Multi-Band Continuous Speech Recognition

Christophe CERISARA, Jean-Paul HATON,
Jean-François MARI, Dominique FOHR
CRIN/CNRS, Bâtiment LORIA, Campus Scientifique, BP 239
54506 Vandoeuvre-les-Nancy CEDEX

Abstract

The problem addressed by this paper is to enhance the continuous speech recognizers robustness to noise. For this purpose, the acoustic signal is filtered into several spectral bands, and independent recognition is achieved in each band. Then, the system recombines the results given by each recognizer and delivers a unique solution. The main advantage of this method is to consider the signal only in the bands which are relevant, and to ignore spectral bands which are corrupted by noise. We are developping a speaker-independent continuous speech recognizer based on this principle.

I. Basic Approach

A. Principle of Multi-Band Systems

The main principle of multi-band systems can be easily represented by the following diagram:

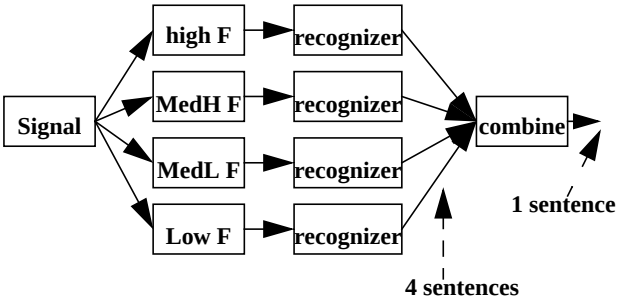

Figure 1: Main Principle of Multi-Band Systems

The acoustic signal is processed through a filterbank which decomposes it into frequency sub-bands (e.g. 4). The choice of the number of bands and of their frequency limits has yet been discussed in [3][4] and we will concentrate rather on the problem of the recombining method.

B. What are the advantages of this system?

Fletcher has worked in the 50's on the way that humans hear meaningless CVC (consonant-vowel-consonant). He concluded that humans decompose the spectral domain into such frequency bands. From his experiments, Fletcher extracted the following formula which shows that an optimal recombination of the sub-band recognition is done in our auditory system.

$$P(E) = \prod_i P(E_i) \qquad (1)$$

$P(E)$ represents the global error-probability, and $P(E_i)$ represents the partial error-probability in the sub-band i. Allen has reactualized his work in [1]. It is of course utopian to hope obtaining such a perfect recombination method, but we are convinced that it is generally interesting to better know the human way of hearing.

Another reason to use such a system is to obtain a better robustness to noise and reverberation. In fact, noise is practically never a phenomenon which corrupts the whole spectrum, but only a limited region of it, and similarly, reverberation differently affects low or high frequencies.

Finally, one could argue that each band contains less information than the whole spectrum. This is true, but we have experimentally shown (see II-*C*) that the *amount of information is more than additive* when using several bands. It means that the total amount of information (in term of phonetic recognition) is superior to the information contained in the Full-Band system. Moreover, a new kind of information is available, i.e., *the combination of phonemes*, which stems from the fact that several recognizers propose several phonemes at the same time.

II. Study of the sub-recognizers

A. Sub-recognizer definition

Each sub-recognizer is a second-order HMM, with three states and a mixture of Gaussian-density-estimators in each state. The Full-Band acoustic vector is composed of 12 MFCC + Δ + $\Delta\Delta$ coefficients, and each sub-band acoustic vector is composed of 6 MFCC + Δ + $\Delta\Delta$ coefficients. For a more detailed presentation of the recognizers and the way they are trained, see [5].

Four frequency bands are used; each of them roughly encompasses one formant. Their limits are [0,901], [797,1661], [1493,2547] and [2298,4000] Hz.

The training corpus is composed of the BREF-80 database [6] and the test corpus is the development database of Aupelf-Uref.

B. Comparison of the sub-recognizers

The first difficulty we had to deal with is the fact that the probabilities returned by the sub-recognizers are not comparable across sub-bands, since the models are trained with the ML-citerion. That means that, for the same piece of signal (before filtering), if recognizer-1 proposes /e/ with probability p_1, and recognizer-2 proposes /i/ with probability p_2, and if $p_1 > p_2$, it does *not* fol-

low that /e/ is the good solution. We can visualize this phenomenon on figure 2. The x-axis represents the log-probability returned by the lowest-frequency recognizer of model /i/ when /i/ and /e/ are pronounced and the y-axis represents the log-probability returned by the medium-low-frequency recognizer of model /e/. The two types of points appear homogeneously in the space: a classifier cannot learn to distinguish between them, simply because the probabilities returned by the recognizers are not comparable.

