A RECOMBINATION MODEL FOR MULTI-BAND SPEECH RECOGNITION

C. Cerisara, J.-P. Haton, J.-F. Mari and D. Fohr
LORIA
BP 239, 54506 VANDOEUVRE-les-NANCY Cedex
FRANCE
email: {cerisara, jph, jfmari, fohr}@loria.fr

ABSTRACT

In this paper, we describe a continuous speech recognition system that uses the multi-band paradigm. This principle is based on the recombination of several independent sub-recognizers, each one assigned to a specific frequency band. The major issue of such systems consists of deciding at which time the recombination must be done. Our algorithm lets each band totally independent from the others, and uses the different solutions to resegment the initial sentence. Finally, the bands are synchronously merged together, according to this new segmentation. The whole system is too complex to be entirely described here, and, in this paper, we will concentrate on the synchronous recombination part, which is achieved by a classifier. The system has been tested in clean and noisy environments, and proved to be especially robust to noise.

1. INTRODUCTION

In multi-band continuous speech recognition, all sub-recognizers[1] can be either totally independent or synchronous at some temporal-unit level. In the first case, the number of phonemes in one sentence and their temporal boundaries differ from one band to another, as illustrated in figure 1. Nevertheless, in order to recombine the bands, the probabilities returned by the sub-recognizers are needed for the same segment of time. A solution to this problem was proposed in [1], in which the sub-recognizers are forced to synchronize when the recombination has to be done. The particularity of our system is that we don't synchronize the sub-recognizers during the Viterbi decoding phase, in order to obtain the optimal solution for each band. Then, the initial sentence is resegmented according to all asynchronous partial solutions. Once this is done, the probabilities returned by the sub-recognizers for each new segment are used by a classifier to recombine synchronously the bands.

[1] In the following sections, the term "sub-recognizer" refers to a HMM-based continuous recognizer applied to a limited frequency band.

This paper is organized in three parts: The first one briefly describes the resegmentation process, the second one analyzes the different types of classifiers used for recombination, and the last one presents the results on a continuous french database.

2. THE RESEGMENTATION PROCESS

It is divided in two parts: the grouping algorithm, which groups the phonemes of different bands together, according to their phonetic similarities, and the resegmentation algorithm, which assigns to each group a new segmentation. The latter also decides which groups will appear in the final solution and which groups will not.

2.1. The Grouping Algorithm

Let assume that all the sub-recognizers are independent, and let call $S_1,\ldots,S_N$ the solutions proposed by each of them. Although these solutions seem to be very different, each of them is an approximation of the input sentence. Thus, a phone in S_i is often temporally and phonetically "close" to another phone in S_j ($i \neq j$). The goal of the resegmentation algorithm is to find the similarities between bands and to group the corresponding phones. The algorithm uses a best-path search through the graph of all possible associations between the phonemes of different bands. The distance used to compare two paths is based on temporal and phonetical similarities between the phonemes in the same group. An example of the solution given by the algorithm is presented in figure 1. This algorithm is fully explained in [3], and the underlying theoretical developments can be found in [2].

2.2. The Resegmentation Algorithm

Once the recognized phonemes in different bands are grouped together, it can be noticed that there are more groups than actually pronounced phonemes. This is due to the fact that sub-recognizers insert phonemes in their solutions. The groups which are likely to have been inserted must then be deleted.

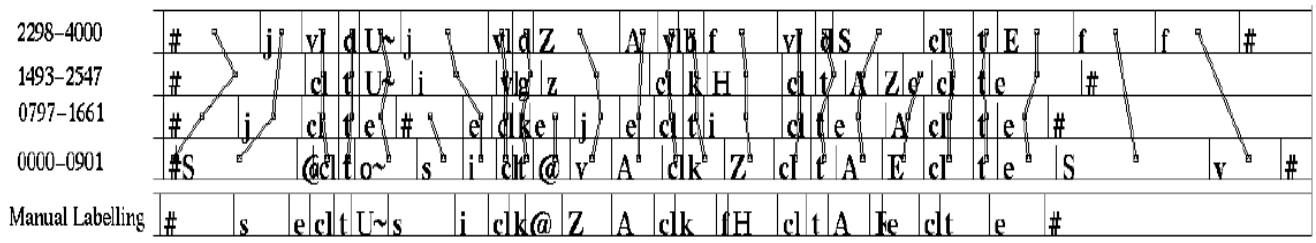

Figure 1. Example of the solutions proposed by the sub-recognizers and result of the resegmentation algorithm for the sentence: "C'est ainsi que Jacques fut arrêté"

