Robust behavior of multi-band paradigm

Christophe Cerisara, Dominique Fohr, Jean-Paul Haton

ROBUST BEHAVIOR OF MULTI-BAND PARADIGM

Christophe Cerisara, Dominique Fohr, Jean-Paul Haton

LORIA / INRIA
Nancy, FRANCE
Email: {cerisara, fohr, jph}@loria.fr

ABSTRACT

In this paper, we have gathered new results obtained with our Multi-Band system on the TIMIT database corrupted by additive noise. The robustness of the Multi-Band system to several kinds of noise is studied in details. The system is robust to spectral-limited noise, as it has already been shown in previous papers, but it is also robust to broad-spectral noise as long as the full-band is added as an additionnal stream to the Multi-Band system. A more robust version of this system is then obtained when training the recombination module in white noise. Thus, robustness to other kinds of noise is increased. Finally, we present the first results obtained with a new training algorithm which globally optimizes the Multi-Band system. Regarding the robustness of the method, these preliminary results are encouraging.

1. INTRODUCTION

When a noise affects only a limited region of the spectrum, the Multi-Band paradigm is beneficial, as it exploits the fact that some frequency bands are not corrupted. However, the recombination module has to decide of the importance of each band. Thus, the system is robust to noise, if and only if the recombination module correctly differientiates between a clean band and a corrupted band. One solution to achieve this is to indicate to the recombination module which band is noisy and which one is not [Berth98]. But this requires the presence of a complex pre-processing step, whose role is to estimate the signal-noise ratio in each band. For the moment, we have prefered to study more carefully the Multi-Band paradigm itself. Thus, we have trained the recombination module in noisy conditions. The choice of the noise added to the training corpus is not straightforward, as the system may develop a dependence on this kind of noise. In addition, one can generally not know what kind of noise the system will have to deal with. It is possible to know it in some specific applications, but this is not the goal of this paper.

This paper is concerned with this important problem of the training of a multi-band system in noise. It is organized as follows. In section 2, we briefly recall the basics of our multi-band system. In section 3 are presented two series of experiments: In the first one, the system is trained in clean speech whereas in the second one, it is trained in noisy speech. Finally, in section 4 we introduce a novel idea related to the global training of a Multi-Band recognizer.

2. BRIEF OVERVIEW OF THE MULTI-BAND SYSTEM

All the experiments which are described here have been carried out with the last version of our Multi-Band system. For a more detailed description of the system, please refer to [Ceri98]. We briefly present below an overview of the system, more precisely describing its new characteristics.

At first, the speech signal is filtered into four sub-bands. The limits of these sub-bands are respectively [0 ... 538 Hz], [461 ... 1000 Hz], [923 ... 2823 Hz] and [2374 ... 7983 Hz]. In each sub-band, three MFCC are computed. Δ and $\Delta\Delta$ parameters are added to these three coefficients and c_0 is deleted. Finally, the vectors in each sub-band have 8 coefficients. The full-band, which corresponds to the [0 ... 7983 Hz] range, is similarly represented by 35 coefficients. Our Multi-Band system is composed of five streams, i.e. the four sub-bands, plus the full-band.

Each of these streams is processed through a second-order Hidden Markov Model. The likelihoods which are then returned by these sub-recognizers are passed to the recombination module. Two kinds of recombination module are tested in this paper: The first one linearly recombines the likelihoods of the sub-recognizers. It is composed of 240 coefficients (5 streams times 48 phones) which are trained with the Minimum Classification Error (MCE) algorithm. This algorithm simply applies a gradient descent procedure to an approximation of the classification error. Please refer to [Juan97] for a general description of this algorithm. The second recombination module consists of a Multi-Layer Perceptron. Its goal is to approximate the best recombination function possible, whatever this function is, linear or not.

3. EXPERIMENTS

3.1 Experimental settings

All the tests have been done on the TIMIT database. Training of the HMM is first achieved on the training part of the corpus, and then the recombination module is trained on the same training part of the corpus. The tests are then done on the coretest part of the corpus. The confidence intervals are given in table 1 for all the results.

Accuracy	10 %	20 %	40 %	60 %	80 %	90 %
Confidence Interval	±0.6 %	±0.8 %	±1.0 %	±1.0 %	±0.8 %	±0.6 %

Table 1: Confidence intervals for all the results

The tests are done in *isolated phoneme mode*. This means that we assume that the segmentation of the sentence into phonemes is known. Actually, the segmentation given with the corpus is used. Each segment is then recognized, and accuracy simply represents the percentage of correct phones recognition. 48 phonetic classes are modeled, but accuracy is computed using only 39 phones.

