2): 127-130.

Anhalt, I. (1984). *Alternative voices: Essays on contemporary vocal and choral composition*. Toronto, University of Toronto Press.

Aoki, N. and T. Ifukube (1996). *Two 1/f fluctuations in sustained phonation and their roles on naturalness of synthetic voice*. Proceedings of the Third IEEE International Conference on Electronics, Circuits, and Systems, ICECS '96., Rhodos, Greece.

Apple, W. and R. M. Krauss (1977). "Effects of pitch and speech rate on personal attributions." *Journal of personality and social psychology* **37**(5): 715-727.

Arons, B. (1993). *Techniques, Perception, and Applications of Time-Compressed Speech*. Proc. Conf. American Voice I/O Society.

Arslan, L. M. and D. Talkin (1997). *Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum*. Proceedings of Eurospeech'97, Rhodes, Greece.

Assmann, P. F. and T. M. Nearey (2003). *Frequency Shifts and Vowel Identification*. Proceedings of the 15th International Congress of Phonetic Sciences, ICPhS, Barcelona.

Assmann, P. F., T. M. Nearey, et al. (2002). *Modelling the perception of frequency shifted vowels*. Proceedings of the 7th Int. Conference of Spoken Language Perception, ICSLP, Denver, Colorado.

Atal, B. S. (2006). "The History of Linear Prediction." *IEEE Signal Processing Magazine*(March): 155-161.

Atal, B. S. and S. Hanauer (1971). "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave." *The Journal of the Acoustical Society of America* **47**(2B): 637-655.

Banse, R. and K. R. Scherer (1996). "Acoustic profiles in vocal emotion expression." *J. Pers. Soc. Psychol.* **70**: 614-636.

Banziger, T. and K. R. Scherer (2005). "The role of intonation in emotional expressions." *Speech Communication* **46**(3-4): 252-267.

Barrett, N. (2003). "Spatio-musical composition strategies." *Organised sound* **7**(3): 313-323.

Barrière, J.-B. (1984). ""Chréode":the pathway to new music with the computer." *Contemporary Music Review* **1**: 181-201.

Barthes, R. (1991). "The Grain of the Voice." In *The responsibility of forms*(ed.). Berkeley, University of California Press**:** 267-277.

Battier, M. (2003). "A constructivist approach to the analysis of electronic music and audio art - between instruments and faktura." *Organised Sound* **8**(3): 249-255.

Bayle, F. (1989). "Image-of-sound, or i-sound: Metaphor/metaform." *Contemporary Music Review* **4**: 165-170.

Belin, P., S. Fecteau, et al. (2004). "Thinking the voice: neural correlates of voice perception." *Trends in cognitive sciences* **8**(3): 129-35.

Belin, P., R. J. Zatorre, et al. (2000). "Voice-selective areas in human auditory cortex." *Nature* **403**(6767): 309-312.

Bella, S. D., J.-F. Giguere, et al. (2007). "Singing proficiency in the general population." *The Journal of the Acoustical Society of America* **121**(2): 1182-1189.

Bergsland, A. (1999). *Vævet av "Roser" - : Olav Anton Thommessens "Vævet av stængler" som lesning av Obstfelders "Roser"*. **MA**. Trondheim, NTNU.

Berti, S., U. Roeber, et al. (2004). "Bottom-Up Influences on Working Memory: Behavioral and Electrophysiological Distraction Varies with Distractor Strength." *Experimental Psychology* **51**(4): 249-257.

Berti, S. and E. Schröger (2003). "Working memory controls involuntary attention switching: evidence from an auditory distraction paradigm." *European Journal of Neuroscience* **17**(5): 1119-1122.

Bey, C. and S. McAdams (2002). "Schema-Based Processing in Auditory Scene Analysis." *Perception & psychophysics* **64**: 844-854.

Bigand, E. (1993). "Contributions of music to research on human auditory cognition." In *Thinking in Sound - The Cognitive Psychology of Human Audition*. S. McAdams and E. Bigand (ed.). Oxford, Oxford University Press.

Blesser, B. and L.-R. Salter (2007). *Spaces Speak, Are You Listenting? : Experiencing Aural Architecture*. Cambridge, Mass, MIT Press.

Bolter, J. D. and R. Grusin (2000). *Remediation. Understanding New Media*. Cambridge, Mass., MIT Press.

Bonada, J., O. Celma, et al. (2002). *Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models*. Proceedings of the International Computer Music Conference, Havana, Cuba.

Bonnel, A. M., F. Faita, et al. (2001). "Divided attention between lyrics and tunes of operatic songs: Evidence for independent processing." *Perception & Psychophysics* **63**: 1201-1213.

Bosma, H. (1995). *Male and Female Voices in Computer Music*. International Computer Music Conference 1995, Banff, Canada.

Bosma, H. (2003). "Bodies of evidence, singing cyborgs and other gender issues in electrovocal music." *Organised Sound* **8**(1): 5-17.

Bossis, B. (2004). "Reflections on the analysis of artificial vocality: representations, tools and prospective." *Organised Sound* **9**(1): 91-98.

Bossis, B. (2005). *La voix et la machine: La vocalité artificielle dans la musique contemporaine*.

Boulanger, R., Ed. (2000). *The Csound Book*. Cambridge, Mass., The MIT Press.

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge, Mass., MIT Press.

Bregman, A. S. (1993). "Auditory scene analysis: hearing in complex environments." In *Thinking in Sound. The Cognitive Psychology of Human Audition*. S. McAdams and E. Bigand (ed.). Oxford, Clarendon Press.

Bregman, A. S., P. A. Ahad, et al. (2000). "Effects of time intervals and tone durations on auditory stream segregation." *Perception & Psychophysics* **62**(3): 626-636.

Bregman, A. S., P. A. Ahad, et al. (2001). "Stream segregation of narrow-band noise bursts." *Perception & Psychophysics* **63**: 790-797.

Bricker, P. D. and S. Pruzansky (1966). "Effects of Stimulus Content and Duration on Talker Identification." *Journal of the Acoustical Society of America* **40**(6): 1441-1449.

Broening, B. (2006). "Alvin Lucier's *I am sitting in a room*." In *Analytical Methods of Electroacoustic Music*. M. Simoni (ed.). New York, Taylor & Francis**:** 89-110.

Brown, B. L., W. J. Strong, et al. (1973). "Perceptions of personality from speech: effects of manipulations of acoustical parameters." *The Journal of the Acoustical Society of America* **54**(1): 29-35.

Brown, J. W. S., R. J. Morris, et al. (1996). "Comfortable effort level revisited." *Journal of Voice* **10**(3): 299-305.

Bruce, V. and A. Young (1986). "Understanding face recognition." *British Journal of Psychology* **77 (Pt 3)**: 305-27.

Brückl, M. and W. F. Sendlmeier (2003). *Aging female voices: An acoustic and perceptive analysis*. Proceedings of VOQUAL, Geneva, Switzerland.

Buccino, G., F. Lui, et al. (2004). "Neural Circuits Involved in the Recognition of Actions Performed by Nonconspecifics: An fMRI Study." *Journal of Cognitive Neuroscience* **16**(1): 114-126.

Burkhardt, F. and W. F. Sendlmeier (2000). "Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis." In *Proceedings, ISCA workshop on Speech and Emotion, Belfast*(ed.)**:** 151-156.

Byrne, M. (1999). *Speech-Based Computer Music: Selected Works by Charles Dodge and Paul Lansky*. Proceedings of the International Computer Music Conference (ICMC 1999), Beijing.

Børset, B. (2006). *Støy og stemmer : radioteknologi og stemme i Nathalie Sarrautes pièces radiophoniques*. **Dr.Art.** Trondheim, NTNU.

Cahn, J. (1990). "The generation of affect in synthesized speech." *Journal of the American Voice I/O Society* **8**: 1-19.

Calvo-Merino, B., D. E. Glaser, et al. (2005). "Action Observation and Acquired Motor Skills: An fMRI Study with Expert Dancers." *Cerebral Cortex* **15**(8): 1243-1249.

Calvo-Merino, B., J. Grezes, et al. (2006). "Seeing or Doing? Influence of Visual and Motor Familiarity in Action Observation." *Current Biology* **16**(19): 1905-1910.

Campion, T. (1966). "Observations in the art of English poesie, 1602." In *A defence of ryme against a pamphlet entituled Obseruations in the art of English poesie, 1603 / Samuel Daniel. Observations in the art of English poesie, 1602 /Thomas Campion.* G. B. Harrison (ed.). New York, Barnes & Noble.

Carterette, E. C. and R. A. Kendall (1995). "Convergent research methods in music cognition." In *Music and the Mind Machine: The Psychophysiology and Psychopathology of the Sense of Music*. R. Steinberg (ed.). Berlin, Springer**:** 3-18.

Cascone, K. (2000). "The Aesthetics of Failure: Post-Digital Tendencies in Contemporary Computer Music." *Computer Music Journal* **24**(4): 12-18.

Causton, R. (1995). "Berio's 'Visage' and the Theatre of Electroacoustic Music." *Tempo*(194): 15-21.

Cherry, E. C. (1953). "Some Experiments on the Recognition of Speech, with One and with Two Ears." *The Journal of the Acoustical Society of America* **25**(5): 975-979.

Chiba, T. and M. Kajiyama (1958). *The vowel: Its nature and structure*. Tokyo, Phonetic society of Japan.

Childers, D., W. Ke, et al. (1987). *Factors in voice quality: Acoustic features related to gender*. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '87.

Childers, D. G. and K. Wu (1991). "Gender recognition from speech. Part II: Fine Analysis." *J. Acoust. Soc. Am.* **90**(4): 1841-1856.

Chion, M. (1983). *Guide des objets sonores: Pierre Schaeffer et la recherche musicale*. Paris, INA/GRM Buchet/Chastel.

Chion, M. (1988). "Du son á la chose: Hypothèses sur l'objet sonore." *Analyse musicale* **11**: 52-58.

Chion, M. (1991). *L'art des sons fixés - ou La Musique Concrètement*. Fontaine, Editions Metamkine / Nota Bene / Sono Concept.

Chion, M. (1994). *Audio-vision: sound on screen*. New York, Columbia University Press.

Chion, M. (1999). *The voice in cinema*. New York, Columbia Univerity Press.

Chion, M. (2005). "Pour une musique concrète 'médiatiste'." In *Portraits Polychromes no.8: Michel Chion*. INA/GRM (ed.). Paris, INA/GRM. **8:** 93-97.

Chowning, J. (1999). "Perceptual fusion and auditory perspective." In *Music, cognition, and computerized sound : an introduction to psychoacoustics*. P. R. Cook (ed.). Cambridge, Mass., MIT Press.

Cisinsky, M. (2007). "Le compositeur, l'outil et la trace historique," http://www.ac-rennes.fr/pedagogie/musique/dswmedia/sud_contexte_cisinsky.html (Retrieved 25.01, 2007).

Clarke, E. F. (2005a). *Ways of Listening : An Ecological Approach to the Perception of Musical Meaning*. Oxford, Oxford University Press.

Clarke, M. (2005b). "Jonathan Harvey's *Mortuos Plango, Vivos Voco*." In *Analytical Methods of Electroacoustic Music*. M. Simoni (ed.). New York, Routledge**:** 111-143.

Cleveland, T. F. (1977). "Acoustic properties of voice timbre types and their influence on voice classification." *The Journal of the Acoustical Society of America* **61**(6): 1622-1629.

Coath, M. and S. L. Denham (2005). "Robust sound classification through the representation of similarity using response fields derived from stimuli during early experience." *Biological Cybernetics* **93**(1): 22-30.

Coath, M., S. L. Denham, et al. (2007). "An auditory model for the detection of perceptual onsets and beat tracking in singing.". *Neural Information Processing Systems, Workshop on Music Processing in the Brain*, Vancouver.

