EXPLOITING TRANSITIONS AND FOCUSSING ON LINGUISTIC PROPERTIES FOR ASR

Jacques Koreman, William J. Barry & Bistra Andreeva

University of the Saarland, FR 8.7 Phonetics
P.O.Box 151150, D-66041 Saarbrücken, Germany

ABSTRACT

This paper describes three cross-language ASR experiments which use hidden Markov modelling. The first one shows that consonant identification improves when vowel transitions are used. In particular, the consonants' place of articulation is identified better, because the vowel transitions contain formant trajectories which depend on the consonant's place of articulation. The second experiment compares consonant identification results when acoustic parameters belonging to the consonant itself (no vowel transitions are used in the second experiment) are used as input to hidden Markov modelling directly with identification rates when acoustic-phonetic mapping is performed before applying hidden Markov modelling. It is shown that acoustic-phonetic mapping greatly improves consonant identification rates. In the third experiment, the acoustic parameters from the vowel transitions are also mapped onto consonantal (not vocalic) features, as are the acoustic parameters belonging to the consonants. The additional use of vowel transitions does not lead to further improvements in the consonant identification, however. This is probably due to undertraining of the vowel transitions in the Kohonen network.

1. INTRODUCTION

A lexicon and language model can compensate for many phone identification errors at the acoustic level by excluding impossible words and word sequences. For spontaneous speech recognition, we must however also optimise phonetic decoding of the acoustic signal. We shall therefore present several controlled experiments on the enhancement of consonant identification which do not use a lexicon or language model.

One of the main challenges to ASR is the coarticulation between sounds. Many ASR systems based on hidden Markov modelling deal with coarticulation by using generalised triphones as the basic unit for recognition. Triphone models can cater for context-sensitivity at least to immediately neighbouring sounds; generalised triphones exploit the known acoustic similarities between some transitions and thus maximise the training data [1,2,3,4]. The use of triphones thus helps the system to cope with the lack of homogeneity in data belonging to different realisations of the same phoneme. But besides making the acoustic parameters for a phone less homogenous, context-sensitivity also means that information about the identity of a sound is available in its immediate context. We shall try to address this information explicitly by using the acoustic parameters belonging to vowel transitions in addition to those belonging to the consonants to identify the consonant. This is the aim of the first experiment (section 2).

The second aim of this paper is to address linguistic information available in the signal explicitly. The assumption is that if we manage to do this properly, it will be easier to distinguish phonemes – which is the ultimate goal of a speech-to-text ASR system. In a second experiment, the acoustic parameters belonging to consonants (without surrounding vowel transitions) are therefore mapped onto phonetic features by two Kohonen networks before applying hidden Markov modelling. In this mapping, (only) linguistically relevant phonetic features, like [±plosive] and [±uvular], are defined on the basis of the acoustic parameters. The results are compared to a baseline experiment in which the acoustic parameters are used as input for hidden Markov modelling directly, i.e without acoustic-phonetic mapping (section 3).

In a third experiment, we shall combine the use of vowel transitions and acoustic-phonetic mapping. The vowel transitions are used for "relational processing", i.e. the acoustic parameters belonging to the vowel transitions are mapped onto phonetic features of the neighbouring consonants, as will be explained in section 4. The results are compared to a baseline experiment in which only the consonant is used for identification after mapping onto (non-relational) consonantal phonetic features.

All the experiments which are presented are cross-language, which stresses the generality of the phonetic principles which are the subject of this paper and at the same time allows us to maximise the training data.

2. EXPERIMENT 1: VOWEL TRANSITIONS

In the first experiment, consonant identification rates are compared in two subexperiments. In the first or baseline experiment, consonants are identified on the basis of acoustic parameters belonging to what is traditionally labelled as a consonant. In the second subexperiment, the surrounding vowel transitions are used additionally. Since the formant trajectories in the vowel transitions depend on the place of articulation of the neighbouring consonant, it is expected that the information in the vowel transition enhances consonant identification (as is the case for human listeners [5,6,7,8]). To enhance the difference between the two subexperiments, only intervocalic consonants, which are flanked by a vowel offset and a vowel onset transition, are identified.

