# SPEAKER RECOGNITION USING DECISION FUSION

M. Chenafa, D. Istrate
*RMSE, ESIGETEL, 1 Rue du Port de Valvins, 77215 Avon-Fontainebleau, France*

V. Vrabie, M. Herbin
*CReSTIC, Université de Reims Champagne-Ardenne, Chaussée du Port, 51000 Châlons-en-Champagne, France*

Keywords: Biometrics, Speaker recognition, Speech recognition, Decision fusion, GMM/UBM.

Abstract: Biometrics systems have gained in popularity for the automatic identification of persons. The use of the voice as a biometric characteristic offers advantages such as: is well accepted, it works with regular microphones, the hardware costs are reduced, etc. However, the performance of a voice-based biometric system easily degrades in the presence of a mismatch between training and testing conditions due to different factors. This paper presents a new speaker recognition system based on decision fusion. The fusion is based on two identification systems: a speaker identification system (text-independent) and a keywords identification system (speaker-independent). These systems calculate the likelihood ratios between the model of a test signal and the different models of the database. The fusion uses these results to identify the couple speaker/password corresponding to the test signal. A verification system is then applied on a second test signal in order to confirm or infirm the identification. The fusion step improves the false rejection rate (FRR) from $21,43\%$ to $7,14\%$ but increase also the false acceptation rate (FAR) from $21,43\%$ to $28,57\%$. The verification step makes however a significant improvement on the FAR (from $28,57\%$ to 14.28%) while it keeps constant the FRR (to $7,14\%$).

## 1 INTRODUCTION

Biometric recognition systems, which identify a person on his/her physical or behavioral characteristics (voice, fingerprints, face, iris, etc.), have gained in popularity among researchers in signal processing during recent years. Biometric systems are also useful in forensic work (where the task is whether a given biometric sample belongs to a given suspect) and law enforcement applications (Atkins, 2001). The use of the voice as a biometric characteristic offers the advantage to be well accepted by users whatever his culture. There are two categories in voice-based biometric systems: speaker verification and speaker identification. In identification systems, an unknown speaker is compared to the N known speakers stored in the database and the best matching speaker is returned as the recognition decision. Whereas in verification systems, an identity is claimed by a speaker, so the system compares the voice sample to the claimed speaker's voice template. If the similarity exceeds a predefined threshold, the speaker is accepted, otherwise is rejected. For each system two methods can be distinguished: text-dependent and text-independent. In the first case, the text pronounced by the speaker is known beforehand by the system, while in the second case the system does not have any information on the pronounced text (Kinnunen, 2003).

It is well known that the performances of voice-based biometric systems easily degrade in the presence of a mismatch between the training and testing conditions (channel distortions, ambient noise, etc.). One method that can be used to improve the performances of these systems is to merge various information carried by the speech signal. Several studies on information fusion were led to improve the performances of automatic speakers recognition system (Higgins et al., 2001)(Mami, 2003)(Kinnunen et al., 2004). However, the results are less successful compared to biometric systems based on other modalities (fingerprint, iris, face, etc).

In this paper a new fusion approach is proposed by using two kinds of information contained in the speech signal: the speaker (who spoke ?) and the keyword pronounced (what was said ?). The aim of this method is to use a first test signal to identify a couple speaker/password corresponding to this signal. This step is done by combining two identification systems based on likelihood ratio approach: a speaker identification system (text-independent) and a speech iden-

Chenafa M., Istrate D., Vrabie V. and Herbin M. (2008).
SPEAKER RECOGNITION USING DECISION FUSION.
In *Proceedings of the First International Conference on Bio-inspired Systems and Signal Processing*, pages 267-272
DOI: 10.5220/0001065502670272

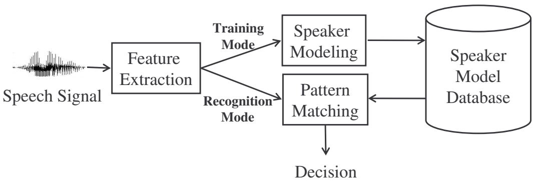

Figure 1: Components of a speaker recognition system.

tification system (speaker-independent). The speaker identified by this fusion is then verified by a classical verification text dependent system using a second test signal. In practical situations, the two test signals can be viewed as a composed password. The system provides good improvements on the two types of error usually computed for biometric systems: the false rejection rate (FRR) and the false acceptation rate (FAR). The experiments presented in this study use the platform ALIZE developed by the LIA laboratory (Bonastre et al., 2005).