Code, D. L. (1990). "Observations in the Art of Speech. Paul Lansky's Six Fantasies." *Perspectives of New Music* **28**(1): 144-169.

Coleman, R. O. (1976). "A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice." *Journal of speech and hearing research* **19**(1): 168-80.

Collins, S. A. (2000). "Men's voices and women's choices." *Animal Behaviour* **60**(6): 773-780.

Collins, S. A. and C. Missing (2003). "Vocal and visual attractiveness are related in women." *Animal behaviour* **65**(5): 997-1004.

Compton, A. J. (1963). "Effects of Filtering and Vocal Duration upon the Identification of Speakers, Aurally." *Journal of the Acoustical Society of America* **35**(11): 1748-1752.

Compton, R. J. (2003). "The Interface Between Emotion and Attention: A Review of Evidence from Psychology and Neuroscience." *Behav Cogn Neurosci Rev* **2**(2): 115-129.

Cone, E. T. (1974). *The composer's voice*. Berkeley, Cal., University of California Press.

Connor, S. (2001). "The Decomposing Voice of Postmodern Music." *New Literary History* **32**: 467-483.

Cook, N. and N. Dibben (2001). "Musicological approaches to emotion." In *Music and emotion: theory and research*. P. Juslin and J. A. Sloboda (ed.). New York, Oxford University Press**:** 45-70.

Cook, P. R. (1996). "Singing Voice Synthesis: History, Current Work, And Future Directions." *Computer Music Journal* **20**(3): 38-46.

Cook, P. R. (1999). "Articulation in speech and sound." In *Music, cognition, and computerized sound : an introduction to psychoacoustics*. P. R. Cook (ed.). Cambridge, Mass., MIT Press.

Cooke, M. and D. P. W. Ellis (2001). "The auditory organization of speech and other sources in listeners and computational models." *Speech Communication* **35**(3-4): 141-177.

Couper, M. P., E. Singer, et al. (2004). "Does Voice Matter? An Interactive Voice Response (IVR) Experiment." *Journal of official statistics* **20**(3): 1.

Coward, S. W. and C. J. Stevens (2004). "Extracting Meaning from Sound: Nomic Mappings, Everyday Listening, and Perceiving Object Size from Frequency." *Psychological Record* **54**(3): 349-364.

Cusack, R. and B. Roberts (2000). "Effects of differences in timbre on sequential grouping." *Perception & psychophysics* **62**(5): 1112-1120.

Cutler, A., D. Dahan, et al. (1997). "Prosody in the comprehension of spoken language: a literature review." *Language and speech* **40 (Pt 2)**: 141-201.

Darwin, C. J. (1981). "Perceptual grouping of speech components differing in fundamental frequency and onset-time." *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology* **33**(2): 185-207.

Darwin, C. J. (1997). "Auditory grouping." *Trends in Cognitive Sciences* **1**(9): 327-333.

Davis, M. H. and I. S. Johnsrude (2003). *Hierarchical Processing in Spoken Language Comprehension*. **23:** 3423-3431.

Davis, M. H. and I. S. Johnsrude (2007a). "Hearing speech sounds: Top-down influences on the interface between audition and speech perception." *Hearing Research* **229**(1-2): 132-147.

Davis, M. H. and I. S. Johnsrude (2007b). "Hearing speech sounds: Top-down influences on the interface between audition and speech perception." *Hearing Research* **In Press, Corrected Proof**(doi:10.1016/j.heares.2007.01.014).

Davis, M. H., I. S. Johnsrude, et al. (2005). "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences." *Journal of experimental psychology. General* **134**(2): 222-241.

Debruyne, F. and W. Decoster (1999). "Acoustic differences between sustained vowels perceived as young or old." *Logopedics, phoniatrics, vocology* **24**(1): 1-5.

DeCasper, A. J. and W. P. Fifer (1980). "Of human bonding: newborns prefer their mothers' voices." *Science* **208**(4448): 1174-6.

deCharms, R. C., D. T. Blake, et al. (1998). "Optimizing Sound Features for Cortical Neurons." *Science* **280**(5368): 1439-1444.

Delalande, F. (1998). "Music Analysis and Reception Behaviours : Sommeil by Pierre Henry." *Journal of New Music Research* **27**(1-2): 13-66.

Deliège, I. (2001). "Similarity Perception <-> Categorization <-> Cue Abstraction." *Music Perception* **18**(3): 233-243.

Deutsch, D. (1999). "Grouping mechanisms in music." In *The Psychology of Music*. D. Deutsch (ed.). San Diego, Academic Press. **2:** 299-348.

Dodge, C. (1989). "On Speech Songs." In *Current Directions in Computer Music Research*(ed.). Cambridge, MASS., MIT Press**:** 9-17.

Dreßen, N. (1982). *Sprache und Musik bei Luciano Berio : Untersuchungen zu seinen Vokalkompositionen*. Regensburg, Bosse.

Driver, J. (2001). "A selective review of selective attention research from the past century." *British Journal of Psychology* **92**(1): 53-78.

Dudley, H. (1939). "The vocoder." *Bell labs. Rec.* **17**: 122.

Dyson, F. (1994). "The genealogy of the Radio Voice." In *Radio Rethink: Art, Sound and Transmission*. D. Augaitis and D. Lander (ed.). Banff, Walter Phillips Gallery.

Eagleton, T. (1983). *Literary Theory*. Oxford, Blackwell.

Eco, U. (1989). *The open work*. Cambridge, Mass., Harvard University Press.

Ekeberg, F. (2002). *Space in Electroacoustic Music: Composition, Performance and Perception of Musical Space*. **PhD**. London, City University.

Elliott, L. L. (1971). "Backward and Forward Masking." *International Journal of Audiology* **10**(2): 65 - 76.

Emmerson, S. (1986). "The relation of language to materials." In *The language of electroacoustic music*. S. Emmerson (ed.). London, Macmillan**:** 17-39.

Emmerson, S. (1998). "Acoustic/electroacoustic: The relationship with instruments." *Journal of New Music Research* **27**(1): 146 - 164.

Emmerson, S. (2000). "'Losing touch?':the human performer and electronics." In *Music, Electronic Media and Culture*(ed.). Aldershot, Ashgate**:** 194-216.

Emmerson, S. (2007). *Living electronic music*. Aldershot, Ashgate.

Emmerson, S. and D. Smalley (2009). "Electro-acoustic music," *Grove Music Online* http://www.oxfordmusiconline.com/subscriber/article/grove/music/08695 (Retrieved 15.09.2009.

Erickson, M., S. Perry, et al. (2001). "Discrimination Functions Can They Be Used to Classify Singing Voices?" *Journal of Voice* **15**(4): 492-502.

Erickson, M. L. (2003). "Dissimilarity and the Classification of Female Singing Voices: A Preliminary Study." *J. Voice* **17**: 195-206.

Erickson, M. L. and S. R. Perry (2003). "Can Listeners Hear Who Is Singing? A Comparison Three-note and Six-note Discrimination Tasks." *J. Voice* **17**: 352-368.

Eriksson, A. and H. Traunmüller (2002). "Perception of vocal effort and distance from the speaker on the basis of vowel utterances." *Perception & psychophysics* **64**(1): 131-9.

Eriksson, E. J. (2007). *That voice sounds familiar: factors in speaker recognition*. **PhD**. Umeå, Umeå University,.

Escera, C., K. Alho, et al. (1998). "Neural Mechanisms of Involuntary Attention to Acoustic Novelty and Change." *Journal of Cognitive Neuroscience* **10**(5): 590-604.

Escera, C. and M. J. Corral (2007). "Role of Mismatch Negativity and Novelty-P3 in Involuntary Auditory Attention." *Journal of psychophysiology* **21**(3/4): 251-264.

Fadiga, L., L. Craighero, et al. (2002). "Speech listening specifically modulates the excitability of tongue muscles: a TMS study." *The European journal of neuroscience* **15**(2): 399-402.

Fairbanks, G. and F. Kodman (1957). "Word Intelligibility as a Function of Time Compression." *The Journal of the Acoustical Society of America* **29**(5): 636-641.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague, The Netherlands, Mouton.

Farnetani, E. (1999). "The Handbook of phonetic sciences." In *The Handbook of phonetic sciences*. W. J. Hardcastle and J. Laver (ed.). Cambridge, Mass., Blackwell**:** 371-404.

Feinberg, D. R. (2004). "Fundamental frequency perturbation indicates perceived health and age in male and female speakers." *The Journal of the Acoustical Society of America* **115**(5): 2609.

Feinberg, D. R., B. C. Jones, et al. (2005). "Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices." *Animal behaviour* **69**: 561-568.

Fellowes, J. M., R. E. Remez, et al. (1997). "Perceiving the sex and identity of a talker without natural vocal timbre." *Perception & psychophysics* **59**(6): 839-49.

Ferrand, C. T. (2002). "Harmonics-to-Noise Ratio: An Index of Vocal Aging." *Journal of Voice* **16**(4): 480-487.

Ferrara, L. (1984). "Phenomenology as a Tool for Musical Analysis." *The Musical Quarterly* **70**(3): 355-373.

Fiske, J. (1990). *Introduction to communication studies*. London, Routledge.

Fitch, W. T. and J. Giedd (1999). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging." *The Journal of the Acoustical Society of America* **106**(3): 1511-1522.

Fletcher, H. (1922). "The Nature of Speech and Its Interpretation." *J.Franklin Instit.* **193**(6): 729-747.

Foreman, C. G. (1999). *Dialect Identification From Prosodic Cues*. Proceedings of ICPhS99, San Francisco.

Foulke, E. and T. G. Sticht (1969). "Review of research on the intelligibility and comprehension of accelerated speech." *Psychological bulletin* **72**(1): 50-62.

Friend, M. and M. J. Farrar (1994). "A comparison of content-masking procedures for obtaining judgments of discrete affective states." *The Journal of the Acoustical Society of America* **96**(3): 1283-90.

Fruhstorfer, H., P. Soveri, et al. (1970). "Short-term habituation of the auditory evoked response in man." *Electroencephalography and Clinical Neurophysiology* **28**(2): 153-161.

Føllesdal, D. (1989). "Fenomenologien - en tilnærming til det subjektive." In *Spor etter mennesket*. L. Bliksrud and A. Aarnes (ed.). Oslo, Aschehoug**:** 291-304.

Gallese, V. (2006). "Intentional attunement: a neurophysiological perspective on social cognition and its disruption in autism." *Brain research* **1079**(1): 15-24.

Gallese, V., L. Fadiga, et al. (1996). "Action recognition in the premotor cortex." *Brain* **119 (Pt 2)**: 593-609.

Ganong, W. F. (1980). "Phonetic categorization in auditory word perception." *Journal of experimental psychology. human perception and performance* **6**(1): 110-25.

Gardiner, J. M. (1983). "On Recency and Echoic Memory." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **302**(1110): 267-282.

Gaver, W. W. (1993). "What in the world do we hear? An ecological approach to auditory event perception." *Ecological psychology* **5**(1): 1-29.

Genette, G. (1980). *Narrative discourse*. Oxford, Basil Blackwell.

Genette, G. (1997). *Paratexts. Thresholds of interpretation*. Cambridge, Cambridge University Press.

Georgaki, A. (1998). *Problems techniques et enjoux esthetiques de la voix synthese dans la recherche et creation musicales*. **Thèse de doctorat**. Paris, Ecole des Haute Etudes en Sciences Sociales IRCAM.

Georgaki, A. (1999). "Proteïc voices in the computer music repertory (1972-1997)." In *ICMC Proceedings 1999*(ed.)**:** 553-556.

Gerratt, B. R., K. Precoda, et al. (1988). "Source characteristics of diplophonia." *The Journal of the Acoustical Society of America* **83**(S1): S66-S66.

Gibson, E. (1998). "Linguistic complexity: locality of syntactic dependencies." *Cognition* **68**(1): 1-76.

