CAN WE USE THE LINGUISTIC INFORMATION IN THE SIGNAL?

Jacques Koreman & Bistra Andreeva

Institute of Phonetics, University of the Saarland, Saarbrücken, Germany
{jkoreman, andreeva}@coli.uni-sb.de, http://coli.uni-sb.de/~koreman

Abstract

This article discusses the use of phonetic features in automatic speech recognition. The phonetic features are derived from acoustic parameters by means of Kohonen networks. Behind the use of phonetic features instead of standard acoustic parameters lies the assumption that it is useful to help the system to focus on linguistically relevant signal properties. Previous experiments using very simple hidden Markov models to represent the phones (with only one mixture for each state and without a lexicon or language model) have indeed shown that the phoneme identification rates on the basis of phonetic features were considerably higher than on the basis of acoustic parameters. When eight mixtures per state are used in hidden Markov modelling, the phoneme identification rates for three different sets of phonetic features were found to be lower than those obtained from a system in which the acoustic parameters are modelled directly. It is suggested that the results are still good enough, however, to further explore the use of phonetic features in a complete automatic speech recognition system: if each phone sequence representing a word in the lexicon is replaced by a sequence of underspecified phonetic feature vectors, the use of phonetic features in the acoustic decoding *may* have certain advantages.

Phonus 5, Institute of Phonetics, University of the Saarland, 2000, 47-58.

1. Introduction

In most standard automatic speech recognition (ASR) systems, phone models are computed from a spectral representation of the signal, usually mel-frequency cepstral coefficients (MFCC's), energy and the corresponding delta (change) and delta-delta (velocity or rate of change) parameters by means of hidden Markov modelling (HMM). We increasingly find statements in the literature (e.g. Bourlard et al., 1996; Pols, 1999) which express the belief that the limits of these purely stochastic ASR systems (in terms of correct recognition rates) have been reached. Despite the high correct recognition rates in applications for any particular task, or rather because there still is a (small) gap between human and machine performance, the applicability of ASR systems in large-vocabulary, speaker-independent environments where spontaneous speech is used, is still limited. The reason for this is that even if an ASR system makes only few errors, they prevent an efficient user interaction with the system.

One way to try and counteract this is by *reducing* the active lexicon by using restrictive dialogue management (e.g. in a train information system which asks „Where would you like to go?" instead of „Can I help you?", or which uses checks on the processed information as in „So you want to travel to Saarbrücken next Tuesday?"). Also, an *extension* of the lexicon by adding pronunciation variants (see other articles in this volume) can enhance the probability of finding the correct word candidate, because pronunciation variants allow for the effects of phonological processes found in spontaneous speech (segment deletion and insertion as well as change of phone class by assimilation or reduction), so that alternative lexical entries are closer to the actual realisations of the words. In this article, however, we shall focus our attention on attempts to improve the acoustic decoding.

If the main problem of acoustic decoding in an ASR system is the variability in the signal, the use of linguistic knowledge to reduce this variability may help to improve the performance of standard ASR systems. In most ASR systems using HMM, variation in the signal is modelled in phone or phoneme models by using multiple mixtures per state to deal with allophonic and other systematic (gender, speaker, etc.) variation, while random variation is handled by data descriptions in terms of continuous density functions. This paper compares such modelling with a more linguistically oriented modelling strategy. In our approach, variation in the acoustic signal is reduced by mapping the acoustic parameters onto a set of phonetic features, which are then used as input to HMM. We shall present results for three phonetic feature sets.

2. The derivation of phonetic features

As in Koreman et al. (1999), two 50x50 Kohonen networks are used to derive phonetic features from the acoustic parameters. The acoustic parameters are computed from the microphone signals for read texts in the Eurom0 database spoken by 2 male and 2 female speakers for each of the four languages German, English, Italian and Dutch (29 minutes in total, not counting silent portions). A 15-msec Hamming window is used with a pre-emphasis of 0.97 and a step size of 5 msec to compute 12 MFCC's, energy and the corresponding delta parameters, which are a weighted mean of 5 frames centered around the current frame. The HTK package (Young et al., 1995) is used for the derivation of the acoustic parameters as well as for HMM. Figure 1 shows the complete system.

