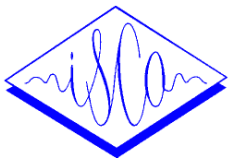

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

# STRUCTURE OF ALLOPHONIC MODELS AND RELIABLE ESTIMATION OF THE CONTEXTUAL PARAMETERS

*D. Jouvet, K. Bartkova & A. Stouff*

France Télécom, CNET Lannion, LAA/TSS/RCP,
Route de Trégastel, 22300 Lannion, FRANCE

## ABSTRACT

This paper presents a contextual modeling of phonemes, and describes a new technique that renders a reliable estimation of contextual parameters. Using this approach the modeling of all of the acoustic realizations of a given sound is integrated into a single complex unit, for which each entry and exit state is assigned to a specific context.

Context clustering trees are defined and used in order to provide a reliable estimation of the contextual parameters. Using hand-made trees, a 12 % reduction in the error rate is achieved on a 250-word vocabulary set, which is distinct from the training vocabulary. Finally, an automatic context clustering procedure is presented and applied in order to automatically generate these clustering trees. Using this automated approach the reduction in the error rate is comparable to that of the hand-made trees.

## I. INTRODUCTION

For automatic speech recognition based on Hidden Markov Models, several kinds of basic units may be used. Although word-based models lead to high recognition performance for isolated word recognition, this approach rapidly becomes intractable as the size of the vocabulary increases. In this case, phoneme-based models are often used. However, a detailed modeling of the acoustic realizations of the phonemes is necessary in order to achieve efficient recognition. This is generally obtained using contextual units (triphones) which model the acoustic realization of a sound in a specific left and right context [1]. This approach results in a large number of contextual units. In terms of real tasks, the finite size of the training data base makes it impossible to estimate all of contextual units. Moreover, the estimation of some parameters may not be reliable.

Consequently, there must be a trade-off between a large number of parameters which provide a detailed modeling, and a small number of parameters wherein a reliable estimation can be obtained. A reasonable number of parameters is usually achieved by limiting the inventory of contextual units. Some theoretical phonetic knowledge may be used in order to specify the allophonic units which are relevant to the current recognition task [2]. Another way consists of training all possible contextual units, and then merging "similar" contextual models [3].

Another problem may arise if the vocabulary used in the training and testing are different. Some contextual units assigned to contexts which are not present in the training corpus, and thus untrained, may still be required for the test vocabulary. In this case, less specific models are used (left or right context dependence only), or a decision tree may be used in order to determine which trained contextual model is most relevant to this context [4]. The decision tree is usually obtained using a classification and regression tree methodology [5].

In order to achieve a detailed contextual modeling with a limited number of parameters, an integrated modeling of all the acoustic realizations of sounds has been developed [6]. This approach will be briefly reviewed herein. As with the standard triphone approach, if a context is not present in the training vocabulary, there is no way of accurately estimating the corresponding parameters. However, an approximated value may be obtained by enlarging the assigned context so that enough training examples fall in it. The value of the contextual parameters assigned to this enlarged context will provide an approximation for the untrained parameters. The procedure developed in order to obtain these approximations is based on context clustering trees, and will be described later.

Following a brief overview of the speech recognition system and the speech data base used in this study, the allophonic modeling will be presented. The body of this paper is devoted to the procedure that was recently developed in order to obtain an approximation of the values of untrained contextual parameters. This procedure is based on context clustering trees that can either be theoretically defined using phonetic knowledge, or automatically generated using a context clustering procedure.

## II. SYSTEM OVERVIEW AND SPEECH DATABASE

The speech recognition system used in this study, is based on a Markov modeling approach. This system was developed for telecommunications applications [7], and is used in interactive vocal servers.

