SPEAKER RECOGNITION BY GAUSSIAN MIXTURE MODEL (GMM)

Meglouli H[1] ,Khebli A[1],
[1] *Laboratory of the electrifications of industrial companies, University of Boumerdes Algeria,*

Email: hmeglouli@yahoo.fr, Khebli_m@yahoo.fr

SUMMARY

Since the first works dedicated to speaker recognition, many approaches have been proposed in the literature: vector approach, connectionist, predictive, statistical, etc. This wide range, only the statistical approach remains at the forefront of systems of automatic speaker recognition in recent years. It is to represent a sequence of acoustic vectors from the parameterization by a finite number of statistical parameters. In this article, we use the Gaussian mixture model (GMM) for speaker recognition in independent mode of the text.

Key WORDS: speaker recognition, GMM, verification, speaker identification,MFCC,PLP.

I. INTRODUCTION

Automatic speaker recognition is interpreted as a particular task pattern recognition. This area includes issues related to the identification or verification based on information contained in the speaker's acoustic signal. The scope is very broad, ranging from home applications to military applications, through judicial applications [2, 3, 6].

In this work, based on the statistical approach, we focus on the speaker recognition in independent mode of the text by the Gaussian mixture. In this context, a sequence of acoustic training vectors is represented by a mixture of Gaussian ie a weighted sum of Gaussians. Each of them is characterized by a mean vector and a covariance matrix.

When learning, customers model parameters are usually estimated using the EM algorithm (Expectation-Maximization) coupled to the approach Maximum likelihood estimation (MLE). When the speaker recognition, the similarity measure between a client and a model sequence of test vectors based on the Maximum A Posteriori approach (MAP).

II. THE GAUSSIAN MIXTURE

Mixtures of Gaussian (sum of weighted Gaussian) are used to model the speech signal of a particular speaker. This method is the most used regarding speaker recognition in independent mode of the text [1, 5, 9, 13].

The use of a GMM model is essentially justified by appealing to the interpretation of classes mixture: it is certain that the vectors will be distributed differently depending on the characteristics of the speech sound considered (voiced sound, unvoiced, or more finely according to the phoneme). Each component will model the underlying sets of acoustic class, each class representing acoustic events (vowels, nasal, ..., etc.). Thus, the spectral shape of the ith class can be represented by the mean and the covariance matrix of the ith component. These classes characterize own acoustic space for each speaker.

II.1 MIXTURE MODEL

A Gaussian mixture model is a weighted sum of M Gaussian densities (see Figure 1). S is a speaker and an acoustic vector dimension D; the Gaussian mixture is defined as follows [11, 12]:

$$p(x/\lambda_s) = \sum_{m=1}^{M} \pi_m^s b_m^s(x) \qquad (1)$$

where represent the Gaussian densities, parameterized by a mean vector and a covariance matrix.

$$b_m^s(x) = \frac{1}{(2\pi)^{D/2} \left|\sum_m^s\right|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_m^s)' \Sigma_m^{s\,-1}(x-\mu_m^s)\right] \qquad (2)$$

Represent the weight of the mixture with.

The speech signal of a speaker is thus modeled by a set of parameters noted .lambda.s [8, 12, 13]:

$$\lambda_s = (\pi_m^s, \mu_m^s, \Sigma_m^s) \qquad (3)$$

This model may take several forms, including as regards the covariance matrices. You can assign a covariance matrix for each Gaussian, or use a common global covariance matrix for all Gaussian.

II.2 LEARNING MODEL

It is during the learning phase, all λ estimate the parameters of a GMM speaker model. The conventional method is that of maximum likelihood (ML) whose

purpose is to determine the model parameters that maximize the likelihood of the training data

II.2.1 EXPECTATION-MAXIMIZATION ALGORITHM

The EM algorithm (Expectation Maximization) can be considered as a special case of the gradient algorithm [4, 11]. It involves both observations and missing variables X (the index of the Gaussian m = 1, ..., M). This algorithm maximizes, iteratively, the likelihood function. This maximization is not direct. It involves the auxiliary function which is defined as the expectation of the joint likelihood logarithm (including the observed variables and hidden variables) on the full set of training variables, calculated on the basis of the current settings [4] . This auxiliary function is expressed as follows:

$$Q(\theta,\theta^{(t)}) = \sum_{m=1}^{M}\sum_{n=1}^{N} p(m/x_n,\theta^{(t)}) \log p(x_n,m/\theta) \quad (4)$$

The model parameters are estimated by canceling the partial derivative of the auxiliary function in relation to each of them.