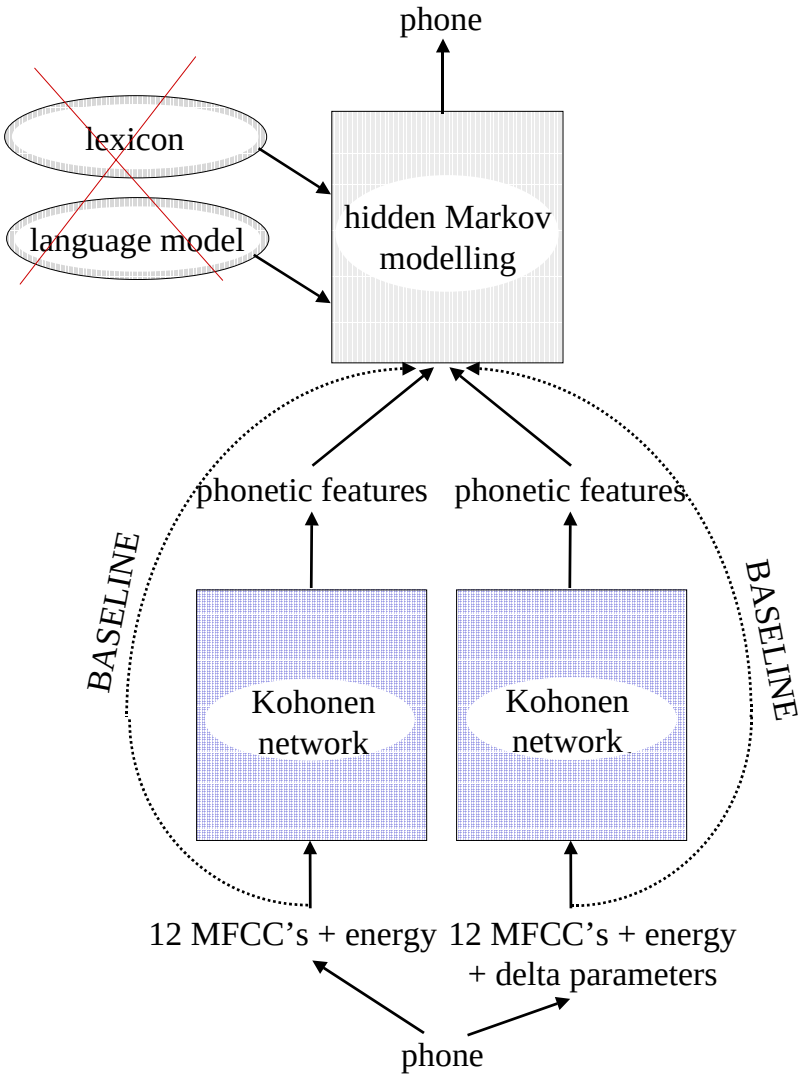

Figure 1. Architecture of a hybrid consonant identification system which uses acoustic-phonetic mapping

Of the two Kohonen networks which we use, the first, „static“ Kohonen network is trained with the MFCC's and energy parameters of all frames, and should provide the best mapping for the steady states of phones, since no delta parameters (which should be close to zero, since ideally there is little or no spectral variation during the steady states) lead to unnecessary noise in the system. Since in the case of transitions – which, as the name indicates, reflect spectral change – the delta parameters are much more important than the MFCC and energy values from which they are derived, a second, „dynamic“ Kohonen network is trained with MFCC's, energy *and* the corresponding delta parameters (again, from all the frames). After the phonotopic organisation, the Kohonen networks are calibrated for phonetic features by offering an acoustic vector as input to the system and storing the corresponding phonetic feature vector with the winning neuron. The phonetic feature vector is derived from the segment label to which the acoustic frame belongs and consists of features with a pre-defined value of 1 (present) –1 (not present) or 0 (not relevant). The vector size depends on the phonetic feature set which is used. After presenting all the data for calibration, an average across all phonetic feature vectors stored for a neuron is calculated for each neuron. In the acoustic-phonetic mapping, the output of the Kohonen networks consists of the averaged phonetic feature vector for the neuron activated by the acoustic input vector (in fact, a weighted average for the K-nearest neurons is used). In Koreman et al. (1997) the training, calibration and mapping procedures are described in more detail. Three different phonetic feature sets are used in the present experiments: IPA features (19 features directly derived from the row and column labels of the IPA charts), a set of 16 features based on SPE and an underspecified version of the latter feature set (derived by setting the value of redundant features to zero). The features are listed in Table 1.

