TEXT INDEPENDENT SPEAKER VERIFICATION USING GMM

Charles B. de Lima[*†], *Abraham Alcaim*[‡], *and José A. Apolinário Jr.*[†]

[†]IME - Department of Electrical Engineering
Praça Gal. Tibúrcio, 80 – Urca
22.290-270 Rio de Janeiro, RJ, Brazil
cborges,apolin@epq.ime.eb.br

[‡]CETUC/PUC-Rio
Rua Marquês de São Vicente, 225 – Gávea
22453-900, Rio de Janeiro, RJ, BRAZIL
alcaim@cetuc.puc-rio.br

ABSTRACT

This paper presents the performance of a text independent speaker verification system using Gaussian Mixture Model (GMM) for the Brazilian Portuguese. The Gaussian components of the GMM statistically represent the spectral characteristics of the speaker, leading to an effective speaker recognition system. The main goal here is a detailed evaluation of the parameters used by the GMM such as the number of Gaussian mixtures, the amount of time for training and testing. Aiming at the definition of the best set of features for a reasonable response, this work helps the comprehension of the model and gives insights for further investigation. We have used 36 speakers in the experiments, all modeled with 15 mel-cepstral coefficients. For 32 Gaussians, 60 seconds of training, and 30 seconds of testing, the system has no failure for a reasonably clean speech signal. The results have shown that the higher the amount of time for training and testing, the better are the results for a given statistical modeling of the speakers. It was interesting to note that when the time of training drops to 10 seconds, increasing the number of Gaussian was irrelevant to the system performance.

1. INTRODUCTION

The recognition of a human being through his voice is one of the simplest forms of automatic recognition because it uses biometric characteristics which come from a natural action, the speech. Speech, being present everywhere from telephone nets to personal computers, may be the cheapest form to supply a growing need of providing authenticity and privacy in the worldwide communication nets [1].

Research in the area of speaker recognition has significantly grown over the last few years due to a vast area of applications where the recognition can be used such as

– Access control: to devices, networks, and data in general;

– Authentication for business transactions as a tool to prevent fraud in: shopping over telephone, credit card validation, transactions over Internet, bank operations, etc.

– Law enforcement: penitentiary monitoring, forensic applications, etc.

– Help to handicapped.

– Military use: classified information requiring speaker identification.

The speech for security purpose can be used with other validation devices such as magnetic cards and passwords. In the future, more and more applications will include man machine iteration: speech operated devices will control the sound and the illumination of public environments and cars, and voice activated e-mail is expected to be used by everyone.

Speaker verification is the task of verifying if a speech signal (elocution) belongs or not to a certain person, which means a binary decision. The decisions are carried out in the so-called speakers open set [2] because the recognition is done in an unknown speakers set (possible impostors). As to text dependency, recognition can be dependent or independent. Systems demanding a pre-determined word or phrase are text dependent. Such systems can offer precise and reliable comparisons between two speech signals with the same text, in phonetically similar environments, requiring only 2 to 3 seconds of speech for training and testing. In text independent systems, such comparisons are not so easy to be obtained. The performance decreases as compared to text dependency. Moreover, in order to obtain reasonable statistics of the signal, it is, in general, necessary from 10 to 30 seconds of speech signal for training and testing [3].

In speaker recognition, the Gaussian Mixture Model (GMM) can be seen as a hybrid between two effective models used in speaker recognition: an unimodal Gaussian classifier and a vector quantization (VQ) codebook [4]. This scheme combines the robustness and smoothing properties

*The author thanks CAPES for partial funding of this work.

of the parametric Gaussian model with the arbitrary modeling capability of a non-parametric VQ. The GMM performs the spatial separation of acoustic classes and its main difference comparing to VQ concerns the fact that distances are not used to separate classes but probabilities from a set of Gaussian probability density functions previously estimated. The GMM can also be understood as an only state HMM (Hidden Markov Model) [5], having as observations mixtures of Gaussian PDFs (probability density functions). These components may model a vast phonetic class to characterize the sound produced by a person [6]. This fact justifies its use in speaker recognition.

This paper is organized as follows. In Section 2, the GMM is reviewed. Section 3 contains details of the system configuration followed by simulation results in Section 4, and conclusions in Section 5.