In the case of the average, we have:

$$\frac{\partial Q}{\partial \mu_m} = \sum_{n=1}^{N} p(m/x_n,\theta^{(t)}) \left[\frac{(x_n - \mu_m)}{\sigma_m^2} \right] = 0 \quad (5)$$

The new estimator of the mean becomes:

$$\mu_m^{(t+1)} = \frac{\sum_{n=1}^{N} p(m/x_n,\theta^{(t)}) x_n}{\sum_{n=1}^{N} p(m/x_n,\theta^{(t)})} \quad (6)$$

Regarding

$$\frac{\partial Q}{\partial \sigma_m} = \sum_{n=1}^{N} p(m/x_n,\theta^{(t)}) \left[\frac{(x_n - \mu_m)^2}{\sigma_m^3} - \frac{1}{\sigma_m} \right] = 0 \quad (7)$$

The new variance estimator becomes

$$\sigma_m^{2(t+1)} = \frac{\sum_{n=1}^{N} p(m/x_n,\theta^{(t)})(x_n - \mu_m^{(t)})^2}{\sum_{n=1}^{N} p(m/x_n,\theta^{(t)})} \quad (8)$$

The estimated weight of the mixture of components is quite simple since it is scalar parameters. It should however take into account the constraint that exists on these parameters. The constrained maximization is solved simply by introducing a Lagrange multiplier associated with the constraint [4, 8, 10, 12]. The function to be maximized becomes:

$$Q^*(\theta,\theta^{(t)}) = Q(\theta,\theta^{(t)}) + \tau\left(\sum_{m=1}^{M} \pi_m - 1 \right) \quad (9)$$

where is the Lagrange multiplier.

By canceling the partial derivative of with respect to (which eliminates the terms containing the means and variances), we get:

$$\frac{\partial Q^*}{\partial \pi_m} = \frac{1}{\mu_m} \sum_{n=1}^{N} p(m/x_n,\theta^{(t)}) + \tau = 0 \quad (10)$$

Summing this expression over all m components, we get that, which gives us so:

$$\pi_m^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} p(m/x_n,\theta^{(t)}) \quad (11)$$

In which values can simply be obtained by Bayes rule:

$$p(m/x_n,\theta^{(t)}) = \frac{p(m/\theta^{(t)})\, p(x_n/\theta^{(t)})}{\sum_{k=1}^{M} p(k/\theta^{(t)})\, p(x_n/\theta^{(t)})} \quad (12)$$

II.2.2 SPEAKER IDENTIFICATION

Or a group of ξ speakers represented by GMM models: λ1, λ2, ..., λξ. The objective of the identification phase is to find, from an observed sequence X, the model which is the maximum a posteriori probability [7, 13]. That is to say:

$$\hat{S} = \arg \max_{1 \le s \le \xi} p(\lambda_s / X) \quad (13)$$

which gives, according to the Bayes rule

$$\hat{S} = \arg \max_{1 \le s \le \xi} \frac{p(X/\lambda_s)}{p(X)} p(\lambda_s) \quad (14)$$

Assuming equal probability of occurrence of speakers, and that the probability of a sequence X is the same for all speakers, the classification of law becomes:

$$\hat{S} = \arg \max_{1 \le s \le \xi} p(X / \lambda_s) \quad (15)$$

By using the logarithm and independence between observations, the identification system calculates the following score:

$$\hat{S} = \arg \max_{1 \le s \le \xi} \prod_{n=1}^{N} p(x_n / \lambda_s) \quad (16)$$

$$\hat{S} = \arg \max_{1 \le s \le \xi} \sum_{n=1}^{N} \log(p(x_n / \lambda_s)) \quad (17)$$

is replaced by its value in the expression (1), equation (17) becomes:

$$\hat{S} = \arg \max_{1 \le s \le \xi} \sum_{n=1}^{N} \log \sum_{m=1}^{M} \left(\frac{\pi_m^s}{(2\pi)^{D/2} \left| \sum_m^s \right|^{1/2}} \exp\left[-1/2 (x_n - \mu_m^s)' \sum_m^{-1s} (x_n - \mu_m^s) \right] \right) \quad (18)$$