This paper is organized as follows. Section 2 provides a general description of a speaker recognition system. Section 3 presents the proposed fusion system. The experiments are discussed in Section 4, followed by conclusions in the last section.

# 2 SPEAKER RECOGNITION SYSTEM

Figure 1 shows the structure of an automatic speaker recognition system. This system operates in two modes (training and recognition) and can be used for both identification or verification tasks. In the training mode, a new speaker (with known identity) is enrolled into the system's database, while in the recognition mode an unknown speaker gives a speech input and the system makes a decision about the speaker identity.

## 2.1 Feature Extraction

Feature extraction is the first component in an automatic speaker recognition system (Furui, 1997). This phase consists of transforming the speech signal in a set of feature vectors called also parameters. The aim of this transformation is to obtain a new representation which is more compact, less redundant, and more suitable for statistical modeling and calculation of distances. Most of the speech parameterizations used in speaker recognition systems relies on a cepstral representation of the speech signal (Lee et al., 1996).

### 2.1.1 MFCC and LFCC Parameters

The *Mel-frequency cepstral coefficients* (MFCC) are motivated by studies of the human peripheral auditory system. Firstly, the speech signal $x(n)$ is divided into $Q$ short time windows which are converted into the spectral domain by a Discret Fourier Transform(DFT). The magnitude spectrum of each time window is then smoothed by a bank of triangular bandpass filters (Figure 2) that emulate the critical band processing of the human ear.

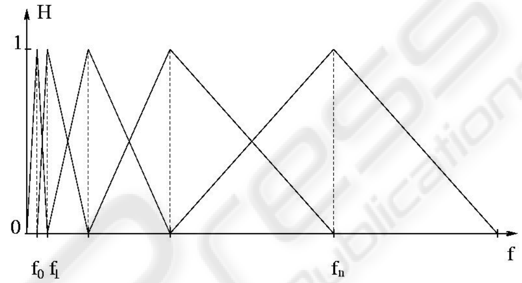

Figure 2: Mel filter bank.

Each one of the bandpass filter $H(k,m)$ computes a weighted average of that subband, which is then logarithmically compressed:

$$X'(m) = \ln\left(\sum_{k=0}^{N-1} |X(k)|H(k,m)\right) \tag{1}$$

where $X(k)$ is the DFT of a time window of the signal $x(n)$ having the length $N$, the index $k$, $k = 0,\dots,N-1$, corresponds to the frequency $f_k = kf_s/N$, with $f_s$ the sampling frequency, the index $m$, $m = 1,\dots M$ and $M << N$, is the filter number, and the filters $H(k,m)$ are triangular filters defined by the center frequencies $f_c(m)$ (Sigurdsson et al., 2006). The log compressed filter outputs $X'(m)$ are then decorrelated by using the Discrete Cosine Transform (DCT):

$$c(l) = \sum_{m=1}^{M} X'(m)\cos(l\frac{\pi}{M}(m-\frac{1}{2})) \tag{2}$$

where $c(l)$ is the $l^{th}$ MFCC of the considered time window. A schematic representation of this procedure is given in Figure 3.

There are several analytic formulae for the Mel scale used to compute the center frequencies $f_c(m)$. In this study we use the following common mapping:

$$B(f) = 2595\log_{10}(1+\frac{f}{700}) \tag{3}$$

The LFCC parameters are calculated in the same way as the MFCC, but the triangular filters use a linear frequency repartition.

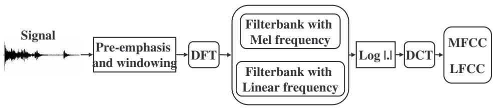
Signal
Pre-emphasis and windowing
DFT
Filterbank with Mel frequency
Filterbank with Linear frequency
Log |.|
DCT
MFCC
LFCC

Figure 3: Extraction of MFCC and LFCC parameters.

#### 2.1.2 $\Delta$ and $\Delta\Delta$ Parameters

After the cepstral coefficients have been calculated and stored in vectors, a dynamic information about the way these vectors vary in time is incorporate. This is classically done by using the $\Delta$ and $\Delta\Delta$ parameters, which are polynomial approximations of the first and second derivatives of each vector (Kinnunen et al., 2004).

