IJoFCS (2009) 1, 9-21

e-forensic
PRESS

The International Journal of
FORENSIC COMPUTER SCIENCE

www.IJoFCS.org

# Automatic Speaker Recognition with Multi-resolution Gaussian Mixture Models (MR-GMMs)

Frederico Q. D'Almeida(1), Francisco A. O. Nascimento(2), Pedro A. Berger(3), and Lúcio M. da Silva(4)

*(1) Brazilian Federal Police -Brazil*
*(2,4) Department of Electrical Engineering at University of Brasilia - Brazil*
*(3) Department of Computer Science at University of Brasilia - Brazil*

**Abstract -** Gaussian Mixture Models (GMMs) are the most widely used technique for voice modeling in automatic speaker recognition systems. In this paper, we introduce a variation of the traditional GMM approach that uses models with variable complexity (resolution). Termed Multi-resolution GMMs (MR-GMMs); this new approach yields more than a 50% reduction in the computational costs associated with proper speaker identification, as compared to the traditional GMM approach. We also explore the noise robustness of the new method by investigating MR-GMM performance under noisy audio conditions using a series of practical identification tests.

## 1. Introduction

Modern automatic speaker recognition (ASR) systems based on Gaussian Mixture Models (GMMs) have proven quite effective at identifying speakers given certain voice segments (Reynolds, 1992). However, high recognition rates require complex models with at least 16 components (Reynolds and Rose, 1995). If a noise-robust system were developed, then the complexity would likely exceed 80 components (D'Almeida *et al.,* 2008; Ming *et al.*, 2007) and carry a proportional increase in computational costs.

In this study, we present a Multi-resolution Gaussian Mixture Model (MR-GMM) speaker recognition technique in which each speaker is represented by at least two distinct models: a low resolution (low complexity) model used to conduct a pre-classification of the speakers, and a high resolution (high complexity) model used to achieve the final classification.

During the first identification stage, a large portion of the modeled speakers are eliminated by submitting the unknown voice signal to all low resolution models. Although such models are too simple to yield a certain match, the process does discard a great majority (up to 85%) of incorrect speakers. During the second identification stage, high resolution models are used to test the best-match results from the first stage. In this manner, calculations using high resolution (high complexity) models are only carried out for a small fraction of the overall number of speakers.

The final result is a two-stage identification process using models of varying levels of complexity (multi-resolution) that yields results similar to traditional GMM systems but at a considerably lower computational cost. This study also investigates the proposed model's sensitivity to noise by conducting simulations with different noise levels and comparing the results with traditional GMM methods.

## 2. Gaussian Mixture Model (GMM)

Gaussian Mixture Models are a useful data modeling tool when variables are distinctly clustered (Reynolds, 1992). This distribution is modeled as the weighted sum of *M* Gaussian distributions, each of dimension *D*, as given by

$$p(\vec{x}\mid\lambda)=\sum_{i=1}^{M}p_i b_i(\vec{x}). \qquad (1)$$

Here $\vec{x}$ is a *D*-dimensional parameter (variable) vector, $b_i(\vec{x})$ are the *M* Gaussian distributions comprising the model, and $p_i$ are the respective weights of each component. Note that *i* ranges from 1 to *M*. Each component of the GMM, $b_i(\vec{x})$, is a *D*-dimensional Gaussian given as

$$b_i(\vec{x})=\frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}}\exp\left\{-\frac{(\vec{x}-\vec{\mu}_i)'\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)}{2}\right\}, \qquad (2)$$

which has an average value of $\vec{\mu}_i$ and a covariance matrix $\Sigma_i$. The weights of the mixture components are appropriately normalized so that their sum total is unity.

In equation , λ represents the complete description of a GMM including its averages, weights and covariance matrices:

$$\lambda=\{p_i,\vec{\mu}_i,\Sigma_i\},i=1,\ldots,M\ . \qquad (3)$$

In ASR systems, the voice of each speaker is modeled by a different GMM which produces a model $\lambda_s$ with *s* ranging from 1 to the total number of modeled speakers *S*. The modeled universe of speakers is represented by *U:*

$$U=\{\lambda_s,s=1,\ldots,S\}. \qquad (4)$$

### 2.1. GMM Training

For GMM training, an audio file is required containing voice recordings of each speaker to be modeled (training files). For each training file, various parameter vectors $\vec{x}_t$ are calculated for different instances in time *t*. The set of these parameter vectors extracted from the training files of a particular speaker is represented by

$$X_s=\{\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_T\}. \qquad (5)$$

Note that the speaker index *s* in equation  has been removed from the right side of the equation for clarity.

The aim of the GMM training is to adjust the parameters of the model of speaker, $\lambda_s$, in order to maximize the probability of occurrence of the set of parameter vectors $X_s$. To simplify the problem, it is assumed that each parameter vector $\vec{x}_t$ is independent of the others, allowing the following notation:

$$p(X_s\mid\lambda_s)=\prod_{t=1}^{T}p(\vec{x}_t\mid\lambda_s). \qquad (6)$$

This is a non-linear function of the parameters of model $\lambda_s$, which does not allow direct maximization. Generally, maximization of  is performed with the *Expectation-Maximization* (EM) algorithm as described by Dempster *et al.* (1977).