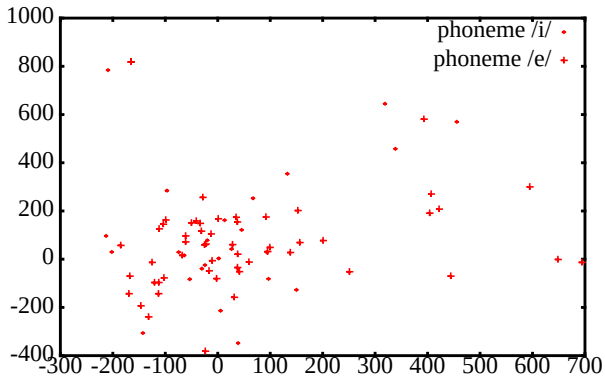

Figure 2: log-probability for /i/ band 0 (x-axis), and /e/ band 1 (y-axis), when /i/ and /e/ are pronounced

Two different ways exist to cope with this problem. The first one is to ignore the probabilities returned by the sub-recognizers and to give a result based only on the labels of the phonemes. The second one is to use a *discriminative training* of the models, in order to make the probabilities comparable. In this paper, we will deal with the former method, using only labels of the phonemes.

C. Potentiality of the Sub-band Methods

The sum of information among the bands is greater than the information of the Full-Band system. We have tried to demonstrate that by building the "best" sentence achievable using only the answers of the sub-recognizers. This sentence is actually created by using all the information given by each sub-recognizer. The results are presented in table 1.

Table 1: Potential accuracy of the 4-band system (in %)

L. F.	ML. F.	MH. F.	H. F.	4-band	FB
45.1	37.5	38.6	38.8	82.4	73.4

These results are not these of our system, but we have computed them only to demonstrate that the problem of recombining the recognizers is worth working on it. The main issue is now to find all the information given by the sub-recognizers, and to eliminate their "bad" answers.

III. The Recombination Method

A. Basic Approach

The major difficulty in a multi-band system is that the different lists of phonemes proposed by the sub-recognizers do not have the same number of phonemes, and the temporal limits of the phonemes are quite different. It is thus very difficult to decide at which time the recombination must be done.

The first solution of this problem is to combine the bands frame by frame, and to construct a new sequence of frames, each one depending on which models the 4 recognizers were at the same time. Then, an algorithm groups this succession of frames into a list of phonemes [2]. The second approach is to use synchrony-points where all the recognizers are due to arrive together. The recombination can then be done in these points [3][4].

We did not want to impose synchrony constraints to our system, because such points prevent each sub-recognizer from finding its optimal path. We propose a new way of recombining the four answers. It consists of grouping the phonemes and extracting a single phoneme out of each group. Each group is composed of one or zero phoneme per band, and carry the propositions of all sub-recognizers which are likely to represent the same pronounced phoneme. In a group, a band can be "empty", because its sub-recognizer may not have "seen" the corresponding phoneme. These groups can thus be composed of 1,2,3 or 4 phonemes, and we respectively call them *k*-group, $k \in \{1, 2, 3, 4\}$.

B. Grouping algorithm

The algorithm which groups the phonemes is an algorithm which finds the longest-path in a graph. Each vertex of the graph is characterized by the 4 current phonemes proposed by the sub-recognizers, and the transition from one vertex to another corresponds to the creation of a *k*-group. Some heuristics are used to speed up the computation time.

The "cost" of a transition corresponds to a score which computes the similarity of the phonemes in the *k*-group created. This score is in fact a mixture of two factors: the first one computes the similarity of the time-segmentation of the phonemes, and the second one their "phonetic" similarity (i.e. the probability that they could have been proposed by different recognizers at the same time). The final formula is given below, and the beginning of the graph which will build the *k*-groups in figure 4 is presented in figure 3.

