Multi filter bank approach for speaker verification based on genetic algorithm

Christophe Charbuillet, Bruno Gas, Mohamed Chetouani, Jean Luc Zarader

Université Pierre et Marie Curie-Paris6, FRE2507
Institut des Systèmes Intelligents et Robotique (ISIR), Ivry sur Seine, F-94200 France
Christophe.Charbuillet@lis.jussieu.fr, Gas@ccr.jussieu.fr,
Mohamed.Chetouani@upmc.fr, jean-luc.zarader@upmc.fr

Abstract

Speech recognition systems usually need a feature extraction stage which aims at obtaining the best signal representation. State of the art speaker verification systems are based on cepstral features like MFCC, LFCC or LPCC. In this article, we propose a feature extraction system based on the combination of three feature extractors adapted to the speaker verification task. A genetic algorithm is used to optimize the features complementarities. This optimization consists in designing a set of three non linear scaled filter banks. Experiments are carried out using a state of the art speaker verification system. Results show that the proposed method improves significantly the system performances on the 2005 Nist SRE Database. Furthermore, the obtained feature extractors show the importance of some specific spectral information for speaker verification.

1. Introduction

Speech feature extraction plays a major role in speaker verification systems. State of the art speaker verification systems front end are based on the estimation of the spectral envelope of the short term signal, e.g., Mel-scale Filterbank Cepstrum Coefficients (MFCCs), Linear-scale Filterbank Cepstrum Coefficients (LFCCs), or Linear Predictive Cepstrum Coefficients (LPCCs). Even if these extraction methods achieve good performances on speaker verification, they do not take into account specific information about the task to achieve. To avoid this draw back, several approaches have been proposed to optimize the feature extractor to a specific task. These methods consist to simultaneously learn the parameters of both the feature extractor and the classifier [1]. This procedure consists in the optimization of a criterion, which can be the Maximization of the Mutual Information (MMI) [2] or the Minimization of the Classification Error (MCE) [3]. In this paper we proposed to used a genetic algorithm to design a feature extraction system adapted to the speaker verification task.

Genetic algorithms (GA) were first proposed by Holland in 1975 [4] and became widely used in various disciplines as a new means of complex systems optimization. In recent years their have been successfully applied to the speech processing domain. Chin-Teng Lin and al. [5] proposed to apply a GA to the feature transformation problem for speech recognition and M. Zamalloa and al. [6] worked on a GA based feature selection for speaker recognition. 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 unsupervised optimization methods. So they can be used as an exploration tool, free to find the best solution without any constraint. In a previous work [7] we used this approach to show the importance of specific spectral information for the speaker diarization task. State of the art speaker verification systems are based on a cepstral feature extraction front end (LFCC, MFCC, LPCC) follow by a GMM [8] or an hybrid GMM/SVM classifier [9]. Nowadays, an alternative an increasingly used approach consists in fusing different systems. This technique can be divided in two main categories depending on the source of this difference. The systems based on a classifier's variety [10] and the systems based on different features. Our study deals with the second principle. We can quote the work of M. Zhiyou and al. [11] which consist of combining the LFCC and MFCC features, or the study of Poh Hoon Thian & al. [12] who proposed to complete the LFCC's with spectral centroids subands features.

In this paper we proposed to fuse three systems based on different feature extractors. A genetic algorithm is used to optimize the feature extractor's complementarities. Figure 1 describes this approach. In the second section, a description of the feature extraction method is given. Afterward, we describe the genetic algorithm we used, followed by its application to complementary feature extraction. Then, the experiments we made and the obtained results are presented.

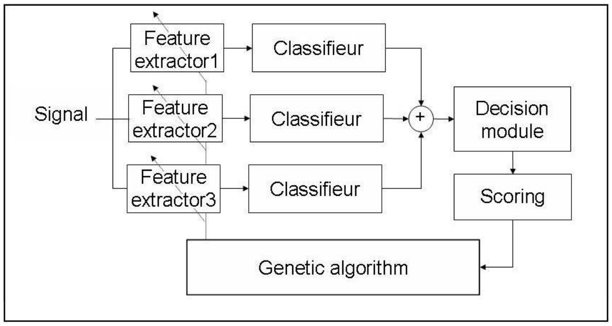


Figure 1: *Feature extraction optimization*

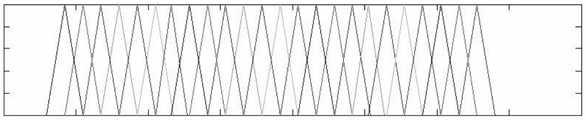


Figure 2: *Linear scaled filter bank*

2. Filter bank based feature extractors

The conventional MFCC and LFCC feature extractor process mainly consists of modifying the short-term spectrum by a filter bank. This process has four steps:

- Compute the power spectrum of the analyzed frame;
- Sum the power spectrum for each triangular filter of the bank;
- Apply the log operator to the obtained coefficients;
- Compute the Discrete Cosine Transform (DCT).