### 2.2. Speaker Identification

To identify the speaker belonging to the test voice file, one must determine which of the models $\lambda_s$, of universe *U*, present the greatest *a posteriori* probability of a set of parameters calculated from the given test file. That is,

$$\tilde{s}=\arg\max_{\lambda_s\in U}p(\lambda_s\mid Y)=\arg\max_{\lambda_s\in U}\frac{p(Y\mid\lambda_s)p(\lambda_s)}{p(Y)}, \qquad (7)$$

where *Y* is the set of parameter vectors calculated from the test file, $Y = \{\vec{y}_1, \vec{y}_2, ..., \vec{y}_T\}$, and Bayes' rule is applied.

Assuming that all speakers are equally probable, $p(\lambda_s) = cte., s = 1, ..., S$, and considering that *p(Y)* is a constant depending solely on the tested recording (and therefore the same for all speakers in the universe), then identifying the speaker is as simple as calculating

$$\tilde{s} = \arg\max_{\lambda_s \in Y} p(Y \mid \lambda_s). \tag{8}$$

If one also assumes independence among the elements of the test parameter vector, as formulated in  for the training parameters, and maximize the logarithm of the probability instead, then equation  becomes

$$\tilde{s} = \arg\max_{\lambda_s \in U} \sum_{t=1}^{T} \log p(\vec{y}_t \mid \lambda_s). \tag{9}$$

Note that using the logarithm helps to avoid numerical problems since the probabilities involved in equation  are extremely small.

Since the length of times for each speaker's audio file are not exactly equal, equation  is normalized with respect to time as

$$\tilde{s} = \arg\max_{\lambda_s \in U} \frac{\sum_{t=1}^{T} \log p(\vec{y}_t \mid \lambda_s)}{T}. \tag{10}$$

It is assumed that correct identification takes place when the speaker who maximizes equation , $\tilde{s}$, is in fact the correct speaker $\hat{s}$ :

$$\tilde{s} = \hat{s}. \tag{11}$$

## 2.3. Computational Cost

The computational cost associated with identifying the speaker is a function of several factors. For example, it depends on the number of speakers in the universe *S,* since all models need to be simulated to find the specific model that maximizes equation ; the duration of the test file used since the voice parameter vectors $\vec{y}_t$ are extracted at fixed time intervals; the dimension of the Gaussian components *D* used in the model; and the number of components *M* of the models

All parameters were studied when minimizing the computational cost associated with speaker identification except the number of speakers in the universe, which cannot be altered for a given application (Reynolds and Rose, 1995). For the number *M* of model components, Reynolds and Rose (1995) determined that at least 8 to 16 components are required for good system performance with noiseless audio; a result confirmed during this study as well. When developing multi-conditional systems robust to noise, the minimum number of components increases to between 64 and 128 according to D'Almeida *et al.* (2008) and Ming *et al.* (2007).

From the definition of the GMM in equation , the total computational cost *W* of an identification task (during the test phase only) is approximately proportional to the number of Gaussians *M* in the speaker models $\lambda$. In reality, for each new component introduced into the models, it is necessary to calculate a new Gaussian defined by equation , multiply it by the coefficient of the mixture, and then add the new product to the renormalization sum in equation . The computational costs for temporal normalization and maximum identification (the *argmax* function) expressed in equation  are independent of the number of model components. These costs are of little overall significance since the calculations are executed only once for each speaker and test file, as compared to the calculations for equations and  which are executed once for each parameter vector $y_i$. Thus, even for short test files lasting only 1 second, there are still 50 calculations□ carried out for equations  and  when evaluating equation . The cost for calculating the logarithm is also independent of the number of model components, and even though it must be performed for each parameter vector $y_i$, it is still of little relevance since it consists of a single scalar operation.

We also consider the cost of extracting the parameters $y_i$ from the questioned audio file. The parameters which are most frequently used in ASR systems are the Mel Frequency Cepstral Coefficients (MFCC) which offer the best identification performance (D'Almeida and Nascimento, 2006). In this case, FFT and other calculations are needed, which carry relatively high computational complexities. However, since this calculation is carried out just once for the whole procedure, then this cost will not be as significant in terms of the total identification effort with a sufficiently large speaker universe.

Therefore, under certain conditions normally met by ASR systems, it can be stated with certain precision that the identification cost for GMM models is proportional to the quantity of model components:

$$W_{GMM} \propto M \quad . \tag{12}$$

## 3. Multi-Resolution Gaussian Mixture Models (MR-GMM)

There have been several experiments aimed at minimizing the number of GMM components without compromising speaker identification (e.g., Reynolds and Rose, 1995), but one issue that has not been fully addressed in the literature is how optimizing models with less than 16 components (for noiseless audio) significantly reduces the performance of the system in terms of the number of correctly identified speakers. However, in this situation, it is expected that the correct model still receives one of the highest scores (on equation (10) classification). Thus, modeling with a fewer number of components may not be capable of exactly determining the correct speaker, but must be capable of separating, within the universe, a subgroup containing the correct speaker.

