ACOUSTIC PARAMETERS VERSUS PHONETIC FEATURES IN ASR

Jacques Koreman*, Bistra Andreeva* & Helmer Strik†

** Institute of Phonetics ,University of the Saarland, Saarbrücken, Germany*
† A²RT, Dept. of Language and Speech, University of Nijmegen, Nijmegen, The Netherlands

ABSTRACT

By mapping acoustic parameters onto phonetic features, it is possible to explicitly address the linguistic information in the signal. For the experiments presented in this paper, we mapped cepstral parameters onto two sets of phonetic features, one based on the IPA chart and the other on SPE. As a result, the phoneme identification rates in a hidden Markov modelling framework increase from 15.6% for the cepstral parameters to 42.3% and 31.7% for the IPA and SPE features, respectively. Furthermore, for phonetic features the resulting confusions between phonemes are often less severe from a phonetic point of view. The theoretical implications of the differences are addressed.

1. INTRODUCTION

In most current automatic speech recognition (ASR) systems, the acoustic signal is recognised on the basis of hidden Markov models (HMM's) in conjunction with a lexicon and a language model. It is well-known that if the top-down restrictions of lexicon and language model are not used and acoustic-phonetic decoding is done with the HMM's alone, phone accuracy is generally relatively low. The goal behind our work is to improve the speech recognition results (i.e. decrease the word error rate, WER) by improving the phone accuracy.

Several articles have recently appeared in which the microphone signal is preprocessed in order to optimise the extraction of linguistic information in the signal, e.g. [1,2,4,5]. In [4,5] it was shown that the use of linguistic information can lead to a substantial improvement in the phone accuracy. The identification of pre-segmented intervocalic consonants in a hidden Markov modelling system was shown to improve from 13.2 to 52.0% by mapping the spectral representation of the signal onto IPA-based features by means of a Kohonen network.

Given these encouraging results, we decided to pursue this approach further. We extended the phonemes studied from only intervocalic consonants in [4,5] to all consonants and vowels. Furthermore, besides the IPA features (which were used in [4,5]) we also employed SPE features for the experiments described in the current paper. Within an HMM framework, we shall compare phoneme identification rates for cepstral parameters, IPA features and SPE features.

2. MATERIAL AND METHOD

2.1. Material

The speech material consisted of English, German, Italian and Dutch read passages from the Eurom0 database (2 male and 2 female speakers per language, 2 – 3.5 minutes per speaker). Eurom0 is manually segmented and labelled with SAMPA symbols. For the current experiment some of the labels had to be adapted. The reason is that some phonemes (represented by the same SAMPA symbol) have very different acoustic realisations. For example, the Italian /r/ is an apical tap or trill, while the English /r/ is a (post-)alveolar approximant.

Further, our system requires that the closure phase of plosives and affricates be labelled separately from the rest of the sound, so that additional labels had to be invented ([p0] and [b0] were used to label all voiceless and voiced closures, respectively). A full description of the names of the labels used in this paper are described in [6]. The label names for vowels follow the normal SAMPA conventions, except /{/, which had to be replaced by /AE/ to be accepted by HTK as a possible label (also, numbers are not allowed at the beginning of a label name); /Uschwa/ indicates the Dutch rounded central vowel, which sounds much like a stressed /@/.

2.2. Input Data to HMM

Three different sets of input data were used in our hidden Markov modelling experiments. They are described below.

2.2.1. Acoustic Parameters

For our baseline experiment, 26 acoustic parameters were computed from the 16 kHz microphone signal using HTK [7] (with a 15-ms Hamming window, a step size of 5 ms and pre-emphasis of 0.97): 12 mel-frequency cepstral coefficients (MFCC's), energy and the corresponding 13 delta parameters.

2.2.2. Phonetic Features: IPA & SPE

In a second experiment 19 IPA-based features were used (see Table 1), as in [4,5,6]. We shall refer to this experiment as the IPA experiment. The first 13 IPA features in Table 1 are only used for consonants, the last 5 only for vowels, while the feature [voiced] is used for both consonants and vowels.

