Chapter 1

Phonetic Knowledge in Speech Technology
- and phonetic knowledge from speech technology?

William J. Barry
Institut für Phonetik, Universität des Saarlandes, Germany

Wim A. van Dommelen
Department of Language and Communication Studies, NTNU, Trondheim, Norway

Jacques Koreman
Institut für Phonetik, Universität des Saarlandes, Germany[1]

Abstract: The contributions to this volume are considered within the framework of the question: “What sort of phonetic knowledge is relevant to speech technology?” This discussion throws light on the existing and the potential relationship between speech technology and the phonetic sciences, the possibilities for mutual gain and the need, ultimately, for researchers to emerge who combine the interest and expertise needed in both areas.

Key words: Phonetic knowledge, speech synthesis, speech recognition

1. INTRODUCTION

Although speech recognition and speech synthesis started out as much the territory of phoneticians and other linguists as of engineers, the linguistic approach soon lost terrain, in recognition applications at least, to (non-linguistically orientated) engineers who were less concerned with formal linguistic insights, treating the signal as a pattern just like any other, and this with outstanding success. But the successes of engineering approaches are seen to have limits, most clearly in the challenges of spontaneous speech recognition and expressive speech synthesis, and once more the question arises whether the inclusion of additional linguistic, and more specifically phonetic knowledge is warranted. Of course, it is the degree of success so far which raises our sights to higher targets and exposes the limitations of techniques which were devised for the tasks already (more or less) accomplished. With continuous (read or rehearsed) speech recognition systems commercially available, the drastic drop in performance found with *spontaneous speech* suggests that a ceiling may have been reached with the current processing methods. Similarly, with the intelligibility of speech synthesis systems no longer causing basic problems, the interest in and the calls for increased *naturalness* and *expressivity* in speech synthesis have become stronger. Here too there exists the realisation that the most successful approach to commercial synthesis, namely (fixed or variable) speech unit concatenation, while impressively natural within restricted domains, cannot provide the *flexibility of expression* together with the naturalness that is ultimately required. It is therefore practically limited. Finally, from the research point of view, i.e., in terms of learning how the production and perception of speech works, it is not theoretically satisfying.

The call for the inclusion of phonetic knowledge, however, presupposes both that the required knowledge is available, and that it exists in a form which is exploitable by the speech technology application. The question whether it is correct to make that assumption can be answered for both aspects with “partly”, and the degree to which it is correct varies with the application being considered. In other words, there is certainly a lot of phonetic knowledge available which is not being used and which is relevant to speech recognition or synthesis (Pols 1999), but much of it does not exist in a form in which it can be used immediately. But of course there is also a great deal about the phonetic structuring of speech which is *not* understood, which could be of help to those speech technologies, and which has come to the notice of phoneticians as a result of contact to the field of speech technology. Thus, a simple answer to the question in our sub-title is “yes”.

Before expanding on this issue we consider the individual contributions to this volume in terms of the way they include phonetic knowledge in the application they are presenting or considering. Alternatively, in the case of the discussion papers we consider their stance with regard to the potential integration of phonetic knowledge.

2. PHONETIC KNOWLEDGE IN THIS VOLUME

The six papers presented at the symposium comprise five which were selected from among the abstracts submitted because they appeared to promise studies that represent different approaches to the theme of integrating phonetic knowledge in speech recognition or speech synthesis. The sixth paper, by panel member Steve Greenberg, brings together empirical results and general discussion. The final form of the papers does in fact reveal a diverse understanding of that theme. Four of them are reports of experimental applications. But Batliner & Möbius present a discussion of principle rather than details of a practical application (though they point to an example instantiation reported elsewhere), and Greenberg presents both a discussion of principle and concrete analysis examples from his own work. Among the discussants too, there was a healthy spread of opinion on the issue. The reworking and expansion of the papers and the post-hoc formulation by the panel members of the opinions developed during the ESE discussion session and presented in this volume underline the different perspectives. The following review of the chapters does not strictly separate the papers presented at the symposium from the panel members' discussions, but rather progresses from the purely empirical analyses to the theoretical discussions, removing the line between speakers and discussants.

Two of the empirically orientated papers were concerned with decoding the linguistic structure from the acoustic signal (Carson-Berndsen & Walsh; Christensen et al.) and two addressed the question of relating the decoded structure to representations in the lexicon, namely the problem of multiple pronunciations (Gravier et al.; Pastor & Casacuberta).

