RESULTS OBTAINED IN SPEAKER RECOGNITION USING GAUSSIAN MIXTURE MODELS

Marieta GÂTA*, Gavril TODEREAN**

* Department of Mathematics and Computer Science, University of Baia Mare,
Faculty of Science, Baia Mare, Romania
** Department of Communications, Technical University of Cluj Napoca,
Faculty of Electronics and Telecommunications

Corresponding author: Marieta Gâta,
marietag@ubm.ro

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 by our database. Models are calculated with EM algorithm for GMM. We study the influences of different 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), and numbers of iterations.

Key words: Gaussian mixture models, speaker identification, MFCC, covariance matrix.

1. INTRODUCTION

This paper presents a speaker identification system based on Gaussian Mixture Models (GMM) which attains excellent recognition performance for text-independent speech. System based on GMM is robust which results from fact that GMM works with statistically based representations of speaker identity. For comparison, the system based on GMM is tested with different approach. There are three main goals of these experiments: 1) comparison 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. We use GMM because in speaker recognition, whether text dependent or independent, the model consisting of a simple GMM for each speaker is very successful [3].

The Gaussian mixture speaker model was introduced in [4] 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, a Gaussian mixture density models the distribution of feature vectors extracted from a person's speech. We use a probability density function consisting of maxim of 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 M speakers, the objective is to find the speaker model m^* which has the maximum posterior probability for the input feature vector sequence $X = (x_1, x_2 \ldots x_t \ldots x_T)$ [2].

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. Number of speakers was 200 (123

male and 77 female) from different classes of age (student from different faculties, from the first to the fourth year of study, which means the age 18-22). Each speaker uttered 4 sentences, 2 for testing and 2 for training. The training data used for learning was not used for testing [1]. 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 22.05 kHz, 16 bit and mono. Training sentences were selected with view to approximately doubling previous length of speech (from 4 to 10 seconds). The training sentences are: „*Un număr de telefon este format din cifrele zero unu doi trei patru cinci sase şapte opt noua zece*" and "Principalele operații matematice sunt adunarea scăderea înmulțirea şi împărțirea" and the testing sentences are: „*Numărul meu de telefon este patru zero doi sase doi unu doi trei patru cinci*" and „*Automobilul meu atinge o viteza de o suta optzeci de km pe oră*". The feature vectors used in systems were 12th order MFCC (obtained from 20 mel-wrapping filter banks) with no regression. The next sections presents results obtained in four experiments. In the first experiment we prospect for some relation between the number of mixture components and recognition performance of our system based on GMM. The second experiment present relation between the different amount of training data (id the length of the wav files), number of mixture components and recognition performance. The third and fourth experiments presents the relation between the number of iterations, number of mixture components and recognition performance.

3. RELATION BETWEEN NUMBER OF MIXTURE COMPONENTS AND RECOGNITION PERFORMANCE

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).

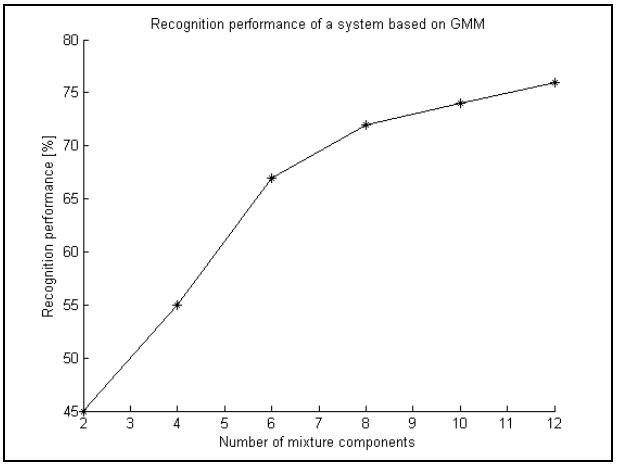

Figure 1. GMM with full covariance matrix.

We can predict that GMM models with more components will surpass GMM models with less components. We obtained better results with larger models without the harmful 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. RELATION BETWEEN DIFFERENT AMOUNT OF TRAINING DATA, NUMBER OF MIXTURE COMPONENTS AND RECOGNITION PERFORMANCE

A GMM is tested on model sizes extracted from wav file of 4, 6 and 10 seconds. We verify that the recognition results increase along with the number of mixture components and with the amount of training data. 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

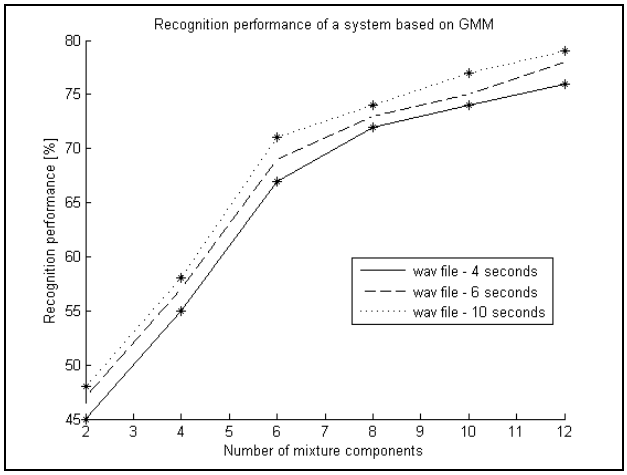

Figure 2. Performance curves obtained from GMM models that were trained with different amount of data.