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

Table 1. IPA and SPE features used in the experiments

SPE-based features (see Table 1) were used in a third experiment, the SPE experiment. All 16 SPE features are used

for both consonants and vowels.

All segments in our speech material were labelled with the corresponding IPA and SPE features. The features have a value 1 for “present”, -1 for “not present” and 0 for “not relevant”. The zero value was used for all vocalic features in the specification of consonants (and vice versa) in IPA, while in SPE all features are fully specified (-1 or 1).

2.3. Kohonen Networks

The phonetic features mentioned above were calculated from the acoustic parameters by means of Kohonen networks [3]. Two parallel 50 x 50 Kohonen networks are used for this acoustic-phonetic mapping. The first network is trained on the 13 static parameters (12 MFCC's and energy); the second is trained on the 13 corresponding delta parameters. Thus, the first Kohonen network models static information in the acoustic signal, while the second network models dynamic information (although this is not explicitly addressed in this paper).

The training phase of the Kohonen networks consists of three parts. (1) First, the Kohonen networks are allowed to self-organise on the basis of the acoustic parameters in all frames. The result is a so-called phonotopic map. (2) In the second step, the same acoustic parameters are fed into the networks again. The winner-takes-all principle is applied. Thus, for each frame the corresponding phonetic features are assigned to the most active neuron. (3) Finally, for each neuron average phonetic feature values are computed.

During mapping phonetic features are calculated in the following way. The acoustic parameters of one frame are fed into the Kohonen network. Then a weighted sum is calculated of the average phonetic feature values of the winning (i.e. most active) neuron and its K-nearest neighbours.

2.4. Identification and Evaluation

Each of the three feature sets was used to train hidden Markov models (HMM’s), which in turn were applied to identify segments. The same material was used for training and identification.

For each of the phones, a 3-state left-to-right HMM is trained with a single probability density function per state; no states are allowed to be skipped. In total, 53 different HMM’s were trained, 32 for consonants and 21 for vowels. Some of the consonantal HMM’s represent subphonemic units. For plosives and affricates separate models are trained for the closure phases, one HMM for voiced and one for voiceless closures (with labels [b0] and [p0], respectively; see also section 2.1).

The 53 HMMs do not represent the recognition units. During recognition an allophone dictionary is used consisting of 56 units (35 consonants + 21 vowels). In this dictionary the plosives and affricates contain an optional closure symbol.

For evaluation, allophones are pooled into phoneme categories, since only phonemic distinctions are relevant to distinguish between words in the lexicon (not used here, see section 1). More specifically, Italian [r] ([ralv]), English [ɹ] ([rret]) and German [ʀ] ([Ruvu]) are pooled into one /r/; also, dental [t] ([tden], in Italian) was pooled into one /t/ class with alveolar [t]. Thus, 53 phonemic units (32 consonants and 21 vowels) are discerned during evaluation.

In order to evaluate our results we always started by calculating a confusion matrix. On the basis of this confusion matrix two evaluation scores were obtained:

(1) Ident = total of all correct identification numbers / total number of classes to be identified

(2) ACIS = total of all correct identification percentages / total number of classes to be identified

The correct identification rate (Ident) is simply the sum of the numbers on the diagonal divided by the sum of all numbers in the confusion matrix. For computing the average correct identification score (ACIS, cf. [6]), we first normalise each row in the confusion matrix: each number in a row is divided by the sum of the numbers in that row (see e.g. Tables 3a and b). The resulting numbers on the diagonal indicate the percentage of correctly identified segments for that class. ACIS is then calculated by summing the precentage numbers on the diagonal, and dividing it by the number of classes. Thus, the difference between Ident and ACIS is that ACIS compensates for the number of occurrences of each phoneme.