Figure 2 presents the linear scaled filter bank used for the LFCC's computation. This feature extractor is known to be the most robust for the short band signals representation. The purpose of our study is to find a set of three cepstrum based feature extractors design for high level fusion. To this end, we propose to use a genetic algorithm to optimize, the number of filters on the bank, the scaled of the filter bank and the number of cepstral output coefficients.

3. Genetic algorithm

A genetic algorithm is an optimization method. Its aim is to find the best values of the system's parameters in order to maximize its performance. 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 feature extractor defined by its genes. Genes consists in a condensed an adapted representation of the feature extractor's operational parameters.

3.1. Gene encoding

Parameter's encoding plays a major role in a genetic algorithm. By an adapted parameter representation, this method can strongly increases the speed convergence of the algorithm. Moreover it reduces the over fitting effect by reducing the parameters dimention. The parameters we chose to optimize are:

- Nf: Number of filters in the bank;
- Nc: Number of cepstral coefficients;
- C_i: Center frequency of the i^{th} filter in the bank;
- B_i: Band width of the i^{th} filter in the bank.

Parameters C and B are encoded with two polynomial functions described by the equations (1) and (2). This encoding method reduces the parameter's dimention from 50 to 12 (in the average case) and guaranties the filter bank's regularity. The parameter Nf and Nc are not encoded and will be directly muted.

$$C_i = gc_0 + gc_1 \cdot \frac{i}{Nf} + gc_2 \cdot (\frac{i}{Nf})^2 + ... + gc_N \cdot (\frac{i}{Nf})^N \quad (1)$$

$$B_i = gb_0 + gb_1 \cdot \frac{i}{Nf} + gb_2 \cdot (\frac{i}{Nf})^2 + ... + gb_N \cdot (\frac{i}{Nf})^N \quad (2)$$

Where $\{gc_0,...,gc_N\}$ and $\{gb_0,...,gb_N\}$ are the genes relative to the parameters $\{C_0,...,C_{Nf}\}$ and $\{B_0,...,B_{Nf}\}$; N is the polynomial order; Nf represent the number of filter.

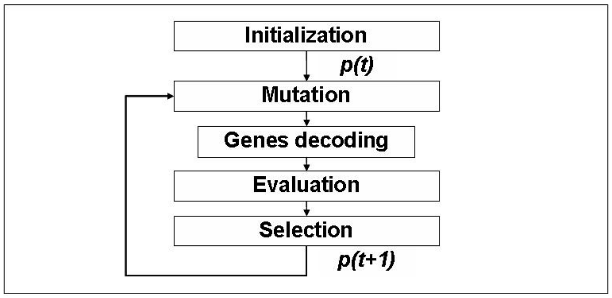


Figure 3: *Genetic algorithm*

3.2. Genetic algorithm description

The algorithm we used is made of four operators: **M**utation, **D**ecoding, **E**valuation and **S**election (M, D, E, S). These operators are applied to the current population $p(t)$ to produce a new generation $p(t+1)$ by the relation:

$$p(t+1) = S \circ E \circ D \circ M(p(t)) \quad (3)$$

Figure 3 represent this algorithm. The first step consists on a random initialization of the feature extractor's genes. Then, the operators are iteratively applied.
The *Mutation* operator consists in a short random variation of the genes.
The *Decoding* operator aim at decode the genes to obtain the operational feature extractor's parameters.
The *Evaluation* operator's goal is to evaluate each feature extractor performances. The evaluation criterion we used is defined on the section 3.3.
The *Selection* operator selects the Ns better feature extractors of the current population. These individuals are then cloned according to the evaluation results to produce the new generation $p(t+1)$ of Np feature extractors. As a consequence of this selection process, the average of the performance of the population tends to increase and in our application adapted feature extractors tend to emerge.

