FILTER BANK DESIGN FOR SPEAKER DIARIZATION BASED ON GENETIC ALGORITHMS

C. Charbuillet, B. Gas, M. Chetouani, J. L. Zarader

Laboratoire des Instruments et Systèmes d'Ile-de-France
Université Pierre et Marie Curie, Paris, FRANCE

ABSTRACT

Speech recognition systems usually need a feature extraction stage aiming at obtaining the best signal representation. In this article we propose to use genetic algorithms to design a feature extraction method adapted to the speaker diarization task.
We present an adaptation of the common MFCC feature extractor which consists in designing a filter bank, with optimized bandwidths.
Experiments are carried out using a state-of-the-art speaker diarization system. The proposed method outperforms the original filter bank based on the Mel scale one. Furthermore, the obtained filter bank reveals the importance of some specific spectral information for speaker recognition.

1. INTRODUCTION

Speech feature extraction plays a major role in speaker recognition systems. Current speech feature extraction methods are based on speech production for the Linear Predictive Coding (LPC) or perception for the Mel Frequency Cepstral Coefficients (MFCC). However, traditional methods in speech feature extraction do not take into consideration the specific information about the task to achieve. To overcome this drawback, several approaches have been proposed by Katagiri [1], Torkkola [2], and Chetouani [3]. The main idea of these methods is to simultaneously learn the parameters of both the feature extractor and the classifier. This procedure consists in the optimization of a criterion, which can be the Maximization of the Mutual Information (MMI) or the Minimization of the Classification Error (MCE).

This paper presents a new framework for the optimization of the filter bank of a MFCC based feature extractor. The MFCC feature extractor is considered as a standard for most of the tasks such as speech and speaker recognition or language identification.

The MFCC feature extraction process mainly consists in the modification of the short-term spectrum by a filter bank. The central frequencies and the bandwidths are inspired from the human auditory system [4]. The used non-linear scale is the Mel. The following figure (Fig.1.) presents the Mel triangular filter bank.

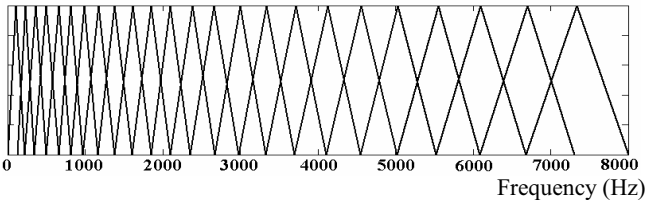

Fig.1. The Mel scaled triangular filter bank

We propose, in this paper, to design a filter bank for the speaker diarization task. For this purpose, we use genetic algorithms.

Genetic algorithms (GA) were first proposed by Holland in 1975 [5] and became widely used in various disciplines as a new mean of complex systems optimization. GAs most attractive quality is certainly their aptitude to avoid local minima. However, our study relies on another quality which is the fact that GAs are an unsupervised optimization method. So they can be used as an exploration tool, free to find the best solution without any constraint.

In the first part, the speaker diarization task is presented and the clustering system is described. Afterwards, we present the genetic algorithm we used, followed by the experiment we made and the obtained results. In the end, we will present an analysis of the obtained filter bank on which some spectrum properties for speaker discrimination will be showed.

2. SPEAKER DIARIZATION

Speaker diarization is composed of two successive phases: segmentation and clustering. The aim of the segmentation phase is to detect the speaker changes. The clustering phase associates all the segments produced by the same speaker (i.e. the clusters). Speaker diarization is currently gaining importance as a mean of indexing voluminous spoken data accumulated, for archival use [6]. For both segmentation

and clustering, the Bayesian Information Criterion (BIC) is used. Speaker change detection is obtained during an abrupt change of the BIC. This segmentation phase forms the initialization of our adaptation process. Indeed, features are extracted with the MFCC and compared by the BIC.
In the second phase, a hierarchical clustering also based on the BIC allows to merge the segments having the highest BIC difference.
This difference is defined as:

$$dBIC(C_i, C_j) = -D_r(C_i, C_j) + \frac{\lambda}{2}\left(d + \frac{d(d+1)}{2}\right)\log\left(n_i + n_j\right)$$

With:

$$D_r(C_i, C_j) = \frac{n_i + n_j}{2}.\log\left|\Sigma_{ij}\right| - \frac{n_i}{2}.\log\left|\Sigma_i\right| - \frac{n_j}{2}.\log\left|\Sigma_j\right|$$

Where C_i, C_j are two sequences of feature vectors representing two clusters to merge; n_i is the size of Ci and Σi the covariance matrix estimated from the cluster data Ci.; λ is a penalty coefficient usually fixed to 1.5.

The key idea of the proposed work is to optimize the filter bank for a global improvement of this clustering phase.