Table 1. IPA and SPE features used in the experiments

IPA	labial, dental, alveolar, palatal, velar, uvular, glottal, plosive, fricative, nasal, lateral, approximant, trill, voiced, mid, open, front, central, rounded
SPE	consonantal, syllabic, nasal, sonorant, low, high, central, back, rounded, anterior, coronal, continuant, voiced, lateral, strident, tense

The output vectors of the two Kohonen networks are simply concatenated for each frame and modelled by means of hidden Markov models for each of the 53 phones (not phonemes). The reason for training and identifying phones is that, also since we used four different languages, it is not useful to create one hidden Markov model for the clearly different realisations of for instance /r/ as a retroflex approximant or an alveolar or uvular trill. Each phone is therefore modelled separately. The hidden Markov models are simple 3-state, left-to-right models, in which no state can be skipped; only the diagonal covariances are used. The same models are used in the baseline system, in which hidden Markov models are trained on the acoustic parameters directly, i.e. without mapping onto phonetic features. In none of the experiments do we use a lexicon or language model, since we want to evaluate the effect of using phonetic features on the acoustic decoding per se. Using a lexicon and language model would prevent phone confusions which lead to non-words or sequences of words which the language model does not allow; of course, conclusions about the usefulness of acoustic-phonetic mapping in a complete ASR system must be drawn with extra care, since it is at least theoretically possible that an improvement at the acoustic decoding level does not change the performance in a complete system. This is the case if all possible phone confusions are prevented by the lexicon and/or language model. Another reason for interpreting our results with care is that we perform a phone identification task (same test and training data) with pre-segmented phones, so that the results are an indication of, but not immediately generalisable to a real ASR situation. The reason for performing identification instead of recognition is the small amount of training data. Control experiments in which a recognition task is carried out with the TIMIT database are underway. Phoneme identification results are obtained by pooling different allophones of each identified phoneme.

3. Relation to previous experiments

In Koreman et al. (1999), all hidden Markov models consist of 1 mixture per state, both in the experiments using phonetic features and in the one using acoustic parameters. The phoneme identification results of these experiments are shown in Table 2.

Clearly, the use of phonetic features leads to an improvement of the phoneme identification rates in our experiments. Not only do we find better phoneme

Table 2. Phoneme identification results for acoustic parameters and three sets of phonetic features

Acoustic parameters	15.6 %
IPA features	42.6 %
SPE features	36.2 %
Underspecified SPE features	46.1 %

identification results, we also find in comparable experiments in which only consonants were identified (Koreman et al., 1998) that the confusions between phones are phonetically much more acceptable (since the phones differ in fewer phonetic features) than if we use acoustic parameters for HMM. This is exemplified in Table 3, which shows consonant confusions for acoustic parameters and for phonetic features based on the IPA consonant chart.

Table 3. Some consonant confusions on the basis of acoustic parameters versus IPA features. Only the main confusions are shown.

C	r	j	m	n
Ac. Par.	g (61%) ʎ (16%) w (13%)	ʎ (53%) j (18%) ɲ (12%)	w (23%) ʎ (18%) m (16%) ɲ (13%) ʀ (10%)	w (28%) ɲ (18%) ʀ (16%) ʎ (12%)
IPA feat.	r (84%) ʀ (5%) l (4%)	j (94%) z (6%)	m (63%) n (11%) ŋ (10%)	n (26%) m (21%) ŋ (20%) r (6%)

Table 2 shows that the phone [r] is correctly identified in 84% of the cases when IPA features are used as input to HMM, whereas when we perform HMM directly on the acoustic parameters it is not even present in the list of the most often identified consonants. We further find that acoustic-phonetic mapping leads to confusions with phonetically similar phones. In the case of [r] confusions with another trill, namely

[ʀ], and with another alveolar liquid ([l]) make up the majority. If no mapping is applied, confusions with [ɡ], [ʎ] and [w] are most frequent. These show little phonetic similarity to [r]. Similar patterns are found for other phones, so that we concluded that phone confusions are phonetically less severe when acoustic-phonetic mapping is performed. This was more objectively confirmed by the average phonetic misidentification score (APMS), which reflects the phonetic severity of the consonant confusions (see Koreman et al., 1998). Further details are also given in Section 4.2.

4. Modelling variability

The Kohonen networks fulfill two functions simultaneously. First, since the acoustic space is organised phonotopically in a Kohonen network, it can represent different realisations of a phoneme in different parts of the phonotopic map. When the phonotopic map is calibrated, the same vector of phonetic features can be attached to different neurons, so that different variants of a phoneme can still have the same (or very similar) representations in terms of phonetic features. Second, the Kohonen network is used to map the acoustic parameter space onto the phonetic feature space, which represents only the distinctive properties of the phones.