Several methods have been tested to eliminate these group; we will only present the best one in section 4. In each remaining group, a unique segmentation is computed. This segmentation is needed by the classifier used in the recombination part. Actually, it is not possible to recombine the probabilities returned by the sub-recognizers for different segments of speech, and a unique segmentation has so been computed for each group. For the time being, we decided to associate to each group the segmentation of the most likely band presents in the group. This probability corresponds to the probability that the model M proposed by the band *i* is correct ($P(M|\text{band } i)$), and is computed from the confusion matrix of each sub-recognizer.

3. THE RECOMBINATION PART

As pointed out above, the recombination uses synchronous information from the sub-recognizers. It is achieved by means of a classifier, whose inputs are the normalized probabilities returned by the sub-recognizers for the current segment of speech and for every phone model. A supervised training of the classifier is done using manual labeled speech. Two kind of classifiers have been tested: a linear classifier with discriminative training of its weights, and a Multi-Layer Perceptron (MLP).

3.1. Linear Classifiers

Linear classification consists of computing for each phone model a score $S(M,X)$, which is the linear combination of the probabilities of the sub-recognizers for this model. For each phone,

$$S(M,X) = \sum_{i=1}^{N} \alpha_{i,M} P(X|M, \text{band } i)$$

where N is the number of bands, $P(X|M, \text{band } i)$ is the normalized probability returned by the sub-recognizer i for the model M and the acoustic vectors X, and $\alpha_{i,M}$ is a coefficient to estimate. The model with the maximum score is chosen to represent the current group.

Three kinds of linear classifiers have been tested. In the first one, training of coefficients $\alpha_{i,M}$ is done discriminatively using the Minimum Classification Error (MCE) algorithm. This algorithm attempts to minimize the classification error rate by applying a gradient descent method to the classification error. Actually, it is not directly possible to compute the gradient of the classification error, since it is not continuous. This is why the MCE algorithm replaces the error rate by a continuous cost function. The cost function we used is:

$$l_k(d_k) = \frac{1}{1+e^{-\xi d_k}}$$

where

$$d_k(x) = -g_k(x) + \left[\frac{1}{M-1} \sum_{j,j\neq k} g_j(x)^{\eta} \right]^{1/\eta}$$

represents the misclassification measure, (it intuitively enumerates how likely a class-*k* observation is misclassified as any other class observation), and

$$g_j(x) = S(j,x)$$

is the discriminant measure, i.e. the score of the association of input x to class j. k represents the actual class of input x and ξ and M are parameters of the MCE algorithm. The principle of the algorithm is fully explained in [6] and [7].

In section 4, this classifier is refered as the MCE-linear classifier. The *mean linear* and *the reduced linear* classifier refers to the two other linear classifiers. The former is of the same type as the linear classifier described before, except that all the coefficients $\alpha_{i,M}$ are set equals (no MCE training of the coefficients is achieved). In fact, it simply computes the mean of the probabilities assigned to a model. The latter is even simpler, as all the coefficients are set equals, and the full-band recognizer is not considered any more: only sub-recognizers are used to classify the segments.

3.2. Multi-Layer Perceptron

We also tested a non linear classifier under the form of a Multi-Layer Perceptron. We used a three-layered perceptron, with 175 inputs (35 phonemes for 4 frequency-limited bands plus 1 full-band for the whole spectrum), 40 neurons in the hidden layer, and 35 outputs, one for each possible phoneme. Classical back-propagation algorithm is used for training.

4. EXPERIMENTS

Two kinds of experiments have been carried out: The first one was done on isolated phoneme recognition, and was aiming at characterazing more carefully the classifier used in the recombination part, whereas the second one was done with the final system on continuous speech.

For all experiments, the database used to train the models was the BREF database [5]. The test database is the development database of Aupelf-Uref [4]. It has been split into two parts: the first one is used to train the classifiers and the second part is used for final tests. The sub-recognizers are second-order Hidden Markov Models [8] of phonemes. 35 phonemes were used to label the databases. The spectrum has been split into four bands, each one roughly encompassing one formant. The acoustic vectors are made up with 12 MFCC + Δ + ΔΔ coefficients for the full-band recognizer and of 6 MFCC + Δ + ΔΔ coefficients for the sub-recognizers.