### 2.2 Speaker Modeling

The training phase uses the acoustic vectors extracted from each segment of the signal to create a speaker model which will be stored in a database. In automatic speaker recognition, there are two types of methods that give the best results of recognition: the deterministic methods (dynamic comparison and vector quantization) and statistical methods (Gaussian Mixture Model - GMM, Hidden Markov Model - HMM), these last ones being the most used in this domain. In this paper, we have chosen to use a system based on GMM-UBM. This choice was motivated by two reasons: modeling by GMM is very flexible with regard to the type of the signal and using the GMM gives a good compromise between performances and the complexity of the system.

#### 2.2.1 GMM-UBM

In this research, the method used for speaker modeling is the GMM using the universal background model (UBM). The UBM has been introduced and successfully applied by (Reynolds, 1995) in speaker verification. This model is created by using all recording speech of the database, the aim being to have a general model of speakers which will be then used to adapt each speaker model.

The matching function in GMM is defined in terms of the log likelihood of the GMM (Bimbot et al., 2004) given by:

$$p(X|\lambda) = \sum_{q=1}^{Q} log\ p(x_q|\lambda) \tag{4}$$

where $p(x_q|\lambda)$ is the Gaussian mixture density of the $q^{th}$ segment in respect to the speaker $\lambda$:

$$p(x_q|\lambda) = \sum_{i=1}^{G} p_i f(x_q|\mu_i^{(\lambda)}, \Sigma_i) \tag{5}$$

with the mixing weights constrained by:

$$\sum_{i=1}^{G} p_i = 1 \tag{6}$$

In these expressions $x_q$ is the D-dimensional acoustic vector corresponding to the $q^{th}$ time window of the input signal, $p_i$, $\mu_i^{(\lambda)}$ and $\Sigma_i$ ($i = 1, \ldots, G$) are the mixture weight, mean vector, and covariance matrix of the $i^{th}$ Gaussian density function (denoted by $f$) of the speaker $\lambda$, while $G$ denotes the number of GMM used by the model.

The speaker model $\lambda$ is thus given by:

$$\lambda = \left\{ p_i, \mu_i^{(\lambda)}, \Sigma_i | i = 1, \ldots, G) \right\} \tag{7}$$

the UBM model having the same form:

$$UBM = \left\{ p_i, \mu_i^{(UBM)}, \Sigma_i | i = 1, \ldots, G \right\} \tag{8}$$

The mean vectors of speaker model $\mu_i^{(\lambda)}$ are adapted to the training data of the given speaker from the UBM, i.e. $\mu_i^{(UBM)}$, by using the Maximum a Posteriori (MAP) adaptation method (Gauvain and Lee, 1994), the covariance matrices and mixture weights remaining unchanged.

### 2.3 Pattern Matching and Decision

Given a segment of speech, $Y$, and a hypothesized speaker, $S$, the task of speaker recognition system is to determine if $Y$ was spoken by $S$. This task can be defined as a basic hypothesis test between

$$\begin{cases} H_0\text{: Y is from the hypothesized speaker S} \\ H_0\text{: Y is not from the hypothesized speaker S} \end{cases}$$

To decide between these two hypotheses, the optimum test is a likelihood ratio given by:

$$\frac{p\,(Y|H_0)}{p\,(Y|H_1)} \begin{cases} \geq & \theta \quad Accept \quad H_0 \\ < & \theta \quad Reject \quad H_0 \end{cases} \tag{9}$$

where $p(Y|H_i)$ is the probability density function for the hypothesis $H_i$ evaluated for the observed speech segment $Y$, also referred to the likelihood of the hypothesis $H_i$. The decision threshold for accepting or rejecting $H_0$ is $\theta$. A good technique to compute the values of the two likelihoods, $p(Y|H_0)$ and $p(Y|H_1)$ is given in (Doddington, 1985).

## 3 PROPOSED SYSTEM

In this paper a new method for automatic speaker recognition based on fusion information is proposed. The architecture of this method is described in Fig. 4.

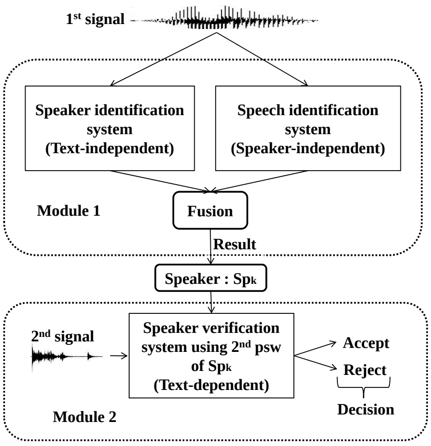
1st signal
Speaker identification system (Text-independent)
Speech identification system (Speaker-independent)
Module 1
Fusion
Result
Speaker : Spk
2nd signal
Speaker verification system using 2nd psw of Spk (Text-dependent)
Accept
Reject
Decision
Module 2

Figure 4: Global system architecture.