3.2 First Experiment

3.2.1 Description

In this first experiment, the system is trained in clean speech and tested in noisy speech. The noise is generated or recorded independentely of the signal and amplified at different levels before being added to the signal. Several Signal Noise Ratio (SNR) can thus be tested. The SNR values indicated in the following tests have been computed using the power of the signal and of the noise in windows of 64 ms length.

The goal of this first experiment is to study the robustness of our Multi-Band system to filtered noise. In the following, we call *filtered noise* a white noise which has been processed through a pass-band filter whose limits are [2900 ... 5300 Hz]. This noise has been added to each sentence of the test corpus, and recognition is achieved on these sentences.

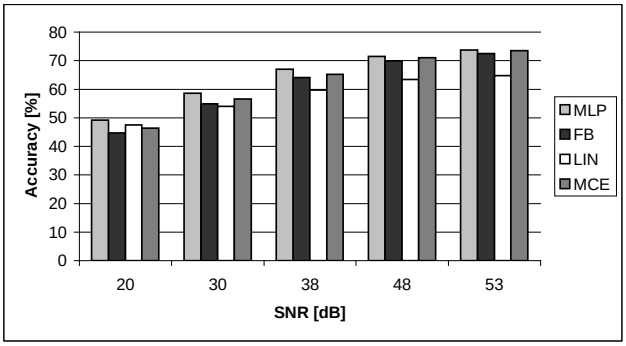

Figure 1: Accuracy of the Multi-Band systems in filtered noise

FB denotes the full-band system, which is also the reference system, and **MLP** denotes the Multi-Band system when the recombination is achieved by a MLP. Two linear systems have also been tested: the first one, called **LIN**, consists of simply averaging the four sub-bands. It is important to note that, with this recombination, no training of the recombination module is done ! Moreover, the full-band is not considered any more. Actually, we have done previous experiments which have suggested that the full band is one of the less robust to noise. We have thus tested this hypothesis with our system on the TIMIT database. The second linear system that we have tested is called **MCE**. It makes use of the five streams and it learns the recombination coefficients with the MCE algorithm. Results are presented in figure 1.

3.2.2 Discussion

The Multi-Band system with the neuronal recombination is the best system in clean speech as well as in noisy speech. However, one can note that the more powerful the noise is, the better the simple average of the four sub-bands is. This confirms the hypothesis which has yet been exposed, i.e. that the full-band is especially sensitive to noise. Table 2 shows it even more clearly.

	B1	***B2***	***B3***	***B4***	***FB***
Clean speech	39 %	37 %	48 %	40 %	73 %
Filtered noise SNR=20 dB	38 %	34 %	47 %	06 %	45 %

Table 2: Accuracy of the five streams at SNR = 20 dB

Since we are dealing with filtered noise, only the fourth band is affected by it. However, the full-band is also seriously affected: the third band alone outperforms it, even with a signal-noise ratio of 20 dB, which is still perfectly understandable by humans.

3.3 Second experiment

3.3.1 Basic idea

As explained in section 1, we have decided to train the recombination module in white noise. At first, one may think that the training of the recombination module in white noise should give the same results as the training in clean speech. Actually, as all the bands are a priori equally affected by the noise, it seems impossible to correctly adjust the contribution of each band. Nevertheless, each phone is differently affected by the noise, depending on which band is considered. For example, the energy which constitutes a formant of a vowel is much more important than the energy produced by the noise in the same frequency band. At the opposite, it might be very difficult to differentiate between a fricative and the surrounding noise in the high frequencies. These are the kinds of behaviours that we are trying to model when training the recombination module in white noise.

3.3.2 Description

The goal of this experiment is to increase the robustness of the Multi-Band system to noise by training the recombination module in noisy speech. The models are still trained in clean speech, and only the recombination

module is trained in noisy speech. In fact, we want to take into account the robustness of each phone in each band, and not to modelize the noisy phones themselves. Doing that would make the models too much dependent of the noise used during training. Whereas by training only the recombination module in white noise, the importance of each phone in each band is modified depending on its robustness to noise, and we think that this kind of training is less dependent to noise.