Gilbert, A. N., R. Martin, et al. (1996). "Cross-modal correspondence between vision and olfaction: The color of smells." *The American Journal of Psychology* **109**(3): 335-351.

Gilboa-Schechtman, E., W. Revelle, et al. (2000). "Stroop Interference following Mood Induction: Emotionality, Mood Congruence, and Concern Relevance." *Cognitive Therapy and Research* **24**(5): 491-502.

Gobl, C. and A. Ní Chasaide (2003). "The role of voice quality in communicating emotion, mood and attitude." *Speech Commun.* **40**: 189-212.

Godøy, R. I. (1997). *Formalization and Epistemology*. Oslo, Universitetsforlaget.

Godøy, R. I. (2006). "Gestural-Sonorous Objects: embodied extensions of Schaeffer's conceptual apparatus." *Organised sound* **11**(02): 149-157.

Goggin, J. P., C. P. Thompson, et al. (1991). "The role of language familiarity in voice identification." *Memory and Cognition* **19**(5): 448-458.

González, J. (2004). "Formant frequencies and body size of speaker: a weak relationship in adult humans." *Journal of phonetics* **32**(2): 277-287.

Grandjean, D., D. Sander, et al. (2005). "The voices of wrath: brain responses to angry prosody in meaningless speech." *Nature Neuroscience* **8**(2): 145-146.

Grey, J. M. (1977). "Multidimensional perceptual scaling of musical timbres." *The Journal of the Acoustical Society of America* **61**(5): 1270-1277.

Griffiths, P. (2009). "Sprechgesang," *Grove Music Online. Oxford Music Online* http://www.oxfordmusiconline.com/subscriber/article/grove/music/26465 (Retrieved 05.11.2009.

Grossberg, S., K. K. Govindarajan, et al. (2004). "ARTSTREAM: a neural network model of auditory scene analysis and source segregation." *Neural Networks* **17**(4): 511-536.

Gunji, A., S. Koyama, et al. (2003). "Magnetoencephalographic study of the cortical activity elicited by human voice." *Neuroscience letters* **348**(1): 13-6.

Günzburger, D. (1995). "Acoustic and perceptual implications of the transsexual voice." *Archives of Sexual Behavior* **24**(3): 339-348.

Halford, G. S., W. H. Wilson, et al. (1998). "Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology." *Behavioral and Brain Sciences* **21**(6): 803-831.

Hall, E. T. (1979). "Proxemics." In *Nonverbal Communication: Readings with Commentary*. S. Weitz (ed.). New York, Oxford University Press.

Hall, J. W. and J. H. Grose (1990). "Comodulation masking release and auditory grouping." *The Journal of the Acoustical Society of America* **88**(1): 119-125.

Handel, S. (1989). *Listening. An Introduction to the Perception of Auditory Events*. Cambridge, Mass., MIT Press.

Handel, S. and M. L. Erickson (2001). "A Rule of Thumb: The Bandwidth for Timbre Invariance Is One Octave." *Music Perception* **19**(1): 121-126.

Hanson, H. M. and E. S. Chuang (1999). "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data." *The Journal of the Acoustical Society of America* **106**(2): 1064-1077.

Haraway, D. (1991). *Simians, cyborgs, and women: the re-invention of nature*. London, Free Association.

Haraway, D. J. (1997). *Modest_Witness@Second_Millennium.FemaleMan©_Meets_OncoMouseTM*. New York, Routledge.

Harding, S. (1992). "After the neutrality ideal: Science, politics and" strong objectivity." *Social research* **59**(3): 567-587.

Harnad, S. (1987). "Category Induction and Representation." In *Categorical Perception: The Groundwork of Cognition*. S. Harnad (ed.). New York, Cambridge University Press**:** 535-565.

Haueisen, J. and T. R. Knosche (2001). "Involuntary Motor Activity in Pianists Evoked by Music Perception." *Journal of Cognitive Neuroscience* **13**(6): 786-792.

Hawkins, S., S. Heid, et al. (2000). "Assessment of naturalness in the ProSynth speech synthesis project." In *IEE: State-of-the-art in speech synthesis. 13/1-13/6*(ed.).

Heeter, C. (2003). "Reflections on Real Presence by a Virtual Person." *Presence: Teleoperators & Virtual Environments* **12**(4): 335-345.

Hickok, G., B. Buchsbaum, et al. (2003). "Auditory-Motor Interaction Revealed by fMRI: Speech, Music, and Working Memory in Area Spt." *Journal of Cognitive Neuroscience* **15**(5): 673-682.

Hillenbrand, J. M. (2003). "Some effects of intonation contour on sentence intelligibility." *The Journal of the Acoustical Society of America* **114**(4): 2338.

Holeckovaa, I., C. Fischera, et al. (2006). "Brain responses to a subject's own name uttered by a familiar voice." *Brain Research* **1082**(1): 142-152.

Hollien, H. (1974). "On Vocal Registers." *Journal of Phonetics* **2**: 125-143.

Honorof, D. N. and D. H. Whalen (2005). "Perception of pitch location within a speaker's F0 range." *The Journal of the Acoustical Society of America* **117**(4): 2193-2200.

Huang, A., F. Lee, et al. (2001). *Can Voice User Interfaces Say "I"? An Experiment with Recorded Speech and TTS*. Stanford, California, Stanford.

Hugdahl, K., T. Thomsen, et al. (2003). "The effects of attention on speech perception: An fMRI study." *Brain and Language* **85**(1): 37-48.

Hultberg, T. (1994). "Från hätila ragulpr på fåtskliaben till hej tatta gôrem." In *Fylkingen : ny musik & intermediakonst : rikt illustrerad historieskrivning & diskussion för radikal & eksperimentell konst 1933-1993*. T. Hultberg and C. Bock (ed.). Stockholm, Fylkingen**:** 61-81.

Huron, D. (2006). *Sweet anticipation - music and the psychology of expectation*. London, MIT Press.

Ihde, D. (1976). *Listening and Voice. A phenomenology of Sound.* Athens, Ohio, Ohio University Press.

Ihde, D. (1993). *Philosophy of technology*. New York, Paragon House.

Ingemann, F. (1968). "Identification of the speaker's sex from voiceless fricatives." *The Journal of the Acoustical Society of America* **44**(4): 1142-4.

Ishii, K., J. A. Reyes, et al. (2003). "Spontaneous Attention to Word Content Versus Emotional Tone: Differences Among Three Cultures." *Psychological Science* **14**(1): 39.

Jensenius, A. R. (2007). *Action - Sound: Developing Methods and Tools to Study Music-related Body Movement*. **PhD**. Oslo, University of Oslo.

Jerger, J. F. (1957). "Auditory Adaptation." *The Journal of the Acoustical Society of America* **29**(3): 357-363.

Jones, D. E. (1987). "Compositional Control of Phonetic/Nonphonetic Perception." *Perspectives of new music* **25**(1/2): 138.

Jones, M. R. and W. Yee (1993). "Attending to auditory events: The role of temporal organization." In *Thinking in sound: The cognitive psychology of human audition*. S. McAdams and E. Bigand (ed.). Oxford, Clarendon Press**:** 69-112.

Juslin, P. and J. A. Sloboda, Eds. (2001). *Music and emotion: theory and research*. New York, Oxford University Press.

Juslin, P. N. and P. Laukka (2003). "Communication of emotions in vocal expression and music performance: different channels, same code?" *Psychological bulletin* **129**(5): 770-814.

Kappas, A., U. Hess, et al. (1991). "Voice and emotion." In *Fundamentals of nonverbal behavior*(ed.). Cambridge and New York, Cambridge University Press**:** 123-148.

Karlsson, I. (1992). "Evaluations of acoustic differences between male and females voices: a pilot study." *STL-QPSR* **3**(1): 19-31.

Karydis, I., A. Nanopoulos, et al. (2007). "Horizontal and Vertical Integration/Segregation in Auditory Streaming: A Voice Separation Algorithm for Symbolic Musical Data". *SMC'07, 4th Sound and Music Computing Conference*, Lefkada, Greece.

Kashino, M. (2006). "Phonemic restoration: The brain creates missing speech sounds." *Acoustical Science and Technology* **27**(6): 318-321.

Katz, M. (2004). *Capturing Sound*. Berkeley & L.A., CA, University of California Press.

Kayser, C., C. I. Petkov, et al. (2005). "Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map." *Current Biology* **15**(21): 1943-1947.

Keller, E. (2002). "Towards Greater Naturalness: Future Directions of Research in Speech Synthesis." In *Improvements in Speech Synthesis: Cost 258: The Naturalness of Synthetic Speech*. E. Keller, G. Bailly, A. Monaghan, J. Terken and M. Huckvale (ed.). New York Wiley**:** 3-17.

Keller, E., G. Bailly, et al., Eds. (2002). *Improvements in Speech Synthesis: Cost 258: The Naturalness of Synthetic Speech*, John Wiley and Sons.

Kent, R. D. and L. L. Forner (1980). "Speech segment durations in sentence recitations by children and adults." *Journal of phonetics* **8**(2): 157-168.

Kent, R. D. and C. Read (2002). *The acoustic analysis of speech*. San Diego, Singular Publishing Group.

Kessinger, R. H. and S. E. Blumstein (1997). "Effects of speaking rate on voice-onset time in Thai, French, and English." *Journal of Phonetics* **25**(2): 143-168.

Kisilevsky, B. S., S. M. Hains, et al. (2003). "Effects of experience on fetal voice recognition." *Psychological science* **14**(3): 220-4.

Kitayama, S. and K. Ishii (2002). "Word and voice: Spontaneous attention to emotional utterances in two languages." *Cognition & Emotion* **16**(1): 29-59.

Klabbers, E. and R. Veldhuis (2001). "Reducing audible spectral discontinuities." *IEEE Transactions on Speech and Audio Processing,* **9**(1): 39-51.

Klatt, D. H. and L. C. Klatt (1990). "Analysis, synthesis, and perception of voice quality variations among female and male talkers." *The Journal of the Acoustical Society of America* **87**(2): 820-57.

Klumpp, R. G. and J. C. Webster (1961). "Intelligibility of Time-Compressed Speech." *J. Acoust. Soc. Am.* **33**(3): 265-267.

Knudsen, E. I. (2007). "Fundamental Components of Attention." *Annual Review of Neuroscience* **30**(1): 57-78.

Ko, S. J., C. M. Judd, et al. (2006). "What the Voice Reveals: Within- and Between-Category Stereotyping on the Basis of Voice." *Personality and Social Psychology Bulletin* **32**(6): 806-819.

Kotlyar, G. M. and V. P. Morosov (1976). "Acoustical correlates of the emotional content of vocalized speech." *Sov. Phys. Acoust.* **22**: 208-211.

Kramer, L. (1984). *Music and poetry: the nineteenth century and after*. Berkeley and Los Angeles, University of California Press.

Kramer, L. (1990). *Music as Cultural Practice, 1800-1900*. Berkeley, University of California Press.

Kreiner, D. S., N. A. Altis, et al. (2003). "A test of the effect of reverse speech on priming." *The Journal of psychology* **137**(3): 224-32.

Kuwabara, H. (1996). "A perceptual experiment on voice individuality by altering pitch and formant frequencies." *The Journal of the Acoustical Society of America* **100**(4): 2600.

Kuwabara, H. and Y. Sagisaka (1995). "Acoustic characteristics of speaker individuality: Control and conversion." *Speech Communication* **16**: 165-173.

Lacasse, S. (2000) "Voice and Sound Processing: Examples of Mise en Scene of Voice in Recorded Rock Music." *Popular Musicology Online*(5), http://www.popular-musicology-online.com/issues/05/lacasse.html (Retrieved 21.11.2009).

Ladd, D. R., K. E. A. Silverman, et al. (1985). "Evidence for the independent function of intonation contour type, voice quality, and F0 in signalling speaker affect." *J. Acoust. Soc. Am.* **78**: 435-444.

