CREATIVE MATH. & INF.
16 (2007), 124 - 129

Dedicated to Professor Ioan A. RUS on the occasion of his 70^{th} anniversary

Comparative study on some parameters of a text independent speaker identification system for Romanian language based on GMM with MFCC vectors

MARIETA GÂTA AND GAVRIL TODEREAN

ABSTRACT. The speaker recognition technique used here is based on GMM. This approach consists in three phases: parameterization, model training and classification. We compare a model of a speech extracted from an unknown speaker with the models of speakers contained in our database. Models are calculated with EM (Expectation Maximization) algorithm for GMM (Gaussian Mixture Models). We study the influences of several parameters: different texts in the training process and in the testing process, numbers of Gaussians, number of speakers, amount of training data (length of the wav file in seconds), numbers of iterations.

1. INTRODUCTION

This paper presents a speaker identification system based on GMM which attains excellent recognition performance for text-independent speech. The system based on GMM is robust which results from the fact that GMM works with statistically based representations of speaker identity. For comparison, the system based on GMM is tested in different approaches. There are three main goals of these experiments: 1) comparison of system performances on different number of mixture components 2) testing the hypothesis that for a given amount of training data a speaker model has an optimum number of components, 3) to find out the influence of the number of iterations in training process on GMM's system performance.

The Gaussian mixture speaker model was introduced in [5] and has demonstrated high text-independent recognition accuracy for short test utterances. The basis for the recognition system is the GMM used to represent speakers. More specifically, the distribution of feature vectors extracted from a person's speech is modeled by a Gaussian mixture density. We use a probability density function consisting of maximum 12 mixtures.

The density $b(x)$ is a weighted linear combination of M component uni-modal Gaussian densities, each parameterized by a mean vector xm, covariance matrix Σm and weight of mixture cm. The identification system is a straight-forward maximum-likelihood classifier. For a reference group of Mspeakers, the objective is to find the speaker model $m*$ which has the maximum posterior probability for

Received: 20.09.2006. In revised form: 19.02.2007
2000 *Mathematics Subject Classification.* 68T10, 68T50, 60G15.
Key words and phrases. *Gaussian Mixtured Models, Speaker Identification, MFCC, Covariance matrix.*

the input feature vector sequence

$$X = (x_1, x_2, ..., x_t, ..., x_T).$$

2. SPEECH DATABASE

The systems were evaluated on speech database in Romanian. The speakers uttered two different sentences. Individual sentences were chosen to be plentiful for phonemes. The number of speakers was 200 (123 male and 77 female) and different classes of age (student from different faculties, from the first to the fourth year of study, which means the age 18-22) were represented. Each speaker uttered 4 sentences, 2 for testing and 2 for training. Speakers were recorded in two or three sessions (the time among sessions was not longer than approximately one month). The speech was clean (laboratory background), recorded using one microphone and sampled at 22.05kHz, 16bit and mono. Training sentences were selected with view to approximately doubling previous length of speech (from 4 to 10 seconds). The feature vectors used in systems were 12'th order MFCC (Mel Frequency Cepstral Coefficients), obtained from 20 mel-wrapping filter banks, with no regression.

3. EXPERIMENT 1 – RECOGNITION PERFORMANCE OF A SYSTEM BASED ON GMM ON DIFFERENT NUMBER OF MIXTURE COMPONENTS

A GMM with a full covariance matrix is the most complex of the mentioned models. A simplified form, popular in practical speaker recognition, has each component consisting of a mean, the diagonal of the covariance matrix and a weight. These models were tested and results are illustrated in Figure 1 (in training process we used 10 seconds of speech for creating models).

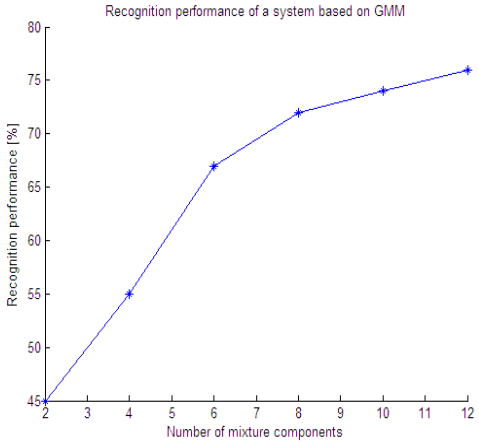

Figure 1. GMM with full covariance matrix

We can predict that GMM models with more components will surpass GMM models with less components, because they can benefit from the better results that come with larger models, without the detrimental effects of inaccurate variances

and weights. It is suggested here that GMM is optimal when a larger number of parameters (and therefore a larger number of components) are used.

4. Experiment 2 – Relation between recognition performance and different amount of training data

A GMM is tested on model sizes, extracted from wav file of 4, 6 and 10 seconds, to verify that the initially recognition results are improved as the number of components in the model increases, and afterwards, that they are degraded as the statistics (primarily the variances and weights) became less and less accurate. The results are shown in Figure 2.