In order to test this hypothesis, we have trained the MLP-recombination module in white noise with SNR=30 dB, and we have tested the resulting system (which is called **MLP-train**) in three different noises: the filtered noise that we have used in the first experiment, a noise recorded from a hair-dryer and a babble noise extracted from the NOISE-ROM-0 CD. Results are presented in table 3.

	MLP-train	***FB (ref)***	***MLP***
Filtered noise SNR=20 dB	53.23 %	44.82 %	49.19 %
Hair-dryer SNR=5 dB	34.41 %	29.53 %	32.19 %
Bable noise SNR=20 dB	37.42 %	36.06 %	38.45 %

Table 3: Recognition rates of the MLP-trained system in different kinds of noises.

3.3.3 Discussion

When the MLP is trained in white noise, its robustness to other kinds of noise, such as the filtered noise described above, is greatly increased. However, the filtered noise is still somehow similar to white noise. That is why we have also tested these systems in a hair-dryer noise and in babble noise. The MLP-trained recombination is still the best system with the hair-dryer noise, but its performance is worse than the MLP recombination in babble noise. A possible explanation of this fact is that, even if the HMMs are not trained in noise, the MLP becomes a little bit dependent on the noise used during its training. One can assume that the main difficulty of babble noise is that it is a non-stationnary noise, whereas all the other kinds of noise that have been used so far are stationnary. It may be this dynamic characteristic of the noise that the MLP-trained recombination does not correctly recognize.

However, one can note that the Multi-Band system is always clearly better than the full-band, whatever kind of noise is used. This contradicts the conclusions of [Tibr97], saying that the full-band system is better when the noise is affecting all frequencies. But there is one major difference between our systems: in this paper, the full-band is added as a fifth stream to the Multi-Band system, whereas in [Tibr97] it is not. This surely partly explains the observed differences in our results. This use of the sub-bands and of the full-band together has recently been applied to clean speech by [Ceri98], and then [Mirg98].

3.3.4 Analysis of the trained MLP

In this section, we show that, after its training in noise, the MLP give more importance to the sub-bands than to the full-band. This is in fact the underlying hypothesis which has motivated our experiment. To test this hypothesis, we have presented artificial patterns to the MLP, before and after its training. Each of these patterns corresponds to the activation of one model in one band (i.e., the input of one model in one band is set to 1 whereas the other inputs are set to 0). The outputs of the MLP give then the influence of the considered model and band.

For each model *i* of the first band, we have saved these outputs in B(*i*) and A(*i*), respectively before and after the training of the MLP. Figure 2 reports the differences A(*i*) – B(*i*). Similarly, figure 3 reports the same differences for the full-band.

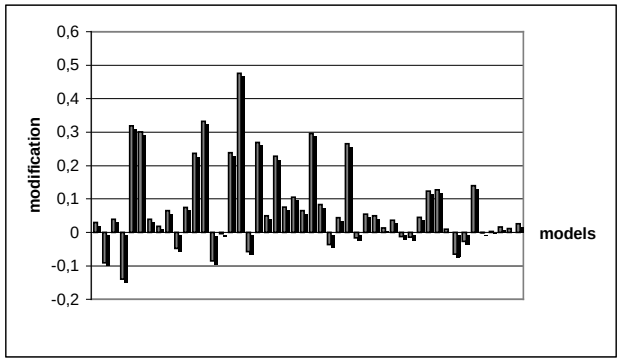

Figure 2: Modification of the influence of the first band after the MLP training.

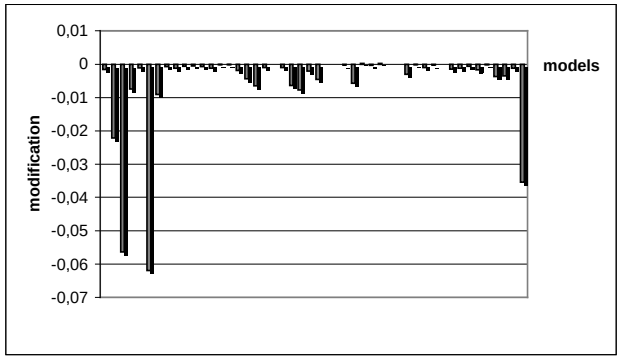

Figure 3: Modification of the influence of the full-band after the MLP training.

