MODELIZATION OF ALLOPHONES IN A SPEECH RECOGNITION SYSTEM

K. Bartkova & D. Jouvet

Centre National d'Etudes des Télécommunications
LAA/TSS/RCP - Route de Trégastel - 22 300 Lannion - France

ABSTRACT

This paper describes a new approach for modelling allophones in a speech recognition system based on Hidden Markov Models (HMM). This approach allows a detailed modelization of the different acoustical realizations of the sounds with a limited amount of parameters by integrating left and right context dependent transitions as well as acoustical targets. Phonetical knowledge is used in the definition of the structure of the models, and a standard HMM training procedure determines the optimal value of the parameters. The efficiency of the approach is demonstrated both in a multispeaker mode, on a 500 word vocabulary, and in a speaker independent mode on several other databases recorded over telephone lines.

1 INTRODUCTION

The hidden Markov modelling approach is now a widely used technique in automatic speech recognition. Although it allows the optimal parameters of a model for a given training corpus (known words or sentences) to be automatically determined, the structure of the models still remains to be defined manually, and the choice of the "best" basic units is difficult.

Basic word units are very suitable for small size vocabularies. But, when the vocabulary size increases, basic sub-word units lead to more compact models. Although phoneme units would be a good theoretical choice, they do not work well in practise, as they do not account for the coarticulation effect due to the context influence. To cope with this problem we had previously developed the pseudo-diphone units [2] which consist of the central part of phonemes, of the transitions between phonemes, and also of some strongly coarticulated sound sequences treated as single units. As an alternative, context dependent units, modelling the acoustical realization of the corresponding sound in a specific left and right context [4], can be used. However, such an approach leads to a large number of models that must be reduced in size, in order to achieved a reliable estimation of the parameters. This can be done a priori, using phonetical knowledge [1], or a posteriori, using some clustering algorithms [3].

In the new approach described in this paper, the Markov models are defined in such a way that they can share as many parameters as possible for modelling the different acoustical realization of any sound. This sharing, based on some a priori phonetical knowledge, allows detailed phonetic distinctions to be introduced in the models, with a limited amount of free parameters that are later determined by an automatic procedure (standard HMM training).

2 MODELLING ALLOPHONES

The new modelization of allophones consists in modelling together, in a single basic unit, all the possible acoustical realizations of a given sound. Each sound is thus represented by a single model, having several entry states and several exit states, and allowing the tying of the probability density functions (pdf's). An example of such a model is represented in figure 1. Each entry or exit state is associated to a specific context, that is, to a class of left or right phonemes having the same acoustical influence on the sound. In this approach, every path from one entry to one exit corresponds to an allophone.

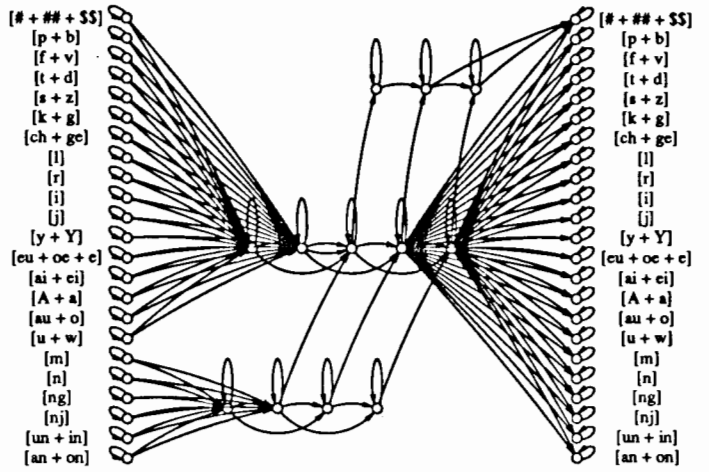

Figure 1 - Structure of the acoustical models used for the vowels, and contexts associated to the entry and exit states.

A typical model for the vowels would consist in a shared central portion representing the acoustical "target", and transitions from each entry to the "target", and from the "target" to each exit. However, if necessary, several acoustical targets may be defined and the number of left and right contexts can be increased as much as necessary. Because of the integrated modelization of all the acoustical realizations of any sound, and of the sharing the gaussian pdf's whenever it is possible, a detailed modelization is obtained with a small number of parameters. Thus, they can be reliably determined using a standard HMM training procedure.