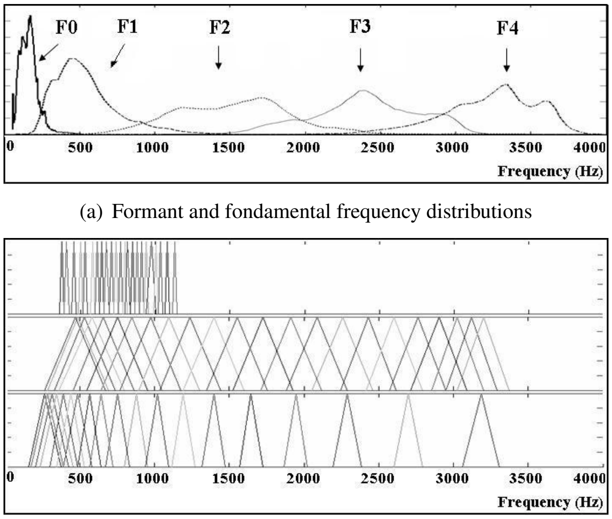
3.3. Application to complementary feature extraction

The objective is to obtain a set of three complementary feature extractors. The main idea is to evolve three isolated populations of feature extractors and to select the best combination. At each generation, the fusion is done for all combination of feature extractors and the resulting Equal Error Rate (EER) is memorized. At the end of this process, the fitness of an individual is defined as the lower EER obtained (e.i. the EER corresponding to the best combination including this feature extractor). As a consequence of this process, each population tends to specialize on specific feature, complementary with the others.

4. Experiments and results

4.1. Data bases

The databases used for the evolution phase and for the test are extracted from the 2005 Nist SRE corpus [13]. This corpus is composed of conversational telephone speech signals passed through different channels, (landline, cordless or cellular) and sampled to 8 kHz. We used 10 male and 10 female with one utterance of 2 min 30s per speaker for the evolution phase. The number of tests between model and test signal involved for each



(a) Formant and fondamental frequency distributions

(b) Obtained filter banks for C1 (top) C2 (midle) and C3 (bottom)

Figure 4: Spectral analyse and obtained solutions

feature extractor evaluation was of 2052. For the test database, we used 50 males and 50 females whose not appear on the training base. The number of tests involved was of 116942.

4.2. Speaker verification system

All experiments we made are based on a state of the art GMM-UBM speaker verification system. This system, called LIA SpkDet [14] was provided by the University of Avignon, France. We used a system with 16 gaussian per mixture, with diagonal covariance matrix.

4.3. Genetic algorithm parameters

The genes $\{gc_0,...,gc_N\}$ and $\{gb_0,...,gb_N\}$ which code for the centres frequencies and the band widths are initialized with a Gaussian normalized random. The parameter Nf are initialized to 24, and Nc to 16.
The parameter we used for the feature extractor's evolution are:

- Population size Np : 20;
- Number of selected individuals Ns : 5;
- Polynomial order for the genes encoding N : 5;
- Mutation method for the polynomials coefficients: Gaussian random variation of ± 0.1;
- Mutation method for Nf : uniform random variation of ± 5;
- Mutation method for Nc : uniform random variation of ± 3;

4.4. Results

In this section, obtained feature extractors are presented and analysed. Figure 4.b presents the obtained filter banks. In order to interpret the obtained solution, a statistical analysis of the fundamental frequency and formants was done on a database composed of 20 male and 20 females. Figure 4.a presents the probability distributions of these mesures.

Table 1: *Comparative results*

Feature extractor	Nf	Nc	F_{min}	F_{max}	EER%
LFCC	24	16	300	3400	**14.44**
MFCC	24	16	300	3400	**14.88**
C1	23	15	360	1145	22.90
C2	25	20	266	3372	14.79
C3	19	19	156	3309	16.07
C1+C2+C3	- -	- -	- -	- -	**12.69**

Table 2: *Fusion analysis*

Feature extractor	Correlation	EER obtained by fusion
C1+C2	0.51	13.21%
C2 + C3	0.83	13.45%
C1 + C3	0.64	15.39%

Table 1 details both the feature extractor's characteristics and the results obtained on the test base. The combination method used is an arithmetic fusion, as illustrated by the figure 1.

Table 2 presents the correlation coefficients between the compared system and the EER obtained by fusion. The correlation is based on the log-likelihood outputs of the compared systems for the whole tests of the test database. A test consists to measure the log-likehood between a speaker model and test signal. The r correlation coefficient is defined by:

$$r = \frac{\sum_{i=1}^{Nt}(S1_i - \bar{S1}) \cdot (S2_i - \bar{S2})}{\sqrt{\sum_{i=1}^{Nt}(S1_i - \bar{S1})^2} \cdot \sqrt{\sum_{i=1}^{Nt}(S2_i - \bar{S2})^2}} \qquad (4)$$

Where $S1_i$ represent the log-likelihood obtained by the system 1 on i^{th} test; Nt is the number of test.

