Efficient Training of GMM Based Speaker Recognition System

Snani Cherifa
Laboratoire d'Automatique et Signaux d'Annaba (LASA)
Université Badji Mokhtar de Annaba, B.P 12
Annaba, 23000 Algeria
sn_cherifa@yahoo.com

ABSTRACT: *Automatic speaker recognition (ASR) is based on speech feature vectors, models, and classifiers. To improve the speaker recognition performance, we must affect at least one of these modules. In this paper we propose to use subband spectral centroids (SSCs) as a complementary features with the traditional MFCC features, and a new GMM training algorithm, with the ultimate goal to search the better mixture component number N for each speaker model, which is fixed in the most speaker recognition systems based on GMM without any priori information, and all speaker models have the same number of components. In experiments, we compared the performance of the proposed scheme with the conventional GMM to show its robustness.*

Keywords: Speaker Recognition System (SRS), Gaussian Mixture Model (GMM), Expectation-maximization (EM) Algorithm, Mel Frequency Cepstral Coefficients (MFCC), Subband Spectral Centroids (SSC)

Received: 21 January 2013, Revised 1 March 2013, Accepted 9 March 2013

1. Introduction

The Gaussian mixture modeling (GMM) approach has become one of the mainstays modeling techniques with its advantage of high recognition rate, easy training. Its superiority over other modeling techniques in text independent speaker recognition systems context has been demonstrated and widely accepted by the research community [1] [2] [3]. A GMM is composed of a joint probability distribution function (PDF) described by the weighted sum of several multivariate Gaussian PDFs, each multivariate Gaussian PDF is termed as a Mixture Component, whose parameters (weighted coefficient, mean vector and covariance matrix) are usually obtained by the Expectation-Maximum (EM) iterative algorithm [4], which converges to the ML estimate of the mixture parameters. The Mixture Component Number (N) which affects the overall performance of EM is not known exactly before training phase in this case all speakers have a GMM model with the identical mixture number example (16, 64, 128). In this paper we propose a new method used training GMM algorithm to search the better mixture Component number N for each speaker model to improve the performance of speaker recognition system.

Speaker recognition is a branch of biometric authentication which refers to the automatic identity recognition of individuals using certain intrinsic characteristics of the person like speech. Speaker recognition can be divided into verification and identification tasks. The verification (Speaker Detection) task is to decide whether or not an unlabeled voice belongs to a claimed speaker. There are only two possible decisions: either to accept the voice as belonging to the claimed speaker or to reject

it as belonging to an impostor. The identification task is to classify an unknown voice with one from a set of enrolled speakers. Speaker identification task is further classified into two cases: The first case called Closed Set where a reference model for the unknown speaker may not exist. The second called Open Set where an additional decision alternative, "*the unknown does not match any of the models*", is required [5] [4].

Speaker recognition can also be classified into text-dependent and text-independent recognitions. In text-dependent recognition, the system knows exactly the spoken text which could be either fixed phrase or prompted phrase. In text-independent recognition, the system does not know the text of the spoken utterance, which could be user selected keywords or conversational speech.

The rest of the paper is organized as follows. In the next session, we present the speaker recognition system this is followed by the description of the feature extraction method (SSCs complementary features with MFCC) chosen within the framework of this paper, Section 3 present the Gaussian Mixture Speaker Model (GMM) parameters. Section 3, we introduce a new technique speaker modeling based on GMM. Section 4 reports experimental results, and Section 5 gives overall conclusions on the work.

2. Speaker Recognition System

A speaker recognition system consists of three components [2], the first one is feature extraction, while the second one is referred to the enrolment or training phase, the last one is referred to as the operational or testing phase. A block diagram of speaker recognition is shown in Figure 1, showing the basic elements discussed in this section. The input speech is sampled and converted into digital format. Feature vectors are extracted from the input speech in the form of Mel-Frequency Cepstral Coefficients (MFCCs) [6] [7] [8] [9] [10] combined by concatenation with Spectral Subband Centroids (SSCs) [11] [12] [13]. The system then branches into two separate phases: training and classification. In the training phase, each registered speaker has to provide samples of their speech so that the system can train reference models for that speaker, whilst in the classification phase the input speech is matched with the stored reference models and the identification is made.