III. EVALUATION

III.1 INFLUENCE OF MODEL ORDER

In this experiment, we study the impact of some of the models (or number of Gaussian) on speaker identification performance in the event that there is very little training data (two seconds of speech). Figure 2 plots the variation identification error rate of 33 speakers according to the order of the models for the PLP coefficients and MFCC delta MFCC.

2 shows that the increase in the number of Gaussians in the representation of the speaker provides improved performance. However, the gain is not significant beyond 512 Gaussian coefficients for delta-MFCC and MFCC and the computing time increases dramatically see Table 1

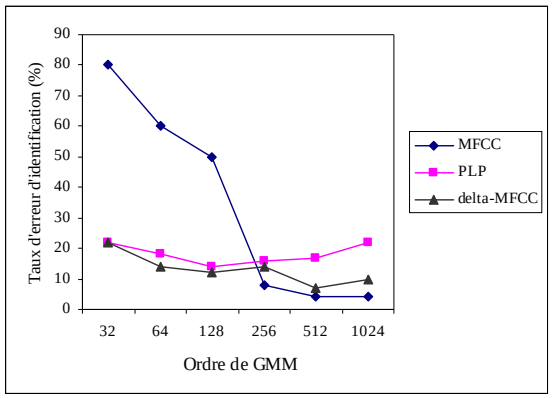

Figure 2: Identification Error Rate

depending on the model of order

The choice of the model order depends on its fineness and quantity of training data. Choose a too low order will affect the accuracy of the model. Select components generate much greater computational load. In general, 256 components are enough to represent a speaker with very little training data (speech 2seconde).

Figure 2 shows that the identification performance is better in the case of MFCC coefficients for a number of model too high, but for a small number of model are used detta-MFCC coefficients or PLP

Ordre du Modèle / Coefficients	32	64	128	256
MFCC	80%	60%	50%	8%
Delta –MFCC	22%	14%	12%	14%
PLP	22%	18%	14%	16%
Temps d'exécution moyen(s)	33.408	44.0916	65.290	114.008

Table 1: identification error rate and average execution time depending on the model order.

III.2 INFLUENCE OF LEARNING TIME

The purpose of this experiment is to evaluate the identification and verification of performance of the speakers 11 according to the amount of learning. The

speakers are modeled by Gaussian and 256 models are trained with a maximum likelihood estimator that fits all parameters.

Figures 3 and 4, we have depicted the variations of the identification error rate and the false acceptance rate obtained with the GMM modeling, depending on the amount of learning.

The correct identification rate significantly decreases up to 30 seconds of speech. Beyond this value, the performance tends to saturate.

These figures also show that, to achieve error rates lower identification to 5%, it is necessary to have at least 30 seconds of the training data. As for the check, just 20 seconds of speech for an error rate close to 20%. However, with 30 seconds of learning, GMM recognition system achieved a misidentification rate and a false acceptance rate of zero (0%).

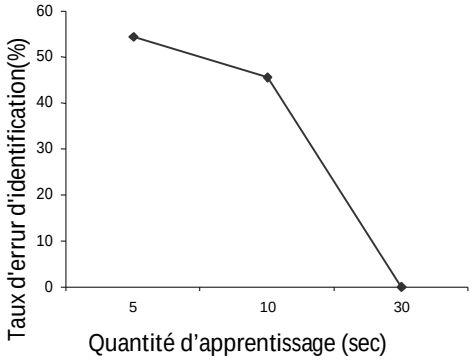

Figure 3 : Identification performance GMM (Models 256 Gaussian GMM)

