DO PHONETIC FEATURES HELP TO IMPROVE CONSONANT IDENTIFICATION IN ASR?

Jacques Koreman, Bistra Andreeva & William J. Barry

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

ABSTRACT

The hidden Markov modelling experiments presented in this paper show that consonant identification results can be improved substantially if a neural network is used to extract linguistically relevant information from the acoustic signal before applying hidden Markov modelling. The neural network – or in this case a combination of *two* Kohonen networks – takes 12 mel-frequency cepstral coefficients, overall energy and the corresponding delta parameters as input and outputs distinctive phonetic features, like [±uvular] and [±plosive]. Not only does this preprocessing of the data lead to better consonant identification rates, the confusions that occur between the consonants are less severe from a phonetic viewpoint, as is demonstrated. One reason for the improved consonant identification is that the acoustically variable consonant realisations can be mapped onto identical phonetic features by the neural network. This makes the input to hidden Markov modelling more homogenous and improves consonant identification. Furthermore, by using phonetic features the neural network helps the system to focus on linguistically relevant information in the acoustic signal.

1. INTRODUCTION

The work presented in this article is related to other work in the area of automatic speech recognition [1,2,3], which also uses phonetic features to recognise speech. The reason for using phonetic features is that, as Bitar & Espy-Wilson put it, by using phonetic features we "directly target the linguistic information in the signal and ... minimize other extra-linguistic information that may yield large speech variability" [1] (p. 1411). In contrast to Bitar & Espy-Wilson, who use a knowledge-based event-seeking approach for extracting phonetic features from the microphone signal on the basis of acoustic cues, the mapping from the acoustic to the phonetic-feature domain is performed by neural nets in the experiments which we shall describe below. We shall compare the consonant identification results from two experiments, one of which uses acoustic-phonetic mapping and another one which does not. To put an end to any doubts you may have about the answer to the question in the title immediately, the answer is "yes".

2. A HYBRID CONSONANT IDENTIFICATION SYSTEM

Two consonant identification experiments were carried out in which segments were modelled by simple left-to-right 3-state hidden Markov models, with each state having only a single Gaussian function to model the observation probabilities [4]. In one experiment, which we shall call the baseline experiment, 12 mel-frequency cepstral coefficients, overall energy and the corresponding delta parameters were used as input to hidden Markov modelling directly. In the second experiment, these acoustic parameters were used as input for two parallel 50x50 Kohonen networks [5] which map the acoustic parameters onto 14 phonetic features, like [±uvular] and [±plosive]. The phonetic features for the consonants are derived from the three dimensions of the IPA chart (place, manner, voicing). The first network maps the mel-frequency cepstral coefficients and energy onto phonetic features, while the second network maps the delta parameter onto (the same!) phonetic features; the output vectors from the two neural network were concatenated and used for hidden Markov modelling. The reasons for using two neural networks instead of one are not relevant for the following discussion and are explained in [6,7]. No lexicon or language model were used.

3. DATA

The acoustic parameters were computed from read passages from the Eurom0 database for English, German, Italian and Dutch (2 male and 2 female speakers for each language). A SAMPA transcription [8] was available for each of the read passages. This SAMPA transcription was adapted to cater for our needs in several ways:

- Plosives and affricates were labelled by two segments: one for the closure ("p0" = voiceless closure; "b0" = voiced closure) and one for the burst-plus-aspiration, ("p", "t", "k") or friction part ("f", "s", "S", "z", "Z"). This was necessary, because the place of articulation of the plosive or affricate cannot be determined from its closure, so that the neural nets cannot determine a feature value for the consonant's place of articulation on the basis of frames belonging to the closure.
- Although the SAMPA transcription symbol for the English approximant /r/ is the same as for the alveolar trill in Italian and Dutch (by some speakers and in

some regions), its realisation is very different. We used /rapr/ to label the English approximant, reserving /r/ for the alveolar trill.