In their "Phonetic Time Maps", Julie Carson-Berndsen & Michael Walsh indicate how ASR can be made more robust by implementing phonetic and phonological constraints in a computational linguistic speech recognition model. The constraints can be used to guide the interpretation of multilinear event representations and to provide top-down predictions. The *Time Map Model* contains representations of the phonotactic constraints in a language. A special feature of the *Time Map Model* is that the phonotactic automata are defined with respect to the *syllable* domain. *Phonetic Time Maps* model phonetic details like the realisation of plosives (e.g., with/without release) or neutral vowels (e.g., elision before a nasal). The knowledge invoked in this approach is, in the first instance, of the type derived from traditionally established observations about allophonic variants and post-lexical modifications to the phonetic string that are captured in context-sensitive statements on assimilation and elision processes. What is particularly interesting about their processing framework is that it can operate both at the level of categorical, constraint-based representations of this knowledge and with a quantitative, probabilistic input to determine the ranking of such constraints.

Heidi Christensen, Børge Lindberg & Ove Andersen describe an ASR system for which the central issue is the exploration of multi-source recognition, which they term "heterogeneous processing". That is, the extraction of complementary phonetic information in different processing streams to provide more robust decoding. So-called “Expert MLPs” supplement the core (fullband; multiband) MLP systems; these are a “*voicing* expert” and a “*broad class* expert". The phonetic insight behind this approach is similar to that which motivates Carson-Berndsen & Walsh, namely the contextually determined change in the segmental identity of an underlying phonetic string. It also rests on the fact that a coarser definition of a segment can be more helpful for lexicon access than an incorrect decision at the phonemic level. In addition it appeals explicitly to parallels with human processing, which has recourse to different temporal and frequency granularities in order to cope with signal degradation.

There is no top-down component in the system other than the choice of “expert”; the whole process is data-driven. A number of other experts could have been chosen, but the two “experts” that were defined are plausible candidates in that the phonological voiced-voiceless opposition is extremely varied in its phonetic realisation, and changes to the phonetic properties of phonemic categories often result in shifts within the same broad class. It is presumably in this sense that the experts are seen as complementary to the *stem* system. In common with all stochastically orientated models, of course, the broad-class decisions and the voiced-voiceless decisions are as dependent on global probabilities as the phoneme decisions made by the stem system. It has no means of specifying the different contextual factors that are known to influence the changes, though it might be argued that this is catered for in the 7-frame (~ 100ms) time base used for training.

Guillaume Gravier, François Yvon, Bruno Jacob & Frédéric Bimbot model contextual constraints on the phonetic forms of words at the search level to limit the search space to permissible pronunciation sequences. Using existing French lexicon resources containing pronunciation variants, they derive morpho-syntactically and phonologically context-sensitive rules to predict liaison, mute-e deletion and liquid consonant truncation.

A slight improvement in performance is found, a success in the light of the reduced search space that the approach offers. More interesting than this modest applicational success within the frame of this volume is, however, the concluding discussion of possible reasons why the results were not more convincing. It highlights the interactions between phonetic factors (production task and speaking style), phonetic modelling complexity, the lexicon resource and the constraint definition.

Moisés Pastor & Francisco Casacuberta derive word-pronunciation variants using stochastic finite state automata to relate the phoneme output of a recognizer to the canonical pronunciation. Pronunciation alternatives are chosen on the basis of three different criteria: number of pronunciations, cumulative percentage, and threshold percentage. The results support the viability of the threshold-percentage criterion. Rather than theoretically discussing the possible use of phonetic knowledge in speech recognition, the authors experimentally show that pronunciation modeling should take into account articulatory reality. Whereas canonical models fail to do justice to the strong pronunciation variation due to deletions, assimilations and reductions, etc., modeling of frequently occurring pronuncation variants can (as also found by others) lead to improved recognition rates. In terms of the added value from this result, either for or from phonetic knowledge, the study confirms that multiple use of the same word will result in a variety of forms, and that the more a word is used, the more likely it will be to deviate from the canonical form.

The four contributions discussed so far all take “phonetic knowledge” at the general level of contextually based phonetic variation into consideration, but they vary considerably in the degree to which differentiated phonetic observations are or can be included. Also, they are all primarily and explicitly involved with the automatic *recognition* process, although contextually differentiated word forms may be one of the crucial aspects, so far neglected, for achieving more natural speech *synthesis*.