2. THE GAUSSIAN MIXTURE MODEL

A mixture of Gaussian probability densities is a weighted sum of M densities, as depicted in Fig. 1, and is given by

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{1}$$

where $\vec{x}$ is a random vector of dimension D, $b_i(\vec{x})$, $i = 1, ..., M$, are the density components, and p_i, $i = 1, ..., M$, are the mixtures weights. Each component density is a D variate Gaussian function of the form

$$b_i(\vec{x}) = \frac{e^{\left(-\frac{1}{2}(\vec{x}-\vec{\mu})' K_i^{-1}(\vec{x}-\vec{\mu})\right)}}{(2\pi)^{\frac{D}{2}}\sqrt{|\mathbf{K}_i|}} \tag{2}$$

with mean vector $\vec{\mu}_i$ and covariance matrix $\mathbf{K}_i$.

Note that the weighting of the mixtures satisfy $\Sigma_{i=1}^{M} p_i = 1$. The complete Gaussian mixture density is parameterized by a vector of means, covariance matrix, and a weighted mixture of all component densities (λ model). These parameters are jointly represented by the following notation.

$$\lambda = \{p_i, \vec{\mu}_i, K_i\} \qquad i = 1, ..., M. \tag{3}$$

The GMM can have different forms depending on the choice of the covariance matrix. The model can have a covariance matrix per Gaussian component as indicated in (3) (nodal covariance), a covariance matrix for all Gaussian components for a given model (grand covariance), or only one covariance matrix shared by all models (global covariance). A covariance matrix can also be complete or diagonal [2].

Since Gaussian components jointly act to model the probability density function, the complete covariance matrix is usually not necessary. Even being the input vectors not statistically independent, the linear combination of the diagonal covariance matrices in the GMM is able to model the correlation between the given vectors. The effect of using a set of M complete covariance matrices can be equally obtained by using a larger set of diagonal covariance matrices [6].

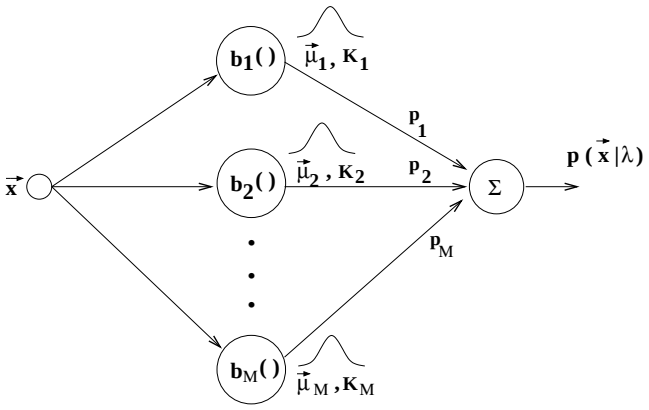

Fig. 1. M probability densities forming a GMM.

For a set of training data, the estimation of the maximum likelihood is necessary. In other words, this estimation tries to find the model parameters that maximize the likelihood of the GMM, The algorithm presented in [4] is widely used for this task. For a sequence of independent T vectors for training $X = \{\vec{x}_1, ..., \vec{x}_T\}$, the likelihood of the GMM is given by

$$p(X|\lambda) = \prod_{t=1}^{T} p(\vec{x}_t|\lambda) \tag{4}$$

The likelihood for modeling a true speaker(model λ) is directly calculated through

$$\log p(X|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log p(\vec{x}_t|\lambda) \tag{5}$$

The scale factor $\frac{1}{T}$ is used in order to normalize the likelihood according to the duration of the elocution (number of feature vectors). The last equation corresponds to the normalized logarithmic likelihood which is the λ model's response.

