Improving Text-Independent Speaker Identification Performance Using Gaussian Mixture Speaker Models

Basel Shbita

under the direction of
Dr. David Palmer
Autonomy Virage, Inc.

Research Science Institute
July 28, 2009

Abstract

Systems that automatically recognize a speaker are increasingly important in human-computer interaction because speech communication has always been and will continue to be the dominant mode of human social bonding and information exchange. This paper investigates the use of Gaussian mixture models (GMMs) for robust text-independent speaker identification. The experiments performed in this research examine several aspects and parameters of GMM usage: algorithmic issues, amount of training data, modeling different languages, and small and large population performance. We found that increasing the amount of training data and decreasing the number of speakers improved the accuracy of text-independent speaker identification using statistical models based on Gaussian mixture models. There also appears to be a maximum number of Gaussian mixture components needed to adequately model speakers and achieve good identification performance for different amounts of training data.

1 Introduction

From human prehistory to the new media of the future, speech communication has been and will be the dominant mode of human social bonding and information exchange. In addition to human-human interaction, this human preference for spoken language communication finds a reflection in human-machine interaction as well. Most computers operating systems and applications depend on a user's keyboard strokes and mouse-clicks, with a display screen as feedback. Today's computers lack the fundamental human abilities to speak, listen, understand, and learn. And even before speech based interaction reaches full maturity, applications in home, mobile and office segments are incorporating spoken language technology to change the way we live and work.

As speech interaction with computers becomes more pervasive in activities, the usage of systems that automatically recognize a speaker increases. Primarily, the speech signal conveys words or message being spoken. The signal also conveys information about the identity of the person talking. While the area of speech recognition is concerned with extracting the message spoken, the area of speaker identification is concerned with extracting the identity of the person speaking [1]. Success in both tasks, recognition and identification depends on extracting, modeling and improving the speaker-independent characteristics of the speech signal which can effectively distinguish one speaker from another.

In this paper, we address the problem of speaker identification. The experiments performed are concerned with text-independent speaker identification which is totally unconstrained with respect to the content of the speech, unlike text-dependent speaker identification systems that require the speech to be a known and specific phrase.

2 Background

2.1 Speech Recognition

A source-channel mathematical model is often used to formulate speech recognition problems [2]. As illustrated in Figure 1, the speaker chooses the source word sequence **W** that is delivered to his/her Speech Generator which produces the speech waveform and performs the speech signal processing component of the speech recognizer [3]. Finally, the speech decoder aims to decode the acoustic signal **X** into a word sequence **W'**, which is hopefully close to the original word **W**.

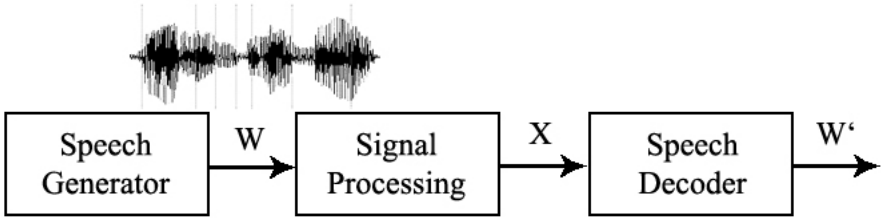

Figure 1: A source-channel model for a speech recognition system [4].

2.2 Speaker Identification

The system determines the identity of a user with statistical models based on Gaussian mixture models (GMMs) [5]: probabilistic models for density estimation which use a mixture distribution. The Gaussian mixture speaker model is an implicit segmentation approach to speaker recognition, providing a probabilistic model of the underlying sounds of a person's voice. The use of Gaussian mixture density for speaker identification is motivated by two interpretations. First, the individual component Gaussians in a speaker-dependent GMM are interpreted to represent broad acoustic classes. These acoustic classes reflect general speaker-dependent vocal tract configurations that are useful for modeling speaker identity. Second,

a 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.

2.2.1 Open and Closed Data Sets

In speaker identification, the reference set of known speakers can be of two types: closed or open [6]. This distinction refers to whether the set contains unknown speakers or not. This paper is concerned only with closed data sets: those including only known speakers. Closed set speaker identification is an easier task than open set speaker identification. With closed set speaker identification the speaker is identified using a nearest neighbor approach, so no thresholding is needed. In open set speaker identification, on the other hand, the closest known speaker is not necessarily the actual speaker, so one has to use a pre-determined threshold to identify samples that are close enough to be deemed to be from the same speaker.