3. RESULTS

In this section, we shall present the results from our three experiments. Acoustic parameters are compared with phonetic features in section 3.1; in section 3.2 the results for the two phonetic feature sets are compared.

3.1. Acoustic Parameters versus Phonetic Features

The phoneme identification rate for the baseline experiment with acoustic parameters is 15.6%, i.e. 15.6% of the phonemes is identified correctly. Acoustic-phonetic mapping raises the phoneme identification rate to 42.3% for IPA features and 31.7% for SPE features. In both cases, the improvement is substantial.

Comparing the confusion matrices from the three experiments (which are available in files 0697_01.GIF, 0697_02.GIF and 0697_3.GIF in the CD-ROM version of these proceedings), it becomes clear that phones are identified better in the mapping experiments if they occur in all languages. The more language-specific phones are often identied more successfully in the baseline system (in particular, /tden, g, p0f, T, C, D, Z, rret, w, Y, V, 2 Uschwa, 3, AE, A, V, 6/) than in both mapping systems. Their effect on the overall phoneme identification is small, however, since the number of realisations for these phones is less than average for the overall corpus (in part because they are language-specific). The deterioration of phoneme identification results for more language-specific phones is an inherent disadvantage of focussing on more abstract, linguistic properties of the signal, which is achieved at the expense of the redundancy which is characteristic of the acoustic signal.

The fact that the variability in the acoustic parameters is retained in the baseline experiment, instead of replacing it by more homogeneous phonetic features, also shows itself in the confusion results. In general, the confusions between phonemes are more severe from a phonetic point of view in the baseline experiment than in the mapping experiments. As an example, let

us consider labial consonants. Labiality clearly has different acoustic properties in plosives than it has in the nasal /m/, for example. The greater variation in the acoustic parameters, which form the input to hidden Markov modelling in the baseline experiment, makes it much more likely that confusions occur – also with non-labials – than if we map all the different acoustic realisations of consonants onto the same feature [labial]. If the varying acoustic parameter values for labial phones are mapped onto a phonetic feature [labial], labials are more likely to be confused among themselves than with other places of articulation. For a further phonetic discussion of the effects of acoustic-phonetic mapping, the reader is referred to [4,6].

3.2. Two Phonetic Feature Sets

The phoneme identification results in the two mapping experiments are 42.3% for the IPA features and 31.7% for the SPE features (see Table 2), as was already mentioned above. Of course, with 16 SPE features versus 19 IPA features, one can argue that the information load on the features is lower in IPA. On the other hand, all 16 SPE features are used for both vowels and consonants, whereas only 6 of the IPA features are used to describe vowels, and 14 are used to disinguish among the consonants, with an overlap in the feature [voiced], which is used for both vowels and consonants (see section 2.2.2).

There is also a difference in overlap between the features. In IPA, the sounds are defined along three largely independent axes (place, manner and voicing). There is only partial independence between the features, because some places of articulation (e.g. palatal) only occur for certain manner classes (fricatives or approximants). The same interdependence exists for example between voicing and manner (e.g. nasals are always phonemically voiced in the four languages). In the SPE feature set, the same feature is often used for different types of phonemes, for instance the feature specification [-coronal] is used to define both labial and velar consonants, and most features which are used in SPE to distinguish between different vowels are also used to distinguish consonants. This makes it easier for vowels and consonants to be confused when SPE features are used than if we map onto IPA features.

IPA	SPE	Explanation
42.3	31.7	Ident for 53 phonemes
81.3	83.5	Ident: V identified as V
87.9	82.5	Ident: C identified as C
35.8	31.3	Ident for the vowels only
53.4	53.0	ACIS for the vowels only
44.4	36.8	Ident for consonants only
53.0	47.1	ACIS for consonants only
53.6	43.0	Ident for cons. place of art.
72.8	65.8	ACIS for cons. place of art.