3. GENETIC ALGORITHM

A genetic algorithm is an optimization method. Its aim is to find the best values of system's parameters in order to maximize its performances. The basic idea is that of "natural selection", i.e. the principle of "the survival of the fittest". A GA operates on a population of systems. In our application, each individual of the population is a filter bank defined by its bandwidth and center frequency parameters.
The algorithm used is called the Selection Evaluation Variation (SEV) [8] and is illustrated in figure 2. To evolve the desired filter bank, we consider a population **p(t)** of **Np** filter banks undergoing a variation-evaluation-selection loop, i.e. **p(t+1) = S E V p(t)**. First, a random initialization is done for each individual of the population **p(0)**. The *variation operator* **V** consists in a random variation of each filter bank's parameters. The *evaluation operator* **E** is defined problem specific, and is usually given in terms of a fitness function. It consists in evaluating the performances of each individual of the population. In our application, the performance of a filter bank is given by the clustering error rate of the entire clustering system using it. After evaluating the performance of each individual filter bank, the *selection operator* **S** selects the **Ns** best individuals. These individuals are then cloned according to the evaluation results to produce the new generation **p(t+1)** of **Np** filter banks. Consequently of this selection process, the average of the performance of the population tends to increase and in our application adapted filter banks tend to emerge.

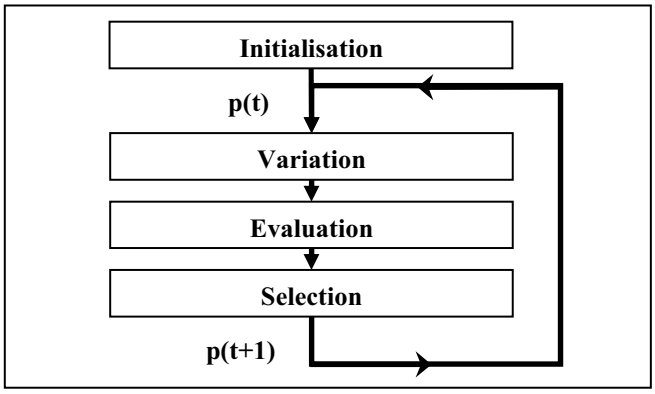

Fig.2. The S.E.V. algorithm

4. EXPERIMENTS AND RESULTS

This section presents the experiments we made and the results we obtained. Our proposed feature extraction method was evaluated on the speaker diarization task of the French ESTER [9] Broadcast News Evaluation.

4.1. Database

The ESTER [9] corpus is composed of 40 hours of audio data recorded from four different French radio broadcast news shows. This corpus is very varied, containing natural speech, telephonic interventions, speech on musical background and all that with different recording qualities. The audio files' length is either of 1 hour or of 20 minutes. The number of speaker involved in each file varies from 1 to 39.
The performance measure used for the speaker diarization task is the official RT 2003 NIST[1] metric. It is based on an optimum "one to one" mapping of reference speaker IDs to system output speaker IDs.

4.2. Evolution data

We used two different databases for the evolution simulation. The first one is called "*evolution base*" and was used for the filter banks' evaluation and selection. It is composed of four hours of representative data. The second one, called "*cross validation base*", was used to evaluate the generalization capability of our algorithm. It is composed of eight hours of radio emission.

4.3. Simulation parameters

In this experiment, we used the **S.E.V.** algorithm to optimize the bandwidth parameters of the filter banks. The overlapping between filters was preserved by using the following rule: "*each filter begins at the middle of the left adjacent one*". Therefore the bandwidth parameters and the previous rule impose the center frequency parameters to be as follows:

[1] http://www.nist.gov/speech/tests/rt/rt2003/spring/

$$C_0 = \frac{B_0}{2} \quad \text{and} \quad C_{i+1} = C_i + \frac{B_{i+1}}{2}$$

C_i and B_i being respectively the center frequency and the bandwidth of the i^{th} filter in the bank.

The bandwidth and center frequency are discrete parameters varying from 0 to 255 and coding the [0 8000] Hz frequency domain. Before starting the **S.E.V.** algorithm, a random initialization of the bandwidths is done using a uniform random distribution of 25 units (~ 780Hz).

For evolving the desired filter bank, the following algorithm parameters have been used:

- *Population size* **Np**: 50
- *Number of selected individuals* **Ns:** 15
- *Variation operator*: uniform random variation of 3 units (~100Hz) for each bandwidth.
- *Evaluation method*: clustering error rate.
- *Number of filters in the bank*: 24
- *Feature vector dimension*: 24

4.4. Evolution simulation

The evolution simulation was done using the previously defined parameters. Figure 3 reports the evolution of the clustering error rate on the evolution and cross validation bases.