2.1 Context Influence

Given that some phonetic environments induce the same coarticulation effects on the adjacent sounds, the entry and exit contexts were defined, for each class of sounds, by grouping together phonemes inducing the same acoustical influence. For instance, consonants sharing the same articulation feature tend to affect the following sound in a similar way. As far as vowels are concerned, the similarity between tongue positions will closely affect the vowel transition towards the neighbouring sounds.

Vocalic contexts for every allophone : As the tongue position in a semi-vowel production is very similar to that of a vowel production, the vocalic contexts involve vowels as well as semi-vowels. According to point and manner of articulation, 10 relevant vocalic contexts were defined : */i/, /j/, high-front-rounded, high-mid-rounded, low, mid-front, mid-back, high-back, front-nasal* and *back-nasal.*

Consonantic contexts for vowel, semi-vowel and liquid allophones : Because of the formant transitions they induce on the vowels, as well as on the semi-vowels, the consonants were grouped, according to the place of articulation, in 9 homogeneous contexts : *labial, labio-dental, dental, alveolar, palato-alveolar, palatal, velar, /r/* and */l/*. However, the nasal consonants */m/, /n/, /nj/* and */ng/* were treated as separate contexts as they may induce a nasalization of the following vowel.

Consonantic contexts for consonant allophones : The transition between two adjacent consonants is less obvious than between two adjacent vowels. On the other hand, consonants assimilate acoustic features (nasality, voicelessness...) easier than vowels. Thus, the merging of consonantic contexts for consonant allophones was slightly different from that used for vowel allophones. 7 relevant contexts were defined according to acoustic features : *voiceless plosives, voiced plosives, voiceless fricatives, voiced fricatives, nasals, /r/* and */l/*.

2.2 Possible "Targets"

The inner part of the models represent the acoustical targets. Thus, in order to take into account the possible assimilation of some of the acoustic context features, several targets, representing "standard" pronunciations as well as modified ones, were modelized. The structure of the targets was carried out in order to allow the modelization of even a rather short duration of the overall sound.

Vowel targets : The following acoustical realizations were possible for the vocalic targets : *voiced, partially devoiced, partially nasalized* or *partially aspirated* (not represented on figure 1). The loss of the voiced feature at the beginning or at the end of the sound could occur only in a left or right pause context. In the same way, the nasalized target was accessible only from a left nasal context.

Consonant targets : In the consonantal target modelization, a difference was made between a *"normal"* non assimilated target, a *devoiced* and a *partially devoiced target* (valid only for voiced consonants) and a *nasalized target.* The partially devoiced target was accessible only in a right pause context and the nasalized target (partially or completely) was valid only after a left nasal context.

Semi-vowel and liquid targets : The structure of the models used for semi-vowels and liquids were very similar to that used for vowels. Nevertheless some specificities separate these two sound classes. One of the main differences consists in the length of target modelizations. As liquids and semi-vowels are sounds realized most of the time with short or even very short duration, and thus are strongly coarticulated with the adjacent sounds, fewer states were attributed to the modelization of their sound targets. Thus 4 "short" targets were used to modelize : a *"normal"* acoustic realization without any assimilation effect, a *devoiced target,* a *partially devoiced target,* and a *partially nasalized target.*

2.3 Phonological Rules

Besides the coarticulation effects between adjacent sounds treated by the allophone models, the system can handle phonological rules in order to modify the "standard" phonetic descriptions of the vocabulary words. These rules were used not to predict a specific pronunciation (as in phonology), but rather to tolerate several pronunciations that might occur in a speaker independent mode. Thus, each application of a rule increased the number of possible pronunciations of the words. These explicit phonological rules were the following : 1) each word ending with a consonant and followed by a pause can be pronounced with a neutral schwa like vocalic sound after the consonant ; 2) a voiced fricative preceded by a pause can begin with a very short schwa like vocalic sound ; 3) a succession of sounds containing a sonorant and the liquid /r/ can be realized with an epenthetic neutral schwa like vowel between them (especially in a slow speaking rate) ; 4) a voiced stop can loose its voiced feature when followed by a voiceless consonant ; 5) a voiced stop, followed by a nasal consonant and sharing with it the same point of articulation, can assimilate its nasal feature.

3 EXPERIMENTS

In order to validate this new approach we tested it on several databases recorded over telephone lines. A 500 word vocabulary was used to study the influence of the structure of the models (number of contexts, usefulness of the targets, etc). This vocabulary was recorded 3 times by 10 speakers, 2 repetitions were used for computing the optimal parameters of the HMM, and the third one was used for testing the recognition performances (in a multispeaker mode). The modelization described above has then been applied (for speaker independent recognition) to several other vocabularies, recorded mainly over long distance telephone lines, by several hundreds of speakers from different parts of France, thus having different accents.