To take advantage of this idea, loosening the success condition might allow high positive identification rates even for models with only 2 or 4 components. This can be expressed mathematically as:

$$\hat{s} \in \tilde{U} , \tag{13}$$

where $\tilde{U}$ is a subset of the speaker universe $U$ given by

$$\tilde{U} = \left\{ \lambda_s \in U, \left[ \frac{\sum_{t=1}^{T} \log p(\vec{y}_t \mid \lambda_s)}{T} \right] \geq \xi \right\}. \tag{14}$$

Here ξ is the $C$-th greatest value of the expression between brackets in equation  calculated for all models in the universe. The value of $C$ is defined based on the model order used and the required performance.

Of course, loosening the system's success condition as described by equations  and  leads to a new problem: the resulting identification is no longer a single speaker, but instead a set of $C$ speakers. Naturally, this does not produce a precise speaker identification. However, the precise identification can then be determined through a new GMM system, now with 16 components, and using a standard identification process as expressed in equation  and applied to the restricted set of speakers $\tilde{U}$.

In this study, we propose a systematic speaker identification process in successive stages through GMM models of various resolutions (number of components). This new modeling approach is termed Multi-resolution Gaussian Mixture Models (MR-GMMs).

### 3.1. Mathematic Formulation and Training

The MR-GMMs are essentially extensions of the traditional GMM approach. The main difference between these two techniques is that, for a given speaker, the MR-GMM approach has two or more distinct GMM models with different degrees of complexity (number of components). Analogous to equation (3), the MR-GMM approach may be formulated as

$$\Lambda = \left\{\lambda_k, k = 1, \ldots, K\right\} = \left\{ \begin{bmatrix} \left\{p_{1,i_1}, \vec{\mu}_{1,i_1}, \Sigma_{1,i_1}\right\} \\ \left\{p_{2,i_2}, \vec{\mu}_{2,i_2}, \Sigma_{2,i_2}\right\} \\ \vdots \\ \left\{p_{K,i_K}, \vec{\mu}_{K,i_K}, \Sigma_{K,i_K}\right\} \end{bmatrix}, \begin{matrix} i_1 = 1, \ldots, M_1; \\ i_2 = 1, \ldots, M_2; \\ \vdots \\ i_K = 1, \ldots, M_K \end{matrix} \right. , \qquad (15)$$

where $M_k > M_{k-1}$. Note that the subscript $k$ for model $\lambda_k$ does not index different speakers as index $s$ does in equation . Rather, $k$ refers to the submodels comprising a single MR-GMM. Consequently, all models $\lambda_k$ are of a single speaker. The set of all MR-GMM models or the universe of modeled speakers $U$ is defined similar to equation as

$$U = \left\{\Lambda_s, s = 1, \ldots, S\right\}. \qquad (16)$$

The training for each submodel $\lambda_k$ of a single MR-GMM model $\Lambda$ is carried out as the training of a normal and independent GMM model. The different submodels of a single speaker may be trained from the same audio segment; this has no effect on the global model since the aim of the MR-GMM is to use models of different resolutions to minimize the overall computational cost during speaker identification. In reality, training the submodels with the same audio segment is a more natural alternative since it eliminates the need to alter the existing databases.

### 3.2. Speaker Identification

There is a significant difference between the MR-GMM and GMM approaches during the speaker identification phase (test phase). For GMMs, the model of each speaker is evaluated according to the tested audio segment in order to find the most likely match according to equation . On the other hand, the MR-GMM approach does not conduct the test in a single evaluation. Successive test stages are carried out, each using a model with a resolution greater than its predecessor, and speakers are gradually selected until the best candidate is determined in the final stage.

The fundamental idea behind MR-GMMs is that they can reduce the computational cost of identifying the speaker by reducing the average complexity of the models used while not sacrificing overall performance. This requires a gradual speaker selection process using models of increasing complexity, using high-complexity models only for a limited number of speakers in the universe.

During the first test phase, less complex models for each speaker, $\Lambda_{s,1}$, are first used to select the $C_1$ models from universe $U$ which best matches the test audio. Note that we use $\Lambda_{s,k}$ as the GMM submodel $\lambda_k$ associated with the MR-GMM $\Lambda_s$. The result of the first test phase is a subset of the speaker universe, denoted as $U_1$, containing $C_1$ models of the best-matching speakers at low resolution:

$$U_1 = \left\{ \Lambda_s \in U, \left[ \frac{\sum_{t=1}^{T} \log p\left(\vec{x}_t \mid \Lambda_{s,1}\right)}{T} \right] \geq \xi_1 \right\}, \qquad (17)$$

where $\xi_1$ is the $C_1$-th greatest value of the expression between brackets in equation , evaluated for all $\Lambda_{s,1}$ models in $U$.