2.2.2 Extraction of Mel-Frequency Cepstral Coefficients (MFCC)

In sound processing, the Mel-Frequency Cepstral Coefficients (MFCC) collectively make up a representation of the short-term power spectrum of a sound. MFCCs are extracted from the speech as follows: First, the speech is sampled at 16 kHz sampling rate and is segmented into overlapping 20 ms frames every 10 ms, providing 320 samples in a frame. Second, frequency analysis is done on each frame using a discrete Fourier transform (DFT).

Next, Mel-scale cepstral feature vectors are extracted from the speech frames [7] by taking the logarithms of the powers at each of the frequencies and then taking the discrete cosine transform (DCT) of the Mel logarithm powers, as if it were a signal. These transforms are important to numerous applications from lossy compression of audio (where small high-frequency components can be discarded). Finally, we get the Mel-Frequency Cepstral Coefficients (MFCCs) which are now the amplitudes of the resulting spectrum [8]. This

process is illustrated in Figure 2.

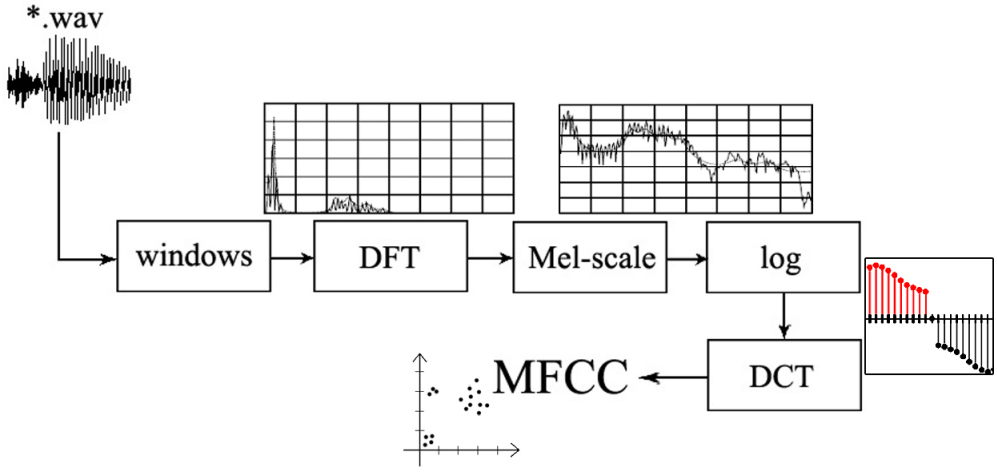

Figure 2: Extraction of Mel-Frequency Cepstral Coefficients (MFCC).

2.2.3 Training and Testing

GMMs are classic parametric models used in many pattern recognition applications. A Gaussian mixture density is a weighted sum of M component densities, as given by the equation

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

where $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x}), i = 1, ..., M$, are the component densities and $p_i, i = 1, ..., M$ are the mixture weights. 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 $\sum_{i=1}^{M} p_i = 1$.

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 the notation

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

For speaker identification, each speaker is represented by a GMM and is referred to by his/her model λ.

The Gaussian mixture model can have several different forms depending on the choice of covariance matrices. The covariance matrix has three different types, including one covariance matrix per Gaussian component as indicated in Equation 3 (nodal covariance), one covariance matrix for all Gaussian components (grand covariance), or a single covariance matrix shared by all speaker models. In addition, the covariance matrix can be full or diagonal. In this paper, nodal, diagonal covariance matrices are primarily used for speaker models.

There are several methods for estimating the parameters of a GMM, but the maximum likelihood (ML) [9] estimation is the most well-established one. The goal of ML estimation is to derive the optimum model parameters that can maximize the likelihood of a certain GMM. Generally, the model parameters are extracted by the Expectation-Maximization (EM) algorithm [10].