Every 16 ms, the acoustic analysis computes a set of 8 Mel frequency cepstral coefficients plus the energy of the frame. The 9 first order, and 9 second order temporal

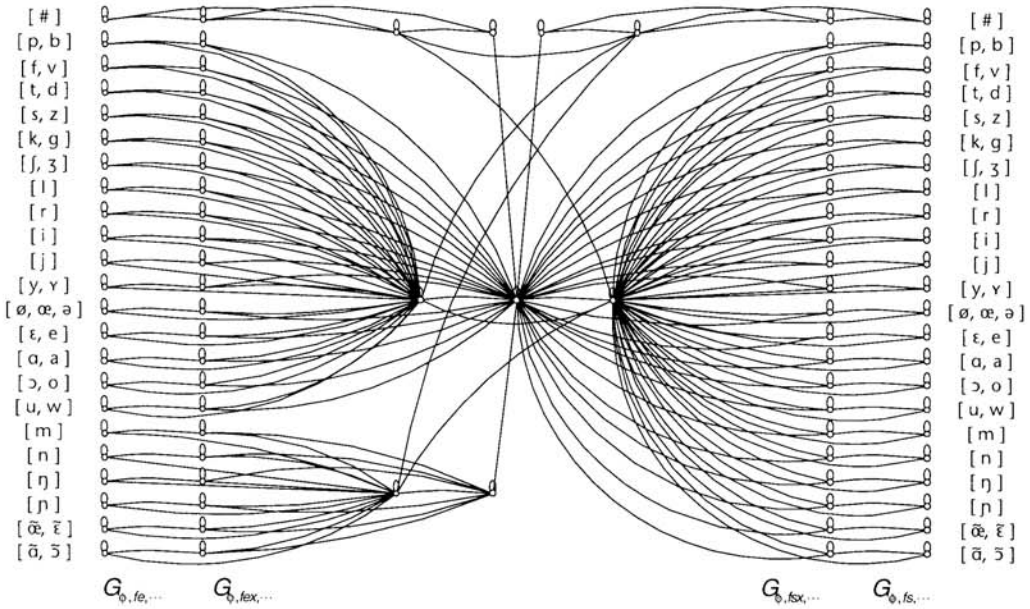


**Figure 1 - Structure of the Allophonic Model Used for the Vowels.**

derivatives are then estimated using a 5 frame window which is centered on the current frame. Consequently the acoustical vectors have 27 components.

The speech recognition system uses a fully compiled network which includes a model for start and end silences, and accepts various kinds of basic units (words, phonemes, allophones, etc.). The probability density functions are Gaussian, and are assigned to the transitions of the network.

The speech corpus used in these experiments was collected over the telephone. The 500 most frequently spoken French words were recorded using 10 speakers. Each speaker uttered the whole vocabulary 3 times. For the experiments described in this paper, the corpus was split into two parts. Repetitions corresponding to the first 250 words were used in training, and repetitions of the other 250 words were used in testing. Some contexts which appeared in the test set, did not necessarily appear in the training set.

## III. ALLOPHONIC MODELS

The allophonic modeling that was developed integrates the modeling of all the contextual realizations of a sound into one complex unit. As such, a large number of parameters can be pooled between the modelings of the different contextual realizations. The complex unit has several entry and exit states (see Figure 1). Each of these states is assigned to a specific left or right context. The contexts were theoretically defined, by grouping phonemes which induce the same, or nearly the same, influence on the acoustic realization of the sound.

The central part of the unit models the "target" which is supposed to be shared between all the contextual realizations. However, because the central part of the sound may be realized in different ways, several "targets" were defined. For example, two "targets" were defined for the vowels, in order to model a devoicing before and/or after a silence (see top of Figure 1). Another "target" was defined in order to model a nasalization (see bottom of Figure 1) which may occur after a nasal sound.

This approach was evaluated on several databases [8] for speaker-independent speech recognition. For isolated word recognition, the allophonic models yield results which are slightly inferior to those obtained using the best word-based models. However, for connected word recognition (as two digit numbers, from 00 to 99) this allophonic approach yields results superior to those using the word-based approach. This is due to a more efficient modeling of the context influence at word boundaries.