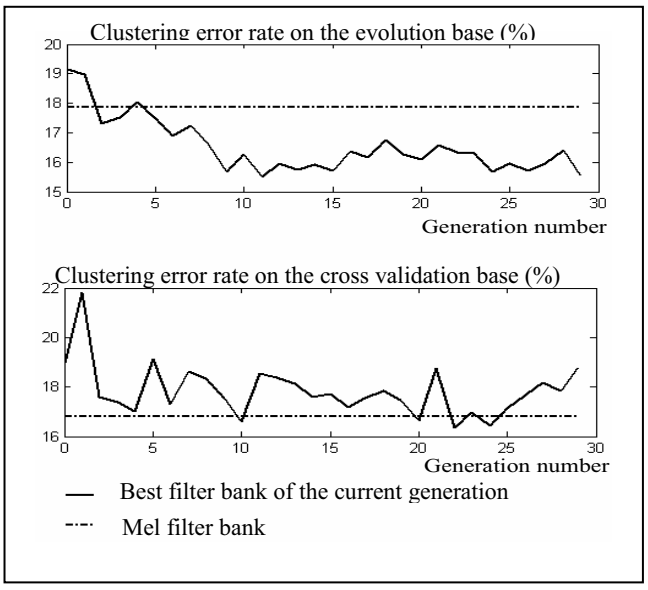

Fig.3. Evolution of the clustering error rate

We can observe that the clustering error rate decreases on the cross validation base until the 24^{th} generation. The degradation of the generalization capacity which follows can be explained as an over adaptation phenomenon. The filter bank obtained in the 24^{th} generation is depicted in Figure 4. We can observe that this filter bank appears very different from the Mel-scaled one (Figure 1). An analysis of this filter bank will be presented further on in the article (see section 7).

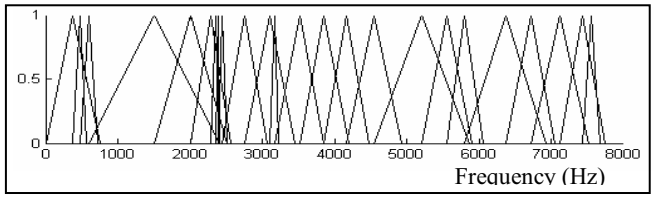

Fig.4. Filter bank obtained at the 24^{th} generation

4.5. Comparative results

The evaluation of the obtained filter bank's capacities was done on two data bases used on the ESTER campaign. The first one, called "*Development Base*" is given to all participants to self evaluate their system. It's composed of eight hours of radio emissions. The second one, called "*Evaluation Base*", is also composed of eight hours of radio emissions and its particularity is the fact that two hours of this base come from two unknown radios. Results on this base are evaluated by an independent organism. The following table (Table 1) reports the clustering error rates obtained from three filter banks: a linear scaled filter bank, a Mel-scaled one and our filter bank obtained with the S.E.V. algorithm.

	Clustering error rate on the Development Base (%)	Clustering error rate on the Evaluation Base (%)
Linear scaled filter bank	17.64	24.11
Mel scaled filter bank	17.80	23.38
Obtained filter bank	**16.98**	**22.74**

Table.1. Comparison of filter bank performances

As we can notice from this table, the obtained filter outperforms the Mel-scaled filter bank and the linear scaled one.

5. FILTER BANK ANALYSIS

In order to interpret the obtained results an additional experiment was made. It consists in repeating the evolution simulation with different initializations of the population **p(0)** and in comparing the obtained filter banks. For all simulations, the initialization of the bandwidths of each filter bank of the population was done using a uniform random distribution of 25 units (~ 780Hz). The simulation parameters and conditions were the same as those described in section 4.3.

Figure 5 presents the filter banks obtained with three different simulations. Their clustering error rates on the

Development Base are respectively **17.54 %**, **17.14 %** and **16.98 %**. They all outperform the Mel-scaled filter bank and the linear scaled one. In order to compare them, Figure 6 presents the bandwidth according to the center frequency of the filters for each filter bank. We can observe similarities between all of the obtained filters. Especially concerning the presence of large filters centered on 1500, 5100, and 6500Hz.