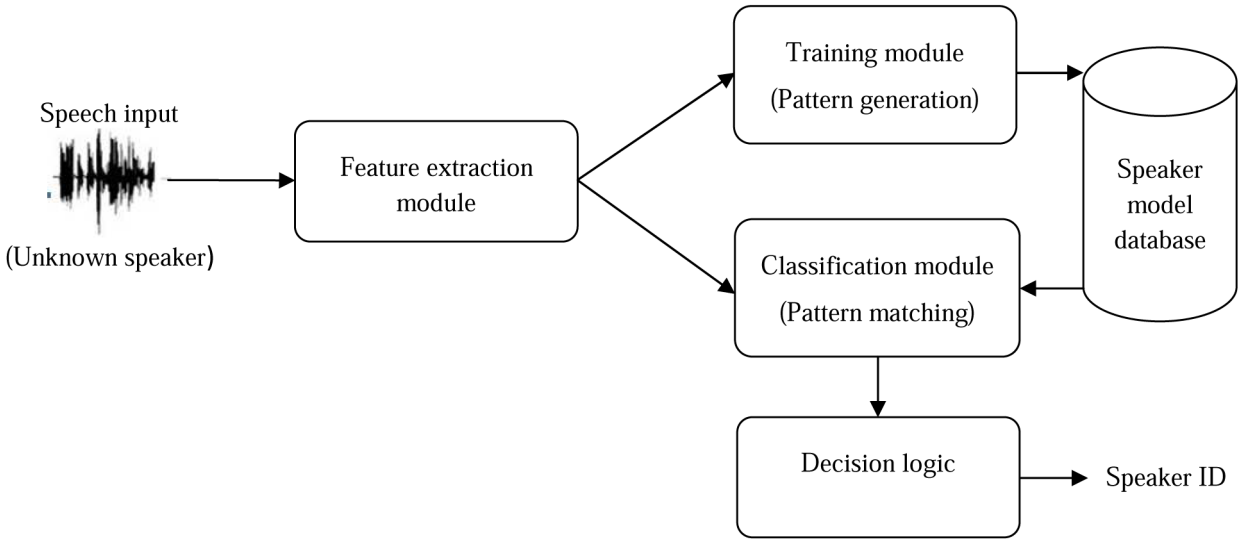

Figure 1. Speaker recognition system

2.1 The feature extraction stage

Feature processing for speaker recognition systems consists of extracting speaker dependent information in a form which can be effectively and efficiently used for model building and recognition. Signal parameterization techniques used for speech recognition are based on extracting information from the short term power spectrum estimates of speech. However, they utilize only amplitude information provided by power spectrum, while the frequency information is left unexplored. For example in MFCC, we use only the information on the total power in each subband, but we do not keep track of the dominant subband frequencies.

Several attempts [11] [12] [13] have recently been made to incorporate the frequency information from the power spectrum in the speech feature vectors. They are based on computing subband spectral centroids (SSC) and using them as additional features in the MFCC-based front-end. The SSCs are closely related to position of spectral peaks (formants) of speech sounds. Since

spectral peak positions remain practically unaffected in presence of additive noise, it is expected that an SSC-based front-end would have a potential of improving the robustness of automatic speech recognition (ASR) systems[11] [12] [13].