4.1. Isolated Phoneme Recognition

All these experiments were achieved on isolated phoneme recognition. That means that manual labeling was used to segment the test corpus, and that these segments of speech were used as the input of the sub-recognizers. The probabilities returned by the sub-recognizers were then passed to the classifiers, and the count of phonemes properly recognized by the classifiers was divided by the number of initial segments to compute accuracy. Table 1 shows the results on clean speech of the two classifiers and of the reference system (full-band), in the same conditions of testing. The confidence interval for these results is +/- 0.4 %.

	Accuracy
Full-Band system	78.9 %
Linear Merging	79.5 %
MLP	82.5 %

Table 1. Results of the different systems on clean speech and isolated phoneme recognition

We have also tested these systems in noisy conditions. The noise was a natural noise (recorded in a subway station at pic hours) added to the speech signal, with SNR ranging from 10 dB to –15 dB. Figure 2 gives the accuracy of respectively the full-band system in black, the MCE-linear classifier in white, and the MLP in gray. The recognition rate of the linear classifier is higher than the accuracy of the full-band system of 1.8 % (on average) for each signal-noise ratio (SNR). Similarly, the MLP accuracy is higher than the full-band system's one of 9.4 %. All differences between accuracy rates in figure 2 and 3 are significant.

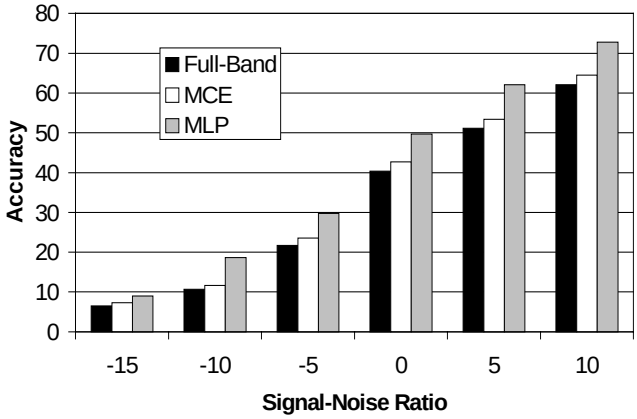

Figure 2. Accuracy of the full-band, MCE-linear and MLP systems in a noisy environment.

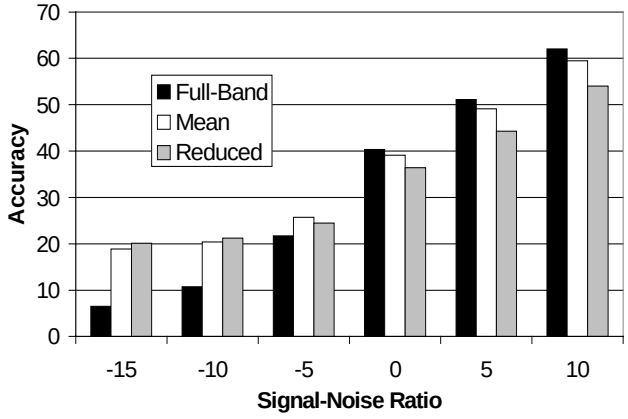

Figure 3. Accuracy of the full-band, mean linear and reduced linear systems in a noisy environment.

We can see in figure 3 that at –15 dB SNR, the reduced linear (in gray) and the mean linear system (in white) are 14.6 % and 12.4 % better than the full-band system (in black) whose accuracy is 6.5 %. This clearly proves that multi-band systems, even with a very simple recombination scheme, are more robust to noise than a full-band system. All theses results are summarized in table 2. Confidence intervals are less than +/- 0.4 %.