Coming now to the two more discussion-orientated and reflective of the six papers, Anton Batliner & Bernd Möbius address the question of knowledge integration in both recognition and synthesis. They are specifically concerned with the different demands placed on *prosodic* knowledge in automatic speech *understanding* (ASU) and text-to-speech *synthesis* (TTS). They introduce the distinction between phonetic-phonological knowledge and phonetic-phonological models and argue for the use of prosodic knowledge rather than prosodic models within ASU. Their standpoint is that models are an abstraction from phonetic reality and therefore introduce a quantisation error into the relationship between the phonetic form and the syntactic or semantic function. Rather than using subtle theoretical concepts, clear and stable prosodic markers need to be identified in order to define phrase boundaries and intonationally (and thus also informationally) important elements.

For synthesis, phonetically detailed prosodic events need to be generated (such as timing of tonal peaks in accented words dependent on consonant features, number of syllables, etc.), but though these events are clearly functional in demarcative, sentence-modal and information-structural terms, there seems to be no way of circumventing the intermediate phonological representation. Different ideologies behind these representations are also seen as a problem, as is the relationship between text and information- or discourse-structure which determines the prosodic form. With regard to a unified solution for intonation modeling in ASU and TTS, which is seen as ultimately desirable, the authors conclude that a common basis is not yet in sight. However, they do go on to discuss the sort of activities that are necessary in the phonetics and speech-technology community to move towards this goal.

In a more generally orientated discussion paper, Steve Greenberg discusses the fundamental importance of the two-way relationship between speech science and technology, i.e., of melding phonetic insight with speech technology to improve both the applications and the basic science. He sees the study of the large, naturally produced speech databases used in speech technology as a way to correct the largely unrealistic picture of speech and language projected by traditional phonetic and linguistic research, which is based largely on small-scale, carefully controlled and read material. In other words, speech and language science can and will improve. But he also sees that the successes in speech technology applications rest, in part, on imperfect scientific foundations, and that increasing demands, driven by the successes so far, are uncovering the limitations.

Greenberg illustrates this conviction with analyses of the Switchboard spontaneous speech database which uncover systematic relations between the prosodic-phonological category of stress accent and acoustic phonetic properties like duration and amplitude. The analysis results are presented both as a relationship of quantitative-phonetic properties to phonological categories, i.e., in terms of enhanced scientific insights, and as technologically exploitable facts. He presents the dramatic effect of stress-accent differences on the recognition performance (in terms of deleted words) of eight different recognition systems. Other relationships which are shown are those between word error rate and syllable structure on the one hand, and between stress accent and vowel identity on the other, which can also be of applicational importance. Importantly, within the framework of this volume, Greenberg illustrates both the gains in phonetic knowledge that come from asking phonetic questions of large databases – the relationships he uncovers are by no means predictable from current phonetic or phonological theory –

and the vital role that the careful phonetic labelling of such databases plays in that process.

Two of the panel members (Jan van Santen and Helmer Strik) take the fundamentally separate worlds of Speech Technology and Phonetics as their point of departure, van Santen concentrating on the implications for speech synthesis, while Strik's discussion is implicitly directed towards speech recognition.

Van Santen presents a relatively optimistic picture of the potential for integrating phonetic knowledge in speech synthesis, particularly with respect to making text-to-speech domain independent, even though there is little evidence of real cross-fertilization to date. On the contrary, developments in speech synthesis technology over the decades indicate a steady divergence from the level of phonetic theory: rule based methods gave way to fixed inventory concatenative techniques, and these appear to be in the process of being superceded by large-corpus based, variable-unit TTS; i.e., with apparently less emphasis on phonetic knowledge. However, linguistics may provide the type of knowledge that is needed to handle unseen unit types, which are still a problem in concatenative systems. This is illustrated by van Santen with reference to the different parts of the Bell Labs text-to-speech system that have been informed by phonetic knowledge: text analysis (computing phonemes; prosodic tags); duration modelling; intonation modelling; signal processing (special coarticulatory facts; segment lengthening details, etc.). He identifies the types of phonetic knowledge as: speech production/perception studies; architectural design; language dependent details (phonotactics, coarticulation, etc.); parameterized mathematical models.

One area in which van Santen particularly sees the need for phonetic support is in the perceptual evaluation of concatenation and signal manipulation techniques, e.g., thresholds for spectral and F0 discontinuities; subsegmental timing; vowel reduction; JND's for pitch contours. But there is also a clear knowledge deficit in the multidimensional modelling of prosodic features, particularly in relation to the definition of the properties covarying in emotional speech. Like others, he sees the potential for a phonetic contribution in a modified paradigm for phonetic research, in the development of a bridging field for research between phonetics and speech technology which he terms *computational phonetics*.