During the second phase of the test, the initial procedure is repeated but now with models with the second lowest resolution $\Lambda_{s,2}$ in universe $U_1$ from the previous stage. In this phase, the result is a subset $U_2 \subset U_1$ given by

$$U_2 = \left\{ \Lambda_s \in U_1, \left[ \frac{\sum_{t=1}^{T} \log p\left(\vec{x}_t \mid \Lambda_{s,2}\right)}{T} \right] \geq \xi_2 \right\}, \qquad (18)$$

where $\xi_2$ is the $C_2$-th greatest value of the expression in brackets evaluated for models $\Lambda_{s,2}$, $s \in U_1$.

The testing process gradually reduces the speaker universe according to

$$U_{k+1} = \left\{ \Lambda_s \in U_k, \left[ \frac{\sum_{t=1}^{T} \log p\left(\overline{x}_t \mid \Lambda_{s,k+1}\right)}{T} \right] \geq \xi_{k+1} \right\}, \quad (19)$$

until the last stage *K* where the model that best adjusts to the audio test (thus, where $C_K$ is always equal to 1) is determined by the expression analogous to equ

$$\hat{S} = U_K = \left\{ \Lambda_s \in U_{k-1}, \frac{\sum_{t=1}^{T} \log p\left(\overline{x}_t \mid \Lambda_{s,K}\right)}{T} \geq \xi_K \right\} = \arg \max_{s \in U_{k-1}} \frac{\sum_{t=1}^{T} \log p\left(\overline{x}_t \mid \Lambda_{s,K}\right.}{T}. \quad (20)$$

### 3.3. Computational Advantage

The computational advantage of using MR-GMM is from the possibility of reducing the average complexity of the models used during speaker identification. Thus, it is essential to determine the parameters $M_k$, the number of components of each of the submodels $\Lambda_{s,k}$, and the quantity of speakers classified to the next stage $C_k$.

The total computational cost for speaker identification with MR-GMM is given by

$$W_{MR-GMM} \propto M_1 + M_2 \frac{C_1}{S} + \ldots + M_K \frac{C_{K-1}}{S} = \sum_{k=1}^{K} M_k \frac{C_{k-}}{S}, \quad (21)$$

where, for simplicity, we define $C_0=S$, indicating that all speakers are considered in the first evaluation. Note that since more than one test is conducted for each speaker (speakers tested and classified in stage *k* are tested again in stage *k+1*), a poor choice of $M_k$ and $C_k$ may cause an increase in the total computational cost compared to traditional GMM models. For example, assuming a MR-GMM model with only two resolutions $M_1=8$ and $M_2=16$ (that is, the first phase of the test is conducted with an 8-component model, and the second and last phase is conducted with a 16–component model) and with $C_1=S/2$ (50% of the best models in the universe pass on to the second training phase), then the total cost *W* of the process is

$$W_{MR-GMM} \propto M_1 + M_2 \frac{C_1}{S} = 16 = M_2 = W_{GMM}, \quad (22)$$

and thus there would be no reduction in the computational cost over a 16-component GMM model.

In addition, as long as the values of $M_k$ and $C_k$ are adequately adjusted, one expects that a MR-GMM model conducting its last cycle of tests comprised of $M_K$ components would have a speaker identification performance equivalent to a GMM model with $M_K$ components. For this reason, computational cost comparisons must be made using a GMM model of this complexity.

In order to effectively reduce the total computational cost of the process, parameters $M_k$ and $C_k$ must be selected such that

$$W_{MR-GMM} \propto \sum_{k=1}^{K} M_k \frac{C_{k-1}}{S} < M_K = W_{GMM}. \quad (23)$$

The relative reduction of the computational cost of the identification process can then be calculated as

$$G = 1 - \frac{W_{MR-GMM}}{W_{GMM}} = 1 - \frac{\sum_{k=1}^{K} M_k \frac{C_{k-1}}{S}}{M_K} = 1 - \sum_{k=1}^{K} \frac{M_k}{M_K} \frac{C_{k-1}}{S}. \quad (24)$$

## 4. Performance Evaluation and Results

Simulations used to analyze the performance of MR-GMM models were conducted using the voice database described by D'Almeida *et al.* (2007). The characteristics are this dataset are described below.

### 4.1. Description of the Audio Database

The voice database consisted of 30 different speakers (*S* = 30), half male and half female. Each speaker was recorded while reading identical pre-defined texts, and each recording was broken into 21 files with the same starting and ending points for all speakers. Thus 21 files were generated for each speaker, indicated by $A_{n,s}$ where $s$ indicates the speaker and $n$ indicates the segment of the recorded file. The first segment of the 21 audio clips for each speaker was used to train the MR-GMM models, and the remaining 20 segments were used to carry out the 20 different identification tests.

All recordings were made in acoustically prepared environments with professional microphones and audio capture cards. Files were acquired at a sampling rate of 22 kHz, 16-bit quantization, in monaural mode. From this initial database, versions of the audio files were generated with other sampling frequencies and codifications: 16 kHz/16 bits, 11 kHz/16 bits, 11 kHz/8 bits-μ-law, 8 kHz/8 bits-μ-law, and 8 kHz/8 bits. Identification tests were conducted using all of the above file versions.

Before carrying out the tests, all audio files were normalized such that their peak amplitude corresponded to 100% of the maximum quantization value. Silent segments were then excluded from the files by an automatic silence detector based on the signal energy measured in 20 ms windows using a 15 ms window overlap (thus 5 ms increments) and a manually defined silence threshold (a single value for all files) based on practical tests.