Ladefoged, P. (2005). *Vowels And Consonants: an introduction to the sounds of languages*. Oxford, Blackwell.

Ladefoged, P. and N. P. McKinney (1963). "Loudness, Sound Pressure, and Subglottal Pressure in Speech." *The Journal of the Acoustical Society of America* **35**(4): 454-460.

Lakoff, G. (1973). "Hedges: A study in meaning criteria and the logic of fuzzy concepts." *Journal of philosophical logic* **2**(4): 458-508.

Lakoff, G. (1987). *Women, fire, and dangerous things: what categories reveal about the mind*. Chicago, Ill., University of Chicago Press.

Lakoff, G. and M. Johnson (1980). *Metaphors we live by*. London, The University of Chicago Press.

Landy, L. (1993). "Sound Transformations in Electroacoustic Music." *Composers Desktop Project Quarterly*.

Landy, L. (1994). "The "Something to Hold on to Factor" in Timbral Composition." *Contemporary Music Review* **10**(2): 49-60.

Landy, L. (2006). "The Intention/Reception Project." In *Analytical Methods of Electroacoustic Music*. M. Simoni (ed.). New York, Routledge**:** 29-54.

Landy, L. (2007). *Understanding the art of sound organisation*. Cambridge, Mass., MIT Press.

Lane, C. (2006). "Voices from the Past: compositional approaches to using recorded speech." *Organised Sound* **11**(1): 3-11.

Lang, A. (2000). "The limited capacity model of mediated message processing." *The Journal of Communication* **50**(1): 46-70.

Lansky, P. (1989). "Compositional Applications of Linear Predictive Coding." In *Current Directions in Computer Music Research*. M. V. Mathews and J. R. Pierce (ed.). Cambridge, Mass, MIT Press**:** 5-8.

Lass, N. J., K. R. Hughes, et al. (1976). "Speaker sex identification from voiced, whispered, and filtered isolated vowels." *The Journal of the Acoustical Society of America* **59**(3): 675-678.

Lass, N. R. O. (1980). "The Effect of Filtered Speech on Speaker Race and Sex Identifications." *Journal of phonetics* **8**(1): 101-112.

Lattner, S., B. Maess, et al. (2003). "Dissociation of human and computer voices in the brain: Evidence for a preattentive gestalt-like perception." *Human Brain Mapping* **20**(1): 13-21.

Laures, J. S. and K. Bunton (2003). "Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions." *Journal of Communication Disorders* **36**(6): 449-464.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge Cambridge University Press.

Laver, J. (2003). "Three semiotic layers of spoken communication." *Journal of Phonetics* **31**: 413-415.

Laver, J. and P. Trudgill (1979). "Phonetic and linguistic markers in speech." In *Social Markers in Speech*. K. R. Scherer and G. H. (ed.). Cambridge, Cambridge University Press.

Lavie, N. (2005). "Distracted and confused?: Selective attention under load." *Trends in Cognitive Sciences* **9**(2): 75-82.

Lavner, Y., I. Gath, et al. (2000). "The effects of acoustic modifications on the identification of familiar voices speaking on isolated vowels." *Speech Communcation* **30**: 9-26.

Lee, K. M. (2004a). "Presence, Explicated." *Communication theory* **14**(1): 27.

Lee, K. M. (2004b). "Why Presence Occurs: Evolutionary Psychology, Media Equation, and Presence." *Presence: Teleoperators and Virtual Environments* **13**(4): 494-505.

Leech, G. N. (1981). *Semantics: the study of meaning*. Harmondsworth, Penguin.

Lessiter, J., J. Freeman, et al. (2001). "Really hear? The effects of audio quality on presence". *Presence 2001*, Philadelphia, PA.

Levman, B. G. (1992). "The Genesis of Music and Language." *Ethnomusicology* **36**(2): 147-170.

Levy, D. A., R. Granot, et al. (2001). "Processing specificity for human voice stimuli: electrophysiological evidence." *Neuroreport* **12**(12): 2653-7.

Levy, D. A., R. Granot, et al. (2003). "Neural sensitivity to human voices: ERP evidence of task and attentional influences." *Psychophysiology* **40**(2): 291-305.

Liberman, A. M., K. S. Harris, et al. (1957). "The discrimination of speech sounds within and across phoneme boundaries." *Journal of Experimental Psychology* **54**(5): 358.

Liberman, A. M. and D. H. Whalen (2000). "On the relation of speech to language." *Trends in cognitive sciences* **4**(5): 187-196.

Lieberman, P. (1984). *The biology and evolution of language*. Cambridge, Mass., Harvard University Press.

Lieberman, P. (1991). *Uniquely Human. The Evolution of Speech, Thought and Selfless Behaviour*. London, Harvard University Press.

Liénard, J. S. and M. G. Di Benedetto (1999). "Effect of vocal effort on spectral properties of vowels." *The Journal of the Acoustical Society of America* **106**(1): 411-22.

Link, S. (2001). "The Work of Reproduction in the Mechanical Aging of an Art: Listening to Noise." *Computer music journal* **25**(1): 34.

Linville, S. E. (1996). "The sound of senescence." *Journal of Voice* **10**(2): 190-200.

Logan, J. S., S. E. Lively, et al. (1991). "Training Japanese listeners to identify English /r/ and /l/: A first report." *The Journal of the Acoustical Society of America* **89**(2): 874-886.

Lombard, M. and T. Ditton (1997) "At the heart of it all: The concept of presence." *Journal of computer-mediated communication* **3**(2), http://jcmc.indiana.edu/vol3/issue2/lombard.html (Retrieved 20.4.2007).

MacLeod, C. M. (1991). "Half a Century of Research on the Stroop Effect: An Integrative Review." *Psychol. Bull.* **109**(2): 163-203.

Marin, C. M. H. and S. McAdams (1991). "Segregation of concurrent sounds. II: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width." *The Journal of the Acoustical Society of America* **89**(1): 341-351.

Martin, K. D. (1999). *Sound-Source Recognition: A Theory and Computational Model* **Doctoral thesis**. Massachusets Institute of Technology.

Mathiesen, J. (2006). "Song - Ancient," *Grove Music Online* www.grovemusic.com (Retrieved 18.12, 2006).

Mattingly, I. G. (1974). "Speech Synthesis for Phonetic and Phonological Models." In *Current Trends in Linguistics*. T. A. Sebeok (ed.). The Hague, Mouton**:** 2451-2487.

Mayo, C., R. A. J. Clark, et al. (2005). "Multidimensional scaling of listener responses to synthetic speech." In *Proc. Interspeech 2005, Lisbon, Portugal, September 2005*(ed.). Lisbon.

McAdams, S. (1984). "The auditory image: A metaphor for musical and psychological research on auditory organization." In *Cognitive processes in the perception of art*. W. R. Crozier and A. J. Chapman (ed.). Amsterdam, Elsevier Science Publishers B.V.

McAdams, S. (1989). "Psychological constraints on form-bearing dimensions in music." *Contemporary Music Review* **4**(1): 181-198.

McAdams, S. (1993). "Recognition of sound sources and events." In *Thinking in Sound - The Cognitive Psychology of Human Audition*(ed.). Oxford, Oxford University Press**:** 146-198.

McAdams, S. and D. Matzkin (2003). "The Roots of Musical Variation in Perceptual Similarity and Invariance." In *The cognitive Neuroscience of Music*(ed.). Oxford, Oxford University Press**:** 79-94.

McAdams, S., S. Vieillard, et al. (2004). "Perception of Musical Similarity Among Contemporary Thematic Materials in Two Instrumentations." *Music Perception* **22**(2): 207-238.

McClelland, L. J. and J. L. Elman (1986). "The TRACE Model of Speech Perception." *Cognitive psychology* **18**: 1-86.

McMullen, E. and J. R. Saffran (2004). "Music and Language: A Developmental Comparison." *Music Perception* **21**(3): 289-311.

Mendes, A. P., H. B. Rothman, et al. (2003). "Effects of vocal training on the acoustic parameters of the singing voice." *Journal of Voice* **17**(4): 529-43.

Mertens, W. (2004). "Basic Concepts of Minimal Music." In *Audio Culture: Readings in Modern Music*. C. Cox and D. Warner (ed.). London, Continuum**:** 307-312.

Middleton, R. (2006) "'Last Night a DJ Saved My Life': Avians, Cyborgs and Siren Bodies in the Era of Phonographic Technology." *Radical Musicology* **1**, http://www.radical-musicology.org.uk (Retrieved 08.11.2009).

Miller, G. A. (1956). "The magic number seven, plus or minus two: Some limits on our capacity for processing information." *Psychological review* **63**(2): 81-97.

Mithen, S. (2005). *The Singing Neanderthals*. London, Weidenfeld & Nicholson.

Molino, J. (1990). "Musical fact and the semiology of music." *Musical analysis* **9**(2): 113-156.

Moore, B. C. J. (2003). *An Introduction to the Psychology of Hearing, Fifth Edition*. Amsterdam, Academic Press.

Moore, B. C. J. and C. Tan (2003). "Perceived naturalness of spectrally distorted speech and music." *J. Acoust. Soc. Am.* **114**(1): 408-419.

Mozziconacci, S. J. L. (1998). *Speech Variability and Emotion: Production and Perception*. **PhD**. Technische Universiteit Eindhoven.

Muller-Gass, A. and E. Schröger (2007). "Perceptual and cognitive task difficulty has differential effects on auditory distraction." *Brain Research* **1136**: 169-177.

Murdock, B. B. j. (1962). "The Serial Position Effect of Free Recall." *Journal of Experimental Psychology* **64**(5): 482-488.

Murphy, T. A., M. Matlin, et al. (2003). *Curvature covariation as a factor in perceptual salience*. Neural Engineering, 2003. Conference Proceedings. First International IEEE EMBS Conference on.

Murphy, T. S. (1999) "Music After Joyce: The Post-Serial Avant-Garde." *Hypermedia Joyce Studies* **2**(1), http://hjs.ff.cuni.cz/archives/v2/murphy/ (Retrieved 15.03.2009).

Murray, I. R. and J. L. Arnott (1993). "Toward a simulation of emotion in synthetic speech: A review of the literature on human vocal emotion." *J. Acoust. Soc. Am.* **93**(2): 1097-1108.

Murray, I. R. and J. L. Arnott (1996). "Synthesizing emotions in speech: Is it time to get excited?" In *Proc. International Conf. on Spoken Language Processing*(ed.)**:** 1816–1819.

Murray, I. R., J. L. Arnott, et al. (1996). "Emotional stress in synthetic speech: Progress and future directions." *Speech Communication* **20**: 85-91.

Murry, T. and S. Singh (1980). "Multidimensional analysis of male and female voices." *The Journal of the Acoustical Society of America* **68**(5): 1294-300.

Nash, E. B., G. W. Edwards, et al. (2000). "A Review of Presence and Performance in Virtual Environments." *International Journal of Human-Computer Interaction* **12**(1): 1-41.

Nass, C. and S. Brave (2005). *Wired for speech. How Voice Activates and Advances the Human-Computer Relationship*. Cambridge, Mass., MIT Press.

Nass, C. and L. Gong (2000). "Speech interfaces from an evolutionary perspective." *Communications of the ACM* **43**(9): 36.

Nass, C. and K. M. Lee (2001). "Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction." *Journal of experimental psychology. Applied* **7**(3): 171-81.

Nass, C., Y. Moon, et al. (1997). "Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices." *Journal of Applied Social Psychology* **27**(10): 864-876.

Nejime, Y. and B. C. J. Moore (1998). "Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss." *The Journal of the Acoustical Society of America* **103**(1): 572-576.

Newbrook, M. and J. M. Curtain (1998). "Oates' theory of Reverse Speech: a critical examination." *Forensic Linguistics* **5**(2): 174-192.

Norman, K. (1996). "Real-world music as composed listening." *Contemporary Music Review* **15**(1): 1-27.