- The German SAMPA symbol /v/ is used to describe a sound which is normally realised as a labiodental approximant (IPA symbol /ʋ/). We have therefore labelled it /vapr/, to distinguish it from English, Italian and Dutch fricative /v/.
- In Dutch, the SAMPA symbol /w/ is used to describe a labiodental approximant, and here, too, we have adopted the label /vapr/ to describe it. The transcription symbol /w/ is reserved for the bilabial approximant, which only occurs in English and Italian.
- Further, Dutch has both an alveolar and a uvular “r”-sound, which we shall transcribe /r/ (as in Italian) and /R/ (as in German), respectively.
- In some dialects of Dutch, there is no /G-x/ opposition. Since there were only few realisations of the Dutch voiced velar fricative /G/, we replaced it by /x/.
- In order to have more training data for the HMMs, we pooled Italian geminate consonants with non-geminates.

Each segment was modelled by a hidden Markov model (HMM). In the identification step consonants, not the segments modelled by hidden Markov models, were identified. The consonant /t/ for example was defined in the phoneme dictionary as the (optional) HMM for "p0" followed by the HMM for "t". Different language- and speaker-specific variants of /r/, such as /rapr/, /r/ and /R/, were modelled by different HMM's; otherwise, the same labels were used for different allophones of the same sounds, as for instance for dark and clear /l/. Only intervocalic consonants were used in this experiment (for reasons explained in [6,7]). Note that in the hidden Markov modelling identification step, no restrictions were imposed on the consonant which can be identified, so that for example in German, Italian and Dutch /T/ and /D/ could be identified despite the fact that these phonemes only occur in English.

4. THE EFFECT OF MAPPING ON CONSONANT IDENTIFICATION

In the experiment in which acoustic parameters are mapped onto phonetic features, 52.00% of the consonants (with 32 possible consonants) are identified correctly. In the baseline experiment, the percentage of correctly identified consonants is only 13.17. In the mapping experiment all consonants are identified better, except /D/ (in English only) and /C/ (in German only). Without acoustic-phonetic mapping, the majority of the other consonants are misidentified more often than not; many of them are *never identified correctly.*

The overall percentages reported above are influenced strongly by the number of realisations for each consonant. Since we are interested in a phonetic analysis of error patterns across the consonants regardless of the number of occurrences in the database, we have computed the percentage of correct identifications for each consonant separately. The average correct identification score (ACIS) is then computed as

$$\text{ACIS} = \frac{\text{total of all correct identification percentages}}{\text{number of consonants to be identified}}$$

where the multiple is the total of the percentages along the diagonal of the confusion matrix and the denominator is the number of rows in the confusion matrix. We thus compensate for the consonants' actual number of occurrences and give each consonant equal weight. The ACIS is 68.47% when mapping is applied and only 31.22% when it is not. The reason why ACIS with mapping is only double that without mapping, while the correct identification rate is almost four times as high, is that many infrequent consonants were already identified well in the baseline experiment. We will try to offer a phonetic explanation for this finding in the following section.
The consonant misidentifications also show an interesting tendency. An incorrectness coefficient, which we shall call the average phonetic misidentification score (APMS) was computed as

$$\text{APMS} = \frac{\text{phonetic misidentification coefficient}}{\text{sum of the misidentifications}}$$

The phonetic misidentification coefficient is the sum of all the products of the misidentification percentage (all percentages in non-diagonal cells of the confusion matrix) times the number of incorrectly identified phonetic categories (place and manner of articulation, and voicing). This gives a measure of the severity of the error in terms of phonetic features, with possible values between 1 (*either* place *or* manner *or* voicing wrong) and 3 (place, manner *and* voicing wrong). This score went down from 1.79 when no mapping is applied to 1.57 when it is. This indicates that after mapping, the incorrectly identified consonant is on average closer to the phonetic identity of the consonant which was produced. The number of confusions on 2 or 3 phonetic categories is reduced substantially.

5. A PHONETIC INTERPRETATION OF THE TWO MEASURES

If we look into the phonetic detail which hides behind the ACIS measures for the two experiments, we find some interesting patterns. Although at first sight there seems to be a negative correlation between the number of realisations of a consonant and its correct identification percentage in the baseline experiment, this correlation is not borne out by the data. Although all consonants with $n > 100$ have low identification rates in the baseline experiment (15.8% or less), not all "rare" consonants ($n < 100$) are identified well. This is shown in table 1, which lists all consonants with $n < 100$.