Table 1. GMM identification performance for different amounts of training data and model orders

Amount of training speech	Model order	[%] correct
4 seconds	M=2	45
	M=4	55
	M=6	67
	M=8	72
	M=10	74
	M=12	**76**
6 seconds	M=2	46
	M=4	57
	M=6	69
	M=8	73
	M=10	75
	M=12	**78**
10 seconds	M=2	48
	M=4	58
	M=6	71
	M=8	74
	M=10	77
	M=12	**79**

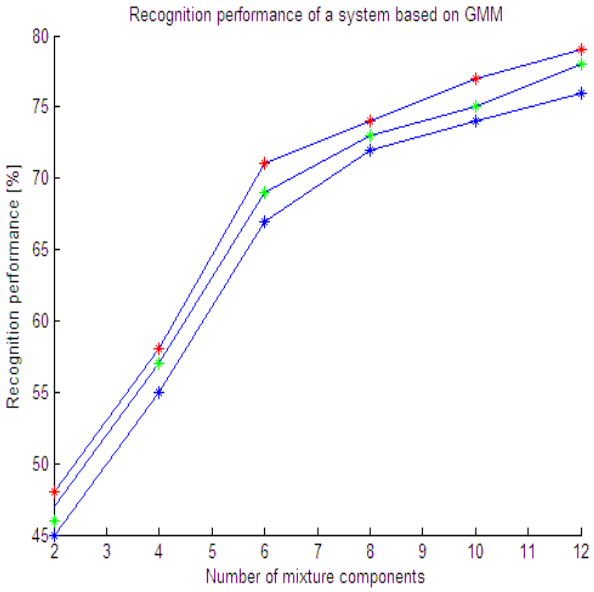

Figure 2. Performance curves obtained from GMM models that were trained with different amount of data

After comparing maximums of these curves (least recognition errors with given amount of data) we can see that the growing length of training data drifts to best recognition results in the right of the figure (higher number of components in the model). Together with this drift the recognition error is reduced. The selection of actual recognition scores can be read in Table 1. Marked cells indicate best recognition results for a concrete method and the amount of training data. In GMM case all best results were achieved by employing most of the components. Small amount of training data (4 and 6 seconds) are not ideal for GMM method (insufficient statistics), the best results were achieved with 12 components of GMM and the size of the wav files of 10 seconds .

5. Experiment 3 – Trials with number of iterations using EM algorithm

The EM (Expectation Maximization) algorithm is used in GMM training process (generally in hidden Markov models and other learning techniques). It detects model parameters by maximizing the log-likelihood of incomplete data and iteratively maximizing the expectation of log-likelihood from complete data. In this section an importance of EM iterations for improving recognition score will be demonstrated. These experiments were performed on GMM with diagonal covariance matrix and this could be a reason why the improvement after 20 iterations (50 iterations are recommended) is not so expressive. Obtained results are illustrated in Figure 3 and Figure 4.

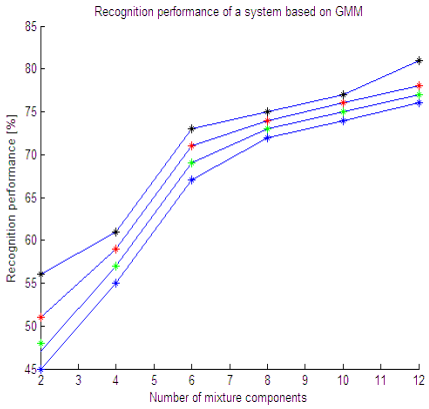

Figure 3. Influence of EM iterations on recognition performance obtained from GMM for models with 4, 6, 8 respectively 12 mixture components

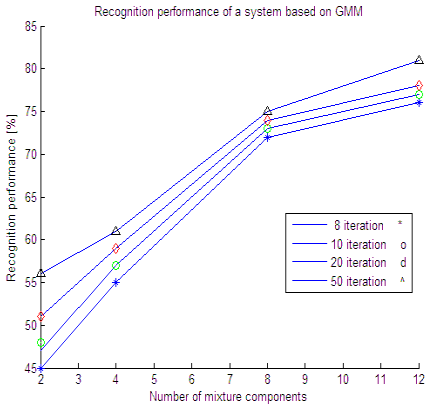

Figure 4. As for figure 3 but for models with 4, 8 respectively 12 mixture components

6. CONCLUSIONS

According to the expectations, the performance of the method is very good. It is not only worth of this method but these good results are also supported by the quality of the signal and also by other technical aspects. It is concluded that maximizing the use of speaker data, which is translated into maximizing the size (number of components) of the model, is important to improve speaker recognition. But on the other hand, if the size of the model is too large and we don't have enough training data, it can markedly reduce the performance of the recognition system. The best performance was obtained with 12 mixture components of GMM and 50 iterations of the process.

REFERENCES

[1] Cassidy S., *Speech Recognition,* Speech Hearing and Language Research Centre, Macquarie University, 2001

[2] Fredouille C., Pouchoulin G., Bonastre J.F., Azzarello M., Giovanni A. and Ghio A., Application of Automatic Speaker Recognition techniques to pathological voice assessment, *Proceedings of the Interspeech'2005-Eurospeech, 9^{th} European Conference on Speech Communication and Technology,* Lisbon, Portugal, September 2005

[3] Jin Q. and Waibel A., Application of LDA to Speaker Recognition, *Proceedings of the ICSLP-00,* Beijing, China, October 2000

[4] Morris A., Wu D. and Koreman J., *GMM based clustering and speaker separability in the Timit speech database,* IEICE Transactions Fundamentals, Communications, electronics, Informatics & Systems, Vol. E85, 2005.

[5] Reynolds D.A., *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification,* Ph.D. Thesis, Georgia Institute of Technology, September 1992

MARIETA GÂTA
NORTH UNIVERSITY OF BAIA MARE
DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
VICTORIEI 76
430122 BAIA MARE, ROMANIA
E-mail address: marietag@ubm.ro

GAVRIL TODEREAN
TECHNICAL UNIVERSITY OF CLUJ NAPOCA
FACULTY OF ELECTRONICS AND TELECOMMUNICATIONS
DEPARTMENT OF COMMUNICATIONS
GEORGE BARIŢIU 26-28
400027 CLUJ-NAPOCA, ROMANIA
E-mail address: Gavril.Toderean@com.utcluj.ro