Norman, K. (2000). "Stepping outside for a moment: narrative space in two works for sound alone." In *Music, Electronic Media and Culture*. S. Emmerson (ed.). Aldershot, Ashgate**:** 217-244.

Norman, K. (2004a). *Sounding Art. Eight Literary Excursions through Electronic Music*. Aldershot, Ashgate.

Norris, D., J. M. McQueen, et al. (2003). "Perceptual learning in speech." *Cognitive Psychology* **47**(2): 204-238.

Noyes, J. (2001). "Talking and writing - how natural in human-machine interaction?" *International Journal of Human-Computer Studies* **55**: 503-519.

Nusbaum, H. C. (1997). "Measuring the naturalness of synthetic speech." *International journal of speech technology* **2**(1): 7.

Nusbaum, H. C. and T. M. Morin (1992). "Paying attention to differences among talkers." In *Speech perception, production and linguistic structure*. Y. i. Tohkura, Y. Sagisaka and E. Vatikiotis-Bateson (ed.). Tokyo, Ohmsha**:** xiv, 463 s.

Ockelford, A. (2004). "On similarity, derivation and the cognition of musical structure." *Psychology of Music* **32**(1): 23-74.

Omori, K., H. Kojima, et al. (1997). "Acoustic characteristics of rough voice: subharmonics." *Journal of Voice* **11**(1): 40-47.

Ondishko, D. (1990). *Six Fantasies on a Poem by Thomas Campion: Synthesis and Evolution of Paul Lansky's Music Compositions.* **Ph.D.** Rochester, New York, Eastman School of Music.

Orr, D. B., H. L. Friedman, et al. (1965). "Trainability of listening comprehension of speeded discourse." *Journal of educational psychology* **56**: 148-56.

Palmeri, T. J., S. D. Goldinger, et al. (1993). "Episodic encoding of voice attributes and recognition memory for spoken words." *Journal of experimental psychology. Learning, memory, and cognition* **19**(2): 309-28.

Palombini, C. (1998). "Technology and Pierre Schaeffer: Pierre Schaeffer's Arts-Relais, Walter Benjamin's technische Reproduzierbarkeit and Martin Heidegger's Ge-stell." *Organised Sound* **3**(01): 35-43.

Parsa, V. and D. G. Jamieson (2001). "Acoustic Discrimination of Pathological Voice: Sustained Vowels Versus Continuous Speech." *J Speech Lang Hear Res* **44**(2): 327-339.

Pashler, H. E. (1999). *The Psychology Of Attention*. Cambridge, Mass., MIT Press.

Payri, B. and J. L. M. Bono (2007). "Auditory scene analysis and sound source coherence as a frame for perceptual study of electroacoustic music language". *Electroacoustic Music Studies - EMS07*, Leicester, UK.

Pereira, C. (2000). "Dimensions of emotional meaning in speech." In *SpeechEmotion-2000*(ed.). Newcastle, Northern Ireland, UK.**:** 25-28.

Perry, T. L., R. N. Ohde, et al. (2001). "The acoustic bases for gender identification from children's voices." **109**(6): 2988-2998.

Peterson, G. E. and I. Lehiste (1959). "Identification of Filtered Vowels." *The Journal of the Acoustical Society of America* **31**(6): 844-844.

Pickering, M. J. and S. Garrod (2007). "Do people use language production to make predictions during comprehension?" *Trends in Cognitive Sciences* **11**(3): 105-110.

Pierce, J. R. (1999). "Hearing in Time and Space." In *Music, cognition, and computerized sound: an introduction to psychoacoustics*. P. R. Cook (ed.). Cambridge, MA,, MIT Press**:** 89-103

Pisoni, D. B., H. C. Nusbaum, et al. (1985). "Perception of synthetic speech generated by rule." *Proceedings of the IEEE* **73**(11): 1665.

Plant, R. L. and R. M. Younger (2000). "The interrelationship of subglottic air pressure, fundamental frequency, and vocal intensity during speech." *Journal of Voice* **14**(2): 170-177.

Polkosky, M. D. and J. R. Lewis (2003). "Expanding the MOS: Development and Psychometric Evaluation of the MOS-R and MOS-X." *International Journal of Speech Technology* **6**: 161-182.

Pollack, I. (1959). "Message Repetition and Message Reception." *The Journal of the Acoustical Society of America* **31**(11): 1509-1515.

Pollack, I., J. M. Pickett, et al. (1954). "On the Identification of Speakers by Voice." *Journal of the Acoustical Society of America* **26**(3): 403-406.

Popova, Y. (2005). "Image schemas and verbal synaesthesia." In *From Perception to Meaning : Image Schemas in Cognitive Linguistics*. B. Hampe (ed.). Berlin, Walter de Gruyter & Co**:** 395-419.

Potter, R. F. (2000). "The Effects of Voice Changes on Orienting and Immediate Cognitive Overload in Radio Listeners." *Media Psychology* **2**(2): 147-177.

Potter, R. F. and J. Choi (2006). "The Effects of Auditory Structural Complexity on Attitudes, Attention, Arousal, and Memory." *Media Psychology* **8**(4): 395-419.

Potter, R. F., A. Lang, et al. (1998). "Identifying Structural Features of Radio: Orienting and Memory for Radio Messages". *Paper presented to the Theory and methodology division of the association for education in journalism and mass communication*, Baltimore, MD.

Press, C., G. Bird, et al. (2005). "Robotic movement elicits automatic imitation." *Cognitive Brain Research* **25**(3): 632-640.

Ptacek, P. H. and E. K. Sander (1966). "Age recognition from voice." *Journal of speech and hearing research* **9**(2): 273-7.

Paas, F., A. Renkl, et al. (2004). "Cognitive Load Theory: Instructional Implications of the Interaction between Information Structures and Cognitive Architecture." *Instructional Science* **32**(1): 1-8.

Ramig, L. L. A. and R. R. L. Ringel (1983). "Effects of physiological aging on selected acoustic characteristics of voice." *Journal of speech and hearing research* **26**(1): 22-30.

Reeves, B. and C. Nass (1996). *The media equation: how people treat computers, television, and new media like real people and places*.

Reich, A. R. and J. E. Duke (1979). "Effects of selected vocal disguises upon speaker identification by listening." *The Journal of the Acoustical Society of America* **66**(4): 1023-1028.

Remez, R. E., J. M. Fellowes, et al. (1997). "Talker identification based on phonetic information." *Journal of Experimental Psychology: Human Perception and Performance* **23**(3): 651-666.

Remez, R. E., P. E. Rubin, et al. (1981). "Speech perception without traditional speech cues." *Science* **212**(4497): 947-9.

Risset, J. C. and D. L. Wessel (1999). "Exploration of timbre by analysis and synthesis." In *The Psychology of Music*. D. Deutsch (ed.). San Diego, CA, Academic Press**:** 113-169.

Ritter, W., H. G. Vaughan, et al. (1968). "Orienting and habituation to auditory stimuli: A study of short terms changes in average evoked responses." *Electroencephalography and Clinical Neurophysiology* **25**(6): 550-556.

Rizzolatti, G., L. Fogassi, et al. (2001). "Neurophysiological mechanisms underlying the understanding and imitation of action." *Nature reviews. Neuroscience* **2**(9): 661-70.

Roads, C., J. Strawn, et al. (1996). *The computer music tutorial*. Cambridge, Mass. & London, UK, MIT Press.

Rodet, X. (2002). "Synthesis and processing of the singing voice." In *Proc.1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*(ed.). Leuven, Belgium**:** 99-108.

Rosch, E. (1975). "Cognitive representations of semantic categories." *Journal of experimental psychology. General* **104**(3): 192-233.

Rosch, E. (1978). "Principles of Categorization." In *Cognition and Categorization*. E. Rosch and B. B. Lloyd (ed.). Hillsdale, New Jersey, John Wiley & Sons.

Rosch, E. and C. B. Mervis (1975). "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive psychology* **7**(4): 573-605.

Roy, S. (1996). "Form and referential citation in a work by Francis Dhomont." *Organised Sound* **1**(1): 29-41.

Roy, S. (1998). "Functional and Implicative Analysis of Ombres Blanches." *Journal of New Music Research* **27**(1-2): 165-184.

Roy, S. (2003). *L'analyse des musiques électroacoustiques : Modèles et propositions*. Paris, L'Harmattan.

Ruz, M. and J. Lupiáñez (2002). "A review of attentional capture: On its automaticity and sensitivity to endogenous control." *Psicológica* **23**: 283-309.

Saberi, K. and D. R. Perrott (1999). "Cognitive restoration of reversed speech." *Nature* **398**(6730): 760-760.

Schaeffer, N. and N. Eichhorn (2001). "The effects of differential vowel prolongations on perceptions of speech naturalness." *J. Fluency Disord.* **26**: 335-348.

Schaeffer, P. (1952). *A la recherche d'une musique concrète*. Paris, Editions du Seuil.

Schaeffer, P. (2002). *Traité des objets musicaux*. Paris, Le Seuil.

Schaeffer, P. (2004). "Acousmatics." In *Audio culture. Readings in Modern Music*. C. Cox and D. Warner (ed.). New York, Continuum**:** 76-81.

Schafer, R. M. (1994). *Our Sonic Environment and the Soundscape: The tuning of the World*. Rochester, Vermont, Destiny Books.

Scherer, K. R. (1978). "Personality inference from voice quality: the loud voice of extroversion." *European Journal of Social Psychology* **8**(4): 467-487.

Scherer, K. R. (1979). "Personality markers in speech." In *Social markers in speech*. K. R. Scherer and H. Giles (ed.). Cambridge, Cambridge University Press**:** 147-209.

Scherer, K. R. (1986). "Vocal affect expression: A review and a model for future research." *Psychol. Bull.* **99**: 143-165.

Scherer, K. R. (1995). "Expression of emotion in voice and music." *J. Voice* **9**(3): 235-248.

Scherer, K. R. (2003). "Vocal communication of emotion: A review of research paradigms." *Speech Communication* **40**(1-2): 227-256.

Scherer, K. R., R. Banse, et al. (2001). "Emotion inferences from vocal expression correlate across languages and cultures." *J. Cross Cult. Psychol.* **32**(1): 76-92.

Scherer, K. R., D. R. Ladd, et al. (1984). "Vocal cues to speaker affect: Testing two models." *J. Acoust. Soc. Am.* **76**: 1346-1356.

Scherer, K. R. and M. R. Zentner (2001). "Emotional effects of music: Production rules." In *Music and emotion: Theory and research*. P. Juslin and J. A. Sloboda (ed.). New York, Oxford University Press**:** 361-392.

Schirmer, A. and S. A. Kotz (2006). "Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing." *Trends in Cognitive Sciences* **10**(1): 24-30.

Schlauch, R. S., S. E. Miller, et al. (2005). "Examining explanations for fundamental frequency's contribution to speech intelligibility in noise." *The Journal of the Acoustical Society of America* **118**(3): 1933-1933.

Schmidt-Nielsen, A. and K. R. Stern (1985). "Identification of known voices as a function of familiarity and narrow-band coding." *J. Acoust. Soc. Am.* **77**(2): 658-663.

Schröder, M. (2001). *Emotional Speech Synthesis: A review*. Proceedings: Eurospeech 2001, Aalborg, Vol. 1.

Schwartz, M. F. (1968). "Identification of speaker sex from isolated, voiceless fricatives." *The Journal of the Acoustical Society of America* **43**(5): 1178-9.

Schwartz, M. F. and H. E. Rine (1968). "Identification of speaker sex from isolated, whispered vowels." *The Journal of the Acoustical Society of America* **44**(6): 1736-7.

Schwarz, D. (1993). "Listening Subjects: Semiotics, Psychoanalysis, and the Music of John Adams and Steve Reich." *Perspectives of New Music* **31**(2): 24-56.

Schweinberger, S. R., A. Herholz, et al. (1997). "Recognizing Famous Voices: Influence of Stimulus Duration and Different Types of Retrieval Cues." **40**(2): 453-463.