2.1. Data

HTK software [9] was used to compute 12 mel-frequency cepstral coefficients (MFCC's), energy and the corresponding delta parameters from a 16 kHz microphone signal. A 15-ms Hamming window was applied to minimise smearing of the spectral changes in the transitions over time; a step size of 5 ms and preemphasis of 0.97 were used. The acoustic parameters were computed for English, German, Italian and Dutch read passages from the Eurom0 database.

Vowel transitions were defined to be 35 ms long (cf. [10]), but can be shorter if the vowel is less than 70 ms. In that case, the vowel onset and offset transitions coincide with the first and second half of the vowel.

Since the acoustic parameters are very similar for the transitions of a vowel into consonants which share the same place of articulation, labelnames were generalised across consonantal place of articulation. Eight places of articulation were distinguished, namely labial (lab), dental (den), alveolar (alv), alveolo-palatal (alp), palatal (pal), velar (vel), uvular (uvu) and glottal (glo). The transitions were thus labelled as "i:_lab", "O_vel", "alp_u:", etc. (vowels in SAMPA notation).

Consonants were labelled in adapted SAMPA. It was necessary to adapt standard SAMPA notation because, although all SAMPA labels are phonemic within a language, there is overlap when different languages are used. To give one example, the SAMPA symbol /r/ is used to represented the acoustically very different realisations of that phoneme in English, where it is an alveolar approximant, and in Italian, where it is an alveolar trill. Further, our system requires that the closure phase of plosives and affricates be labelled separately from the rest of the sound, so that extra labels had to be invented. The labelnames used in this paper are described elsewhere in these proceedings ([11]) and also in [12].

2.2. Hidden Markov modelling

For each of the labels, a 3-state left-to-right hidden Markov model (HMM) was trained with a single probability density function per state (also using HTK). There were 280 different HMM's for vowel transitions; the number of consonant HMM's was 30. To maximise the training data, all consonants, not only intervocalic ones, were used. For testing, only the signal belonging to intervocalic consonants was used in the first (baseline) subexperiment. In the second subexperiment, vowel off- and onset transitions were used in addition. The signal belonging to a single consonant (plus transitions) was offered to the system for identification without any further context.

2.3. Results and discussion

The use of vowel transitions leads to an increase in consonant identification (for 32 consonants) by 2.66 percent points compared to the baseline experiment. Although the effect of adding vowel transitions may not seem very large at first sight, it must be pointed out that the identification of the consonants' place of articulation (8 places) improved by 18.21 percent points (table 1). The results show that the information available in vowel transitions about the place of articulation of the neighbouring consonant can be used successfully in an ASR system (cf. [13]).

Table 1. Identification rates for 32 consonant categories without/with use of vowel transitions, as well as for 8 place-of-articulation and 7 manner-of-articulation categories

	no V transitions	**V transitions**
consonant	13.17%	15.83%
place	26.57%	44.78%
manner	46.79%	41.97%

3. EXPERIMENT 2: ACOUSTIC-PHONETIC MAPPING

In a second experiment the signal belonging to the consonant (the same consonants as in the first experiment) is first mapped onto distinctive phonetic features like [±labial] and [±nasal], which are then fed into a HMM procedure. The results are compared with those from the first subexperiment of Experiment 1, in which hidden Markov modelling is carried out on the basis of acoustic parameters directly. It is expected that by applying acoustic-phonetic mapping, the system can focus on linguistically relevant signal properties [11,14,15].

3.1. Data

The input signals are the same as in Experiment 1, as are the consonantal labels (except that no vowel transitions were used in the present experiment). The phonetic features which are used [12] were derived directly from the 3 dimensions of the IPA chart (manner, place and voicing) and are thus closely related to the articulatory properties of the consonants rather than being more abstract phonological features.

3.2. Mapping and hidden Markov modelling

Acoustic-phonetic mapping was performed separately for MFCC’s and energy and for the delta parameters by two parallel Kohonen networks ([16]; see section 4.2), after which the output vectors, containing phonetic features, were concatenated and fed into hidden Markov modelling [12]. The consonants were modelled by 30 HMM’s.