Helmer Strik has a rather less optimistic expectation for bringing the two different worlds of Speech Technology and Phonetics together. As negative examples of potentially useful, but in practice unusable phonetic knowledge, he takes segment duration and lexical stress to show the difference between quantitatively supported insights and computationally usable analytic data. More positively, he shows the possibility of a phonetically oriented point of departure in pronunciation variation modelling: Rule knowledge is used, but the essential probabilities have to be derived from the data. This underlines his view that the existing phonetic knowledge is not complete and, above all, that it needs to be presented in probabilistic terms. As further illustration of the incompleteness of phonetic, and more generally linguistic knowledge he points out that prosodic models are rarely used in speech recognition, among other things because of the almost exclusive focus on F0, and that language models are based on written rather than spoken language. His conclusion is that using phonetic/linguistic knowledge in Speech Technology *can* be useful – improvement at the signal-processing level, for example, has been achieved due to knowledge about human auditory perception – but he clearly sees its use restricted to achieving a last few percent improvement.

Bill Ainsworth's discussion takes a long-term view of the Speech Technology scene, registering the divergence over the decades of speech technology methodology from the phonetic foundation, which focused on the facts of production and perception. To underline this he points out that hidden Markov models, the dominant approach in ASR, are very unrealistic models of speech production. Despite the positive point made by Strik (see above), a neglected aspect of speech science knowledge in terms of speech technology exploitation is the human auditory system, though it is partly modeled in modern ASR (multi-layer perceptrons; multi-band processing). Fundamental research into the physiology and neuro-anatomy of hearing has progressed greatly in recent years without its potential for speech signal processing having been exploited. To integrate phonetic knowledge in Speech Technology we need to base recognition and synthesis on realistic models of audition and speech production. Alone among the contributors he stresses the need to develop new mathematical models – though he does not claim to know what form they are likely to take – to cope, e.g. with the crucial fact that the *underlying control gestures* in speech production overlap.

3. QUO VADIS PHONETIC KNOWLEDGE

The contributions to this volume can be viewed both as a reflection of existing limits to the integration of phonetic knowledge in speech technology applications and as pointers towards ways in which more knowledge can be of use in the future. We would like to bring those pointers together to a more general statement, and to give the reader a backdrop against which to consider the message of the individual contributions.

In his San Francisco ICPhS XIV keynote address, Louis Pols (1999) also addressed the question of Phonetics being of use to Speech Technology and vice versa. He took the difference between human and machine decoding as a point of departure, not because machine recognition should orientate itself in terms of processing principles on human recognition, but merely because it highlighted the potential for improvement. Imitation of human functionality, not duplication of human processing should be the aim. Understanding the limitations of the machine system and what makes it less “flexible, robust and efficient” (p. 9) than humans might contribute to improvement of the system. Within the present volume, Ainsworth is most explicit in taking this line of argument and pointing the finger at the Hidden Markov approach as an example of very powerful modelling which diverges fundamentally from the functionality of human speech production (and, one should add, of speech perception). None of the authors points the finger at concatenative synthesis as being perhaps *even further removed* from that functionality, lacking, as it does, the basic independence of the source from the filter characteristics of the system. Without that independence, naturally expressive synthesis is practically impossible. Thus, the message would appear to be that the courage to backtrack and reassess is necessary in both the main areas of speech technology. Pursuing established and hitherto very successful approaches might just be leading into a cul-de-sac.

Understanding what makes human speech decoding flexible, robust and efficient is, in broadest terms, *psycholinguistic* knowledge, part of which is more strictly *phonetic* knowledge. But that knowledge is certainly not normally couched in terms that can be directly integrated into automatic speech recognition, as many people in the past, including several in this volume, have pointed out. For speech technology and phonetics to have direct mutual benefit from each other, comparable data and comparable data representation are necessary. Viewing this from the speech technology vantage, Roger Moore (1995) used the term “computational phonetics” (cf. also van Santen in this volume), but a common data representation is perhaps an illusory aim. For one thing, even within speech technology, many different forms of data representation are required, depending on the task at hand. As Pols (1999, p.9) points out, average formant values for vowels are of no use for vowel recognition in different contexts though they may be sufficient for formant synthesis. One might add that for building a diphone synthesis system not even vowel formant values are necessary, merely the knowledge that the quality of a vowel changes systematically along its time course as a product of local context, making the diphone a sensible building block.