For all simulations, MFCC parameters were used as the modeling parameters since this has produced the best results in ASR applications (D'Almeida and Nascimento, 2006). The parameters were calculated for each 20 ms audio window with no overlapping, using filter banks applied directly to the signal frequency spectrum calculated in the same window. From each window, 12 MFCC parameters were extracted. Parameter normalization was performed for both the model and test phases as outlined in Reynolds and Rose (1995), as a way of increasing system performance (removal of the average values of the coefficients).

## 4.2. Test Procedure and Results

Tests were carried out by comparing the performances (number of correct identifications) of a traditional 16-component GMM system ($W_{GMM} = 16$) with three MR-GMM systems. All MR-GMM models were constructed with only two identification stages (resolution) ($K = 2$), given the relatively small speaker universe available in the database (S = 30). Diagonal covariance matrixes $\Sigma_i$ were used for all models since Reynolds and Rose (1995) demonstrated that this has no negative effect on the overall performance.

The first MR-GMM model, called MR-GMM 1, used parameters $C_0=30$, $C_1=5$, $M_1=6$, and $M_2=16$. The computational cost of this model calculated using equation  is

$$W_{MR-GMM\,1} \propto \sum_{k=1}^{K} M_k \frac{C_{k-1}}{S} = 6 + 16\frac{5}{30} = 8.67 \quad (25)$$

The second MR-GMM model, called MR-GMM 2, used parameters $C_0=30$, $C_1=5$, $M_1=4$, and $M_2=16$ and produced a computational cost of

$$W_{MR-GMM\,2} \propto \sum_{k=1}^{K} M_k \frac{C_{k-1}}{S} = 4 + 16\frac{5}{30} = 6.67 \quad (26)$$

The third MR-GMM model, called MR-GMM 3, used parameters $C_0=30$, $C_1=5$, $M_1=2$, and $M_2=16$ and resulted in a computational cost of

$$W_{MR-GMM\,3} \propto \sum_{k=1}^{K} M_k \frac{C_{k-1}}{S} = 2 + 16\frac{5}{30} = 4.67 \quad (27)$$

The reductions in the computational costs of the MR-GMM models, as calculated by equation , are 45.81%, 58.33% and 70.83%, respectively.

To carry out the comparisons in similar conditions and exclude affects other than the difference in modeling techniques, the submodels of the 16-component MR-GMM ($\Lambda_{s,2}$) were the same for all MR-GMM models and were also used as the GMM models of each speaker ($\lambda_s$). Thus, it was possible to eliminate performance differences from the training of the models so that the alterations in the correct identification rates only reflected the difference between the MR-GMM modeling techniques using progressive identification stages versus the traditional GMM technique.

The results of the tests are presented in Table 1. Note that the MR-GMM 1 and MR-GMM 2 models gave results that closely matched the traditional GMM model, while MR-GMM 3 suffered from significant performance loss. Thus MR-GMM models can provide a computational cost reduction of up to 58% with no relevant losses in system performance for noiseless audio. Also note that for the MR-GMM 1 and MR-GMM 2 models, significant performance losses were only observed for the 8 kHz audio sampling frequency and 8 bit codification. For all other cases, there was no significant performance alteration since the maximum difference in correct identification rates was limited to 0.3 percentage points for the MR-GMM 1 model and to 0.5 percentage points for the MR-GMM 2 model.

### 4.3. Robustness to noise

To analyze the noise-sensitivity of the MR-GMM models, simulations were performed using noiseless audio samples and with the same samples after adding various levels of noise. In these simulations, only the MR-GMM 1 and MR-GMM 2 models were used since the performance of the MR-GMM 3 model was considered unsatisfactory even for noiseless audio.

The noisy audio samples were generated from the “noiseless” audio (the originally captured files) by adding white noise to the files. Even though the noise presence was simulated, practical tests with ASR systems carried out by Ming *et al.* (2007) confirmed that the computational addition of noise is quite similar to the physical (acoustic) addition of noise present during the moment of capture.

The values of the signal-to-noise ratios (SNRs) used in the simulations were 60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 10 dB, 8 dB and 5 dB. The 60 db SNR value corresponds to the audio acquisition system’s intrinsic SNR; this value was estimated from the average energy of the signal during the moments of silence (absence of voice) and during speaking. Thus, no additional noise was inserted for the 60 dB SNR audio. For the remaining cases, noise addition was performed so as to maintain a particular average SNR over the entire audio segment. This was done by calculating the average energy of the signal, $E_s$, in the audio sample according to

$$E_s = \sum_m y_s^2[m], \qquad (28)$$

where $y_s$ is the audio signal vector and $m$ the temporal index of the samples. A noise vector $y_n$ was then generated containing samples with zero mean and Gaussian-distributed amplitudes, and having the same dimension of the signal vector $y_s$. The energy of the noise vector was then calculated as

$$E_n = \sum_m y_n^2[m]. \qquad (29)$$

Subsequently, the amplitudes of the noise vector were adjusted so as to obtain the desired SNR, that is

$$y'_n = y_n.SNR.\sqrt{\frac{E_s}{E_n}}. \qquad (30)$$

Finally, the noise vector with adjusted amplitudes was added to the original signal vector, generating a audio vector $y$ used in the analysis:

$$y = y_s + y'_n. \qquad (31)$$

For the noise robustness analyses, only the following sampling frequencies and codifications were used: 22 kHz / 16 bits, 11 kHz / 16 bits, 8 kHz / 8 bits μ-law, and 8 kHz / 8 bits linear.