## IV. ESTIMATION OF CONTEXTUAL PARAMETERS

The majority of contextual parameters were correctly estimated using the Viterbi training procedure. However, the contextual parameters assigned to contexts that are not present in the training set are not estimated. Moreover, if a context occurs infrequently in the training set, the estimation of the corresponding parameters may not be reliable. Untrained contextual parameters are likely to appear when the test vocabulary is different from the training vocabulary.

Hence the following procedure was applied after the training phase in order to provide a value for the untrained parameters and modify the value of the parameters that were considered to be unreliable.

### IV.1 Description of the procedure

The procedure is based on context clustering trees. The leaves are assigned to the contexts of the allophonic units (more than 20 in this case), and the root is assigned to all the phonemes (i.e., the largest possible context). Moving from the leaves up to the root provides larger and larger contexts. Figure 2 shows a partial example of such a tree, and exhibits the contexts assigned to the leaves and nodes.

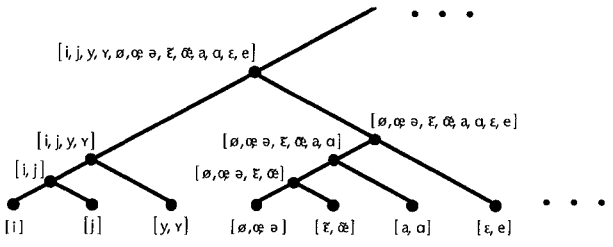

**Figure 2 - Example of a Context Clustering Tree.**

As shown in Figure 1, the contextual Gaussian densities assigned to the phoneme $\phi$ are identified by the context (for example [i]), and the position in the model (entry "*fe*", just after the entry "*fex*", just before the exit "*fsx*" and the exit "*fx*"). As shown in Figure 3, the adjustment of the contextual parameters will be applied separately to each phoneme and each position. The actual Gaussian densities of the model, and their corresponding parameters, are assigned to the leaves. The Gaussian density assigned to a node is the result of the merging of the two Gaussian densities that are assigned to the 2 sub-nodes (or leaves).

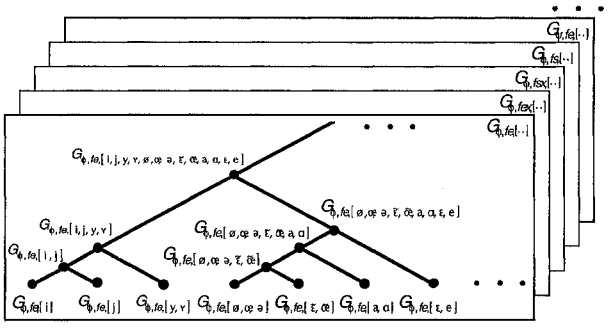

**Figure 3 - Contextual Parameters and Context Clustering Trees.**

The merging procedure was originally developed in order to reduce the number of parameters in a word-based model [9]. Using the parameters of the two Gaussians assigned to the sub-nodes, this procedure provides an estimation of the parameters of the resulting Gaussian. The number of frames from which the Gaussian parameters were estimated is also assigned to each node (and leaf). Consequently, all of the parameters are easily and rapidly computed. These parameters define the contextual Gaussian function assigned to the corresponding context. The Gaussian function linked to the root is thus assigned to a context-independent modeling (because the corresponding context is the full set of phonemes).

From the parameters of all these Gaussian functions, an estimation for the parameters of the actual Gaussian functions of the model can be deduced. Starting from each leaf, the number of frames used to estimate the parameters of the corresponding Gaussian function is compared to a given threshold. If this number is larger than a predefined value (around 25, see *threshold used for re-estimation* on Table 1) the Viterbi estimation is considered to be reliable, and the Gaussian parameters are not modified. If not, the tree must be ascended until a node with a number of frames greater than the threshold, is attained. This Gaussian function is considered to be a good approximation of the untrained (or unreliable) Gaussian function assigned to the leaf, and its parameters are then copied.