Table 1: Correct identification percentages in the baseline (no mapping) and mapping experiment for all consonants with n < 100, ordered according to correct identification percentage in the baseline experiment

Cons	no mapping	mapping	n
C	100.0	75.0	8
J	100.0	100.0	4
L	100.0	100.0	10
D	97.8	91.3	46
w	94.1	100.0	17
rapr	91.2	96.5	57
x	88.2	93.4	76
S	78.1	90.6	32
R	50.0	77.5	80
g	47.6	57.1	21
vapr	35.2	66.7	54
b0Z	28.0	96.0	25
j	17.6	94.1	17
h	6.7	86.7	15
N	3.8	6.2	26
p	1.4	33.3	72
f	1.2	64.6	82
p0f	0.0	100.0	3
p0s	0.0	72.2	54
b	0.0	4.4	84
b0z	0.0	70.3	37

A closer look at table 1 shows that the consonants which are recognised best (correct identification rate > 80%) are mostly language-specific consonants: /C/ (German), /J/ (Italian), /L/ (Italian), /D/ (English), /w/ (English, Italian), /rapr/ (English), /x/ (German, Dutch). It seems that consonants which do not contain cross-language variability are acoustically more homogenous and therefore recognised better in the baseline system than other sounds.

Table 2: Correct identification percentages in the baseline (no mapping) and mapping experiment for all affricates and the corresponding fricatives

	mapping		Corr.	mapping	
Affric.	no	yes	fric.	no	yes
p0f	0.0	100.0	f	1.2	64.4
p0s	0.0	72.2	s	3.1	64.7
p0S	0.0	40.2	S	78.1	90.6
b0z	0.0	70.3	z	10.4	50.5
b0Z	28.0	96.0	Z	no intervocalic real.	

That affricates are not recognised well despite the fact that they are mostly language-specific can be easily understood: the component parts (closure and friction), which are modelled by separate hidden Markov models, occur in all languages and are therefore probably less homogenous. The affricates are recognised much better when mapping is applied, as is shown in table 2. The fricatives which correspond to the affricates are also recognised better.

The identification of voiceless plosives /p/, /t/ and /k/ also improves greatly (the difference between baseline and mapping experiment is 31.9, 32.4 and 58.2 percent points, respectively, with consonant identification rates under 6% in the baseline experiment). Especially /k/, which varies widely from a velar to a pre-velar place of articulation depending on the identity of the surrounding vowels, is identified much better.

Another source of variation in the realisation of voiceless plosives is the presence or absence of aspiration: English and German have aspirated voiceless plosives, whereas Italian and Dutch do not. It seems unlikely that the variation in aspiration can be better handled by the system which uses mapping, because the different spectral properties of the aspiration do not so much depend on the place of articulation of the consonant as on the following vowel (of which, one could say, it is the voiceless realisation). This is corroborated by the confusions which occur after mapping: although the voiceless plosives /p/, /t/ and /k/ are confused with non-plosives far less often than in the baseline system, they *are* confused with each other quite frequently.

As is well-known, the consonant /h/ is more context-sensitive than any other due to its spectral dependence on the neighbouring vowels. Its identification improves by 80 percent points (from 6.7% in the baseline system to 86.7% in the mapping system). This again stresses the ability of the neural network to map acoustically variable signals onto the same phonetic features. The ability of the system which uses acoustic-phonetic mapping to handle allophonic variation so well can be easily understood: the neural network(s) on the one hand respect the acoustic variability, which leads to different allophones being modelled in different parts of the phonotopically organised Kohonen network(s), while on the other hand it outputs the same phonetic feature vector for the different allophones of a phoneme.

It is probably not only this homogenising effect of the mapping which increases the consonant identification rates. It is difficult to explain some of the improvements that we find when mapping is applied. This is for example the case for /f/, for which the constraints on its articulation do not seem to leave much room for variation: not only does /f/ require a precise labiodental articulation, so that the place of articulation is less variable than for /S/, for instance, there is also hardly a resonating cavity before (i.e. downstream) its articulation place, so that no strong

context-dependent variation should be expected on that account either. The improvement which is found must probably be put down to the ability of the neural net to select linguistically distinctive phonetic features which allow for a better separation of the consonants in hidden Markov modelling. This lies behind the APMS measure presented in the previous section and shows in the types of confusions which occur in the baseline and in the mapping system: whereas consonants are confused with phonetically very different consonants in the baseline system, these confusion occur far less often in the system which uses acoustic-phonetic mapping. In the baseline system, /r/ is never identified as itself, but confused with /g/ in 61% of the cases, as well as with /L/ (16%), /w/, (13%) and several other phonemes. After mapping, /r/ is recognised as itself in 85% of its realisations, and next rarely as /R/ and /l/, which are both phonetically close to /r/, being a trill and an alveolar approximant like /r/, respectively.