It must be pointed out that calculating the noiseless audio energy was done after removing the silent segments, since this removal was part of the signal pre-processing. Thus, the average power (total energy per total time) calculated from this signal is greater than it would be if the entire signal (including the silent excerpts) were considered. Consequently, to obtain an established SNR, the average power of the noise signal to be added to the signal, $y'_n$, is likewise greater than it would be for the entire audio file (including silent segments). For the conducted tests, it was found that the SNR values in the present study are equal, on average, to values nearly 10% greater than those that would be obtained if the noise addition process were conducted on the entire audio file.

Note that the SNR measurement methodology used in this study was chosen for several reasons. First, since the speaking rhythm and pause intervals between words and sentences varied according to each individual, the measurement of the global SNR would depend on these periods of silence. Thus, different SNR values would be obtained even if the average power of the signal (in the segments with voice) and the noise remained fixed. Second, the detection of the speech and silence segments is much simpler for noiseless audio files for which a simple energy detector may be used.

The MR-GMM results obtained using the noise-augmented files were compared to results obtained from traditional GMM results using the same files. Just as in the noiseless audio analysis, the submodels of the 16-component MR-GMM, $\Lambda_{s,2}$, were used as GMM models of each speaker, $\lambda_s$, in order to eliminate performance differences resulting from the training of the models. The results of the tests conducted with models MR-GMM 1 and MR-GMM 2 are organized in Tables 2 through 5 below. Visually summaries of Tables 2 through 5 are presented in Figures 1 through 4 as well.

For the 22 kHz audio files, the performance differences between the MR-GMM and GMM approaches were extremely small. For MR-GMM 2, this difference was limited to 2.2 percentage points for the worst case and had an average difference of 0.7 percentage points. For the MR-GMM 1 model, the differences were limited to 0.2 percentage points for the worst case and the model performed on average as well as the traditional GMM approach.

For the 11 kHz audio files, the MR-GMM 2 model again had a worst-case performance degradation of 2.2 percentage points and an average degradation of 0.4 percentage points. The MR-GMM 1 model had no degradation in this simulation.

Simulations using the 8 kHz μ-law audio indicate that the MR-GMM 2 approach has a degradation limited to 1.2 percentage points and an average difference of 0.4 percentage points. The MR-GMM 1 model has a maximum degradation of 0.5 percentage points but performed on the average 0.2 percentage points better than the traditional GMM approach.

Results for the 8 kHz 8-bit linear audio files show a maximum loss of 3.2 percentage points and an average loss of 0.6 percentage points for the MR-GMM 2 model, and a maximum of 1.2 percentage points and average of 0.1 percentage points for the MR-GMM 1 model.

## 5. Conclusion

The Multi-resolution Gaussian Mixture Models, or MR-GMMs, proposed in this study proved to be an effective alternative for developing high–performance automatic speaker recognition systems with significant reductions in computational costs over traditional Gaussian Mixture Models. Our simulations indicate that cost reductions of up to 58.3% for noiseless audio are possible with no significant degradation in the speaker identification performance. For noisy audio, depending on the sampling frequency and type of codification, reductions of as much as 45% to 58% are possible with no relevant losses in the identification rates as compared to the traditional GMM approach. Note that these results hinge on a relatively small sample database of 30 speakers, thus the use of MR-GMM models with more than two identification stages was not feasible. This may have limited the gain in computational cost, and larger databases using more identification stages may yield further improvements in computational efficiencies.

## References

[1] D'Almeida. F.Q. e Nascimento. F.A.O. (2006). Comparação de Desempenho de Parâm. da Fala em Sist. de Rec. Auto. de Locutor. Congresso Brasileiro de Automática – CBA 2006.

[2] D'Almeida. F.Q.. Nascimento. F.A.O. Berger. P.A.. da Silva. L. M. (2007). Efeitos da Codificação MP3 em Sist. de Rec. Auto. de Locutor via GMM. XXV Simpósio Brasileiro de Telecomunicações – SBrT 2007.

[3] Dempster. A.. Laird. N. e Rubin. D. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal Royal Statistical Society. Vol. 39. pp. 1-38.

[4] Ming. J.. Hazen. T.. Glass. J.R. e Reynolds. D.A. (2007). Robust Speaker Recognition in Noisy Conditions. IEEE Trans. on Audio. Speech and Lang. Proc.. vol. 15. pp. 1711-1723.

[5] Reynolds. D.A. (1992). A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification. Ph. D. Thesis. Georgia Inst. of Tech.