Schötz, S. (2003). *Speaker Age: A First Step From Analysis To Synthesis*. Proceedings of the XVth ICPhS, Barcelona.

Schötz, S. (2004). *The Role of F0 and Duration in Perception of Female and Male Speaker Age*. Speech Prosody 2004, Nara, Japan.

Scott, B. L. and R. A. Cole (1972). "Auditory Illusions as Caused by Embedded Sounds." *The Journal of the Acoustical Society of America* **51**(1A): 112.

Scott, S. K. and R. J. S. Wise (2004). "The functional neuroanatomy of prelexical processing in speech perception." *Cognition* **92**(1-2): 13-45.

Segnini, R. and B. Ruviaro (2005). *Analysis of Electroacoustic Works with Music and Language Intersections*. ICMC 2005 Proceedings, Barcelona.

Shannon, R., F.-G. Zeng, et al. (1995). "Speech Recognition with Primarily Temporal Cues." *Science* **270**(5234): 303-304.

Sheffert, S. M., D. B. Pisoni, et al. (2002). "Learning to recognize talkers from natural, sinewave, and reversed speech samples." *Journal of experimental psychology. human perception and performance* **28**(6): 1447-69.

Sherburne, P. (2004). "Digital Dicipline: Minimalism in House and Techno." In *Audio Culture: Readings in Modern Music*. C. Cox and D. Warner (ed.). New York, Continuum.

Shrivastav, R., H. Hollien, et al. (2003). "Shifting perceptions of age in voice." *The Journal of the Acoustical Society of America* **114**(4): 2336-37.

Skipper, J. I., H. C. Nusbaum, et al. (2005). "Listening to talking faces: motor cortical activation during speech perception." *NeuroImage* **25**: 76-89.

Slaney, M. (2004). "The History and Future of CASA." In *Speech Separation by Humans and Machines*(ed.)**:** 199-211.

Sloboda, J. A. and P. Juslin (2001). "Psychological perspectives on music and emotion." In *Music and emotion: Theory and research*. P. Juslin and J. A. Sloboda (ed.). New York, Oxford University Press.

Smalley, D. (1986). "Spectro-morphology and structural processes. ." In *The language of electroacoustic music*. S. Emmerson (ed.). Baisingstoke, The Macmillan Press**:** 61-93.

Smalley, D. (1992). "The listening imagination: Listening in the Electroacoustic Era." In *Contemporary Musical Thought. Volume 1*. J. Paynter, R. Orton and P. Seymor (ed.). London, Routledge**:** 514-554.

Smalley, D. (1993). "Defining Transformations." *Interface* **22**: 279-300.

Smalley, D. (1997). "Spectromorphology: explaining sound-shapes." *Organised Sound* **2**(2): 107-126.

Smalley, D. (2007). "Space-form and the acousmatic image." *Organised sound* **12**(01): 35-58.

Smith, D. R. R. and R. D. Patterson (2005). "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age." *The Journal of the Acoustical Society of America* **118**(5): 3177-3186.

Smith, D. R. R., R. D. Patterson, et al. (2005). "The processing and perception of size information in speech sounds." *The Journal of the Acoustical Society of America* **117**(1): 305-318.

Smith, P. M. (1979). "Sex markers in speech." In *Social markers in speech*. K. R. Scherer and H. Giles (ed.). Cambridge, Cambridge University Press**:** 109-146.

Smith, Z. M., B. Delgutte, et al. (2002). "Chimaeric sounds reveal dichotomies in auditory perception." *Nature* **416**(6876): 87-90.

Stacey, P. (1989a). *Contemporary tendencies in the relationship of music and text with special reference to Pli selon pli (Boulez) and Laborintus II (Berio)*. New York, Garland Publ.

Stacey, P. F. (1989b). "Towards the analysis of the relationship of music and text in contemporary composition." *Contemporary Music Review* **5**(1): 9-27.

Sterne, J. (2003). *The audible past*. Durham & London, Duke University Press.

Stevens, C., N. Lees, et al. (2005). "On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference." *Computer Speech and Language* **19**: 129-146.

Stickney, G. S. and P. F. Assmann (2001). "Acoustic and linguistic factors in the perception of bandpass-filtered speech." *The Journal of the Acoustical Society of America* **109**(3): 1157-65.

Stockhausen, K. (1958). "Actualia." *Die Reihe* **1**(45-51).

Stockhausen, K. (1960). "Speech and Music." *Die Reihe* **6**: 40-64.

Stockhausen, K. (1992). "Gesang der Jünglinge (1955-56) Elektronische Musik (Programmtext zur Uraufführung 1956)." In *Stockhausen Edition no. 3 (Electronic Music 1952 - 1960), Booklet accompanying music CD*(ed.), Stockhausen Verlag**:** 41-73.

Sturm, J. A. and C. H. Seery (2007). "Speech and Articulatory Rates of School-Age Children in Conversation and Narrative Contexts." *Language, Speech, and Hearing Services in Schools* **38**(January issue): 47-59.

Sullivan, K. P. H. and F. Schlichting (1998). "A perceptual and acoustic study of the imitated voice." *The Journal of the Acoustical Society of America* **103**(5): 2894-2894.

Summerfield, Q., J. F. Culling, et al. (1992). "Auditory Segregation of Competing Voices: Absence of Effects of FM or AM Coherence [and Discussion]." *Philosophical Transactions: Biological Sciences* **336**(1278): 357-366.

Sundberg, J., M. N. Thörnvik, et al. (1998). "Age and voice quality in professional singers." *Logopedics, phoniatrics, vocology* **23**(4): 169-176.

Sundberg, J. J. (1987). *The science of the singing voice*. DeKalb, Ill., Northern Illinois University Press.

Sussman, E., I. Winkler, et al. (2003). "Top-down control over involuntary attention switching in the auditory modality." *Psychonomic bulletin & review* **10**(3): 630-637.

Sussman, E. S. (2005). "Integration and segregation in auditory scene analysis." *The Journal of the Acoustical Society of America* **117**(3): 1285-1298.

Tai, Y. F., C. Scherfler, et al. (2004). "The Human Premotor Cortex Is 'Mirror' Only for Biological Actions." *Current Biology* **14**(2): 117-120.

Tempelaars, S. (1996). *Signal Processing, Speech, and Music*. Exton, PA, Swets & Zeitlinger.

ten Hoopen, C. (1992a). "Abstract and mimetic qualities in electroacoustic music." In *Technology Amsterdam*(ed.). Amsterdam, Rodopi**:** 119-132.

ten Hoopen, C. (1992b). *Polarised listening strategies for electroacoustic music*. Secondo Convegno Europeo di Analisi Musicale, Trento.

Thomas, J. C. (2006). "Sur L'expression de la nature dans les musiques electroacoustiques," http://www.ac-rennes.fr/pedagogie/musique/dswmedia/sud_contexte_thomas.html (Retrieved 19.12, 2006).

Thoresen, L. (1985). *Auditive analysis of musical structures. A summary of analytical terms, graphical signs and definitions*. ICEM Conference on Electro-acoustic Music, Stockholm, Sweden, Royal Swedish Academy of Music.

Thoresen, L. (2007a) "Form-building Transformations." *The Journal of Music and Meaning* **4**(Winter), http://www.musicandmeaning.net/issues/showArticle.php?artID=4.3 (Retrieved 16.05.2009).

Thoresen, L. (2007b). "Spectromorphological analysis of sound objects: an adaptation of Pierre Schaeffer's typomorphology." *Organised Sound* **12**(02): 129-141.

Titze, I. R. (1989). "Physiologic and acoustic differences between male and female voices." *The Journal of the Acoustical Society of America* **85**(4): 1699-1707.

Tononi, G., G. M. Edelman, et al. (1998). "Complexity and coherency: integrating information in the brain." *Trends in Cognitive Sciences* **2**(12): 474-484.

Trask, R. L. (1996). *A Dictionary of Phonetics and Phonology*. London, Routledge.

Traunmüller, H. H. and A. A. Eriksson (2000). "Acoustic effects of variation in vocal effort by men, women, and children." *The Journal of the Acoustical Society of America* **107**(6): 3438-51.

Trent, S. A. (1995). "Voice quality: Listener identification of African-American versus Caucasian speakers." *The Journal of the Acoustical Society of America* **98**(5): 2936.

Tro, J. (2000). "Aspets of Control and Perception". *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00)*, Verona, Italy.

Trouvain, J., S. Schmidt, et al. (2006). *Modelling personality features by changing prosody in synthetic speech*. Proceedings of the Conference on Speech Prosody Dresden.

Truax, B. (2001). *Acoustic communication, second edition*. Westport, Connecticut, Ablex.

van Dommelen, W. A. (1990). "Acoustic parameters in human speaker recognition." *Language and speech* **33 (Pt 3)**: 259-72.

Van Lancker, D., J. Kreiman, et al. (1985a). "Familiar voice recognition: patterns and parameters. Part I: Recognition of backward voices." *Journal of phonetics* **13**: 19-38.

Van Lancker, D., J. Kreiman, et al. (1985b). "Familiar voice recognition: patterns and parameters. Part II: Recogntion of rate-altered voices." *Journal of phonetics* **13**: 39-52.

van Wijngaarden, S. J., H. J. M. Steeneken, et al. (2002). "Quantifying the intelligibility of speech in noise for non-native listeners." *The Journal of the Acoustical Society of America* **111**(4): 1906-1916.

Vaughan, M. (1994). "The human-machine interface in electroacoustic music composition " *Contemporary Music Review* **10**(Part 2): 111-127.

Verfaille, V., C. Gustavino, et al. (2005). *Perceptual evaluation of vibrato models*. Proceedings of Conference on Interdisciplinary Musicology, CIM05, Montréal, Canada.

Versfeld, N. J. and W. A. Dreschler (2002). "The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners." *The Journal of the Acoustical Society of America* **111**(1 Part 1): 401-8.

von Kriegstein, K. and A.-L. Giraud (2006). "Implicit Multisensory Associations Influence Voice Recognition." *PLoS Biol* **4**(10): 1809-1820.

von Kriegstein, K., A. Kleinschmidt, et al. (2006). "Voice Recognition and Cross-Modal Responses to Familiar Speakers' Voices in Prosopagnosia." *Cereb. Cortex* **16**(9): 1314-1322.

von Kriegstein, K., A. Kleinschmidt, et al. (2005). "Interaction of Face and Voice Areas during Speaker Recognition." *Journal of Cognitive Neuroscience* **17**(3): 367-376.

Vouloumanos, A. and J. F. Werker (2004). *Tuned to the signal: the privileged status of speech for young infants*. **7:** 270-276.

Vuilleumier, P. (2005). "How brains beware: neural mechanisms of emotional attention." *Trends in cognitive sciences* **9**(12): 585-594.

Wagner, I. and O. Köster (1999). *Perceptual recognition of familiar voices using falsetto as a type of voice disguise*. 14th International Congress of Phonetic Sciences, ICPhS99, San Fransisco.

Walker, J. F., L. M. D. Archibald, et al. (1992). *Articulation Rate in 3- and 5-Year-Old Children*. **35:** 4-13.

Walton, J. H. and R. F. Orlikoff (1994). "Speaker Race Identification From Acoustic Cues in the Vocal Signal." *J. Speech Hear. Res.* **37**: 738-745.

Warren, R. M. (1970). "Perceptual restoration of missing speech sounds." *Science* **167**(917): 392-3.

Warren, R. M. (1982). *Auditory Perception: A New Synthesis*. New York, Pergamon Press.

Warren, R. M. and R. P. Warren (1970). "Auditory illusions and confusions." *Scientific American*(233): 30-36.

Watkins, A. J. (2005). "Perceptual compensation for effects of reverberation in speech identification." *The Journal of the Acoustical Society of America* **118**(1): 249-262.

