Speakers Identification in a Telephone Conversation Using Gaussian Mixture Model

Mohsen Bazyar, Ahmad Keshavarz, and Reza Dyanat

***Abstract*—In this paper, we analyze two unknown speakers identification with having a sample of speech from a finite set of speakers and text-independent speech .Then refers to conventional methods in speaker identification after expression of the existing problems. Cepstral coefficients are evaluated as the most successful feature vectors. Gaussian mixture models, adapted gaussian mixture model and different normalization techniques have been introduced. Classification method of a telephone conversation to the speaker homogeneous segments is expressed. A speaker identification system is implemented and with various experiments was evaluated.**

***Index Terms*—Speaker identification, gaussian mixture model, normalization, DET curve.**

I. Introduction

The human speech signal conveys many levels of information ranging from phonetic content to speaker identity and even emotional status. Speech is one of the natural forms of communication. Recent development has made it possible to use this in the security system. In speaker identification, the task is to use a speech sample to select the identity of the person that produced the speech from among a population of speakers. In speaker verification, the task is to use a speech sample to test whether a person who claims to have produced the speech has in fact done so. This technique makes it possible to use the speakers' voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers. In order to implement the text-independent speaker ID system, one must go through several steps, including feature extraction, feature matching, and finally, identification of the speaker. Feature extraction is a method that takes a small amount of data from the voice signal which can later be used to generate a representation of each speaker. Feature matching involves the actual procedure of using vector quantization to identify the speaker according to the characteristics of the known speakers.

II. Conventional Method in Speaker Identification

Successful method in speaker identification systems is statistical modeling using the feature vectors extracted from the given speech samples. The most important distinction between the speaker identification systems is in type of feature vectors and the statistical model. By having statistical model of the speakers the problem is reduced to identifying the most likely model. By comparing this likelihood with a threshold level, it is clear that this speaker is between well-known speakers or not. A speaker identification system consists of two phases: learning phase and testing phase.

A. Mel Frequency Cepstral Coefficients (MFCC)

The first step of speech signal processing involves the conversion of analog speech signal into digital speech signal. Framing is the process of segmenting the speech samples obtained from the analog to digital conversion into small frames with time length in the range of 20ms to 40 ms. Windowing is performed on each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. Thus for each tone with an actual frequency f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. One approach to simulating the subjective spectrum is to use a filter bank. That filter bank has a triangular band pass frequency response. The number of mel spectrum coefficients K, is typically chosen as 20. In this final step, the conversion of the log mel spectrum back to time. The result is called the mel frequency cepstrum coefficients (MFCC).discrete cosine transform (DCT) is used to convert them back to the time domain.

B. Gaussian mixture model

In GMM, we model the speaker data using statistical variations of the features. Hence, it provides us a statistical representation of how speaker produces sounds. Gaussian mixture density is shown to provide a smooth approximation to the underlying long-term sample distribution of observations obtained from utterances by a given speaker. These are the important motivations for using GMM as a modeling technique.

A Gaussian mixture density is a weighted sum of M component densities and is given by the equation

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

Each component density is a D-variate Gaussian function of the form:

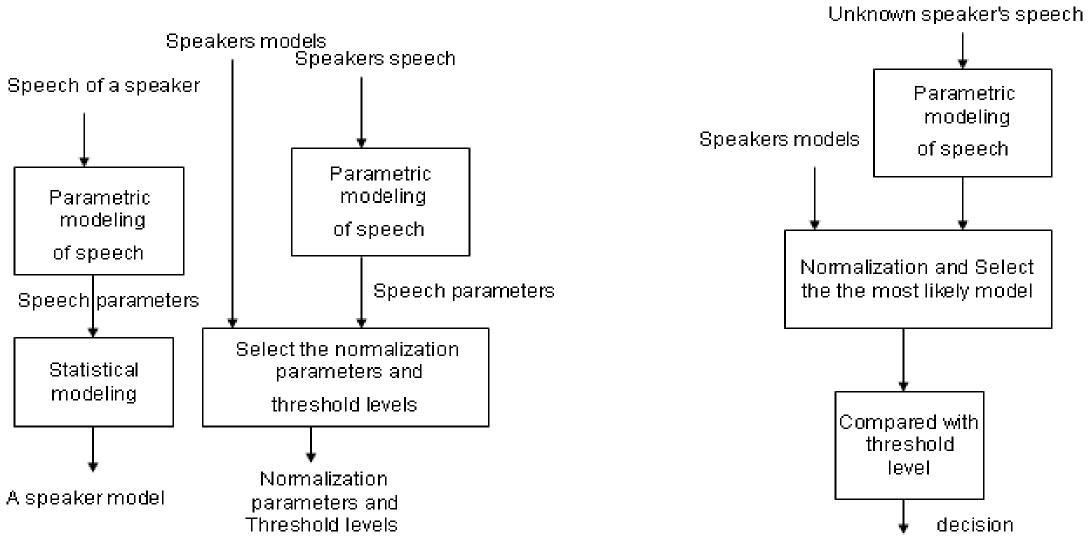
$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)'\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)\right\} \tag{2}$$

with mean vector $\vec{\mu}_i$ and covariance matrix Σ_i . The mixture weights satisfy the constraint that $\sum_{i=1}^{M} p_i$.

Manuscript receive March 16, 2012; revised May 10, 2012.

Mohsen Bazyar is with the Department of Communication, Bushehr Branch, Islamic Azad University, Iran (e-mail: mohsenbazyar114@yahoo.com).

Ahmad Keshavarz and Reza Dyanat are with the Persian Gulf University of Bushehr, Iran.


Speakers models
Speakers speech
Speech of a speaker
Parametric modeling of speech
Parametric modeling of speech
Speech parameters
Speech parameters
Statistical modeling
Select the normalization parameters and threshold levels
A speaker model
Normalization parameters and Threshold levels
Unknown speaker's speech
Parametric modeling of speech
Speakers models
Normalization and Select the the most likely model
Compared with threshold level
decision

(a) (b)

Fig. 1. (a) training phase; (b)testing phase

The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by: $\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\}, i = 1,.,\mathrm{M}$

C. Universal background models

In the GMM-UBM system, we derive the hypothesized speaker model by adapting the parameters of the UBM using the speaker's training speech and a form of Bayesian adaptation Unlike the standard approach of maximum likelihood training of a model for the speaker independently of the UBM, the basic idea in the adaptation approach is to derive the speaker's model by updating the well-trained parameters in the UBM via adaptation. Like the EM algorithm, the adaptation is a two step estimation process. The first step is identical to the expectation step of the EM algorithm. Unlike the second step of the EM algorithm, for adaptation these new sufficient statistic estimates are then combined with the old sufficient statistics from the UBM mixture parameters using a data-dependent mixing coefficient.

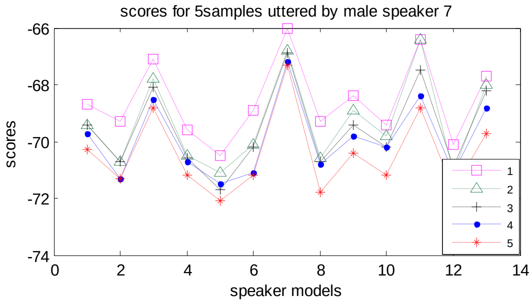


Fig. 2. Raw scores to 5 speech samples from one speaker

III. Score Normalization

This section reviews the various normalization methods tested in this work, i.e.., world model , Z-norm and T-norm for score normalization .The use of score normalization techniques has become important in GMM based speaker identification systems for reducing the effects of the many sources of statistical variability associated with log likelihood ratio scores. With the log-likelihood score $f(y|m)$ for the speaker S and the utterance **X**, and the log likelihood score $f(y|m_w)$ for the world model of speaker S and utterance **X**, the normalized score is then given by

$$L_m(y) = \frac{f(y|m)}{f(y|m_w)} \tag{3}$$

Five parts of speech from one speaker has been given to 13 different models and through them No. 7 was the model corresponding to the input speech. The raw scores are given in figure 3. Figure 4 shows the same scores after normalization by the World model. Zero normalization, Z-norm in short, is one of score normalization methods applied for speaker verification at the score .variations in a given utterance can be removed by making the log-likelihood scores relative to the mean and variance of the distribution of the impostor scores.