This differentiation should highlight the difference between knowledge in the form of an *insight* into a phenomenon, the quantitative *specification* of that phenomenon (which may have led to the insight), and the *format* (e.g., average values, probability density functions or CART-trees) in which the quantitative data can be used for a particular application.

What is presumably meant by mutual benefit to the two different disciplines is (re)presentation in terms of the other discipline’s problems, questions and aims. A phonetician always looks at data in terms of trying to develop an explanatory model for a human’s ability to produce or perceive speech. However, she/he cannot be expected to deliver analysis results in the form required e.g. for a particular recognition algorithm or a particular synthesis system. This would be equivalent to expecting a speech technology engineer to ask phonetic questions of a database to gain his/her own insights. In fact, if the results are a new insight which could be important for speech technology, there is possibly no ready form of representation available; its exploitation might well require a new algorithmic approach. What *can* be expected, however, is that the analysis is carried out on data that is *relevant* for a particular application, and that the observation is at a *level of delicacy* that is relevant for the task. Finding an effect e.g. of a particular contextual factor, when the speech material has been carefully controlled and all possible confounding factors excluded, will not generalize to any realistic ASR task. Finding a robust effect in a large continuous-speech database is something else, however, and there should then also be an interest to communicate the implications of the observation in a manner which members of both communities can understand.

Communicating, on the one hand, what phenomena are clearly functionally important, and, on the other hand, saying how they should be dealt with in a speech technology application are two very different things. While the linguist is implicated in the former task, we suggest that it is the task of the speech technology scientist to undertake the latter. An example of the former is the well established simultaneous global and local importance of duration (cf. Pols 1999, p. 12). Within any given tempo (varying locally within a global frame) there are globally calculable durational differences between phonemically long and short segments. In many languages, there are locally determined allophonically longer and shorter variants and local durational increases related to accentuation (which in turn is related to information structure). Finally, there is the local phrasal function of final lengthening. With regard to dealing with such functionally important variation within ASR, these insights present a strong challenge because they certainly cannot be exploited within present-day stochastical methods, dependent as they are on global probabilities. However, their functional and communicative importance has to be understood and accepted otherwise the challenge will not be recognized.

What emerges very clearly from this and other discussions is the need to understand both sides of the problem. Viewed within the present structures of science, the need for interdisciplinary interest and cross-disciplinary activities is undeniable. The greater access phonetically trained researchers have to the databases and tools used in mainline technology applications, the more likely it is that quantitative answers to phonetic questions can be presented in a way which can be useful for speech technology applications. Conversely, speech technology engineers will be increasingly prepared to look for innovative processing solutions, the more contact they have with quantitatively supported statements about the complex relationships between the relatively simple signal parameters (duration, intensity, frequency and spectral energy distribution, and their derivatives) and the communicative functions they are trying to decode (ASR) or encode (synthesis). What is certainly not to be expected as a rule at present is the phonetician who can develop new processing algorithms or the speech technology engineer who can ask new phonetic questions of a speech database.

However, a certain indication of the developing contact in the two areas of science can be gained from looking at the change in phonetically orientated contributions to Eurospeech conferences during the twelve years from Eurospeech I in Paris, 1989 to Eurospeech Scandinavia in Aalborg 2001. Although weaker than the growth in purely technology-orientated papers, the papers dealing with phonetic questions or integrating phonetics in technological applications grew by a very substantial 45% from 93 to 135. Ultimately, as a product of this increasing contact between the two disciplines more exemplars of the currently rare hybrid scientist should appear: the "linguist speech-technology engineer" and the "speech-technology linguist", or to borrow Roger Moore's and Jan van Santen's term, the *computational phonetician*.

Ultimately, progress towards and in interdisciplinary research, like other human interactions, is the product of the individuals involved. They must be interested and committed. But we echo van Santen's comment (this volume) that changes are sociologically determined, and a framework for contact and interaction is needed. A symposium and a published discussion are a first step in the right direction, inter-departmental courses and inter-disciplinary degrees are a further goal. However, any change of socio-scientific climate must also be triggered and established by individuals.

6. LITERATURE

Moore, R. (1995). "Computational phonetics", *Proceedings ICPhS 1995*, Stockholm, Vol. 4, p. 68-71

Pols, L. (1999). "Flexible, robust, and efficient human speech processing versus present-day speech technology", *Proceedings of ICPhS 1999*, San Francisco,Vol. 1, p. 9-16.