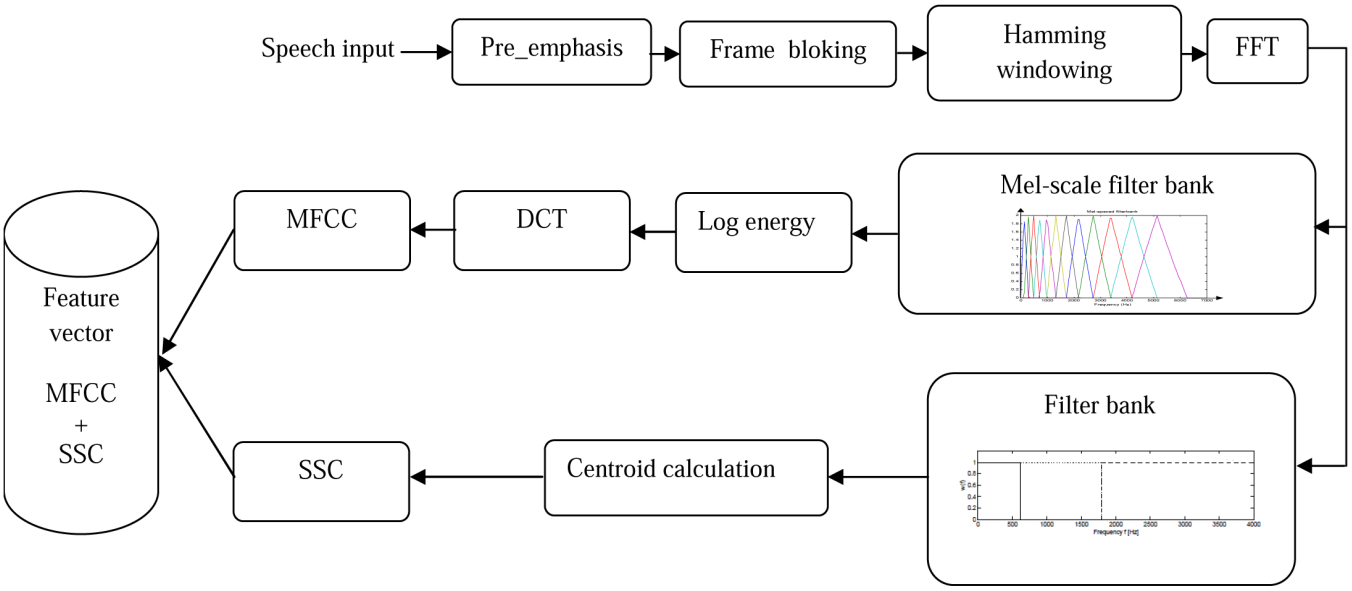

Figure 2. The feature extraction process MFCC + SSC

Figure 2 showing the process of computing feature vectors as is described in more detail next: The feature extraction from the speech samples consists of a filtering process with pre-emphasis and an extraction process of spectral features using a short term analysis [7]. The input to the system is speech sampled at 11025Hz and converted to 16-bit digital format. As showing in figure 2 there are four common steps for SSCs and MFCCs features

1. Pre-emphasis: Its purpose is to flatten the speech spectrum so as to reduce the dynamic range using a first order filter

$$P(z) = 1 - 0{,}95\, z^{-1} \qquad (1)$$

The pre-emphasis makes the upper harmonics of the fundamental frequency more distinct, and the distribution of energy across the frequency range more balanced.

2. A frame blocking: The speech signal continuously changes due to articulatory movements, and therefore, the signal must be broken down in short frames of about 20–30 ms in duration, within this interval, the signal is assumed to remain stationary and a spectral feature vector is extracted from each frame. Where the continuous speech signal is blocked into frames of 256 samples (which is equivalent to ~ 30 msec windowing during which human articulatory conguration does not change dramatically), with adjacent frames separated by 128 samples and overlaps it by 256 – 128 samples.

3. A frame windowing: A Hamming window is applied to each individual frame in order to minimize the signal discontinuities, and consequently the spectral distortion, at the beginning and end of each frame. typically the Hamming window is used, which has the form:

$$w(n) = 0.54 - 0.46 \cos\left[\frac{2\pi\,(n)}{N-1}\right], \; 0 \le n \le N-1 \qquad (2)$$

4. A Discrete Fourier Transform process using a FFT algorithm: Which converts each frame of 256 samples from the time domain into the frequency domain, where the most important speech/speaker information resides, the result obtained is the signal's periodogram

5-1. Subband Sspectral Centroids (SSCs))

The subband spectral centroids features have similarities with the formant frequencies and can be extracted easily and reliably (without any estimation errors) from the power spectrum of the speech signal. In order to define spectral subband centroids, we

divide the frequency band into a fixed number of subbands and compute the centroid for each subband using the power spectrum of the speech signal.

Let us assume that the frequency band [0 to *Fs* /2], where *Fs* is the sampling frequency is divided into *M* subbands. Let the lower and higher edges of mth subband be l_m and h_m, respectively, and its filter shape be $W_m(f)$. We define the m^{th} subband spectral centroid C_m as follows:

$$C_m = \frac{\int_{l_m}^{h_h} f w_m(f) P^{\gamma}(f)\, df}{\int_{l_m}^{h_h} w_m(f) P^{\gamma}(f)\, df} \tag{3}$$

Where $P(f)$ is the power spectrum and γ is a constant controlling the dynamic range of the power spectrum. By setting $\gamma < 1$, the dynamic rage of the power spectrum can be reduced.