This system is composed by two blocks, the first one made up by two classifiers (speaker and password classifiers) and the second one made up by a verification system using the decision result of the first block. Each speaker is identified by two passwords: the first one is used by both speaker and password identification systems, while the second one by the verification system. In practical situations, these two passwords can be viewed as a composed password.

The identification systems (speakers and passwords identification) are used in open-set (no information available on the possible speakers and passwords). Both systems calculate the likelihood ratio on a first test signal by using equation (4). We used here a normalization model UBM, as presented in section 2.2.1. This means that during the creation of the models (speaker, password), each model is adapted by the MAP method to the UBM model.

The verification system is a classical speaker verification system which is used to confirm or infirm the speaker identified previously by using a second test signal (the second password).

Figure 5 shows the fusion between speaker and speech (password) identification systems.

After sorting the *log* likelihood ratios (for the first test signal) calculated with regard to the speakers model $LLK(X|Sp_i)$, $i = 1,N$ (N is the number of speakers stored in the database) and to the passwords model $LLK(X|Psw_i)$,$i = 1,N$ (N is the number of passwords stored in the database), a first test consists of comparing the most likely speaker given by the speaker classifier with the first three identified passwords given by the password classifier. If his password was found between the three identified passwords, a couple (speaker/password) was thus identified. A second test consists of comparing the most likely password with the first three identified speakers. If this password belongs to one of them, another couple (password/speaker) is identified. In the cases where two couples are identified, the couple with the biggest likelihood ratio (Lk_Sp + Lk_P) is retained. The system can reject directly a recording if there are no identified couples.

Once the first test signal is associated to a speaker, a classical verification is then launched using the second test signal pronounced by the speaker identified previously. If the likelihood ratio of this verification is smaller than the smallest likelihood ratio of the first two recordings used in the training phase, the identity of the speaker is confirmed, otherwise the speaker is rejected.

# 4 EXPERIMENTS

## 4.1 Data Base

In order to evaluate the proposed system a corpus of specific keywords has been recorded. This corpus contains the recordings of 15 isolated words (French language) and 11 numbers (from 0 to 10).
The recordings were stored in WAV format, with a sampling rate $f_s = 16$ kHz. The parameterization was realized by using MFCC parameters for the passwords identification system and LFCC for speaker identification and verification systems. We have optimized the acoustic parameter for this application; all the 8 ms the signal is characterized by a vector made up of 16 ceptrals coefficients $c(l)$ (see Eq. (2)) and their derivative $\Delta\Delta$.

## 4.2 Training and Test Data

For both identification systems (speaker and password) the first password recording is used for the

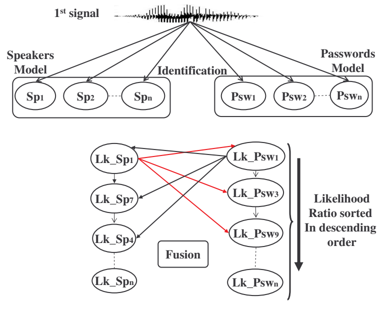

Figure 5: Fusion system architecture(Module 1).

training mode. The verification system uses the second password for the training mode. The speakers database is divided into two equals groups: 7 clients and 7 impostors. Therefore, in the test stage the number positive and negative tests are equals.

1. The speaker identification system (text-independent) uses two recordings of 14 words of the 7 clients for the training phase. For the recognition phase, the system uses one recording of 14 words of the 7 clients and 3 recordings of 14 words of 7 impostors.
2. The password identification system (speaker-independent) uses two recordings of 7 clients for the training phase. For the recognition phase, the system uses one recording of 7 clients and all recordings of the impostors.
3. The verification system uses two recordings of the second passwords of every client for the training phase and a recording of 7 clients as well as all the recordings of 7 impostors for the recognition phase.
4. The reference system uses for the training phase 7 speakers, two recordings of 14 words. For recognition phase we used a recording of 14 speakers passwords and 3 recordings of 14 impostors words.