[6] Reynolds. D.A. e Rose. R.C. (1995). Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Trans. Speech and Audio Proc.. Vol. 3. no. 1. pp 72-83.

## Table 1

Correct identification rates (%) and computational cost reduction

| Sampling Freq. | Model | | | |
|---|---|---|---|---|
| **Codification** | **GMM** | **MR-GMM 1** | **MR-GMM 2** | **MR-GMM 3** |
| **22 kHz / 16 bits** | 100.0 | 100.0 | 100.0 | 99.2 |
| **16 kHz / 16 bits** | 100.0 | 100.0 | 100.0 | 99.0 |
| **11 kHz / 16 bits** | 99.8 | 99.5 | 99.3 | 91.7 |
| **11 kHz / 8 bits μ-law** | 95.7 | 95.7 | 95.2 | 91.0 |
| **8 kHz / 8 bits μ-law** | 96.2 | 96.2 | 96.2 | 93.8 |
| **8 kHz / 8 bits** | 98.7 | 97.5 | 96.8 | 89.8 |
| **Cost reduction (%)** | - | 45.8 | 58.3 | 70.8 |

## Table 2

Correct identification rates (%) for 22 kHz / 16 bits audio

| Noise Test (dB) | Model GMM | MR-GMM-1 | MR-GMM-2 |
|---|---|---|---|
| **60** | 100.0 | 100.0 | 100.0 |
| **50** | 100.0 | 100.0 | 100.0 |
| **40** | 99.3 | 99.3 | 99.3 |
| **30** | 93.7 | 93.7 | 93.7 |
| **26** | 83.0 | 83.2 | 81.8 |
| **20** | 54.7 | 54.7 | 52.5 |
| **16** | 28.5 | 28.8 | 26.5 |
| **10** | 13.8 | 13.7 | 13.2 |
| **8** | 12.0 | 12.0 | 11.3 |
| **5** | 10.0 | 10.0 | 10.2 |

**Table 3**
Correct identification rates (%) for 11 kHz / 16 bits audio

| Noise Test (dB) | Model | | |
|---|---|---|---|
| | GMM | MR-GMM-1 | MR-GMM-2 |
| **60** | 99.8 | 99.8 | 99.8 |
| **50** | 99.8 | 99.8 | 99.8 |
| **40** | 98.5 | 98.5 | 98.5 |
| **30** | 94.3 | 94.3 | 92.2 |
| **26** | 83.5 | 83.5 | 83.5 |
| **20** | 53.7 | 53.8 | 52.0 |
| **16** | 30.2 | 30.2 | 30.2 |
| **10** | 10.7 | 10.7 | 10.7 |
| **8** | 6.0 | 6.0 | 6.0 |
| **5** | 2.8 | 2.8 | 2.8 |

**Table 4**
Correct identification rates (%) for 8 kHz / 8 bits μ-law audio

| Noise Test (dB) | Model | | |
|---|---|---|---|
| | GMM | MR-GMM-1 | MR-GMM-2 |
| **60** | 99.0 | 99.0 | 98.5 |
| **50** | 99.2 | 99.2 | 98.0 |
| **40** | 98.2 | 98.2 | 97.5 |
| **30** | 93.3 | 93.5 | 92.2 |
| **26** | 81.5 | 81.0 | 80.5 |
| **20** | 48.3 | 48.8 | 48.5 |
| **16** | 30.2 | 31.2 | 30.7 |
| **10** | 14.0 | 13.8 | 13.2 |
| **8** | 9.3 | 9.7 | 9.3 |
| **5** | 5.8 | 6.3 | 6.0 |

**Table 5**
Correct identification rates (%) for 8 kHz / 8 bits

| Noise Test (dB) | Model | | |
|---|---|---|---|
| | GMM | MR-GMM-1 | MR-GMM-2 |
| **60** | 95.2 | 95.2 | 95.2 |
| **50** | 96.2 | 96.2 | 96.2 |
| **40** | 97.2 | 97.2 | 97.2 |
| **30** | 96.5 | 96.5 | 96.0 |
| **26** | 93.5 | 93.5 | 93.2 |
| **20** | 71.7 | 71.7 | 69.3 |
| **16** | 46.2 | 45.0 | 43.0 |
| **10** | 11.5 | 11.5 | 11.7 |
| **8** | 8.3 | 8.5 | 8.5 |
| **5** | 5.7 | 5.7 | 5.3 |