Typically, $w_m(f)$ takes the shape of either a rectangular window (ones over the m-th subband and zeros everywhere else) or a triangular window. In this paper, we used the rectangular windows. The use of parameter γ in this function is a design parameter, which can be optimized for a given data, set and task at hand. According to [14], when γ is set to 1, the system can achieve the best performance. And in our experiment the same value was used.

5-2. The MFCC feature extraction process

MFCC [6] [7] is one of the most prevalent feature parameters for speech/speaker recognition are based on the known variation of the human ear's critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech [8] [9] [10] this is expressed in the Mel-frequency scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

a. Mel-scale bandpass filtering: to approximate the frequency resolution of our auditory system.

b. Subbank energy compression: to approximate the nonlinear compression of energy of our auditory perception. More importantly, the log operation makes the subbank energy approximately Gaussian distributed – a requirement for subsequent acoustic modeling.

c. DCT: to transform the spectral information to the cepstral domain in which the energy (information) is dominated by less coefficients.

Denoting the outputs of an M-channel filterbank as $Y(m)$.

$m = 1,..., M$, the MFCCs are obtained as follows:

$$C_n = \sum_{m=1}^{M} [\log Y(m)] \cos\left(\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right) \tag{4}$$

The first component $k = 0$ is excluded from the DCT since it represents the mean value of the input signal which carried little speaker specific information.

6. Extracting Delta and Delta-Delta Features

The dynamic information is typically incorporated by extending the static cepstral vectors by their first and second derivatives [5, 8]. computed as:

The MFCC feature vector has 26 components, including the first 12 cepstral coefficients, the log energy, as well as their first order time derivatives.

$$\Delta C_k = \frac{\sum_{t=-l}^{l} t c_{t+k}}{\sum_{t=-l}^{t} |t|} \tag{5}$$

$$\Delta\Delta C_k = \frac{\sum_{t=-1}^{1} t^2 c_{t+k}}{\sum_{t=-1}^{1} t^2} \quad (6)$$

2.2 The training stage (pattern generation)

For speaker recognition, pattern generation is the process of generating speaker specific models with collected data in the training stage. The most popular generative model used in speaker recognition is the Gaussian Mixture Models (GMM) [15] [16], as is described in more detail in next section.

2.3 The classification stage (pattern matching)

Pattern matching is the task of calculating the matching scores between the input feature vectors and the given models in recognition. This stage is described also in more detail in next section.

The rest of the paper is organized as follows: Section 2 presents the Gaussian Mixture Speaker Model (GMM) parameters. Section 3, we introduce a new technique speaker modeling based on GMM. Section 4 reports experimental results, and Section 5 gives overall conclusions on the work.

3. Implementation based on Gaussian mixture models (GMMs)

The GMM forms the basis for both the training and classification processes. The principle of GMM is to abstract a random process from the speech, then to establish a probability model for each speaker [1], [2]. A Gaussian Mixture density is a weighted sum of M component densities as shown in figure 3.

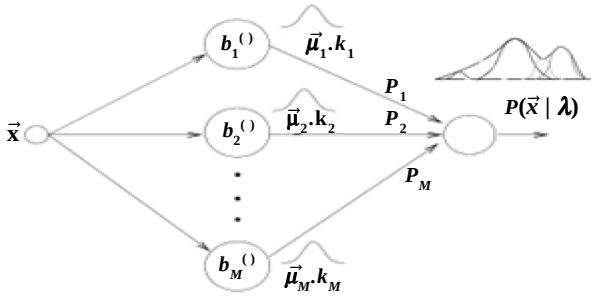

Figure 3. M probability densities forming a GMM

In the GMM model, the features distributions of the speech signal are modeled for each speaker as follows.

$$p(x \mid \lambda) = \sum_{i=1}^{M} p_i b_i(x) \quad (7)$$

Where

$$\sum_{i=1}^{M} p_i = 1 \quad (8)$$

x is a random vector of D-dimension, $p(x \mid \lambda)$ is the speaker model; p_i is the ith mixture weights; $b_i(x)$ is the i^{th} pdf component that is formed by the i^{th} mean μ_t and i^{th} covariance matrix, where $i = 1, 2, 3, \ldots, M$, and M is the number of GMM components [9], each density component is a D-variants Gaussian distribution given equation (10).