### 4.3 Reference System

The results obtained by the global system are compared to a classical verification system (Bimbot et al., 2004). In the training phase of the reference system a speaker model is created from the feature vectors (16 LFCC + $\Delta\Delta$) using two recordings of all the passwords to model speakers; However the recognition phase uses all passwords of the speakers pronounced by impostor and other words.
We have optimized the number of GMM for this application; the optimal value is $G = 16$.

### 4.4 Results and Discussion

Table 1 shows the false rejection rate (FRR) and the false acceptation rate (FAR) of the reference system, the first module of the new system and the global system proposed.

Table 1: performances of different systems.

| Systems | FRR | FAR |
|---|---|---|
| Reference System<br>text dependent<br>(16 LFCC + $\Delta\Delta$) | 21.43% | 21.43% |
| Fusion System between<br>speakers identification<br>(16 LFCC+$\Delta\Delta$)<br>and passwords<br>identification<br>(16 MFCC + $\Delta\Delta$) | 7.14% | 28.57% |
| Verification after fusion<br>(16 LFCC + $\Delta\Delta$) | 7.14% | 14.28% |

The best equal error rate obtained for the reference system is 21.43%, which is high enough but can be justified by the small size of the database. After the fusion of the results between the speaker identification system and the password identification system, we notice that the FAR increases to 28.57% (that is due to the password identification system which increases the chance of impostors to be accepted because the password is well recognized), while the FRR decreases to 7.14%. By using a verification system, which uses the results of this fusion, we improve the FAR (from 28.57% to 14.28%) while the FRR remains the same one (7.14%) because the verification system was adapted to recognize the clients. The global system thus makes an improvement of 43.47% of the FRR and 65.69% of the FAR. Note again that these values are high enough due to the small size of the database.

## 5 CONCLUSIONS

In this paper, we presented several experiments to improve the performances of a voice-based biometric system using decision fusion. The fusion of the speaker identification and the passwords identification was firstly proposed. We show that the fact of modeling the passwords pronounced by the speakers brings improvements in the false reject rate but in the same time it increases the number of the impostors accepted by the system. The second experience proposes an automatic speaker verification using the result (speaker identified) of the first experience. The aim here is to confirm the results returned by the fusion of speaker and password classifiers. This second experience allows us to reduce the number of impostors accepted by the system and improves the results of the fusion by decreasing the FAR from 28.57% to 14.28%. So the global system improves the performances in term of FAR and FRR with regard to the reference system. This study encourages us to continue the experimentation on a corpus with more important size and to consider other kind of fusion such as weigthed ranks.

## REFERENCES

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Chagnolleau, I., Meignier, S., Merlin, T., Garciya, J., Delacrtaz, D., and Reynolds, D. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 4:430–451.

Bonastre, J.-F., Wils, F., and Meignier, S. (2005). Alize, a free toolkit for speaker recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 737–740.

Doddington, G. (1985). Speaker recognition - identifying people by their voices. In *Proc. of the IEEE*, volume 73, pages 1651–1664.

Furui, S. (1997). Recent advances in speaker recognition. In *Proc. of the First International Conference on Audio- and Video-Based Biometric Person Authentication*, volume 1206, pages 237–252.

Gauvain, J.-L. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. In *IEEE Trans. on Speech and Audio*, volume 2, pages 291–298.

Higgins, J. E., Damper, R. I., and Harris, C. J. (2001). Information fusion for subband-hmm speaker recognition. In *International Joint Conference on Neural Networks*, volume 2, pages 1504–1509.

Kinnunen, T. (2003). *Spectral Features for Automatic Text-Independent Speaker Recognition*. PhD thesis, University of Joensuu, Finland.

Kinnunen, T., Hautamki, V., and Fr´anti, P. (2004). Fusion of spectral feature sets for accurate speaker identification. In *9th International Conference Speech and Computer (SPECOM)*, pages 361–365.

Lee, C. H., Soong, F., and Paliwal, K. (1996). *Automatic Speech and Speaker Recognition*. Springer, London, UK, 2nd edition edition.

Mami, Y. (2003). *Reconnaissance de locuteurs par localisation dans un espace de locuteur de reference*. PhD thesis, ENST Paris, France.

Reynolds, D. (1995). Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17:91–108.

Sigurdsson, S., Petersen, K. B., and Lehn-Schiøler, T. (2006). Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. In *Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR)*, pages 286–289.