The speaker verification system requires a binary decision, accepting or rejecting a pretense speaker. Such a system is represented in Fig. 2

The system uses two models which provide the normalized logarithmic likelihood with input vectors $\vec{x}_1, ..., \vec{x}_T$, one from the pretense speaker and another one trying to minimize the variations not related to the speaker (**back-**

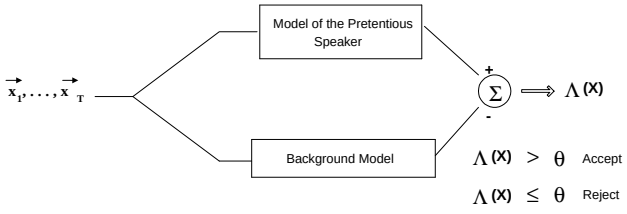

Fig. 2. Speaker verification system using GMM.

ground model), providing a more stable decision threshold [2]. If the system output value (difference between the two likelihood) is higher than a given threshold θ the speaker is accepted; otherwise it is rejected. The background is built with a hypothetical set of false speakers and modeled via GMM (universal background model [7]). The threshold is calculated on the basis of experimental results.

3. SYSTEM CONFIGURATION

This section details the speaker verification system implemented in our experiments. We have used 36 speakers, 23 males and 13 females, from which 5 males (M) and 5 females (F) were selected exclusively to form the background and, therefore, did not participate in the tests. Each speaker read 200 phrases, in Brazilian Portuguese, extracted from [8]. We have used 15 mel-cepstrum coefficients [9] — except in one experiment where 12 coefficients were used in order to evaluate the performance of a lower number of features — and the silence between words were eliminated. The number of Gaussians were 32, 16, and 8. We have used 60, 30, and $10s$ of speech signal for training and 30, 10, and $3s$ for testing. Each background speaker contributed with 6 seconds of speech (without silence). The setting of the decision threshold was established in order to equally minimize the error rate between false acceptance – FA (to accept someone which does not correspond to the true speaker) and false rejection – FR (to reject someone which corresponds to the true speaker). This procedure resulted in an equal error rate (EER) [2] which in our case is assumed to be the mean between the two errors.

4. SIMULATION RESULTS

In the first experiment evaluated the role of the number of mel-cepstrum coefficients (MCC) from 15 to 12 — a reduction of 3 coefficients — using a 32 Gaussians GMM and 60 seconds of training. The results can be seen in Table 1 for 30, 10, and 3 seconds of testing. In this section, all the results (error rates) are given in %.

We note that the performance of the MCC15 is superior to the MCC12 for 30 and 3 seconds of testing. For the case of $10s$, MCC12 was slightly superior. These results indicate that the 3 additional coefficients used in the MCC15 hold information about the speaker which should not be disregarded. In view of results obtained so far, the following tests were carried out with MCC15.

Table 1. Performance of the GMM with a reduction in the dimension of the feature vector.

Test	MCC15			MCC12		
	FR	FA	ERR	FR	FA	ERR
30s	0	0	0	0	0.38	0.19
10s	0.38	0.51	0.44	0.26	0.51	0.38
3s	1.19	1.58	1.38	2.19	1.42	1.80

The results for $60s$ of training varying the number of Gaussians can be seen in Table 2.

Table 2. Variation of the number of Gaussian with 60s of training.

Test	32 G			16 G			8 G		
	FR	FA	ERR	FR	FA	ERR	FR	FA	ERR
30s	0	0	0	0.77	0.38	0.57	2.38	0.38	1.35
10s	0.38	0.51	0.44	0.38	0.77	0.57	1.92	1.67	1.79
3s	1.19	1.58	1.38	1.65	2.23	1.94	3.54	3.62	3.58

We can observe that with 32 Gaussians and $30s$ of test speech, the system provides an excellent result. Reducing the time for testing or the number of Gaussians will cause a decrease in the system performance. Of course, there exists a trade off between the amount of time for testing and the number of Gaussians. For instance, when 32 Gaussians and $3s$ of test are used, the ERR performance is comparable to a system employing 8 Gaussians and $30s$ of testing.

The results for $30s$ of training varying the number of Gaussians can be seen in Table 3.

Table 3. Variation of the number of Gaussian with 30s of training.