Watkins, K. and T. Paus (2004). "Modulation of Motor Excitability during Speech Perception: The Role of Broca's Area." *Journal of cognitive neuroscience* **16**(6): 978-987.

Watkins, K. E., A. P. Strafella, et al. (2003). "Seeing and hearing speech excites the motor system involved in speech production." *Neuropsychologia* **41**(8): 989-94.

Weale, R. (2005). *The Intention/Reception Project: Investigating the Relationship Between Composer Intention and Listener Response in Electroacoustic Compositions*. **PhD**. Leicester, De Montfort University.

Weale, R. (2006). "Discovering How Accessible Electroacoustic Music Can Be: the Intention/Reception project." *Organised Sound* **11**(2): 189-200.

Welch, G. and G. Bishop (2004) "An introduction to the Kalman filter." http://www.loria.fr/~berger/DEA/Documents/kalman_intro.pdf (Retrieved 27.09.2007).

Wendt, L. (1993). "Vocal Neighbourhoods: A Walk through the Post-Sound Poetry Landscape." *Leonardo music journal* **3**: 65-71.

Wesling, D. and T. Slawek (1995). *Literary Voice - The Calling of Jonah*. Albany, State University of New York Press.

Whalen, D. H. and C. E. Hoequist (1993). "Let your synthesizer breathe." *The Journal of the Acoustical Society of America* **93**(4): 2298-2298.

Wickens, C. D. (1991). "Processing resources and attention." In *Multiple-task performance*. D. L. Damos (ed.). London, Taylor & Francis**:** 3-34.

Wikipedia contributors (2007a). "Michael Winslow," *Wikipedia, The Free Encyclopedia* http://en.wikipedia.org/w/index.php?title=Michael_Winslow&oldid=99563121 (Retrieved, 10.01.2007).

Wikipedia contributors (2007b). "Vocal range," http://en.wikipedia.org/wiki/Vocal_range (Retrieved, 10.01.2007).

Wilson, S. M., A. P. Saygın, et al. (2004). "Listening to speech activates motor areas involved in speech production." *Nature neuroscience* **7**(7): 701-702.

Windsor, L. (1995). *A Perceptual Approach to the Description and Analysis of Acousmatic Music*, Unpublished doctoral thesis.

Wingfield, A., J. E. Peelle, et al. (2003). "Speech Rate and Syntactic Complexity as Multiplicative Factors in Speech Comprehension by Young and Older Adults." *Aging, Neuropsychology, and Cognition* **10**(4): 310 - 322.

Winkler, R., M. Brückl, et al. (2003). *The Aging Voice: an Acoustic, Electroglottographic and Perceptive Analysis of Male and Female Voices* Proceedings of the 15th ICPhS, Barcelona, Casual Productions.

Wishart, T. (1986). "Sound symbols and landscapes." In *The Language of Electroacoustic Music*. S. Emmerson (ed.). London, Macmillan**:** 41-60.

Wishart, T. (1988). "The Composition of Vox-5." *Computer Music Journal* **12**(4): 21-27.

Wishart, T. (1996). *On Sonic Art. A new and revised edition. S.Emmerson (ed.)*. London, Routledge.

Wishart, T. (2000a). "Sonic Composition in Tongues of Fire." *Computer Music Journal* **24**(2): 22-30.

Witmer, B. G. and M. J. Singer (1998). "Measuring Presence in Virtual Environments: A Presence Questionnaire." *Presence: Teleoperators and Virtual Environments* **7**(3): 225-240.

Wolfe, V. I., D. L. Ratusnik, et al. (1990). "Intonation and fundamental frequency in male-to-female transsexuals." *The Journal of speech and hearing disorders* **55**(1): 43-50.

Wood, N. and N. Cowan (1995). "The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel?" *Journal of Experimental Psychology: Learning, Memory, and Cognition.* **21**(1): 255-260.

Wouters, J. and M. W. Macon (2001). *Control of spectral dynamics in concatenative speech synthesis*. IEEE Transactions on Speech and Audio Processing.

Wu, K. and D. G. Childers (1991). "Gender recognition from speech. Part I: Coarse analysis." *The Journal of the Acoustical Society of America* **90**(4): 1828-40.

Waadeland, C. H. (2001). ""It Don't Mean a Thing If It Ain't Got That Swing" - Simulating Expressive Timing by Modulated Movements." *Journal of New Music Research* **30**(1): 23 - 37.

Xue, S. A. and D. Deliyski (2001). "Effects of aging on selected acoustic voice parameters: preliminary normative data and educational implications." *Educational Gerontology* **27**(2): 159-168.

Yarmey, A. D., A. L. Yarmey, et al. (2001). *Commonsense beliefs and the identification of familiar voices*. **15:** 283-299.

Yaruss, J. S. (2000). "Converting between word and syllable counts in children's conversational speech samples." *Journal of Fluency Disorders* **25**(4): 305-316.

Young, J. (1996). "Imagening the source: The Interplay of Realism and Abstraction in Electroacoustic Music." *Contemporary Music Review* **15**(I): 73-93.

Zadeh, L. A. (1965). "Fuzzy sets." *Information and control* **8**(3): 338-53.

Zetterholm, E., K. P. H. Sullivan, et al. (2002). *The impact of semantic expectation on the acceptance of a voice imitation*. Proceedings of the 9th Australian International Conference on Speech Science & Technology, Melbourne, Australia.

Musical recordings

Bayle, F. (1998), *Morceaux de Ciels - Théatre d'Ombres*, CD, MAGISON: MGCB 1298.

Berio, L. and B. Maderna (2006), *Acousmatrix 7*, CD, BV Haast Records: BVHaast 9109.

Bodin, L.-G. (1990), *Computer Music Currents 7*, CD, Wergo: WER CD2027-2.

Bodin, L.-G. (2006), *Winter Nightfall*, CD, Firework Edition Records: FER1061.

Coulter, J. (2005), *Shifting Ground*, Film-design for sound, DVD, Griffith University: (no catalog#).

Daoust, Y. (1998), *Musiques Naïves*, CD, empreintes DIGITALes: IMED 9843.

Deutsch, D. (2003), *Phantom Words, and Other Curiosities*, CD with booklet, Philomel Records: PHILOMEL-002.

Dhomont, F. (1996), *Forêt profonde*, CD, (SACEM) / YMX Média (SOCAN), empreintes DIGITALes: IMED 9634.

Dodge, C. (1994), *Any Resemblance is Purely Coincidental*, CD, New Albion Records: NA 043 CD.

Ekeberg, F. (2001), *Intra*, CD, PPP1001.

Eloy, J.-C. (1979), *Shànti*, LP(2), Erato: STU 71205/6.

Lansky, P. (1994a), *Fantasies and Tableaux*, CD, Composers Recordings, Inc.: CRI 683.

Lansky, P. (1994b), *More than idle chatter*, CD, Brigde Records: BCD 9050.

Lansky, P. (1997), *Things she carried*, CD, Bridge Records: BRIDGE 9076.

Lansky, P. (2002), *Alphabet Book*, CD, Bridge Records: BRIDGE9126.

Lejeune, J. (2000), *Messe aux oiseaux*, CD, INA-GRM: INA C2016.

Ligeti, G. (2006), *Requiem ; Aventures ; Nouvelles aventures*, CD, Wergo: WER 6925 2.

Manoury, P. (1998), *En Écho - Neptune*, CD, Accord: 206762 MU 750.

Norman, K. (2004b), *Losing it*, mp3-file, Downloaded at http://www.novamara.com/losingit.html, 18.09.2009.

Olsson, J. (1991), *Cultures électroniques 6 Quadrivium*, CD, Le Chant Du Monde: LDC 278053-54.

Parmerud, Å. (1994), *Osynlig Musik / Invisible Music / Musique invisible*, CD, Phono Suecia: PSCD 72.

Parmerud, Å. (1997), *Grains of Voices*, CD, Caprice: CAP 21579.

Ratkje, M. (2002), *Voice*, CD, Rune Grammofon: RCD 2028.

Reich, S. (1987), *Early Works*, CD, Nonesuch: 7559-79169-2.

Risset, J.-C. (1987), *Sud, Dialogues, Inharmonique, Mutations* CD, INA-GRM: INA C 1003.

Schaeffer, P. (1998), *Pierre Schaeffer - L'œuvre Musicale*, CD(3), INA-GRM: INA c1006-07-08.

Schafer, R. M., B. Truax, et al. (1996), *The Vancouver Soundscape 1973 / Soundscape Vancouver 1996*, CD(2), Cambridge Street Records: CSR-2CD 9701

Schwitters, K. (1992), *Ursonate - Sonate In Urlauten*, CD, Hat Hut Records: ART CD 6109

Stockhausen, K. (2001), *Elektronische Musik 1952-1960*, CD, Stockhausen-Verlag: Stockhausen 3.

Takemitsu, T. (2004), *Complete Takemitsu Edition vol.5 Popular songs, tape-music, music for the theater, radio and TV, addenda*, CD (14), Shogaku-kan: No catalogue #.

Teruggi, D. (2000), *The shining space*, CD, Sargasso (INA/GRM): SCD28033.

Thibault, A. (1990), *Volt*, CD, empreintes digitales DIGITALes: IMED-9003-CD.

Tutschku, H. (1999), *Moment*, CD, empreintes DIGITALes: IMED 9947.

Vande Gorne, A. (1998), *Impalpables*, CD, empreintes DIGITALes: IMED9839.

Various artists (1968), *Extended Voices*, LP, Odyssey: Odyssey 32 16 0156 (stereo).

Various artists (1988), *Cultures électroniques, Vol.3*, CD, Le Chant Du Monde: LDC 278048.

Various artists (1989), *Computer Music Currents 4*, CD, Wergo: WER 2024-50.

Various artists (1990), *Computer music currents 5: Music with computers*, CD, Wergo: WER2025-50.

Various artists (1992), *The Pioneers: Five Text-Sound Artists*, 2 x CD, Phono Suecia: PSCD 063.

Various artists (1993), *Computer Music Currents 1*, CD, Wergo: WE110.

Various artists (1997), *Cultures Electroniques 9 Magisterium*, CD, Le Chant Du Monde: LDC 278062.

Various artists (2005), *50 Years Studio TU Berlin*, DVD, Electronic Music Foundation: EMF DVD054.

Various artists (2006a), *The Composer In The Computer Age III*, CD, CDCM - Centaur: CRC 2213.

Various artists (2006b), *Deep Wireless 3, Radio Art Compilation*, CD(2), New Adventures in Sound Art: (no catalog#).

Various artists (2006c), *Text-Sound Compositions: A Stockholm Festival*, 5xCD, Fylkingen Records: FYCD1024.

Viñao, A. (1994), *Hildegard's Dream*, CD, MUSIDISC: MU 244942.

Westerkamp, H. (1996), *Transformations*, CD, (SOCAN) / YMX Média (SOCAN), empreintes DIGITALes: IMED 9631.

Wishart, T. (1992), *Red Bird - Anticredos*, CD, October Music: Oct 001.

Wishart, T. (2000b), *Voiceprints*, CD, Electronic Music Foundation Ltd.: EMF CD029.

Wishart, T. (2007), *Fabulous Paris: A Virtual Oratorio*, CD, Electronic Music Foundation: OT103.

Zanési, C. (1996), *Arkheion*, CD, INA-GRM: MUSIDISC 245 772.

CD-ROM contents

Folder	Subfolder	Description	File format
Sound examples		See list of sound examples.	.wav
Graphical representations - Six Fantasies	**Acousmographe representations**	These files can be opened with the *acousmographe* software, see Appendix B.	.aks
	Axial representations	These files should be loaded into the acousmographe representations. Cf. Appendix B.	.jpg
	Flash movie representations	These files can be viewed and listened to with most web browsers.	.swf
Six Fantasies Machine		Manual for the SFM instrument	.pdf
		Zipped archive containing the files for the SFM instrument	.zip

List of sound examples

All examples included on the CD-ROM are included with permission from the composers or copyright holders.