After comparing maxim of these curves (least recognition errors with given amount of data) we can see that the growing length of training data generate the best recognition results to the right of the figure (higher number of components in the model). In this part of the figure the recognition error is more reduced then in the left part. The selection of actual recognition scores can be read in table 1. Marked cells indicate best recognition results for concrete method and amount of training data. In GMM case all best results were achieved by using most components. Small amount of training data (4 and 6 seconds) are not ideal for GMM method. The best results were achieved with 12 components of GMM and the size of the wav files of 10 seconds.

5. RELATIONS BETWEEN NUMBER OF ITERATIONS, NUMBER OF MIXTURE COMPONENTS AND RECOGNITION PERFORMANCE

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. We improve the recognition performance after 10 iterations (50 iterations are recommended) but not so excessive. Obtained results are illustrated in figure 3 and figure 4.

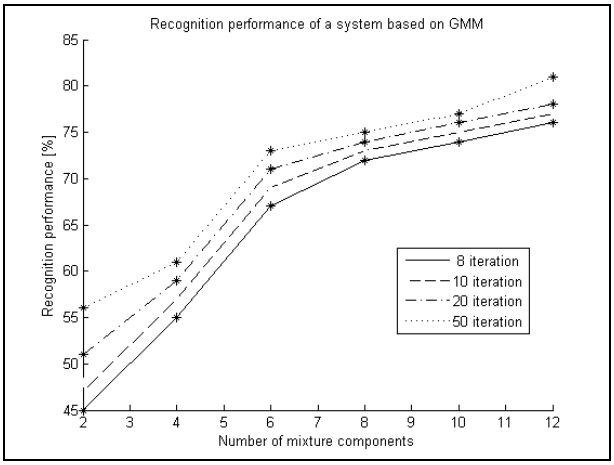

Figure 3. Influence of EM iterations on recognition performance obtained from GMM with diagonal covariance matrix for models with 4,6,8,10,12 number of mixture components.

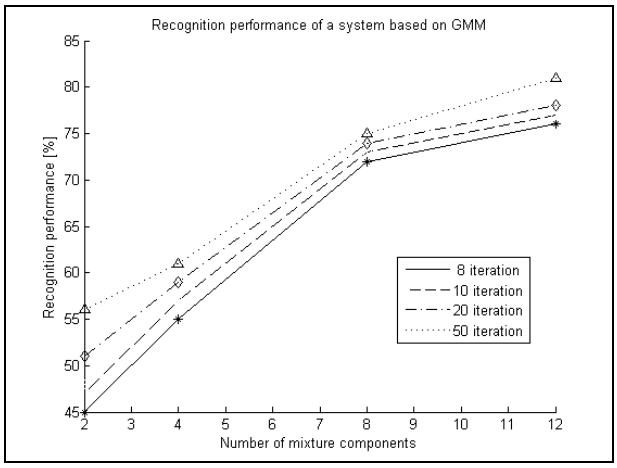

Figure 4. As for figure 3 but for models with number of 4,8,12 mixture components.

6. CONCLUSIONS

According to the expectations, the performance of method is very good. It is not only worth of this method but these good results are also supported by quality of the signal and other aspects. It is concluded that maximizing the use of speaker data, which is translated into maximizing the size (number of components) of the model, improve the 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 performance of the recognition system. The best performance was obtained with 12 mixture components of GMM and 50 iterations of the process.

REFERENCES

1. FREDOUILLE C., POUCHOULIN G., BONASTRE J.-F., AZZARELLO M., GHIO A. G., "Application of Automatic Speaker Recognition techniques to pathological voice assessment", 9th European Conference on Speech Communication and Technology, (Interspeech) Lisboa, September 2005.
2. JIN Q., WAIBEL A., "Application of LDA to Speaker Recognition", Proceedings of International Conference on Spoken Language Processing (ICSLP-2000), Beijing, P.R.China, October 2000.
3. MORRIS A., WU D., KOREMAN J., "GMM based clustering and speaker separability in the Timit speech database", *IEICE Transactions Fundamentals, Communications, Electronics, Informatics & Systems*, **Vol**. E85, 2005.
4. REYNOLDS D.A., A *Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*, Ph.D. Thesis, Georgia Institute of Technology, September 1992.