3.1 Influence of the structure

In the tests reported in table 1, the acoustical analysis computed every 16 ms a set of 8 coefficients : 6 Mel frequency cepstrum coefficients, the logarithm of the total energy, and its temporal variation. The database used is the 500 word vocabulary (the 500 most frequent French words) recorded by 10 speakers. We report only the error rate on the test set.

The first allophone model used a single simple structure for every sound, involving a single target and 13 entry and 13 exit states. Using a single set of 13 contexts, the same for all the sounds, we achieved a 19.1 % error rate. Introducing the contexts defined in the previous section, and several target models for the sounds, the word error rate decreased to 17.0 %. A further improvement, leading to 16.3 % error rate was obtained by shortening the target model for the liquid and semi-vowel. In the preceding tests, there was no loop allowed on the entry and exit states. By adding these loops, represented on figure 1, longer transition between two adjacent sounds have got a better modelization, and further improvement of the recognition score was obtained with a 14.4 % word error rate.

Table 1 - Error rate on the test set of the 500 word vocabulary for different structures of the allophone models.

Structure of the models	Errors
13 contexts & 1 target	19.1 %
More contexts & targets	17.0 %
Liquids & semi-vowels shorter	16.3 %
Loops on entry & exit states	14.4 %

3.2 Efficiency of the Approach

In this section, the allophone modelization is compared to the pseudo-diphone modelization and to the word models. The standard acoustical coefficients computed every 16 ms, were used together with their first and second derivatives.

Using the last modelization, described above, and taking into account the temporal derivatives of the acoustical coefficients we finally obtained a 8.44 % error rate on the 500 word vocabulary, which is significantly better than the 11 % obtained with the pseudo-diphones units on the same database.

Table 2 - Error rate obtained on several databases (for the test set) with different modelizations : Allophone models (All) ; Pseudo-Diphone units (PsD) ; and whole word models (Word).

Error rate	All	PsD	Word
Digits	**0.86 %**	1.33 %	0.69 %
Tregor	**1.00 %**	1.42 %	0.86 %
Numbers	**4.47 %**	5.68 %	——
500-Words	**8.44 %**	11.04 %	——

The other databases used for the comparisons are : Digits (the 10 digits, recorded by 775 speakers), Tregor (36 French words recorded by 513 speakers), and Numbers (French numbers between 00 and 99 recorded by 740 speakers). Each of them was split in two parts : one half for training, and the other half for testing. For these three databases, the speakers were different in the test and the training set, therefore the reported results (error rate on the test set) corresponds to a speaker-independent mode.

As can be seen on the above table, the results achieved by the allophone modelization are significantly better than those obtained with the pseudo-diphones units. Also, even on small vocabularies, the allophone models, which use less gaussian pdf's than the word models, lead to performances which are comparable to those obtained with word models.

4 CONCLUSION

The present study described an efficient way of modelling the allophones by representing in an integrated manner all the different possible acoustical realizations of the sounds. Phonetical knowledge was used for the definition of the structure of the models, whereas a standard HMM training procedure determined the optimal values of the model parameters. The application of the same modelization to different databases led to good performances, demonstrating thus the efficiency of this new approach.

REFERENCES

[1] K. Bartkova & D. Jouvet : ***"Speaker-independent speech recognition using allophones"*** ; Proc. ICPhS 1987, Tallin, USSR, August 1987, Vol. 5, pp. 244-247.

[2] D. Jouvet, J. Monné, D. Dubois : ***"A new network-based speaker-independent connected-word recognition system"*** ; Proc. IEEE Int. Conf. ASSP, Tokyo, Japan, 1986, pp. 1109-1112.

[3] K. F. Lee, H. W. Hon, M. Y. Hwang, S. Mahajan, R. Reddy : ***"The SPHINX speech recognition system"*** ; Proc. IEEE Int. Conf. ASSP, Glasgow, Scotland, 1989, pp. 445-448.

[4] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, J. Makhoul : ***"Context-dependent modeling for acoustic-phonetic recognition of continuous speech"*** ; Proc. IEEE Int. Conf. ASSP, Tampa, Florida, 1985, pp. 1205-1208.