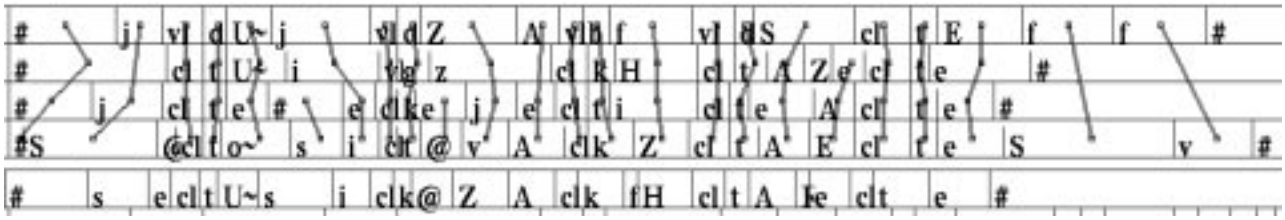

Figure 4: example of the *k*-groups formed by the grouping algorithm

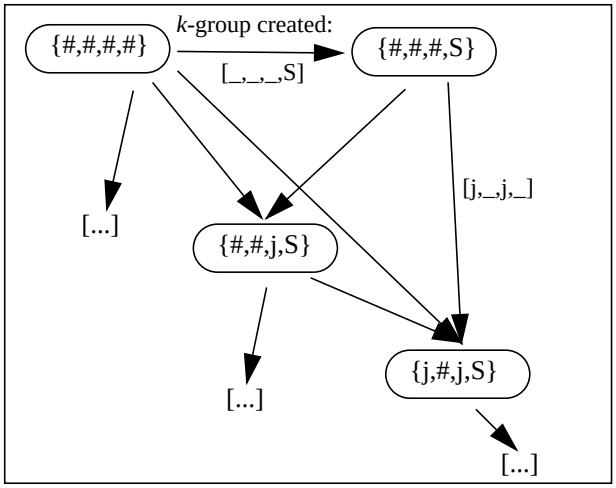

Figure 3: Example of the graph created

$$\frac{\text{intersection}(\text{ph}_{n(1)}, \text{ph}_{n(2)}, \ldots, \text{ph}_{n(k)})^{k}}{\prod_{i=1}^{k} \text{length}(\text{ph}_{n(i)})} \times \sum_{i=1}^{4} \prod_{j=1}^{4} P(\text{ph}_j | \text{ph}_i, \text{band}(j))$$

The cost-function in the longest-path algorithm

In this function, $P(\text{ph}_j|\text{ph}_i, \text{bande}(j))$ represents the probability that the jth band has recognized the model ph_j when the phoneme ph_i has been pronounced, and $intersection(ph_{n(1)}, ph_{n(2)}, \ldots, ph_{n(k)})$ the length of the sequence of frames which are common to all the phonemes of the *k*-group. In this formula, ph_j represents the phoneme proposed by the band j, and the list $(\text{ph}_{n(j)})$ for $1 \le j \le k$ represents all the phonemes accepted in a *k*-group, except the empty-bands.

Figure 4 illustrates the results of the grouping-algorithm on an example. The bottom line is the pronounced sentence, and the four top-lines are the propositions of each sub-recognizer.

C. Building the final answer

There are obviously more groups than phonemes in utterance. That is the reason why we must eliminate some groups, corresponding to phonemes which have been inserted by a sub-recognizer. We have made the assumption that only 1-groups may correspond to inserted phonemes. Actually, *k*-groups with $k \ge 2$ are characterized by the fact that more than two sub-recognizers have proposed the same (or nearly the same) phoneme. Therefore, we guess that they are not inserted phonemes, which is confirmed by experimental studies.

In order to eliminate some of the 1-groups, we compare the likelihood of their phoneme with a pre-defined threshold. If it is greater than the threshold, the 1-group is accepted in the final solution. Otherwise, it is deleted.

All *k*-groups with $k \ge 2$ return one phoneme which is inserted in the final solution. The problem which consists of associating one phoneme to a *k*-group is a classification problem, and we tested different kind of classifiers to solve it.

D. Classifier training

The learning corpus for the classifiers is built as follows: first, the 4 sub-recognizers are used to recognize each sentence of the training database. Their responses are then passed to the grouping algorithm which builds the *k*-groups. Each *k*-group is then associated to a corresponding pronounced phoneme.

Several ways exist to realize this association. Until now, we have used the simplest one: it consists in finding a pronounced phoneme which appears in the *k*-group and in the range of time imposed by the *k*-group (it means that at least one sub-recognizer has recognized it).

If this phoneme exists, it is associated with the current *k*-group, and this association is learnt by the classifier. If no phoneme that respects the two constraints can be found, the current *k*-group is not learnt by the classifier. We could have tried to associate a phoneme to the current *k*-group anyway, even if no sub-recognizer had found the correct answer, but we wanted the classifier to select the good phoneme out of the 4 lists of phonemes and not to build the good answer from the lists. It is just another philosophy, more complex, that will be studied later.