Table 2. Ident and ACIS values for various experiments

Although after separately pooling all vowels and all consonants there is no difference in the classification of vowels (81.3% correct class identification for IPA features; 83.5% correct for SPE features), there is a stronger tendency for consonants to be identified as vowels in the SPE experiment. In the IPA experiment, only 12.1% of the consonants are identified as vowels, against 17.5% in the SPE experiment. This supports our hypothesis that feature sharing across phoneme classes, as is the case in SPE (where some features are used to differentiate between subclasses within both consonants and vowels) increases the scope for confusions.

If we look at the identification of vowels separately, the correct phoneme identification rates (Ident) are 35.82 (IPA) and 31.26 (SPE), respectively. This can be explained by the better identification of the most frequent vowels /@, e, a, i/ in the IPA experiment. If we compensate for the number of occurrences of the vowels (by the using ACIS), this difference disappears. For linguistic interpretation, we prefer the ACIS to an identification rate, because it better reflects how well each of the phonemes is identified (of course, for applications, the average number of realisations phonemes is very important). In the IPA experiment, ACIS is 53.4%, while a value of 53.0% is obtained in the SPE experiment. The similar vowel identification in the two experiments was expected, since the phonetic features which are used to describe the distinctions between the vowels are very similar (see Table 1).

With an ACIS of 53.0% (overal consonant identification rate: 44.4%), the consonants are identified better with IPA features than when SPE features are used (ACIS: 47.1%; overall consonant identification rate 36.8%).

Generalising the data across consonantal place of articulation, we find that in the IPA experiment, the ACIS for consonantal place of articulation is 72.8% (Ident: 53.6%). In the SPE experiment, the ACIS is only 65.8% (Ident: 43.0%). If we compare the confusions between place-of-articulation categories, as shown in figures 3a and 3b[1], we see that dentals are far more often identified as alveolar, uvulars as velar, and glottal consonants as velar in the SPE than in the IPA experiment.
These confusions are for consonant classes which have identical values for the [coronal] and [anterior] values. Labials are more often identified as velars in the SPE experiment than in the IPA experiment; they also share their values for [coronal], as well as for [stridency]. The greater number of confusions between alveolars and velars can be partially explained by their sharing of [stridency]. We suggest that the explanation for the greater number of place confusions in the SPE experiment lies in feature sharing. When we consider the different places of articulation in the IPA feature set, we can see that all places of articulation are equally confusable, since they always differ in one category. Comparing this to SPE, a different relationship between the different places of articulation is implied. Since the place of articulation for consonants is described by a more complex relationship between different feature values than is the case for the IPA features set, more fine-grained distinctions in the relations between the different places of articulation are

[1] The ACIS was derived from this figure by dividing the number for each correct place of articulation by 100 minus the percentage of deletions in the row, summing the outcomes and dividing it by the number of places of articulation (8). This was done, so as to partially correct for consonants which were misidentified as vowels, which appear in the column ”Del”. To get the true ACIS value for this experiment, it should be rerun with consonants only.

possible than when IPA features are used. For example, /s/ and /S/ differ in the features [anterior], whereas /s/ and /x/ differ in the features [anterior], [coronal] and [back]. This results in a higher confusability between the consonants.

	lab	den	alv	alp	pal	vel	uvu	glo	DEL	*n*
lab	58	11	14	0	0	6	0	1	9	*2524*
den	5	83	9	2	0	1	0	0	1	*371*
alv	12	3	57	5	1	6	1	0	14	*6933*
a1p	0	0	6	91	2	0	0	0	1	*320*
pal	4	0	7	1	66	1	0	0	21	*319*
vel	17	7	11	1	0	61	0	0	3	*982*
uvu	9	1	3	0	0	10	41	0	34	*290*
glo	10	4	6	0	3	2	0	52	24	*102*
INS	20	3	36	0	27	3	9	1	1	*1285*