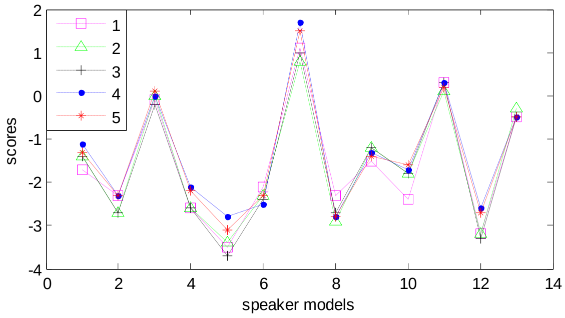


Fig. 3. Score models after the normalization using the world model

$$L_m^z(y) = \frac{L_m(y) - \mu_m}{\sigma_m} \tag{4}$$

where μ_m and σ_m are the mean and standard deviation of the distribution of the impostor scores, which are calculated based on the impostor model. The advantage of this method is that the normalization parameters for each model were calculated in the training phase. T-norm is also called as test-norm because this method is based on the estimation on the test set. Essentially, T-norm can be regarded as a further improved version of Z-norm, as the normalization formula is very similar to that of Z-norm, at least in formality. That is, a normalized score is obtained by:

$$L_m^z(y) = \frac{L_m(y) - \mu_{m-test}}{\sigma_{m-test}} \qquad (5)$$

where μ_{m-test} and σ_{m-test} are the mean and standard deviation of the distribution of the impostor scores estimated on a test set. Experiments have shown that this approach could get better results than z-norm.

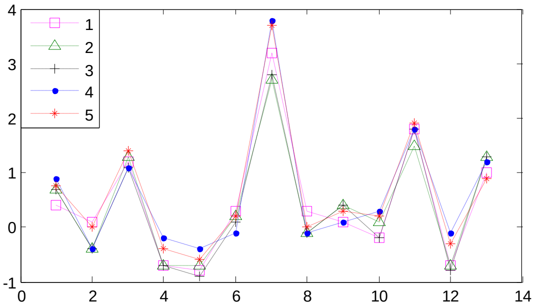


Fig. 4. Score models after the normalization using the world model and z-norm

IV. Experimental Results

A. Comparison of normalization methods

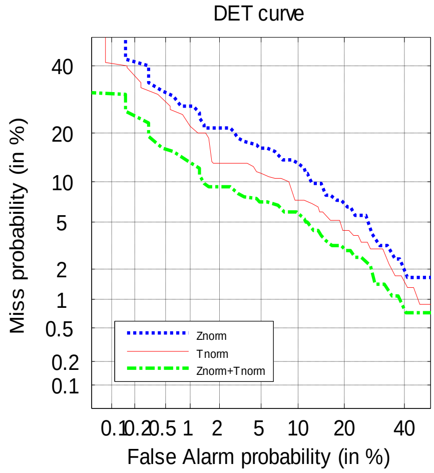


Fig. 5. Comparison of normalization methods

In our experiments, we used feature vectors composed from 12 mel-frequency cepstral coefficients (MFCC) computed using 20 traingular Mel filters. Analysis frame was windowed by 30 milliseconds Hamming window with 10 milliseconds overlapping. The signal was preemphasized by the filter *H(z)=1-0.97 z-1* and silence frame were removed before the feature extraction. As a test materials for our experiments we used the farsdat database. a database of forty speakers, twenty males and twenty females has been prepared to make world model with sampling frequency of 16 kHz and 16 bits per sample. 13 speakers were trained as the target speakers.