Test	32 G			16 G			8 G		
	FR	FA	ERR	FR	FA	ERR	FR	FA	ERR
30s	0.38	1.92	1.15	3.08	2.69	2.88	5.38	4.62	5.00
10s	1.41	1.54	1.47	3.21	3.46	3.33	6.15	4.49	5.32
3s	2.50	3.50	3.00	2.88	5.92	4.40	6.42	6.88	6.65

Again, larger number of Gaussians and amount of time used for testing improve the system performance. In this case, the decrease in the number of Gaussians significantly reduces the performance. Note that, independently of the time for testing, the system with 8 Gaussians does not achieve the performance of the other ones.

The results for $10s$ of training varying the number of Gaussians can be seen in Table 4.

In Table 4, we see that when there is not enough amount of information for training, neither the number of Gaussians nor the time of testing compensate for this lack of statistics.

Table 4. Variation of the number of Gaussian with 10s of training.

Test	32 G			16 G			8 G		
	FR	FA	ERR	FR	FA	ERR	FR	FA	ERR
10s	4.10	4.74	4.42	4.49	5.26	4.87	3.59	5.90	4.74
3s	6.15	7.62	6.88	6.00	7.73	6.86	7.15	7.19	7.17

Throughout the analysis of the results presented here, we can clearly note that the number of Gaussians has a strong influence in the performance. The higher this number the better the modeling obtained by the GMM and, therefore, the better the results will be. The amount of time for training and for testing also have a strong influence for the larger they are the more statistics they are offering and, consequently, the more precise the modeling carried out by the GMM will be. When the statistics provided to train the GMM is poor, the number of Gaussians does not influence the response because there is no data for a more precise modeling, as can be observed in Table 4.

5. CONCLUSIONS

This paper provided a performance evaluation of the model parameters used in a text independent Brazilian Portuguese speaker verification system using GMM. Such parameters were the number of Gaussian densities and the amount of time used in training and testing. Simulations were carried out with a reduced corpus and the results have shown the efficiency of the method (equal error rate - EER - basically equal to 0%) for clean speech with a minimum of 32 Gaussians, 60 seconds of training, and 30 seconds of testing speech signals. Moreover, it was observed that for small time of training speech, e.g. 10 seconds, the number of Gaussians was not very important. It was verified that when the training time is long (60s), a smaller number of Gaussians with a larger testing time yields a performance comparable to a system using a larger number of Gaussians with shorter testing time.

The main purpose of this work was a basic introduction to efficient text independent speaker verification that shall be used as a first step to a more in depth investigation concerning the improvement of this technique in order to increase the robustness of the recognition system for harder and realistic conditions such as small amount of speech and degradation by channel and additive noise.

6. REFERENCES

[1] CAMPBELL, Joseph P.,Jr. *Speaker Recognition: A Tutorial.* Proceedings of IEEE, vol. 85,no. 9, pp. 1437-1462, September 1997.

[2] REYNOLDS, Douglas A. *Speaker Identification and Verification Using Gaussian Mixture Speaker Models.* Speech Communication. vol. 17, pp. 91-108, 1995.

[3] JAYANT M. Naik. *Speaker Verification: A Tutorial.* IEEE Communication Magazine, pp. 42-47, January 1990.

[4] REYNOLDS, Douglas A. *A Gaussian Mixture Modeling Approach to Text Independent Speaker Identification.* PhD Thesis. Georgia Institute of Technology, August 1992.

[5] RABINER, Lawrence R. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* Proceedings of The IEEE, vol. 77, no. 2, February 1989.

[6] REYNOLDS, Douglas A. *Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Model.* IEEE Transactions on Speech and Audio Processing. vol. 3, n. 1, pp. 72-83, January, 1995.

[7] REYNOLDS, Douglas A. Thomas F. Quatieri, and Robert B. Dunn. *Speaker Verification Using Adapted Gaussian Mixture Models.* Digital Signal Processing. vol. 10, pp. 19-41, 2000.

[8] ALCAIM, Abraham, José Alberto Solewicz, and João Antonio de Morais. *Freqüência de ocorrência dos fonemas e listas de frases foneticamente balanceadas no Português falado no Rio de Janeiro*. Revista da Sociedade Brasileira de Telecomunicações, vol. 7, nr 1, December 1992.

[9] DAVIS, Steven B., and Paul Mermelstein. *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences.* IEEE Transactions on Acoustics, Speech, and Signal Processing. vol. ASSP-28, no. 4, August 1980.