**Figure 1**
Correct identification rates
for 22 kHz / 16 bits audio

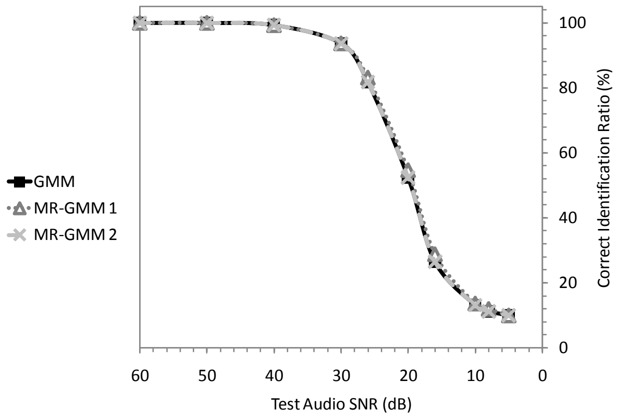


**Figure 3**
Correct identification rates
for 8 kHz / 8 bits μ-law audio

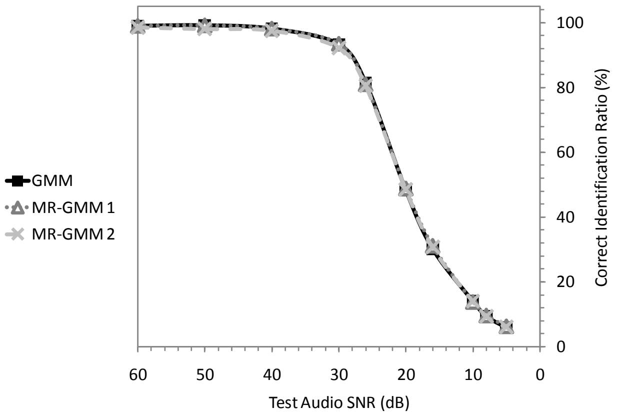


**Figure 2**
Correct identification rates
for 11 kHz / 16 bits audio

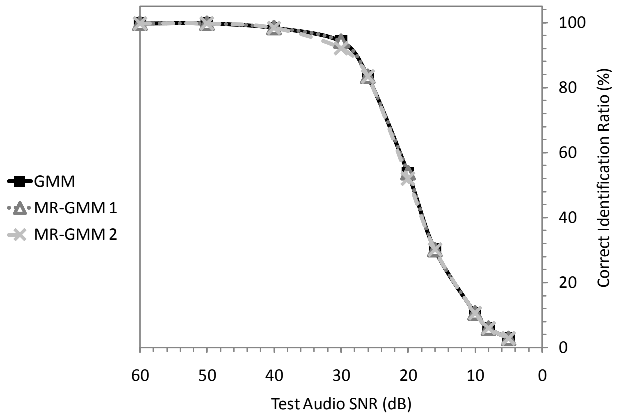
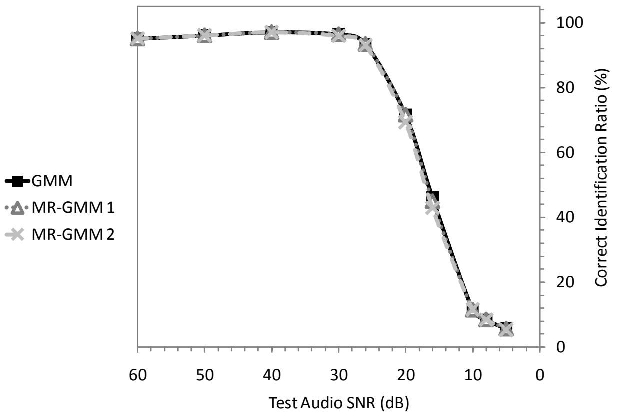


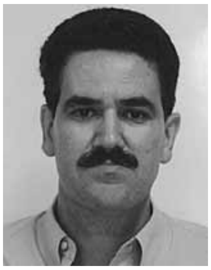
**Figure 4**
Correct identification rates
for 8 kHz / 8 bits



**F.Q.D'Almeida.** He was born in Salvador-BA, Brazil, on January 24, 1978. He graduated in Electrical Engineering, Federal University of Bahia - UFBA, Salvador-BA, Brazil, 2000, got his Master Degree in Electrical Engineering, UFBA, 2003, and graduated in Physics, University of Brasilia - UnB, 2006. He is pursuing his Doctorate Degree in Electrical Engineering at UnB. His field of study is Automatic Speaker Recognition. He also works as a forensic expert at Brazilian Federal Police.

**F. Assis Nascimento** received his B.Sc. in Electrical Engineering from the University of Brasilia in 1982, his M.Sc. in Electrical Engineering from the Federal University of Rio de Janeiro (UFRJ), in 1985, and his Ph.D. in Electrical Engineering from UFRJ in 1988. Currently, he is an Associate Professor at the University of Brasilia and a coordinator of the GPDS (Grupo de Processamento Digital de Sinais).

**Pedro de A. Berger** graduated in Electrical Engineering at Federal University of Ceara - UFC, Fortaleza-CE, Brazil in 1999, earned his M.Sc. in Electrical Engineering at the University of Brasília (UnB) in 2002 and his Ph.D. at UnB in 2006. He has been a professor in the Department of Computer Science at UnB since 2006. His field of study includes digital signal processing, artificial neural networks and biomedical engineering.



**Lúcio M. da Silva** was born in Delfinópolis, MG, Brazil, on April 27, 1958. He received the B.S. in electrical engineering from Pontifical Catholic University of Minas Gerais, in 1981, the M.S. degree from University of Brasilia, in 1989, and the Ph. D. degree from Pontifical Catholic University of Rio de Janeiro, in 1996. He is with the Electrical Engineering Department of University of Brasilia. He is involved in teaching and research activities in speech signal processing and digital transmission systems.