For another sound which behaves very differently in the baseline and the mapping system, namely /j/ (difference in identification 76.5 percent points), a similar behaviour pattern is found. In the baseline system, /j/ is identified as itself for only 18% of its realisations, and confused with /L/ (53%), /J/ (18%), /rapr/, /r/ and /g/ (6% each). Although especially the confusions with /L/ and /J/, both being palatalised consonants, are quite understandable, it compares unfavourably with the identification of /j/ after mapping. In the mapping system, /j/ is recognised as itself in 94% of the cases and confused only with /z/ (6%).

The same is true for all nasal phonemes except /J/, which was already identified correctly in 100% of its realisations. In the baseline system they are not only confused with other nasals, but often also with /R/, /L/, /w/, /rapr/ and /vapr/. In the mapping system, they are mainly confused with other nasals. It seems therefore that the neural net helps to identify the nasality of the consonant.

It is obviously not possible to discuss all the details in the confusion matrices from the two experiments. For that reason, we have included them on the CD-ROM version of the proceedings in [MAPPING.GIF] and [BASELINE.GIF].

6. CONCLUSIONS

As is clear from the consonant identification results for the two hidden Markov modelling experiments presented above, acoustic-phonetic mapping leads to better consonant identification rates, as reflected in the higher ACIS value in the mapping experiment than in the baseline experiment. Furthermore, the confusions that occur in the mapping experiment are less severe than in the baseline experiment from a phonetic viewpoint. This is reflected in the lower APMS value in the mapping experiment. The better performance of the system when acoustic-phonetic mapping is used can be put down to its ability to map acoustically variable consonant realisations to the same phonetic feature vector on the one hand and its ability to select and use only linguistically relevant, distinctive information in the acoustic signal on the other. Thus, the results from our experiments confirm Bitar & Espy-Wilson's assumption quoted in the introduction.

In Bitar & Espy-Wilson's experiments [1], the experimental results were better when a (phonetic) feature-based representation (FBR) was used as input to hidden Markov modelling than when a cepstral-based representation (CBR) was used, but "CBR outperformed FBR when higher mixtures were used. The better performance can be attributed to better modeling of the richer spectral information contained in CBR" [1] (p. 1413). In our mapping system, the rich spectral information contained in the acoustic parameters is not lost, but in fact used by the Kohonen networks when they self-organise. In so far, the effect of using a neural net is similar to that of using multiple mixtures: it allows the system to associate very different (allophonic) spectral characteristics with the same consonant. In our system, an advantage can be seen in the fact that, although the user must define the size of the Kohonen network(s), the requirements on the user to decide on the architecture of the HMM for each of the consonants are relaxed without losing the functional advantages of the use of multiple mixtures.

7. REFERENCES

1. 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.
2. 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.
3. Kirchhoff, K. (1996). Syllable-level desynchronisation of phonetic features for speech recognition. *Proc. ICSLP.*, 2274-2276.
4. Young, S., Jansen, J., Odell, J., Ollason, D. & Woodland, P. (1995). *The HTK Book.* Cambridge: Cambridge University.
5. Dalsgaard, P. (1992). Phoneme label alignment using acoustic-phonetic features and Gaussian probability density functions. *Computer Speech and Language* **6**, 303-329.
6. 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.unisb.de/~koreman/Publications/ Phonus/1997/ph97_Trans.ps.gz.
7. Koreman, J., Barry, W.J., Andreeva, B. (1998). Exploiting transitions and focussing on linguistic properties for ASR. *Proc ICSLP*. (these proceedings).
8. SAMPA symbols, http://www.phon.ucl.ac.uk/home/ sampa.