The test database is composed of 1050 sentences pronounced by 20 French speakers. With this kind of training, the classifier gives the results in table 2.

Table 2: % of *k*-groups recognized by the classifiers

2-groups	3-groups	4-groups
49 %	61 %	81 %

These results are not sufficient to obtain correct accuracy concerning the final phoneme recognition system, but we are working on the classifiers to increase their recognition rate (see IV).

E. Description of several classifiers

Until now, we have tested two different classifiers which give approximately the same results. The first one is a classifier based on the majority vote: it builds a dictionary of the *k*-groups of the training corpus, and in the recognition phase, it returns the majority phoneme proposed by all the closest *k*-groups of the dictionary (using the Hamming distance). The second one makes use of scope classification [7]. Scope classification is an instance-based learning algorithm based on a rule-based classifiers semantics instead of Hamming distance.

The recognition rates of the classifiers are crucial. The sentences are approximately decomposed into 60 % of 4-groups, 14 % of 3-groups, 11 % of 2-groups and 15 % of 1-groups. This decomposition, associated with the preceding results, gives a theoretical final accuracy of 62.5 %, with the hypothesis that no 1-groups are recognized, and that the number of proposed phonemes is equal to the number of pronounced phonemes. Of course, 62 % of final accuracy is not enough, and we know that *k*-groups have enough information to reach higher scores. We are now working on the classifiers to enhance them.

IV. Perspectives

We are studying two training algorithms to achieve the classification task. The first one is based on the general principle: to reach higher scores, training and testing procedures should be the same. The idea is to associate each *k*-group with one phoneme by dynamic programming, as it is done to compute accuracy. The second idea is to create *k*-groups on the 4 bands *plus* the pronounced list of phonemes, ($1 \le k \le 5$) in order to associate a phoneme to a *k*-group using the same criterion that the one used to build *k*-groups. These enhancements of the training-procedure of the classifier should improve the recognition rate of the classifier, which might allow the system to finally reach the desired accuracy.

Another important objective of our future work is to formally define, and then to implement, the discriminative training algorithm for the models. The difference with existing discriminative algorithms is, on one hand, that it applies on second-order HMMs, and on the other hand, that each model should be discriminative with all the other models of all bands, except the models which represent the same phoneme.

V. Conclusion

We have proposed in this paper a recombination method for a multi-band phoneme recognizer. We consider that our method presents several advantages:

1- It does not use any synchrony constraint between the sub-bands.

2- Our main goal is to extract the good phonemes out of the 4 lists of solutions proposed by the sub-recognizers. Rather than selecting the good phonemes, it is easier to eliminate first the "redundant" phonemes between the sub-bands, and then the bad-recognized phonemes. The former is done by grouping the phonemes which are quite similar, and the latter is done by rejecting the 1-groups which are the most often bad-recognized.

This is the basic principles of our system, but it is still at its early stage of development and it needs further investigation to reach its full potentiality.

VI. References

[1] **Jont B. Allen**: *How Do Humans Process and Recognize Speech ?* In IEEE Trans. on speech and audio processing, vol. 2, No 4, October 1994.

[2] **Paul Duchnowsky**: *A New Structure for Automatic Speech Recognition.* PhD. Thesis of the Massachussetts Institute of Technology, September 1993.

[3] **H. Bourlard, S. Dupont**: *A New ASR Approach based on independent processing and recombination of partial frequency bands.* In International Conference on Spoken Language Processing, ICSLP 96 Philadelphia.

[4] **H. Hermansky, S. Tibrewala, M. Pavel**: *Towards ASR On Partially Corrupted Speech.* In International Conference on Spoken Language Processing, ICSLP 96 Philadelphia.

[5] **J.-F. Mari, J.-P. Haton, A. Kriouile**: *Automatic Word Recognition Based on Second-Order Hidden Markov Models.* In IEEE Trans. on speech and audio processing, vol. 5, No 1, January 1997.

[6] **J.L. Gauvain, L.F. Lamel, M. Eskénazi**: *BREF, a Large Vocabulary Spoken Corpus for French*. In Eurospeech 1991, pp 505-508, Genova, Italy.

[7] **N. Lachiche, P. Marquis**: *Rule-based classification without rules thanks to Instance-Based Learning.* In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence Posters, 1997.

[8] **H. Bourlard, S. Dupont**: *Subband-based Speech Recognition.* In ICASSP 97, Munich.