For speaker identification, a group of S speakers is represented by GMMs $\{\lambda_1, \lambda_2, ..., \lambda_s\}$. The objective is to find the speaker model which has the maximum *a posteriori* probability for a given observation sequence. Formally,

$$\hat{S} = \underset{1 \le k \le S}{\operatorname{argmax}} \; Pr(\lambda_k | X) = \underset{1 \le k \le S}{\operatorname{argmax}} \; \frac{p(X|\lambda_k) Pr(\lambda_k)}{p(X)} \tag{4}$$

where the second equation reflects the application of Bayes' rule. Assuming *a priori* equally

likely speakers (i.e., $Pr(\lambda_k) = \frac{1}{S}$) and noting that $p(X)$ is the same for all speaker models, the classification rule simplifies to

$$\hat{S} = \underset{1 \leq k \leq S}{\operatorname{argmax}} \, p(X|\lambda_k). \tag{5}$$

Using the assumed independence between observations, the speaker identification system computes

$$\hat{S} = \underset{1 \leq k \leq S}{\operatorname{argmax}} \sum_{t=1}^{T} \log p(\vec{x}_t|\lambda_k) \tag{6}$$

in which $p(\vec{x}_t|\lambda_k)$ is given in Equation 1.

The test speech produces a sequence of feature vectors $\{\vec{x}_1, ..., \vec{x}_t\}$. The format of the files used is very simple, each one of them just contains one 60-dimensional vector. When testing is performed, vectors' values from the test utterance are compared with all speaker models. The speaker's model with the highest likelihood score in a closed set is the one determined. If the data set is open, a specific percentage value is required.

3 Data Sets

The experiments performed are primarily conducted using WAV-audio files (*.wav) which are organized in folders representing different languages:

Folder name	Language
ARMSA	Arabic / Modern Standard Arabic
DEDE	Deutsch / Deutschland
ENUK	English / United Kingdom
ESES	Español / España
FIPH	Filipino / Philippines
FRFR	French / France
IDID	Indonesian / Indonesia
ITIT	Italian / Italy
JAJP	Japanese / Japan
KOKR	Korean / Korea
MNMN	Mongolian / Mongolia
PLPL	Polish / Poland
PTPT	Portuguese / Portugal
RURU	Russian / Russia
THTH	Thai / Thailand
UGCN	Uyghur / China
VIVN	Vietnamese / Vietnam
ZHCN	Mandarin Chinese / China

Table 1: A table showing the language spoken in each folder.

Each folder in Table 1 contains a couple of sub-folders (5-14). Each one of the sub-folders represents different speaker and contains two WAV-audio files: "train.wav," which has the audio information needed for the training model, and "test.wav," which has the audio information needed for the testing model. The duration of each file is between 30 seconds to 500 seconds, depending on the type of the experiment.

This data set consists 150 speakers, 45% women and 55% men. The speakers' training and test files can be used without being separated by language (language-independent) or can be used within certain language (language-dependent).

4 Experiments

Determining the number of the components in a mixture needed to model a speaker adequately is an important but difficult problem. There is no theoretical way to estimate the number of mixture components *a priori*. For speaker modeling, the objective is to choose the minimum number of components necessary to adequately model a speaker for good speaker identification. Choosing too few mixture components can produce a speaker model which does not accurately model the distinguishing characteristics of a speaker's distribution. Choosing too many components can cause over-fitting, which reduces performance when there are a large number of model parameters relative to the available training data and can also result in excessive computational complexity both in training and testing. The models for each speaker were so specific that they didn't reflect more general characteristics.

The following experiments examine the performance of the GMM speaker identification system for different model orders with respect to the number of Gaussian component densities used, amount of training data and number of speakers within specific languages.

4.1 Language-dependent closed speaker set, varying number of component densities and training speech duration

In the first experiment we assume that the language is already identified and that the problem is to identify the speaker within that language (language-dependent) while it is known that the owner of the unknown voice is one of the known speakers; a closed data set. In this experiment speaker models with 32, 64, 128 and 256 Gaussian component densities were trained using 30, 60, 120 and 180 seconds of speech. For each speaker model within each language, the identification test performance was for 60 seconds of speech.

4.2 Language-dependent closed speaker set, varying number of speakers for different languages

A language-dependent speaker identification within a closed data set is performed in the second experiment. In this experiment speaker models with 1, 2, 4, 8, 16, 32, 64, 128 and 256 Gaussian component densities are trained using 180 seconds of speech. For each speaker model within each language, the identification test performance is for 60 seconds of speech. The experiment is performed on 3 languages with 6 speakers and each time the number of speakers is reduced by one. Next, the same experiment is performed on all languages with 5 speakers.