A statistical model for each speaker in the set is developed and denoted by λ. For instance, speaker s in the set of size S can be written as follows:

$$\lambda_s = \{p_i, \vec{\mu}_i, \Sigma_i\}, i = (1, \ldots M), s = 1, \ldots S, \quad (9)$$

$$b_i(\vec{x}) = \frac{1}{2\pi 2^{D/2} \mid K_i \mid^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)' K_i^{-1} (\vec{x} - \vec{\mu}_i)\right\} \quad (10)$$

3.1 ML Parameter Estimation Steps (training)

To obtain an optimum model representing each speaker we need to obtain a good estimation of the GMM parameters. To this end, the Maximum-Likelihood Estimation (ML) approach, which is a very efficient method, can be used; where for a given of T vectors used for training, $X = (x_1, x_2, ..., x_T)$, the likelihood of GMM can be written as.

$$p(X \mid \lambda_s) = \prod_{t=1}^{T} p(x_t \mid \lambda_s) \tag{11}$$

Since the GMM likelihood of the nonlinear function is impossible that maximizes directly, the ML estimations can be possible by using the EM algorithm iteratively [5], [15], [16].

The training phase consists of two steps, namely initialization and expectation maximization (EM). The initialization step provides initial estimates of the means for each Gaussian component in the GMM model. The EM algorithm recomputed the means, covariances, and weights of each component in the GMM iteratively. Each iteration of the algorithm provides increased accuracy in the estimates of all three parameters. The EM algorithm formulas [5, 16, 17] are the following:

• new estimates of i^{th} weight

$$\overline{p}_i = \frac{1}{T} \sum_{t=1}^{T} p(i \mid x_t, \lambda) \tag{12}$$

• new estimates of mean

$$\vec{\overline{\mu}}_i = \frac{\sum_{t=1}^{T} p(i \mid x_t, \lambda)\, x_t}{\sum_{t=1}^{T} p(i \mid x_t, \lambda)} \tag{13}$$

• New estimates of diagonal elements of i^{th} covariance matrix

$$\overline{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i \mid x_t, \lambda)\, x_t^2}{\sum_{t=1}^{T} p(i \mid x_t, \lambda)} - \overline{\mu}_i^2 \tag{14}$$

• where the likelihood a posteriori of the i^{th} class is given by posterior probability

$$p(i \mid \vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} \tag{15}$$

This process is repeated until convergence is achieved.

3.2 Classification based on GMM

In this stage, after the GMM models for each speaker are estimated, the target is to find the model with the maximum likelihood a posteriori for an observation sequence. The input to the classification system is denoted as

$$X = \{x_1, x_2, x_3, \ldots, x_T\} \tag{16}$$

The rule to determine if X has come from speaker s can be stated as

$$p(\lambda_s \mid X) > p(\lambda_r \mid X) \quad r = 1, 2, \ldots, S\,(r \neq s) \tag{17}$$

Therefore, for each speaker s in the speaker set, the classification system needs to compute and find the value of s that maximizes $p(\lambda_s \mid X)$ according to

$$\hat{S} = \arg \max_{1 \le s \le S} P(\lambda_s \mid X) = \arg \max_{1 \le s \le S} \frac{p(\lambda_s \mid X)\, Pr(\lambda_s)}{p(X)} \tag{18}$$

The classification is based on a comparison between the probabilities for each speaker. If it can be assumed that the prior probability of each speaker is equal, then the term of $p(\lambda_s)$ can be ignored. The term $p(X)$ can also be ignored as this value is the same for each speaker [1], so $p(\lambda_s \mid X) = p(X \mid \lambda_s)$,

Where

$$p(X \mid \lambda_s) = \prod_{t=1}^{T} p(x_t \mid \lambda_s) \quad (19)$$

Practically, the individual probabilities, $\prod_{t=1}^{T} p(x_t \mid \lambda_s)$, are typically in the range $10-3$ to $10-8$, the result $p(x_t \mid \lambda_s)$, will underflow probability for all speakers will be calculated as zero. Thus, $p(X \mid \lambda_s)$ is computed in the log domain in order to avoid this problem. The likelihood of any speaker having spoken the test data is then referred to as the log likelihood. The formula for the log likelihood function is [17].