## IV.2 Hand-made trees

To evaluate this procedure 2 hand-made trees were used. One hand-made tree was applied to the vowels and semi-vowel models, and another hand-made tree was applied to the consonant models. As illustrated in Figure 4, the clustering of vowels and consonants are separate. They only merge at the root of the tree.

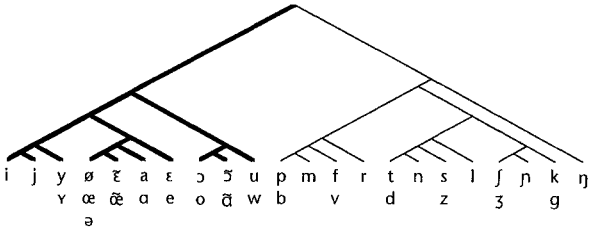

**Figure 4 - Hand-made Context Clustering Tree Applied to Vowels and Semi-vowels.**

The second column in Table 1 illustrates the error rate that was obtained without applying this procedure to the 250-word test vocabulary, and the error rate obtained when applying the procedure with different thresholds. Without any processing, the error rate was 11.47 %, with a 95 % confidence interval of ±0.73 %. When re-estimating the parameters of the Gaussian functions that were estimated with less than 25 or 50 frames, the error rate decreases to about 10 %, which is a 12 % reduction in the error rate. When re-estimating fewer Gaussian functions (threshold equal to 10) or more Gaussian functions (threshold equal to 100) the reduction is not as large. However, the results are still superior to those without processing.

## IV.3 Context clustering

Because the existing data (values of the parameters after the Viterbi training procedure) represents an estimate of the acoustic realization of the sounds, they can be used to automatically generate context clustering trees. In order to do so, it is assumed that a given context clustering tree is shared by several units (phonemes), and that the same context is not always missing in the training set.

The "distance" between two contexts was estimated from the distance between the pairs of Gaussian functions described in [9]. The "distance" between two contexts $\alpha$ and $\beta$ is defined as:

$$D(\alpha,\beta) = K_{\alpha,\beta} \sum_{\phi} \sum_{f.} D\left(G_{\phi,f.,\alpha}, G_{\phi,f.,\beta}\right) \quad \textbf{(Eq. 1)}$$

where $K_{\alpha,\beta}$ is a normalization coefficient, and the summations concern all the phonemes $\phi$ which share the same tree and all the contextual positions (*f.*).

Initially, the list of all the contexts corresponded to the leaves on the tree. The tree was then incrementally built, beginning with the leaves, clustering similar contexts first. In order to do this, the distance between every pair of contexts was computed, and then the two closest ones were merged. This gives the first node on the tree, which

**Table 1 - Recognition Error Rate After the Re-estimation of the Contextual Parameters.**

| Threshold used for re-estimation. | A priori tree | Automatic tree (threshold used for context clustering) | | | |
|---|---|---|---|---|---|
| | | 0 | 25 | 50 | 100 |
| 0 (without re-estimation) | 11.47 % | 11.47 % | 11.47 % | 11.47 % | 11.47 % |
| 10 | 10.47 % | 10.22 % | 10.13 % | 10.39 % | 10.14 % |
| 25 | **10.06 %** | **10.05 %** | **9.94 %** | 10.21 % | 10.20 % |
| 50 | **9.77 %** | **9.99 %** | **10.18 %** | 10.83 % | 10.14 % |
| 100 | 10.44 % | 10.80 % | 10.59 % | 11.18 % | 11.17 % |

corresponds to context [i,j] in Figure 5. Consequently, there is one less context. The list of contexts is updated by removing the two original contexts that were merged (here [i] and [j]) and adding the new context (here [i,j]). The procedure is repeated until only one context remains, (the set of all the phonemes) which constitutes the root of the tree.