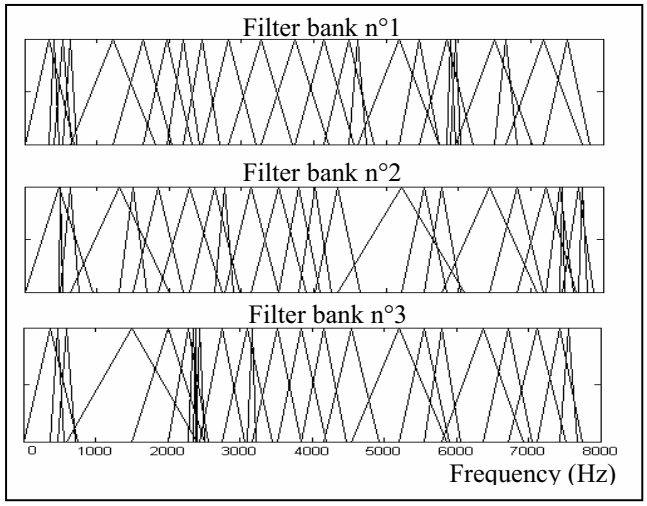

Fig.5. Filter bank obtained with different initial conditions

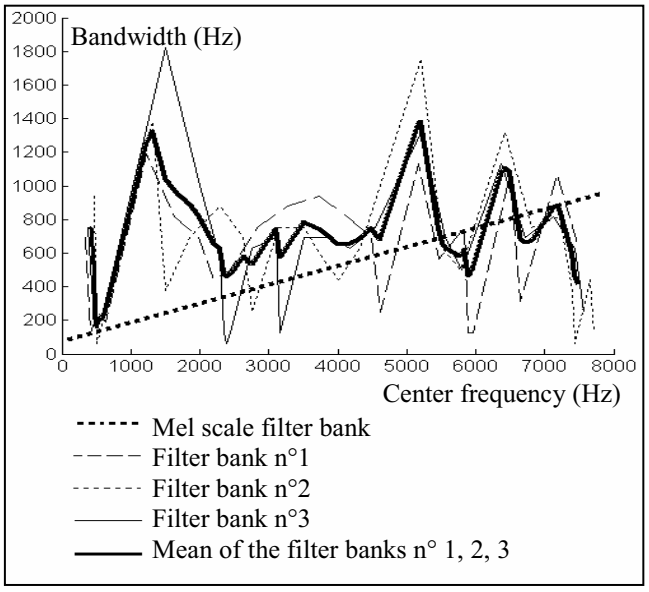

Fig.6. Bandwidth according to center frequency

In addition, the presence of a large filter centered in 1500 Hz makes these filter banks very different from the Mel-scaled one (Figure 1). In fact, MFCCs are more adapted to phoneme discrimination than to speaker discrimination. This task may rely on other speech spectrum properties.

The robustness of the obtained solutions according to the initial conditions allows us to consider that the obtained filter banks are fit on some speaker-discriminative spectral properties.

6. CONCLUSION

In this paper, we proposed to use genetic algorithms to optimize the filter bank of the MFCC feature extractor to the speaker clustering task. The experiments we made showed that the obtained filter bank outperforms the Mel-scaled one. Furthermore, robustness of the obtained solutions according to the initial conditions showed us some spectrum properties which seem to be relevant for the speaker clustering task.
Our work perspectives will consist in evaluating these speaker discriminative spectrum properties. For this we are planning to design a new spectrum-based feature extractor according to this knowledge.

7. REFERENCES

[1] Katagiri, S., *Handbook of Neural Networks for Speech Processing*, Artech House, Boston, 2000.

[2] K. Torkkola, "Feature Extraction by Non-Parametric Mutual Information Maximization", *Journal of Machine Learning Research, Vol. 3*, MIT Press, Cambridge, MA, pp. 1415-1438, 2003.

[3] M. Chetouani, M. Faundez, B. Gas and J.L. Zarader , "Non-linear Speech Feature Extraction for Phoneme Classification and Speaker Recognition", *Nonlinear speech processing : Algorithms and Analysis*, Springer Verlag, 2005.

[4] H. P. Combrinck and E. C. Botha, "On The Mel-Scaled Cepstrum", *Proceedings of the Seventh Annual South African Workshop on Pattern Recognition*, University of Pretoria, Pretoria, 1996.

[5] Holland, J. H., *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI, 1975.

[6] Makhoul, J., Kubala, F., Leek, T., Liu, D., Nguyen, L., Schwartz, R., and Srivastava, A., "Speech and language technologies for audio indexing and retrieval", *Proceedings of the IEEE, Vol. 88, No. 8*, IEEE Press US, Piscataway, pp. 1338-1353, 2000.

[7] Gravier, G. and Betser, M., Audioseg: "Audio Segmentation Toolkit", 2005. http://www.irisa.fr/metiss/guig/index-en.html

[8] F. Pasemann, U. Dieckmann, and U. Steinmetz, "Evolving Structure and Function of Neurocontrollers", *Proceedings of the 1999 Congress on Evolutionary Computation Journal*, IEEE Press US, Piscataway, pp. 1937-1978, 1999.

[9] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.F. Bonastre, and G. Gravier, "The Ester Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News", *Proceedings of Eurospeech/Interspeech'05*, Lisbon, pp.1149-1152, 2005.