4.3 Language-independent closed speaker set, varying number of component densities for a large number of speakers

A language-independent speaker identification within a closed data set is performed in the third experiment. In this experiment speaker models with 1, 2, 4, 8, 16, 32, 64, 128 and 256 Gaussian component densities are trained using 180 seconds of speech and performed on 98 speakers. This experiment is performed in order to give us results for a very large number of speakers compared to the other experiments that are performed on a small number of speakers.

5 Results

5.1 Language-dependent closed speaker set, varying number of component densities and training speech duration

Amount of Training Data	Number of Mixture Components	Accuracy (%)
30 sec	32	98.98
	64	96.94
	128	92.86
	256	77.55
60 sec	32	100
	64	100
	128	100
	256	79.59
120 sec	32	100
	64	98.98
	128	100
	256	100
180 sec	32	100
	64	100
	128	100
	256	100

Table 2: GMM identification performance for different amounts of training data and number of mixture components.

As we expected, with increased training data, identification performance increases (Figure 3). The largest increase in performance occurs when the amount of training data increases from 60 seconds to 120. Increasing the training data to 180 also improves performance but with a smaller increment.

Before the experiment was performed, we assumed that increasing the number of the Gaussian mixture components would make the results more accurate, because more model parameters are given. But, we were surprised to find out that with increased number of Gaussian mixture components, identification performance decreases (Figure 4). This is ap-

Figure 3: Average accuracy as a function of the training time.

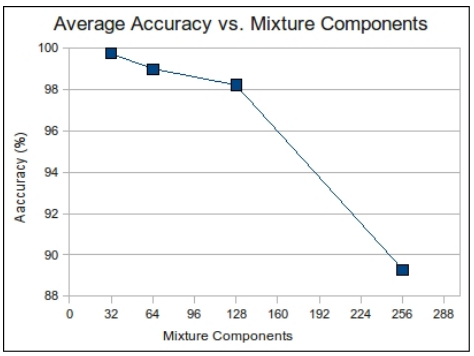

Figure 4: Average accuracy as a function of the number of mixture components.

parently caused by over-fitting. The models for each speaker were so specific that they did not reflect more general characteristics.

As we can see in Figure 5, when the training time is 30 seconds, it affects the performance using any number of Gaussian mixture components. This suggests that more than one minute of conversational speech is necessary to maintain higher speaker identification performance and using more training data improves performance at a decreasing rate.

We can also see that when the number of components is 32 and the training data is at least 60 seconds, it is actually enough to get 100% accuracy for speaker identification within any language in this closed data set.

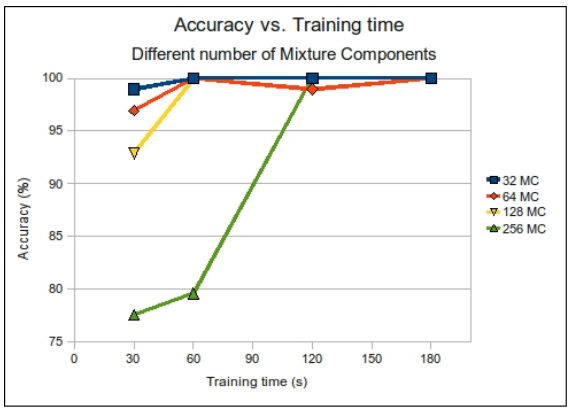

Figure 5: Speaker identification performance as a function of the training time for different number of Gaussian mixture components.

5.2 Language-dependent closed speaker set, varying number of speakers for different languages

Number of speakers	Accuracy(%)		
	FIPH	RURU	ENUK
6	92.59	79.63	88.89
5	93.65	91.11	90.74
4	95.56	91.67	91.11
3	100	100	92.59
2	100	100	100
1	100	100	100

Table 3: GMM identification performance for different number of speakers within different languages.

In this experiment, speaker models with different Gaussian component densities were trained using 180 seconds of speech. As we expected, with increased number of speakers within the data set, identification performance decreases for all the languages examined (Figure 6).