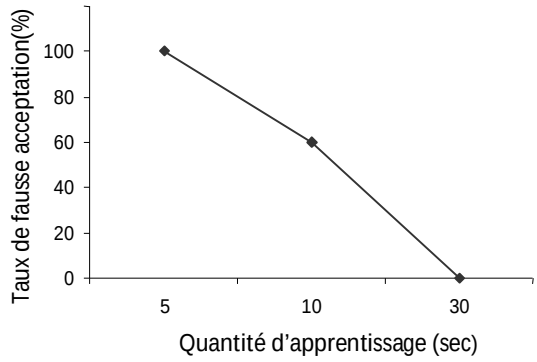

Figure 4 : Audit Performance by GMM

CONCLUSION

In this paper, we presented a system of automatic recognition of speakers in independent mode of the text by the Gaussian mixture model. To model the speech signal of a speaker, we used the Gaussian mixture models (GMM). The model parameters are estimated by the iterative algorithm: Expectation-Maximization (EM). In the identification phase, we used the MAP criterion for identifying an unknown speaker. Mixtures of Gaussian (GMM) provide better recognition rate (96% for a model of order equal to 512) in independent mode of the text. The performance of this method is better when increasing the amount of learning. For example, for a quantity Learning 30s, it was a no identification error rate (0) and a false rejection rate of zero (0). The recognition rate also increases with increasing the model order. For example for a model order equal to 128, it has a recognition rate of 50%, and for an order of the model equal to 256, it has a recognition rate equal to 92%.

BIBLIOGRAPHIE

[1] C. Barras & J. *Gauvain*, Feature and score normalisation for speaker verification of cellular data. *International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), April 6-10, 2003 in Hong Kong SAR, China.

[2] L. Besacier, Un *Modèle Parallèle pour la Reconnaissance Automatique du Locuteur*. Thèse de Doctorat, Université d'Avignon et des Pays de Vaucluse, 1998.

[3] E. Doherty, An evaluation of selected acoustic Parameters for use in speaker identification. J. Phonetics 4: pp 321-326, 1976.

[4] D. Helenca and A. Bonafonte, Estimation of GMM in voice conversion including unaligned data. Dans European *Conference on Speech Communication and Technology* (*EUROSPEECH)*, pp 861-864, 2003, *GENEVA.*

[5] H. Hermansky, M. Narendranath, "Speaker Verification using Speaker specific mappings ", in Proc. Of Speaker *Recognition and its Commercia l and Forensic Applications*, France, April 1998.

[6] Q. Lin, Ea-Ee Jan and J. Flanagan, Microphone Arrays and speaker identification. *IEEE Transaction on speech and audio processing*, vol. 2, no.4, October 1994.

[7] C.-S. Liu, H, C. Wang and C.H. Lee, Speaker Verification using normalized log- likelihood score. *IEEE* Transaction *on speech and audio processing, vol.* 4, no.1, January 1996.

[8] A. Mijail and A. Drygajlo, on the number of Gaussian components in a mixture: An appli-Cation to speaker verification tasks. Dans

European *Conference on Speech Communication and Technology* (*EUROSPEECH)*, pp 2673-2676, 2003, *GENEVA.*

[9] D. A. Reynolds, Model compression for GMM based speaker recognition systems. Dans *European Conference on Speech Communication And Technology* (*EUROSP EECH)*, pp 2005-2008 2003, *GENEVA.*

[10] D. A. Reynolds and C. Richard, Robust text-Independent speaker identification using Gaussian mixture speaker models. *IEEE Transaction on speech and audio processing*, vol. 3, no.1, January 1995.

[11] V. Robbie & al, Dependence of GMM adaptation on feature post- processing for speaker recognition. Dans European *Conference on Speech Communication and Technology* (*EUROSPEECH)*, pp 3013-3016, 2003, *GENEVA.*

[12] Z. Stan et al, Learning to boost GMM based speaker verification. Dans European *Conference on Speech Communication and Technology* (*EUROSPEECH)*, pp 1677-1680, 2003, *GENEVA*

[13] V. Upendra. J. Chaudhari and H, Stéphane. Maes. Multigrained Modeling with Pattern Specific Maximum Likelihood Transformations for Text-Independent Speaker Recognition. *IEEE Transaction on speech and audio processing*, vol. 11, no.1, January 2003.