The first function of the Kohonen networks is in part comparable to that of using multiple mixtures for each state in HMM. Compared to using a single mixture, which leads to a continuous density function with a large variance (since the complete variation for a phoneme must be modelled by only one continuous density function), multiple mixtures can model allophonic variation more adequately by representing different acoustic realisations of the same phoneme by means of several continuous density functions with less variance. A comparison of the phoneme identification results for HMM on the basis phonetic features (see Table 2) with HMM of acoustic parameters, but this time using multiple mixtures, will tell us whether the good results of the previous experiments, both in terms of the phoneme identification scores (Section 4.1) and in terms of the phonetic severity of the confusions that occur (Section 4.2) depend on the modelling of variation or on the different input space (phonetic features instead of acoustic parameters).

4.1. *Comparison of the phoneme identification scores*

In the experiment reported here, hidden Markov models are trained for each phone on the basis of MFCC's, energy and the corresponding delta parameters. Eight instead of one mixture per state are used, except for the phone [ɲ] (only one mixture per state), for which we do not have enough training data to model 8 mixtures (it only occurs in Italian, and even then very infrequently). The phoneme identification rate is 63.7%, which is 17.6 percent points higher than the best identification rate for phonetic features in Table 2 (46.1% for underspecified SPE features).

How can we explain these results? Although we assumed that the phonetic feature vectors should be more homogenous for different realisations of the same phone, the homogenising effect of acoustic-phonetic mapping appears to be limited. Different realisations of the same phoneme *can* be organised in different parts of the phonotopic map and still get the same phonetic feature vector attached to them, but they will normally be affected by the phonetic feature vectors of acoustically similar phones. Take for example the different possible realisations of /l/ in English: in syllable-final position, it is realised as a „dark l", with an [u]-like acoustic quality due to the bulging of the tongue body, while syllable-initially the more neutral (schwa-like) tongue position results in a „clear l". Further, after voiceless plosives the same phoneme can be (partially) devoiced and realised with friction. The example makes clear that the neurons activated by these different allophonic variants may be activated by other, but very different phones as well: the „dark-l neuron" may also be activated by close or mid-close back vowels, whereas neurons activated by „clear l" may also be activated by schwa or other neutral vowels; neurons which are activated by the devoiced „l" may also be activated by fricatives. Depending on the frequency of activation by different phones, the calibration (which averages all the feature values of the activating phonemes) can lead to substantially different phonetic feature values (which are an average across the phonetic features values of all activating frames) for the different allophones. If this actually occurs in our data, using multiple mixtures should also lead to an increase in HMM performance for phonetic feature input. In order to evaluate this, three experiments were carried out (one for each feature set). Except for the size of the input vector, the experiments were exactly the same as the multiple-mixture experiment with acoustic parameters. Table 4 shows the phoneme identification results with 8 mixtures.

Comparison with Table 2 shows that using multiple mixtures in the HMM experiments which use phonetic features indeed leads to an increase in the

performance of the system. Nevertheless they never reach the performance achieved by 8-mixture HMM of the acoustic parameters.

Table 4. Phoneme identification results for acoustic parameters and three sets of phonetic features using 8 mixtures per state in HMM

Acoustic parameters	63.7 %
IPA features	54.2 %
SPE features	54.1 %
Underspecified SPE features	58.4 %

4.2. *Comparison of the phonetic severity of phone confusions*

In previous experiments (see Section 3), we found that the confusions between phones were less severe when phonetic features were used for HMM than when acoustic parameters were used. In order to evaluate this for the 8-mixture experiments presented in Section 4.1 we computed an average phonetic misidentification score (APMS). This measure scores differences between consonants on the phonetic categories *manner*, *place* and *voicing* and between vowels on the categories *degree of opening*, *frontness* and *rounding* by multiplying the percentage of confusions (normalised for each phoneme, to give all phonemes the same weight, irrespective of their number of realisations) by the number of misidentified phonetic categories, and dividing the result by the total of all the misidentification percentages. Confusions between vowels and consonants were given a maximal penalty (3) if they differed in voicing (i.e. if the consonant is voiceless), and otherwise 2. This leads to APMS values between 1 (minimal number of misidentified phonetic categories in a phone confusion) to 3 (maximal number of misidentified phonetic categories).

In Koreman et al. (1998) the APMS value for consonant identification on the basis of IPA features (1.57) was 0.22 lower than for acoustic parameters (1.79). Since the range of APMS values is only 2, this difference is more than 10% of the APMS range. The APMS values for the experiments in Section 4.1 are given in Table 5. The table shows that the advantage of a lower APMS value for phonetic features in comparison to acoustic parameters disappears when 8 mixtures are used in HMM. The

APMS values are very similar, although the confusions between consonants are somewhat less severe for underspecified SPE features than for any other input parameters, in particular acoustic parameters. On the other hand, the confusions between vowels are more severe for phonetic features.