The speaker of the test data is statistically chosen by

$$\hat{S} = \arg \max_{1 \le s \le S} p(X \mid \lambda_s) \xrightarrow{take\ \log} \hat{S} = \arg \max_{1 \le k \le S} \sum_{t=1}^{T} \log p(\vec{x}_t \mid \lambda_s) \quad (20)$$

4. System Configuration

4.1 Classical technique

In this section, we present in first stage the classical technique of training GMM implemented in our experiments where we fixed the number of Gaussian mixture (G) in the beginning of training stage to model the features extracted from each speaker's voice sample. As shown in figure 4, we chose G = 64 mixture. The result of this technique are detailed and compared with new technique in section 4.

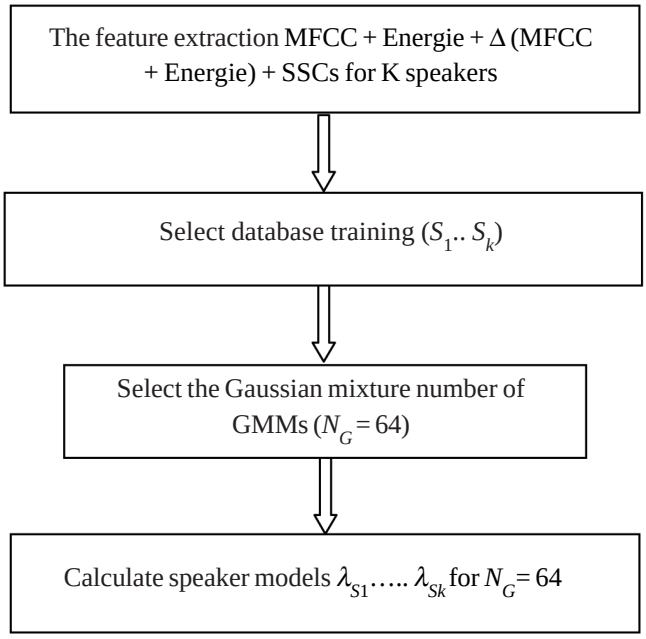

Figure 4. Classical technique to calculate speaker models block diagram

4.2 New techniques description

The new technique of using training GMM algorithm it presented in figure 5, by five steps. we give a short description of all the steps involved in this algorithm

1) For K speakers we fix the max number of Gaussian mixture N_{Gmax} (64 components in our case).

2) Divide a training database into two parts DATA1 and DATA2.

3) Calculate N_{Gmax} models (64 models in our case) for each speaker using first part of training database DATA1.

4) Test the models using the second part training database DATA2.

5) Select the model with the maximum likelihood a posteriori for an observation sequence.