* Sound example not included on CD-ROM due to lacking permission from composer or copyright holder.

† © Copyright 1987 by Hendon Music, Inc. Reproduced by permission.

Example # and description	Duration (s)
Ex 4-1 Parmerud - Les objets obscures II 0'00''- 0'14''	15
Ex 4-2 Vinao - Chant d'Ailleurs 0'00''-0'30''	30
Ex 4-3 Decoust - Interphone 1'43''-3'00''*	76
Ex 4-4 Bodin - GYZO 1'00''-1'25''	25
Ex 5-1 Parmerud - Grains of voices 0'00''-0'23''	23
Ex 5-2 Lansky - Things she carried 2'52''-3'19''	26
Ex 5-3 Bodin - CYBO II 4'01''-4'19''	18
Ex 5-4 Lejeune - Messe-Christe eleison 0'00-1'04''	64
Ex 5-5 Bayle - derriere l'image II 0'11''-1'16''	65
Ex 6-1 Superimposed voices	15
Ex 6-2 Stretched vowel	24
Ex 6-3 Cleaver speech - increasing rate towards noise mode	16
Ex 6-4 Bush speech - looped	41
Ex 6-5 Cage - Solo for voice II 5'32''-6'20''*	47
Ex 6-6 Kaufmann - La voyage au paradis 6'15''-6'32	16
Ex 6-7 Reich - Come out 0-1'15''†	75
Ex 6-8 Bodin - Anima 0'00''-0'22''	22
Ex 6-9 Dodge - In celebration 4'13''-4'21'	8
Ex 6-10 Tutschku - Les invisibles 8'15''-8'27''	11
Ex 6-11 Wishart - Globalalia 0'00''-0'10''	9
Ex 6-12 Coulter-Shifting Ground-1'54''-4'22''	148
Ex 7-1 Vowels without and with microfluctuations	9
Ex 7-2 Vinao-Chant d'Ailleurs 0-0'30''	30
Ex 7-3 Dodge - Speech songs I 0'09''-0'15''	5
Ex 7-4 Ratkje - Chipmunk Party 0'00''-0'16''	16
Ex 7-5 Wishart-VoxV-2'39''-2'45''	6
Ex 7-6 Eloy-Shanti-6'48''-7'05''	17
Ex 7-7 Dodge - In celebration-You want to wave 4'13-4'16''	2
Ex 7-8 Dodge - In celebration-So you wait 6'58''-6'59''	1
Ex 7-9 Dodge - In celebration-Cannot raise your hand 4'15''-4'22''	6
Ex 7-10 Dodge - In celebration-You have seen it 4'06''-4'13''	5
Ex 7-11 Dodge - In celebration-Feeling your lungs 6'28''-6'37''	9

Ex 7-12 Dodge - In celebration-In a 4'26''-4'29''	2
Ex 7-13 Dodge - In celebration-Shei 0'19''-0'26''	6
Ex 7-14 Dodge - In celebration-Bad 3'05''-3'13''	7
Ex 7-15 Dodge - In celebration-You 0'10''-0'11''	1
Ex 8-1 Ratkje - Intro 1'03''-1'06''	2
Ex 8-2 Olsson Up 10'25''-10'36''	11
Ex 8-3 Wishart - Blue Tulips 0'28''-0'45''	16
Ex 8-4 Westerkamp - Kits beach soundwalk 1'13''-1'47''-without and with modification	67
Ex 8-5 Takemitsu - VocalismAI 0'13''-0'21''	7
Ex 8-6 Ekeberg - des cantiques 2'23''-2'52''	28
Ex 8-7 Westerkamp - Kits Beach Soundwalk 0'20''-0'50''	30
Ex 8-8 Zanezi - les mots des Stockhausen 4'24''-4'30''*	5
Ex 8-9 Takemitsu - VocalismAI 3'22''-3'28''	6
Ex 8-10 Bodin - For Jon 15'54''-16'12''	17
Ex 8-11 Lansky - Notjustmoreidlechatter 0'00''-0'20''	20
Ex 9-1 Rapisarda - The day before 0'08''-0'19'' - with and without processing	26
Ex 9-2 Westerkamp - Kits beach soundwalk 0'00''-1'37'	96
Ex 9-3 Norman - Losing it 0'07''-0'48''	41
Ex 9-4 Wishart - Stentor 0'33''-0'48''	19
Ex 9-5 Wishart - The Division of Labour 3'30''-4'10''	40
Ex 9-6 Eloy - Shanti 6'34''-7'00''	25
Ex 10-1 No masking due to frequency separation	25
Ex 10-2 Masking due to activation overlap	25
Ex 10-3 Tail masking from heavy reverberation	4
Ex 10-4 Reverse masking	6
Ex 10-5 Bi-directional masking	4
Ex 10-6 Wishart - Red Bird 19'10''-19'20''	10
Ex 10-7 Lansky - Things she carried 0-1'01''	61
Ex 10-8 Dhomont - A l'oree du conte 0'27''-0'40''	13
Ex 10-9 Coulter - Shifting Ground 1'58''-2'12''	14
Ex 11-1 Seven superimposed vowel sequences transposed with 5hz difference	6
Ex 11-2 Vocal sound with detuned partial	5
Ex 11-3 Vocal sound with partial in separate channel	5
Ex 11-4 Vocal sound with delayed partials	6
Ex 11-5 Word loop speeding up	9
Ex 11-6 Vocal sound with FM partials	5
Ex 11-7 Four superimposed transposed voices	3
Ex 11-8 Sine-wave speech	12
Ex 11-9 Manoury - En echo - section I, event 75	11
Ex 11-10 Bodin - Wonder-void 9'46''10'01''	15
Ex 11-11 Wishart - The Division of Labour 7'28''-7'36''	7
Ex 11-12 Hans Tutschku - Verfall 2'56''-3'00''	4
Ex 11-13 Trevor Wishart - Facets 0'00''-0'06''	6
Ex 11-14 Trevor Wishart - Facets 1'22''-1'30''	8

Ex 12-1 Lansky - her voice 1'16''-1'42''	27
Ex 12-2 Lansky - her presence 0'36''-1'12''	37
Ex 12-3 Lansky - her reflection 2'31''-2'58''	26
Ex 12-4 Lansky - her song 2'25''-3'05''	44
Ex 12-5 Lansky - her ritual 2'00''-2'43''	44
Ex 12-6 Lansky - her self 0'37''-1'05''	29
Ex 12-7 SFM - Recombined words, no accompaniment	3
Ex 12-8 SFM - Phrases with different affective content - AF focus	10
Ex 12-9 SFM - Phrases with different identities	13
Ex 12-10 SFM - Information density - minimal noise mode	8
Ex 12-11 SFM Extreme low speed articulation	41
Ex 12-12 SFM Continuous transformation - increasing naturalness	10
Ex 12-13 SFM Technological artifacts	24
Ex 12-14 SFM Spatial motion - varying presence	14
Ex 12-15 SFM On-off-masking	5
Ex 12-16 SFM Gradually increasing feature salience	7
Ex 12-17 SFM Gradually lowered masking	9
Ex 12-18 SFM Lower to higher stream integration	5

Appendix A: Key to the IPA transcriptions

Vowels:

[iː] – beat (long vowel)

[ɪ] - bit

[ɛ] - bet

[ɜː] – bird

[ə] – about

[æ] – bat

[ɒ] – hot (Received Pronunciation)

[aː] – arm

[ʊ] - put

[ʌ] - but

[uː] – soon

[ɔː] – saw (Received Pronunciation)

Diphtongs

[eɪ] - bait

[aɪ] - bite

[ɔɪ] - boy

[oʊ] – no (General American)

[əʊ] – no (Received Pronunciation)

[juː] - cute

[aʊ] - now

Consonants

[b] – but

[p] – put

[t] – tear

[f] – fit

[g] – gear

[k] – kit

[d] - deal

[v] – veal

[l] – level

[n] – noon

[m] – mum

[ŋ] – thing

[h] – horse

[z] – zip

[ʒ] – pleasure

[ʤ] – fudge

[θ] – teeth

[ʃ] – she

[ð] – there

[ʧ] - chair

[ɹ] – arm

[s] – sit

[w] – we

Appendix B: Guide for using the acousmographe representations

Using the Acousmographe files.

In this appendix I will describe how to install, prepare and use the accompanying **acousmographe** files.

1. Installing Acousmographem, version 3.4

1.1 Download the zipped .exe or .app file from the **acousmographe** web site according to your system: http://www.ina.fr/entreprise/activites/recherches-musicales/acousmographe.html

1.2 Unzip.

1.3 Run the setup file. Follow the instructions. This will install **acousmographe** on your machine. You might also have to install ASIO4All to have the appropriate sound drivers for the program, unless this is already installed on your machine.

1.4 Copy the files from the accompanying CD-ROM to your machine.

2. Open and prepare the .aks file

2.0 Start **acousmographe 3.4**. You will be asked to register by choosing a user name and sending an e-mail to acousmographe.inscription@ina.fr. Please choose register and then check your e-mail for a key code. Type or copy-paste the key in the registration code field to finish the registration. The program will also run if you choose not to register, but you will be prompted to register later.

2.1 Open one of the six .aks files in the “Acousmographe representations” folder by choosing <Open> from the <File> menu or typing ctrl+o, select the file of choice, and click <OK>.

2.2 **Acousmographe** will probably tell you that it can’t find the sound file for this acousmography, and ask you if you want to locate this file. Choose <OK> to locate the file.

2.3 Navigate to the “Sound examples” folder on the CD-ROM or alternatively to the folder where you have copied the sound examples.

2.4 Choose the file corresponding the movement you chose in 2.1. Click <OK> to confirm. **Acousmographe** will now compute a spectrogram from the sound file. This may take some time.

2.5 Make sure that you the whole vertical range of the acousmography, i.e. from 0 to 19000 Hz. If you don't, use the vertical scrolling bar to the right along with the < - > just below the scrolling bar until the view shows the whole range. Note that there are a group of objects in the upper part of the acousmography that shows the text "image" written diagonally in a red font.

2.6 Navigate to the beginning of the acousmography by using the horizontal scrolling bar in the bottom of the window. Make sure that <Multi-selection> is turned on in <Options> in the <Score> menu. Check this option if it is not already checked. Then double-click on the leftmost of the "image" objects. An inspector will pop up showing properties of the image.

2.7 Click <Select image…>. Then navigate to the "Axial representation" folder and the containing folder corresponding to the .aks file chosen in 2.1.

2.8 The name of the picture file should now be prompted in the field where you can type the file name to choose. By clicking <OK> the appropriate picture file will be chosen.

2.9 Repeat the points 2.7 and 2.8 for all of the image objects in the acousmography.

2.10 Save the acousmography file on your hard drive to keep the changes made. You are now ready to start using the file.

2.11 Repeat the points 2.1 to 2.10 for the remainder of the .aks files.

3. Using the acousmography.

3.1 Press the <space bar> or press the play button (green triangle) in the <Player> window (if you can't see it, press F2) to start playing the sound. If you can't hear any sound, check the ASIO4All settings on your system. Feel free to check out the possibilities of playing back different portions of the file, looping playback, playing back at different speeds etc. by using the playback controls in the Player window (F2).

3.2 Choose a view that allows you to see the portions of the acousmography that you want to study by using the vertical and horizontal < + > and < - > buttons. You can also choose <Zoom values> in the <View> menu and then choose the upper and lower values for time and frequency. Choosing 0 to 17000Hz and 0 to 10s will usually give you a good starting point.

3.3 Explore the possibilities of showing and hiding objects in the acousmography by checking or un-checking the buttons for the different layers in the left column of the window. If you want to study the evaluation of one particular premise it is advised that you hide all the others than the one you want to study.

3.4 Saving at any point will keep also your choices for what portions of the acousmography that will be visible and what layers you will show and hide.