Table 5. APMS values across all phones, and for consonants and vowels separately, for 8-mixture HMM of acoustic parameters, IPA, SPE and underspecified SPE features

	acoust. par.	**IPA**	**SPE**	**underspec. SPE**
all phones	1.63	1.65	1.67	1.65
consonants	1.64	1.62	1.63	1.58
vowels	1.63	1.71	1.73	1.75

5. Combined acoustic parameters and phonetic features

The results from the HMM experiments with multiple mixtures in Section 4 raise the question whether the use of phonetic features is at all helpful. In order to test this, we combined the acoustic parameter vector with the vector of underspecified SPE features. Of all previous experiments using phonetic features, underspecified SPE features led to the highest phoneme identification. The correct phoneme identification rate in the experiment using a combined input vector is not only lower than if we use only acoustic parameters, with 56.4% it is even lower than in the experiment where only underspecified SPE features are used (58.4%).

6. Conclusion and discussion

The results so far show that the use of linguistic information in the form of acoustic-phonetic mapping does not lead to an improvement in the phoneme identification rate compared to HMM of the acoustic parameters when multiple mixtures per state are used. The advantage of phonetically less severe confusions between phones when phonetic features are used instead of acoustic parameters that was found in previous experiments (using 1 mixture per state) also disappears.

Although the use of phonetic features does not lead to a better acoustic decoding per se, we can also look at the results more favourably. The fact that there is no (or a small) reduction in the phoneme identification rates (and no consistent difference in the phonetic severity of the phone confusions reflected in the APMS measure) means that phonetic features can still be useful if we exploit them better. In a complete system (with a lexicon and language model), we should not first identify phones on the basis of phonetic features, and then use the phones for comparison with the lexicon. Instead, the phonetic feature vectors themselves should be used directly. In order to achieve this, a single phonetic feature vector must represent each phone at the end of the acoustic decoding. This vector cannot be taken from the output of the Kohonen networks directly, since there each phone, which consists of several frames, is represented as a *sequence* of phonetic feature vectors (one vector for each frame). We can use the Viterbi algorithm in HMM to model the phones (as before), and derive a single phonetic feature vector from the means of the central state of each phone model. This would have the additional advantage of reducing contextual influences.

If we then define the words in the lexicon as sequences of underspecified phonetic feature vectors derived from the distinctions in the phonological system of a language (instead of representing words as sequences of phonemes), we can explicitly ignore the value of a feature derived from the signal if that feature is underspecified in the definition of the word in the lexicon. In this way, coarticulation effects (causing variability within and across words) can be handled more easily, which would lead to better word recognition. This approach is reminiscent of the comparison between the phonetic features derived from the signal and the lexical entries in the FUL (Featurally Underspecified Lexicon) system developed by Lahiri (1999) and Reetz (1998, 1999), in which the exploitation of underspecified phonetic features is also central to the ASR system.

7. References

Bourlard, H., Hermansky, H. & Morgan, N. (1996). Towards increasing speech recognition rates. *Speech Communication* **18**(3), 205-231

Koreman, J., Andreeva, B. & Barry, W.J. (1997). Relational phonetic features for consonant identification in a hybrid ASR system. In: W.J. Barry & J. Koreman (eds.), *PHONUS* **3**, 83-109. Saarbrücken: Institute of Phonetics, University of the Saarland.

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. (ICSLP'98)*, Sydney.

Koreman, J., Andreeva, B. & Strik, H. (1999). Acoustic parameters versus phonetic features in ASR. *Proc. 14th Int. Congress of Phonetic Sciences (ICPhS'99)*, San Francisco, 719-722.

Lahiri, A. (1999). Speech recognition with phonological features. *Proc. 14th Int. Congress of Phonetic Sciences (ICPhS'99)*, San Francisco, 715-718.

Pols, L. (1999). Flexible, robust, and efficient human speech processing versus present-day speech technology. *Proc. 14th Int. Congress of Phonetic Sciences (ICPhS'99)*, San Francisco, 9-16.

Reetz, H. (1998). *Automatic Speech Recognition with Features* (Habilitationsschrift University of the Saarland).

Reetz, H. (1999). Converting speech signals to phonological features. *Proc. 14th Int. Congress of Phonetic Sciences (ICPhS'99)*, San Francisco, 1733-1736.

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