In this experiment normalization methods were compared with each other. First method is using Znorm, second method is using Tnorm and in third method is using Tnorm After applying the Znorm. DET curve from this experiment are shown in figure 6. DET curves are plotted with the false alarm rate on the horizontal axis and the miss probability on the vertical axis. These results are related to the length of 3 to 15 seconds. Therefore 1450 speech samples have been tested. The third method has been much more successful than the other methods. And Tnorm show better results than the Znorm.

B. System Evaluation by Changing the time Length of the Speech

In this experiment, system performance is evaluated according to the input speech. The system was tested for different durations: 1s, 3s, 6s, 9s and 12s. Results are shown in figure 7. System error is reduced by increasing the time length of the input speech and this error reduction is not uniform and by increasing the length of the input speech can be reduced to lower levels.

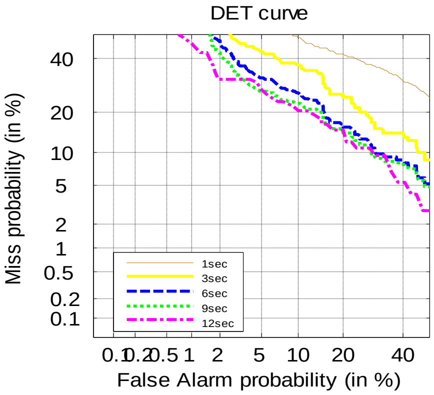


Fig. 6. Comparison of the time length of the speech

TABLE.1: (A)Comparisons of System Accuracy for Depend and Independent Threshold Level (B) Evaluation of the Adaptive Model

accuracy		accuracy		Length of the speech (s)
Adaptive model %	Unadaptive model %	Model-dependent %	Model-independent %	
84.5	74.8	76.9	70.3	3
89.0	78.3	81.4	79.3	6
89.3	82.1	83.8	82.4	9
89.3	83.8	85.2	84.1	12
89.3	81.4	84.8	84.1	15

(a) (b)

C. Model-dependent and model-independent threshold level

Speaker identification system can be used a threshold level for all models or have be a separate threshold level for each model. In this experiment, both methods were compared. In both cases, the threshold levels are calculated in training phase and this threshold are used in testing phase. Table 1 is shown the system accuracy for various length of the speech. Use of model-dependent threshold levels efficiency of the system significantly increases. Adapted model from the world model looks much better than the model was trained with the speaker's speech. In this experiment the performance of these two methods are compared and results are given in Table 1.

V. Conclusion

This paper has introduced and evaluated the use of GMM for text-independent speaker identification. Experimental results show that the proposed system has a good accuracy. The experimental evaluation examined several aspects of using GMM for text-independent speaker identification system:

- Improve the performance of Zero-normalization with increasing speech sample.
- Improve efficiency by increasing the number of replications of training to 13 iterations.
- The GMM maintains high identification performance with increasing number of speakers.
- Increase accuracy with increasing duration of the training speech and suitability for 2 minutes.

References

[1]. X. Lu and J. Dang, “An Investigation of Dependencies between Frequency Components and Speaker Characteristics for Text-Independent Speaker Identification,” *Speech Communication*, no. 50, pp. 312- 322, 2008.

[2]. B., Robert Reynolds, A. Douglas, Quatieri, and F. Thomas, “Approaches to Speaker Detection and Tracking in Conversational Speech,” *Digital Signal Processing 10*, pp. 93-112, 2000.

[3]. Bimbot, Frederic *et al*, “A Tutorial on Text-Independent Speaker Verification,” *EURASIP Journal on Applied Signal Processing 2004:4,* pp. 430-451

[4]. Deller, R. John Jr., Hansen, H. L. John, Proakis, and G. John, *Discrete-Time Processing of Speech Signals*, IEEE Press, Wiley Publishing, 2000.

[5]. S. Young, *et al*., The *HTK Book 3.1,* 2002