The correlation coefficient, which takes value in [-1;1], is a measure of the system's decision similarity. In our application, the classifiers are identical. As a consequence, this measure can be interpreted as the similarity between the information provied by the feature extractors. A correlation of 1 means that the information supplied by the feature extractors are equivalent (i.e. they lead to the same decision). A correlation of 0 means that the information supplied are independent.

Taking into account these different information, we can notice that:

- Information relative to the fundamental frequencies is not used;
- C2 covers a large spectral zone and obtained results similar are to the LFCC or MFCC feature extractors;
- C1 seems to focus exclusively on the first formant;
- C3 presents a high filter density centred on the first formant, while keeping the whole spectre information;
- The de-correlation of the obtained systems are significant.
- The final combination of the three feature extractors improve the system performance of 12% compare to the baseline system.

These results show that the proposed method is reliable. The correlation between the diferent systems and the improvement supplied by the fusion show that the obtained feature extractors are complementary. This improvement seems to be related to the information provied by the first formant. In the final article, a more detailed analyse of these results will be presented.

5. Conclusion

In this paper, we proposed to use a genetic algorithm to optimize a feature extraction system adapted to the speaker verification task. The proposed system is based on a combination of three complementary feature extractors. Obtained results show that the proposed method improves significantly the system performance. Furthermore, the obtained feature extractors reveal the importance of specific spectral information relatives to the first formant.
Our future work will consist in study the robustness of the obtained solutions according to both the initial conditions and the base used for the evolution phase.

6. References

[1] B. G. Mohamed Chetouani, Marcos Faundez-Zanuy and J.-L. Zarader, *Nonlinear Speech Modeling and Applications*. Springer, 2005, ch. Non-linear Speech Feature Extraction for Phoneme Classification and Speaker Recognition, pp. 344–350.

[2] K. Torkkola, "Feature extraction by non parametric mutual information maximization," *The Journal of Machine Learning Research*, vol. 3, pp. 1415–1438, 2003.

[3] C. Miyajima, H. Watanabe, K. Tokuda, T. Kitamura, and S. Katagiri, "A new approach to designing a feature extractor in speaker identification based on discriminative feature extraction," *Speech Communication*, vol. 35, no. 3-4, pp. 203–218, Oct. 2001.

[4] J. H. Holland, "Adaptation in natural and artificial systems," *University of Michigan Press*, 1975.

[5] L. Chin-Teng, N. Hsi-Wen, and H. Jiing-Yuan, "Ga-based noisy speech recognition using two-dimensional cepstrum," in *IEEE Transactions on Speech and Audio Processing*, vol. 8, 2000, pp. 664–675.

[6] M. Zamalloa, G. Bordel, J. L. Rodriguez, and M. Penagarikano, "Feature selection based on genetic algorithms for speaker recognition," in *IEEE Odyssey*, vol. 1, 2006, pp. 1–8.

[7] C. Charbuillet, B. Gas, M. Chetouani, and J. L. Zarader, "Filter bank design for speaker diarization based on genetic algorithms," in *Acoustics, Speech, and Signal Processing, 2006. Proceedings. (ICASSP '06). IEEE International Conference on*, vol. 1, 2006, pp. 673–676.

[8] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, 1995.

[9] S. Fine, J. Navratil, and R. Gopinath, "A hybrid gmm/svm approach to speaker identification," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, vol. 1, 2001, pp. 417–420 vol.1.

[10] K. Farrell, R. Ramachandran, and R. Mammone, "An analysis of data fusion methods for speaker verification," in *Acoustics, Speech, and Signal Processing, 1998. ICASSP '98. Proceedings of the 1998 IEEE International Conference on*, vol. 2, 1998, pp. 1129–1132 vol.2.

[11] M. Zhiyou, Y. Yingchun, and W. Zhaohui, "Further feature extraction for speaker recognition," *IEEE International Conference on Systems, Man and Cybernetics*, vol. 5, pp. 4153–4158, 2003.

[12] N. Poh Hoon Thian, C. Sanderson, S. Bengio, D. Zhang, and K. Jain Anil, "Spectral subband centroids as complementary features for speaker authentication," *Lect. notes comput. sci.*, vol. 3072, pp. 631–639, 2004.

[13] "2005 nist speaker recognition evaluation site." [Online]. Available: http://www.nist.gov/speech/tests/spk/2005/

[14] "Lia spkdet web site." [Online]. Available: http://www.lia.univ-avignon.fr/heberges/ALIZE/LIA RAL