Table. 3a. Consonant confusion percentages and number of realisations in the IPA experiment; all data pooled for place of articulation

	lab	den	alv	alp	pal	vel	uvu	glo	DEL	*n*
lab	38	10	18	1	1	15	1	1	16	*2524*
den	6	67	20	3	0	1	0	0	3	*371*
alv	13	5	45	4	1	11	1	0	20	*6933*
a1p	0	1	6	90	1	0	0	0	1	*320*
pal	2	1	5	1	59	2	0	0	31	*319*
vel	12	5	11	1	0	68	0	0	3	*982*
uvu	8	1	4	0	0	16	28	1	42	*290*
glo	8	0	5	0	4	14	1	49	20	*102*
INS	25	2	30	0	21	5	13	2	2	*1261*

Table 3b. Consonant confusion percentages and number of realisations in the SPE experiment; all data pooled for place of articulation

4. DISCUSSION

In this paper, we have presented an approach to ASR in which linguistic information is extracted from the acoustic signal by acoustic-phonetic mapping. The advantages and disadvantages of acoustic-phonetic mapping on phoneme identification were discussed in section 3.1. It is clear from our experiments that the different feature definitions which can be chosen to represent the linguistic properties of the phonemes have important implications for possible confusions and therefore for the overall phoneme identification results (section 3.2).

The results from our experiments show that Kohonen networks combine several advantages in performing acoustic-phonetic mapping.

First, the phonetic features only reflect those properties of the acoustic signal which are relevant for the distinction between phonemes, and they have a clear linguistic interpretation. Furthermore, when phonetic features are used the resulting confusions among the phonemes are generally less severe from a phonetic point of view and much more systematic, compared to the confusions obtained with cepstral parameters.

Second, the acoustic-phonetic mapping can map different acoustic realisations of a phoneme (allophones) onto the same phonetic features. For instance, different allophones of [l], like clear and dark [l], which can occur in different positions in the syllable, for instance in British English, and partially devoiced versus fully voiced [l] which occur after voiceless plosives and in most other contexts, respectively, can be represented in different parts of the phonotopic map, while at the same time the neurons which model their acoustic properties emit very similar average phonetic feature values. Thus, acoustically heterogeneous realisations of phonemes are replaced by more homogeneous phonetic feature vectors at the input to hidden Markov modelling.

Third, the acoustic-phonetic mapping can be performed automatically and is fast, so that it can be implemented in existing real-time ASR systems.

A disadvantage of Kohonen networks is that they must be trained on segmented and labelled material. However, this has to be done only once. Up till now we have used manually segmented and labelled material. We intend to investigate whether it is possible to use automatically segmented and labelled material. The restriction that manually segmented and labelled data must be used to train the Kohonen networks is not so severe, if we consider that our experiments show that it is possible to generalise these networks across languages, so that already available manually segmented and labelled material from several languages may be used for training.

Our experiments show that Kohonen networks are very advantageous for pre-processing the acoustic parameters. If Kohonen networks are used to map acoustic parameters onto phonetic features, phone accuracy is increased. However, the current experiments have some limitations: training and test data are the same, and we have not checked whether the increases in phone accuracy also lead to reductions in word error rates when a lexicon and language model are used during recognition. Experiments are now underway to evaluate whether the improved phone accuracy obtained by acoustic-phonetic mapping also decreases the word error rate for a complete ASR system with a lexicon and a language model.

REFERENCES

[1] Bitar, N. & Espy-Wilson, C.. 1995. 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] Dalsgaard, P. 1992. Phoneme label alignment using acoustic-phonetic features and Gaussian probability density functions. *Computer Speech and Language* **6**, 303-329.

[4] 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.

[5] Koreman, J., Barry, W.J. & Andreeva, B. 1998. Exploiting transitions and focussing on linguistic properties for ASR. *Proc ICSLP*.

[6] Koreman, J., Andreeva, B. & Barry, W.J. 1998. Do phonetic features help to improve consonant identification in ASR? *Proc. ICSLP*.

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