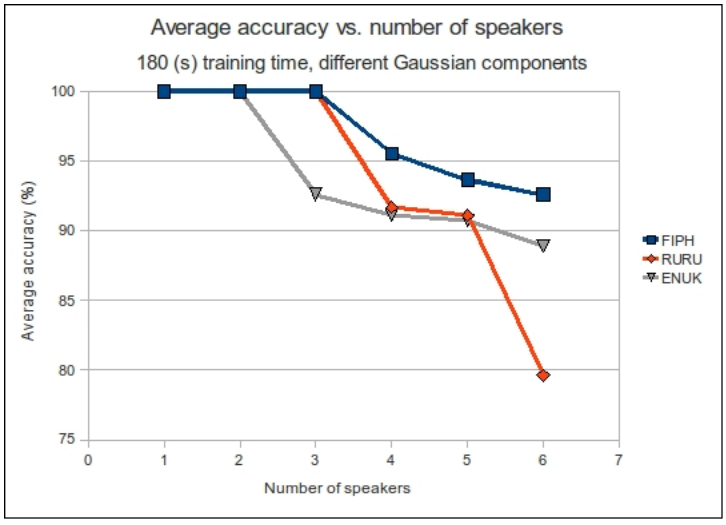

Figure 6: Speaker identification performance as a function of a different number of speakers within different languages.

As we can see in Figure 7, for a constant number of speakers (5 speakers), speaker identification performance was different across languages. Uyghur language (China) got the lowest accuracy percentage from all the languages. English (United Kingdom), Russian, Vietnamese, Portuguese and Filipino didn't get high accuracy percentage compared to the other languages. All of the other languages got 100% accuracy for the same parameters. This might be caused by the difference in the tone of voice, rhythm, style, pace and accent between different languages.

Note that this experiment was performed with 15 languages out of 18. The Mongolian, Thai and Polish languages did not have enough speakers to be used in this experiment.

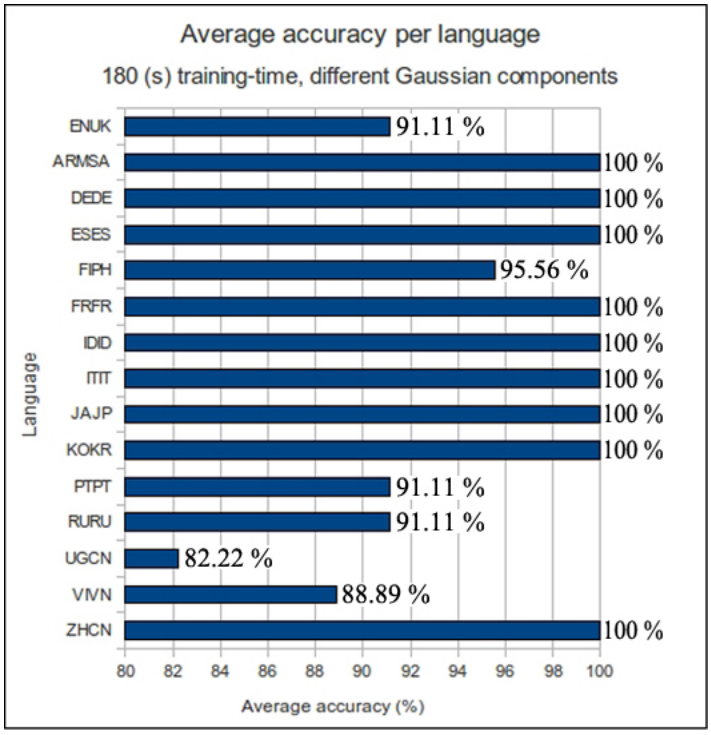

Figure 7: Speaker identification performance for different languages examined with the same number of speakers (5 speakers).

5.3 Language-independent closed speaker set, varying number of component densities for a large number of speakers

Number of Mixture Components	Accuracy (%)
1	98.98
2	98.98
4	98.98
8	98.98
16	100
32	100
64	98.98
128	1.02
256	1.02

Table 4: GMM identification performance for different number of mixture components.

In this experiment speaker models with 1, 2, 4, 8, 16, 32, 64, 128 and 256 Gaussian component densities were trained using 180 seconds of speech and performed on 98 speakers. Before the experiment was performed, we assumed that a small number of the Gaussian mixture components would not be enough to give us high speaker identification accuracy. But, we were surprised to find out that even the 1 Gaussian which includes 60-dimensional Mel-cepstral vectors was enough to give us high accuracy.