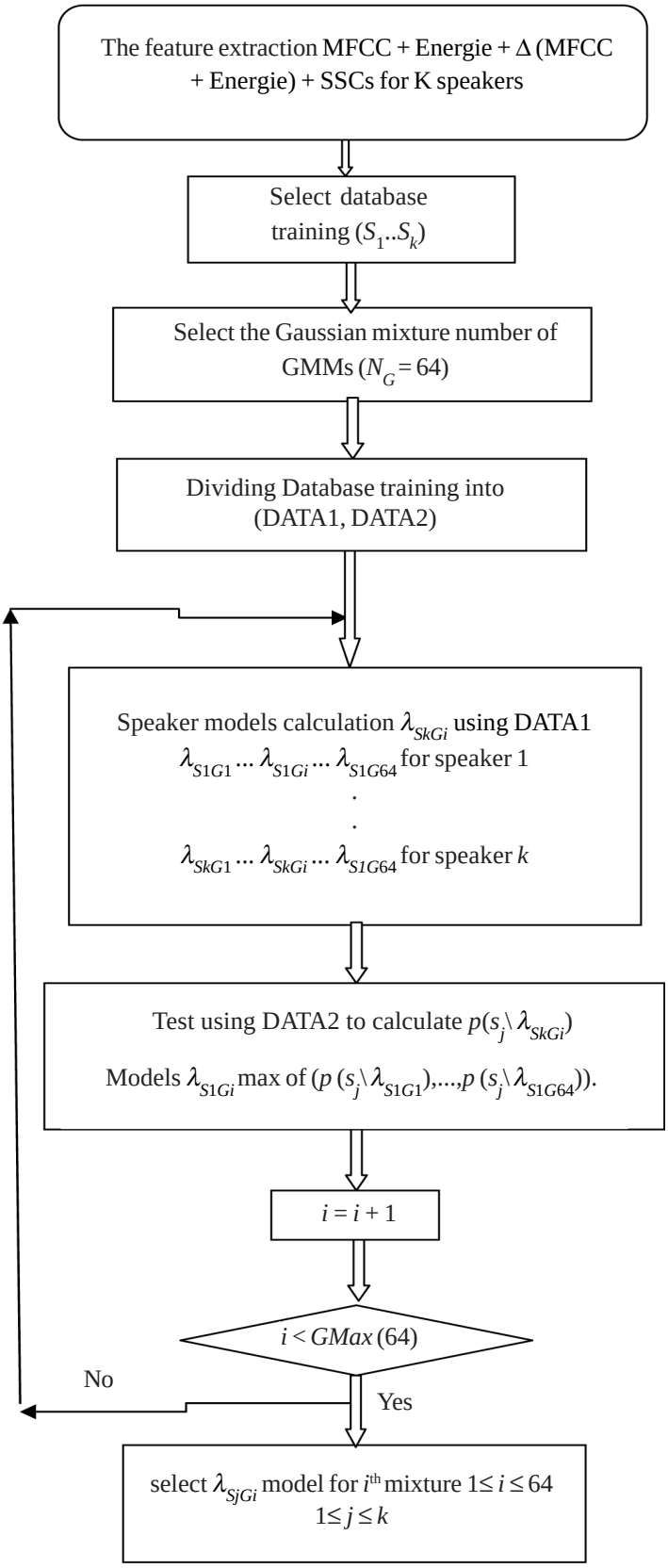

Figure 5. New technique to calculate speaker model block diagram

5. Results and discussion

Different stages of a speaker recognition system are designed and implemented using Matlab 6.5.

Although several experiments aimed for the selection of optimal parameters for the system using the database created in the national laboratory of automatic and signals in Annaba (LASA). From this database, we used 10 words to represent numbers from 0 to 9, ten repetition of each word, which are obtained from 11 speakers, 6 men and 5 women. The words have already been sampled at 11.025 kHz, and digitized with the resolution of 16 bits. End-point detection algorithm has been applied for each word. For MFCC coefficients. We calculate the vector feature extraction from 12 coefficients, complemented by an energy parameter and its delta coefficients complemented by 8 coefficients SSCs. For SSCs features we compute M SSCs using the unsmoothed (FFT) power spectrum. We divide the frequency band [0 Fs / 2]] into M equal-length disjoint subbands and employ a rectangular shape for each subband filter. The lower and higher edge frequencies of M subbands are given by $l_1 = 0$, $h_M = Fs/2$ and $l_{m+1} = h_m = m*Fs / (2 * M)$, for m = 1, 2,…., $M - 1$. We use a ~30 ms long segment with eigth equally-spaced bands, where $M = 8$, and $Fs = 11.025$ kHz

In the flowing we present the results of an experimental study aimed at finding the effect of different parameter value on the recognition performance.

A. **Training phase:** we build a training database BA

a) **Database BA:** containing 11 speakers (6 males and 5 females) each speaker recorded a 5 utterance of each digit the 0 to 4

B. **Testing phase:** we are construct a 3 testing databases BT1, BT2 and BT3

1) **TEST1:** mode text_dependent using BT1 Database

Database BT1: containing 11speakers (6 males and 5 females) each speaker recorded a 10 utterance of each digit the 0 to

2) **TEST2:** mode text_ independent using BT2 Database

Database BT2: containing 11speakers (6 males and 5 females) each speaker recorded a 10 utterance of each digit the 0 to

3) **TEST3:** mode text_independent using BT3 Database

Database BT3: containing 11speakers (6 males and 5 females) each speaker recorded a 5 utterance of each digit the 5 to

Table .1 and figure 6 presents the experimental results concerning the choice of best number of components GMM best_N_G for each speaker using new technique.