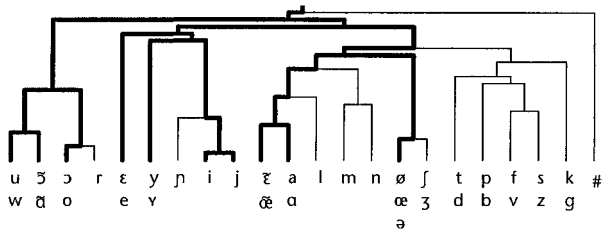


**Figure 5 - Automatic Clustering of the Contexts.**

Figure 5 illustrates the results of the clustering procedure for the vowel units. The position of the horizontal line indicates the rank at which the merging occurred (the first one at the bottom for contexts [i] and [j], the last one at the top).

This approach differs from the classification and regression trees [5] that are used for choosing contextual models in two ways. Firstly, the trees are built beginning with the leaves, using a hierarchical clustering procedure. Secondly, instead of choosing a model for a given context, a method is defined in order to estimate the required contextual parameters if they have not been correctly trained (i.e., from an insufficient number of frames).

## IV.4 Results

Contextual parameters are required in order to compute the distance between two contexts. All of the (Viterbi) estimated parameters can be considered, or only the ones having a "reliable" estimation (i.e., estimated from enough frames). Hence the coefficient $K_{\alpha,\beta}$ in (Eq. 1) which takes into account the actual number of contextual Gaussian functions used in the summation.

Table 1 cites the error rate obtained for different values of the parameters. In each case there was a reduction in the recognition error rate.

The best results are comparable to those obtained using hand-made trees which are defined using phonetic knowledge. Moreover, the threshold used in the clustering procedure doesn't seem to be crucial. Smaller values seem to yield better results. With smaller values, all of the Viterbi estimated parameters are taken into account, even if they are not reliable. However, because the distance between two Gaussian functions accounts for the number of frames from which they have been estimated, this compensates for unreliable estimations.

## V. CONCLUSION

This paper has presented an integrated modeling of the contextual acoustic realizations of sounds in a single complex unit. This approach enables a detailed modeling with a limited number of parameters.

A procedure based on context clustering trees, was then introduced in order to provide a reliable estimation of the contextual parameters of the models. Using hand-made trees a 12 % reduction in the error rate was achieved on a 250 word vocabulary set (distinct from the training set). An automatic estimation of such a tree, based on a hierarchical clustering, yields about the same reduction.

## BIBLIOGRAPHY


[1] 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, Floride, USA, March 1985, pp. 1205-1208.

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

[3] K. F. Lee : "Context-dependent phonetic hidden Markov models for continuous speech recognition" ; *IEEE Trans. on ASSP*, February 1990.

[4] L. R. Bahl et al. : "Large vocabulary natural language continuous speech recognition" ; *Proc. IEEE Int. Conf. ASSP*, May 1989.

[5] Breiman, Friedman, Olshen & Stone : "*Classification and regression trees*" ; Wadsworth, Inc., 1984.

[6] K. Bartkova & D. Jouvet : "Modelization of allophones in a speech recognition system" ; *Proc. ICPhS*, Aix en Provence, France, August 1991, Vol. 4, pp. 474-477.

[7] C. Gagnoulet : "Speech recognition over the telephone : experiments in France" ; *Proc. Voice Systems Worldwide 1990 conference*, London, May 1990, pp. 173-177.

[8] D. Jouvet, K. Bartkova & J. Monné : "On the modelization of allophones in an HMM based speech recognition system" ; *Proc. EUROSPEECH*, Genova, Italy, September 1991, pp. 923-926.

[9] D. Jouvet, L. Mauuary & J. Monné : "Automatic adjustments of the structure of Markov models for speech recognition applications" ; *Proc. EUROSPEECH*, Genova, Italy, September 1991, pp. 927-930.