The difference is quite clear, since nearly all the modifications concerning the first band (which are similar to the modifications in the other sub-bands) are positive, whereas all the modifications of the full-band are negative. This means that, when the MLP is trained in noisy conditions, the influence of the full-band is systematically reduced, whereas the influence of the other bands increases.

A more detailed study of the modifications created by this learning may help us to understand the behavior of each phoneme in noisy conditions, and eventually may lead to modify the training procedure, so that "difficult" noise can also be treated that way. This might constitute an interesting future work for Multi-Band researchers.

4. GLOBAL TRAINING

4.1 Principle

More recently, we have developped a new algorithm whose goal is to globally optimize the Multi-Band system. All the Multi-Band systems that we are aware of use the same two-steps training procedure: First the sub-recognizers are trained, and then the recombination module is trained. It is clearly not optimal. We have developped a new algorithm, based on the MCE criterion, which trains the Multi-Band system in a single step. The main procedure consists of the classical gradient descent of the final error. Parameters of the recombination module are thus directly adjusted during training, and the parameters of the HMMs are simultaneously adjusted using the Corrective Training algorithm developped by Bahl & al. [Bahl88].

This method has yet been succesfully tested in clean speech (the results are on press) and we are now adapting it to noisy speech.

In section 3, we have seen that training the recombination module in noisy speech may help to adapt the Multi-Band system to the environment. A natural extension of this consists of training globally the whole system in noisy speech. As pointed out before, the drawback of this method may be that the system becomes more dependent on the kind of noise that has been added to the training corpus. However, if the results are as satisfying as in clean speech, this system could certainly be used in application-specific tasks.

4.2 Experiments

When global training is applied to clean speech, the system becomes adapted to a clean environment and does not perform quite well in noisy speech. Table 4 reports the robustness of globally trained Multi-Band systems, when they are trained in clean speech. The noise which is added to the test corpus is the filtered noise. The SNR is 20 dB.

	Global Trainining	***FB (ref)***	***MLP***
Filtered noise SNR=20 dB	48.72 %	44.82 %	49.19 %

Table 4: Performance of the globally trained Multi-Band system in filtered noise (SNR=20 dB)

One can note that the globally trained MB-system does still better than the full-band one, but it is also slightly worse than when the two modules of the system have been trained separately. This may be due to the hypothesis formulated above, but these preliminary results must still be confirmed with other experiments.

Experiments concerning the global training in noisy speech are in progress, and results will be available at the time of the workshop.

5. CONCLUSIONS

The work presented in this paper completes the results presented in [Tibr97] on the robustness of Multi-Band to additive noise. The main differences are the recombination level, the use of the full-band stream, the architecture of the system and the database used. We have also attempted to analyze the internal behavior of our Multi-Band system in noisy conditions and to understand where does its robustness come from.

Then, in order to increase the robustness of the system to noise, the recombination module has been trained in white noise. The resulting system has shown to be still robust to other kinds of noise, as long as they were not too much different from white noise.

Finaly, we have also proposed a novel method of global training of a multi-band system in order order to increase its robustness. Experiments are in progress on this point.

6. REFERENCES

[Bahl88] **L. R. Bahl, P. F. Brown, P.V. de Souza and R. L. Mercier**. *A New Algorithm for the Estimation of Hidden Markov Model Parameters*. In Proc. ICASSP'88, April 1988.

[Bert98] **F. Berthommier, H. Glotin, E. Tessier and H. Bourlard.** Interfacing of CASA and partial recognition based on a multistream technique. In Proc. ICSLP'98, Sydney, Australia, 1998.

[Ceri98] **C. Cerisara, D. Fohr and J.-P. Haton**. *A Recombination Model for Multi-Band Speech Recognition*. In Proc. ICASSP'98, Seattle, USA, May 1998.

[Juan97] **B.-H. Juand, W. Chou and C.-H. Lee.** Minimum Classification Error Rate Methods for Speech Recognition. In IEEE Trans on Speech and Audio Processing, 5 (3), pp. 257-265, May 1997.

[Mirg98] **N. Mirghafori and N. Morgan**. *Combining connectionnist Multi-Band and Full-Band Probability Streams for Speech Recognition of natural numbers*. In Proc. ICSLP'98, Sydney, Australia, December 1998.

[Tibr97] **S. Tibrewala and H. Hermansky**. *Sub-band based recognition of noisy speech*. In Proc. ICASSP'97, Münich, Germany, 1997.