Table .2 and figure 7 summarized the results of comparative study between classic technique and a new method using the tree test databases BT1, BT2 and BT3:

We note that the performance of the new technique is superior to the classical technique that for either the text_independent or text_dependent identification. And the SSCs are potential complementary features to conventional features such as MFCCs.

speakers	1	2	3	4	5	6	7	8	9	10	11
best_N_G	61	55	60	40	56	42	63	57	58	64	63

Table 1. BEST number number of components GMM best_N_G

	GMM using MFCC	**BEST_GMM using MFCC**	**BEST_GMM using SSC+MFCC**
average rate % TEST1	95.81	99.09	99.85
average rate % TEST2	83.45	87.63	92.72
average rate % TEST3	71.09	76.18	83.01

Table 2. Comparative study between the classical method GMM and the new method best_GMM

6. Conclusion

To enhance the effectiveness of GMM-based speaker recognition systems we propose in this article a new technique that uses the GMM training algorithm to calculate the best number of mixture components for each speaker model. This number is fixed in the classical methods at the beginning of training phase for all speakers. For feature extraction module subband spectral centroids (SSCs) used as a complementary features with the traditional MFCC features, The comparison shows that the performances of the new technique is superior to the classical technique for either text independent or text dependent identification.

References

[1] Reynolds, D. A. (1995). Speaker Identification and Verification Using Gaussian Mixture Speaker Models. *Speech Commun*., 17 (1-2) 91-108, August.

[2] Kinnunen, T., Li, H. (2010). An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. *Speech Commun*., 52 (1) 12-40, Jan.

[3] Reynolds, Douglas, A., Quatieri, Thomas, F., Dunn, Robert, B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10, 19–41,

[4] Douglas, A., Reynolds, Richard, C., Rose. (1995). Robust text-independent speaker identification using Gaussian Mixture Speaker Models. *In*: IEEE Transactions on Speech and Audio Processing, 3, 72 – 83

[5] Campbell, J. P., Jr. (1997). Speaker recognition: A tutorial. *In*: Proc. IEEE, 85 (9) 1437-1462.

[6] Davis, Paul Mermelstein, Steven, B. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics*, *Speech and Signal Processing*, 28 (4) 357–366.

[7] Douglas, A., Reynolds. (2002). An Overview of Automatic Speaker Recognition Technology. MIT Lincoln Laboratory, Lexington, MA, USA, This paper appears in ICASSP 2002, p. 4072-4075.

[8] Wu, Z. J., Cao, Z. G. (2005). Improved MFCC-Based Feature for Robust Speaker Identification. TSINGHUA Science and Technology, 10, 158-161, Apr.

[9] Vergin, R., O'Shaughnessy, D., Farhat, A. (1999). Generalized mel frequency cepstral coefficients for large-vocabulary speakerindependent continuous-speech recognition. *IEEE Transactions on Speech and Audio Processing*, 7 (5) 525–532.

[10] Hassan, M., Jamil, M., Rabbani, M., Rahman, M. (2004). Speaker identification using Mel frequency cepstral coefficients. *In*: Proceedings of the 3rd International Conference on Electrical & Computer Engineering, p. 565–568.

[11] Paliwal, Kuldip, K. (1998) Spectral subband centroid features for speech recognition. *In*: Proc. ICASSP, May, 2, 6 17-620.

[12] Satoru Tsuge, Toshiaki Fukada, Harald Singer. (1999) Speaker normalized spectral subband parameters for noise robust speech recognition. *In*: Proc. ZCASSP, May.

[13] Dario Albesano, Renato De Mori, Roberto Gemello, Franco Mana. (1999). A study of the effect of adding new dimensions to trajectories in the acoustic space, *In*: Proc. EUROSPEECH, September, 4, 1503-1506.

[14] Bojana Gajiæ, Paliwal, Kuldip, K. (2001). Robust feature extraction using subband spectral centroid histograms. *In*: Proc. ICASSP, May, p. 85-88

[15] Auckenthaler, R. (2001) Test-independent speaker identification with limited resources. Ph.D. thesis, University of Wales.

[16] Reynolds, Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3 (1) 72–83.

[17] EhKan, Allen, T., Quigley, S. F. (2011). FPGA Implementation for GMMBased Speaker Identification. *International Journal of Reconfigurable ComputingVolume*, Article ID 420369, p. 8.