3.3. Results and discussion

Compared to the baseline experiment, in which the acoustic parameters were fed directly into the HMM system, a substantial improvement was found both for the consonant identification rates (38.83 percent points) and for the correct identification of the consonants' place of articulation (39.55 percent points). This shows that it is very effective to help the system focus on linguistically relevant properties by mapping acoustic parameters onto phonetic features. This is further supported by the very considerable increase (30.91 percent points) found in the identification of the consonants' manner of articulation (table 2). A more phonetically oriented interpretation of the results is given in [11].

Table 2. Identification rates for 32 consonant categories without/with use of acoustic-phonetic mapping, as well as for 8 place-of-articulation and 7 manner-of-articulation categories

	no mapping	**mapping**
consonant	13.17%	52.00%
place	26.57%	66.12%
manner	46.79%	77.70%

4. EXPERIMENT 3: ACOUSTIC-PHONETIC MAPPING AND TRANSITIONS

Combining the results from the first two experiments, we used the signal belonging to the consonants together with the preceding and following vowel transitions in the acoustic-phonetic mapping procedure and then for identification of the consonant by applying hidden Markov modelling. Vowel transitions were used for what we shall call "relational processing", i.e. the signal belonging to the vowel transitions was used to extract information about the neighbouring consonant by mapping the acoustic parameters belonging to the transitions onto phonetic features representing the place of articulation of the neighbouring consonant. The results are compared with those from the subexperiment of Experiment 2 which uses mapping of the acoustic parameters belonging to the consonants only.

4.1. Data

The input signals are the same as in the previous experiments. Both consonants and vowel transitions are mapped onto consonantal phonetic features. Labels are different from the ones used in experiment 1 when vowel transitions are mapped: because vowel identity is not relevant for the aim of the experiment, namely consonant identification, all vowels were pooled, resulting in labelnames like "V_uvu" and "den_V". Pooling of all vowels was not possible in Experiment 1, because the vowel transitions *are* different for each vowel and these acoustic differences have to be respected in hidden Markov modelling to ensure sufficient homogeneity of the inputdata for each HMM. The reason why this is not a problem in the mapping experiment which is presented here is that the Kohonen networks map different acoustic variants onto the same phonetic feature vectors, thus creating homogenous input to hidden Markov modelling despite of the variability in the acoustic parameters for different vowels.

4.2. Mapping and hidden Markov modelling

The mapping procedure was the same as in Experiment 2. Hidden Markov modelling was also the same as in that experiment, except that 46 HMM's are used when transitions were mapped (30 consonant HMM's plus 2 x 8 HMM's for vowel onset and vowel offset transitions for 8 places of articulation – cf. 280 transition HMM’s in Experiment 2).

Since Experiment 1 has shown that vowel transitions only enhance identification of the neighbouring consonant's place of articulation, the acoustic parameters belonging to vowel transitions are only mapped onto place-of-articulation features of the neighbouring consonant. For frames belonging to consonants, the acoustic parameters are mapped onto the full set of consonantal phonetic features, as was the case in Experiment 2.

Since different acoustic parameters are relevant for vowel transitions than for the consonants themselves, two Kohonen networks were trained. One was trained with delta parameters only, and should therefore better reflect the spectral changes which are relevant for the formant trajectories in the vowel transitions. Since spectral change is not important for the consonant HMM’s, the other Kohonen network was trained with MFCC’s and energy. By concatenating the output vectors from the two Kohonen networks, the hidden Markov modelling is allowed to select the relevant information.

4.3. Results and discussion

We only found a small improvement in the consonant identification rates (0.23 percent points) as well as in the identification rate of the consonants' place of articulation (1.59 percent points). For manner of articulation, there was a small decrease of 1.03 percent points (table 3).

Table 3. Identification rates for 32 consonant categories after mapping, without/with use of vowel transitions, as well as for 8 place-of-articulation and 7 manner-of-articulation categories

	no V transitions	**V transitions**
consonant	52.00%	52.23%
place	66.12%	67.71%
Manner	77.70%	76.67%

First, of course, it must be noted that the baseline results are already good, so that large improvements are less likely. But analysis of the data also showed that the negligible changes in the identification rates compared to the improvements in the first two experiments probably lie in undertraining of the vowel transitions in the Kohonen networks used for acoustic-phonetic mapping. Experiments for a larger corpus (TIMIT) are underway to verify this assumption.