	-15 dB	-10 dB	-5 dB	0 dB	5 dB	10 dB
Full-Band	6.5	10.7	21.7	40.3	51.1	62.1
Low-F. band	16.1	16.0	17.5	21.8	28.9	35.7
MediumL-F. band	11.1	11.8	12.8	15.8	19.6	26.0
MediumH-F. band	14.4	15.4	16.7	22.2	27.1	34.3
High-F. band	17.9	19.4	22.5	30.3	32.4	37.7
MCE-linear	7.3	11.6	23.6	42.7	53.4	64.5
Mean-linear	18.9	20.4	25.7	39.1	49.1	59.5
Reduced-linear	20.1	21.2	24.5	36.4	44.3	54.0
MLP	9.1	18.6	29.8	49.7	62.1	72.8

Table 2. Results (in %) of all the systems in noisy speech

It is worth noticing that the MLP and MCE-linear classifiers are not as good as the reduced linear classifier in very noisy environments. This is due to the fact that they have been trained with clean speech, and training them with noisy speech would certainly increase their recognition rate.

4.2. Continuous Speech Recognition

In these experiments, inserted groups have to be deleted. For the time being, best results are obtained by removing the groups whose score (returned by the resegmentation part) is lower than a threshold and the groups which have not been recognized by the full-band. The classifiers described before are then applied on the remaining groups.

First results on clean and noisy speech are presented in table 3. The confidence interval is less than +/- 0.4 %.

	Accuracy
Full-Band system (clean speech)	71.7 %
MCE-Linear Merging (clean speech)	72.1 %
Full-Band system (-5 dB)	22.0 %
MCE-Linear Merging (-5 dB)	23.0 %
Mean-Linear Merging (-5 dB)	28.1 %
Reduced-Linear Merging (-5 dB)	25.9 %

Table 3. Results of the final system on continuous speech recognition, in clean and noisy speech

As can be seen in table 3, there is no significant difference between the two systems on clean speech. However, the classifiers used in this part have not yet been finaly tuned, and only the linear merging has been tested. We can reasonably hope to get better results as soon as the classifiers are correctly integrated in the Viterbi algorithm.

5. CONCLUSION

We have presented in this paper a continuous speech recognition system using the multi-band principle. The different classifiers have given best results than a full-band system in clean speech and in noisy speech, but they have proved to be especially efficient in noisy environments. The integration of these classifiers in the final continuous speech recognizer is underway.

A major conclusion of this study is that the classical HMM recognizer is not adapted to noisy speech. Some results have even proved that each sub-recognizer, taken individually, has an accuracy at least two times superior to the full-band one, for a SNR of -15dB.

Finally, each type of recognizer tested in this study appears to be efficient in a given environment, but not as good in another one. It would thus be interesting to combine several classifiers under the control of a system sensitive to the environment, for example a SNR estimator.

6. REFERENCES

[1] H. Bourlard and S. Dupont. Subband-based speech recognition. In *Proc. ICASSP '97*, Munich, Germany, 1997.

[2] C. Cerisara. Dealing with Loss of Synchronism in Multi-Band Continuous Speech Recognition. In *Computational Models of Speech Pattern Processing*, Keith M. Ponting ed., NATO ASI Series F, Springer Verlag, 1997 (to be published).

[3] C. Cerisara, J.-P. Haton, J.-F. Mari, and D. Fohr. Multi-band continuous speech recognition. In *Proc. EUROSPEECH '97*, 3(1), pages 1235-1238 Rhodes, Greece, 1997.

[4] D. Fohr, J.-P. Haton, J.-F. Mari, K. Smaïli, and I. Zitouni. MAUD: Un prototype de machine à dicter vocale. In *1ère JST 1997 FRANCIL de l'AUPELF-UREF*, pages 25-30, 15-16 april 1997, Avignon, France.

[5] J.-L. Gauvain, L.-F. Lamel, and M. Eskénazi. BREF, a large vocabulary spoken corpus for French. In *Proc. EUROSPEECH '91*, pages 505-508, Genova, Italy, 1991.

[6] B.-H. Juang and S. Katagiri. Discriminative Learning for Minimum Error Classification. In *IEEE Trans. on Signal Processing*, 40(12), pages 3043-3054, December 1992.

[7] B.-H. Juang, W Chou, and C.-H. Lee. Minimum Classification Error Rate Methods for Speech Recognition. In *IEEE Trans. on Speech and Audio Processing*, 5(3), pages 257-265, May 1997.

[8] J.-F. Mari, J.-P. Haton, and A. Kriouile. Automatic word recognition based on second-order hidden Markov models. In *IEEE trans. on Speech and Audio Processing*, 5(1), January 1997.