As we can also see in Table 4 and Figure 8, when we used 16 and 32 Gaussian components, speaker identification performance was on 100% accuracy, which was also unexpected for a such a large number of speakers. When the number of Gaussian components was 128 and 256 the accuracy decreased and was very small, which apparently happened because of the large number of model parameters relative to the available training data, resulting in excessive computational complexity both in training and testing. This is a classic example of over-fitting. The models for each speaker were so specific that they did not reflect characteristics useful for prediction.

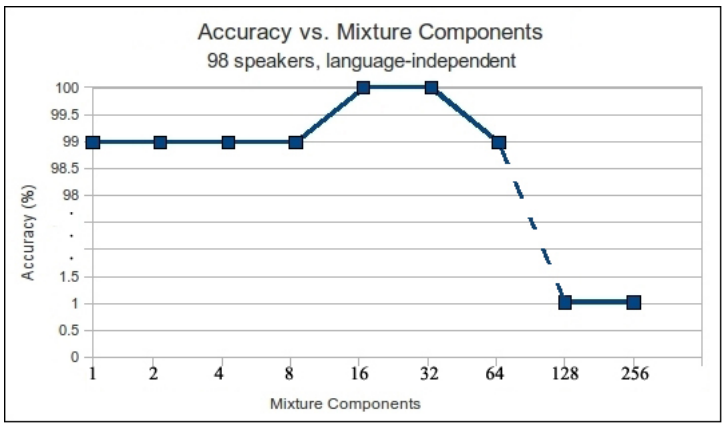

Figure 8: Speaker identification performance as a function of a different number of Gaussian mixture components for 98 speakers.

6 Conclusion

The experimental evaluation in this paper examined several aspects and parameters using Gaussian mixture models for text-independent speaker identification. Some observations and conclusions are:

- Identification performance of the Gaussian mixture speaker model increases with increased amount of training data.
- There appears to be a maximum number of Gaussian mixture components needed to adequately model speakers and achieve good identification performance for different amounts of training data, both for a large number of speakers and for a small number of speakers. Choosing too many components caused over-fitting, where the models for each speaker were so specific that they did not reflect more general characteristics.
- The Gaussian mixture speaker model's identification performance increases with decreased number of speakers within a specific language.
- Speaker identification performance for various languages was different while using constant parameters. It is likely that it is caused by the difference in the tone of voice,

rhythm, style, pace and accent between different languages.

The results in this paper indicate that Gaussian mixture models provide a robust speaker representation for the difficult task of speaker identification.

7 Acknowledgments

I would like to thank Dr. David Palmer, my mentor, and Mr. Mahesh Krishnamoorthy, Mr. Alvin Garcia, and all of Autonomy Virage's employees for giving me the opportunity to work in this field and to use their equipment for this research. I would also like to thank the Center for Excellence in Education and the Massachusetts Institute of Technology for sponsoring me this summer.

I would also like to thank my tutor, Mr. Galen Pickard and Prof. Daoud Bashouty for all the help and the support they provided to me. I offer my sincerest thanks and gratitude to my sponsors, Mr. Steve Shapiro and Mr. Sami Geraisy.

References

[1] C. D. Manning and H. Schuze. *Foundations of Statistical Natural Language Processing.* 6th Ed. Massachusetts Institute of Technology (2003).

[2] X. Huang, A. Acero and H. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice-Hall PTR, Upper Saddle River, New Jersey 07458 (2001).

[3] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition.* Prentice-Hall PTR, Upper Saddle River, New Jersey 07458 (1993).

[4] F. Jelinek. *Statistical Methods for Speech Recognition,* Language, Speech, and Communication. Cambridge, MA, MIT Press (1998).

[5] D. A. Reynolds and R. C. Rose. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Transactions of Speech and Audio Processing, VOL. 3, NO. 1 (January 1995), 72-83.

[6] P. Rose. *Forensic Speaker Identification.* Taylor & Francis, New York (2002).

[7] D. A. Reynolds. Speaker Identification and Verification Using Gaussian Mixture Speaker Models. Speech Communication, VOL. 17 (March 1995), 91-108.

[8] S. Young, G. Evermann, M. Gales, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland. *The HTK Book (for HTK version 3.4).* Cambridge University Engineering Department, Cambridge, UK (2006).

[9] G. McLachuo. *Mixture Models.* Marcel Dokker, New York (1998).

[10] A. Dempster, N. Laird and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc., VOL. 39, (1977), 1-38.