5. CONCLUSIONS

Three cross-language experiments were carried out to show how consonant identification can be improved even if no lexicon or language model is used.

In a first experiment the use of vowel transitions was investigated in a consonant identification experiment based on hidden Markov modelling, using mel-frequency cepstral coefficients, energy and the corresponding delta parameters as input parameters. It was shown that especially the consonants' place of articulation was identified much better when transitions are available to the system.

In a second experiment, the acoustic parameters were mapped onto distinctive phonetic features and then used for hidden Markov modelling. A large improvement in consonant identification, as well as in the identification of place and manner of articulation, was found in comparison to the baseline experiment, in which the acoustic parameters were fed into hidden Markov modelling directly.

In a third experiment it was shown that mapping vowel transitions onto relational phonetic features does not lead to the expected improvement in consonant identification. The most likely reason for this is that acoustic-phonetic mapping in the Kohonen network breaks down due to undertraining of the vowel transitions. The experiment will be repeated with the TIMIT database to overcome this problem.

Given these good results from our experiments, the application of the proposed signal processing is expected to improve the phone recognition rates in a complete ASR system, since we need rely less on the lexicon and language model to correct errors at the acoustic level. An experiment is currently in progress to test the validity of this expectation.

6. REFERENCES

1. Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M. & Makhoul, J. (1985). Context-dependent modeling for acoustic-phonetic recognition of continuous speech. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing.*

2. Derouault, A.-M. (1987). Context-dependent phonetic Markov models for large vocabulary speech recognition. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* 360-363.

3. Deng, L., Lennig, M., Gupta, V. & Mermelstein, P. (1988). Modeling acoustic-phonetic detail in an HMM-based large vocabulary speech recognizer. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* 509-512.

4. Lee, K.-F. (1990). Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. *IEEE Transitions on Acoustics, Speech and Signal Processings,* 599-609.

5. Liberman, A., Delattre, P., Cooper, F. & Gerstman, L. (1954). The role of consonant-vowel transitions in the perception of stop and nasal consonants. *Psychol. Monogr.* **68**(8), 1-13.

6. Delattre, P., Liberman, A. & Cooper, F. (1955). Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am.* **27**(4), 769-773.

7. Delattre, P. (1968). From acoustic cues to distinctive features. *Phonetica* **18**, 198-230.

8. Stevens, K. & Blumstein, S. (1978). Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am.* **64**(5), 1358-1368.

9. Young, S., Jansen, J., Odell, J. Ollason, D. & Woodland, P. (1995). *The HTK Book.* Cambridge: Cambridge University.

10. Furui, S. (1986). On the role of spectral transitions for speech preception. *J. Acoust. Soc. Am.* **80**(4), 1016-1025.

11. Koreman, J., Andreeva, B. & Barry, W.J. (1998). Do phonetic features help to improve consonant identification in ASR? *Proc. Int. Conf. on Spoken Lang. Proc.* (these proceedings).

12. Koreman, J., Barry, W.J. & Andreeva, B. (1997). Relational phonetic features for consonant identification in a hybrid ASR system. *PHONUS* **3**, 83-109. Saarbrücken (Germany): Institute of Phonetics, University of the Saarland *OR* http://www.coli.uni-sb.de/~koreman/Publications/Phonus/1997/ph97_Trans.ps.gz.

13. Cassidy, S & Harrington, J. (1995). The place of articulation distinction in voiced oral stops: evidence from burst spectra and formant transitions. *Phonetica* **52**, 263-284.

14. Bitar, N. & Espy-Wilson, C. (1995a). Speech parameterization based on phonetic features: application to speech recognition. *Proc. 4th European Conference on Speech Communication and Technology,* 1411-1414.

15. Bitar, N. & Espy-Wilson, C. (1995b). A signal representation of speech based on phonetic features. *Proc. 5th Annual Dual-Use Techn. and Applications Conf.,* 310-315.

16. Dalsgaard, P. (1992). Phoneme label alignment using acoustic-phonetic features and Gaussian probability density functions. *Computer Speech and Language* **6**, 303-329.