Robust Speaker Identification Using Fusion of Features and Classifiers

Smarajit Bose, Amita Pal, Anish Mukherjee, and Debasmita Das

***Abstract*—Speaker identification using Gaussian Mixture Models (GMMs) based on Mel Frequency Cepstral Coefficients (MFCCs) as features, proposed by Reynolds (1995), is one of the most effective approaches available in the literature. The use of GMMs for modeling speaker identity is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes, and the capability of mixtures to model arbitrary densities. In this work, we have established empirically how combining two different well-known set of features (MFCCs and Perceptual Linear Predictive Coefficients) and using ensemble classifiers in conjunction with principal component transformation and some robust estimation procedures, can be used to enhance significantly the performance of the MFCC-GMM speaker recognition systems, using the benchmark speech corpus NTIMIT.**

***Index Terms*—Mel frequency cepstral coefficients, Perceptual Linear Predictive Coefficients, Gaussian mixture models, ensemble classifiers, classification accuracy, trimmed means, NTIMIT.**

I. Introduction

Automatic speaker identification/recognition (ASI/ASR) is the generic term applied to the automatic process of inferring the identity of a person from an utterance made by him, on the basis of speaker-specific information embedded in the corresponding speech signal. This technique has important practical applications, e.g., it can be used to verify the identity claimed by persons trying to access secure systems, that is, it enables access control of various services by voice. Other real-life activities where it is immediately applicable and useful include voice dialing, banking over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information, and remote access to computers. Another important application of speaker recognition technology is in forensics.

Speaker recognition, being essentially a pattern recognition problem, can be specified broadly in terms of the features used and the classification technique adopted. From experience gained over the past several years from research going on, it has been possible to identify certain features extracted from the complex speech signal, that carry a great deal of speaker-specific information. In conjunction with these features, researchers have also identified classifiers which perform admirably. Mel Frequency Cepstral Coefficients (MFCCs) and Perceptual Linear Predictive Coefficients (PLPCs) are the popularly used features, while Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), Vector Quantization (VQ), Neural Networks are some of the more successful speaker models/classification tools. Any good review article on speaker recognition (for example, [1], [2]) contains details and citations about more than a few of these features and models. It is quite apparent that much of the research involves juggling various features and speaker models in different combinations to get new ASR methodologies.

Reynolds [3], Reynolds and Rose [4] proposed a speaker recognition system based on MFCCs as features and GMMs as speaker models and, by implementing it on the benchmark data sets TIMIT and NTIMIT, demonstrated that it works almost flawlessly on clean speech (TIMIT) and quite well on noisy telephone speech NTIMIT). This approach is still one of the best available in the literature.

In this paper, we have established empirically, with the help of the benchmark speech corpus NTIMIT, how the classification accuracy of the basic MFCC-GMM speaker recognition system can be further enhanced significantly by

1) *combining the two feature sets (MFCCs and PLPCs):* It was evident that both the feature sets have relevant information regarding the identity of the speaker though MFCCs had a little edge. After combining the two feature sets, the classification accuracy vastly improved.
2) *implementing robust estimation procedures like the trimmed mean, to eliminate the effect of outliers*, that is, observations that are too different from the majority of observations, and may be due to the inherent variability in the data set or to measurement error.

We also use the following two ideas from our previous work [5]:

1) *incorporating into the model the individual correlation structures of the feature sets for each speaker*: This is a significant aspect of the speaker models that Reynolds ignored totally by assuming the MFCCs to be independent. This is achieved by the simple device of the Principal Component Transformation (PCT) [6]), which is a linear transformation derived from the covariance matrix of the MFCC vectors obtained from the training utterances of a given speaker, and is applied to the MFCC vectors of the corresponding speaker to make the individual coefficients uncorrelated. Due to differences in the correlation structures, these transformations are also different for different speakers. The GMMs are fitted on the MFCCs transformed by the principal component transformations instead of the original

Manuscript received September 3, 2017; revised October 17, 2017.

Smarajit Bose and Amita Pal are with the Interdisciplinary Statistical Research Unit, Applied Statistics Division, Indian Statistical Institute, Kolkata, India (e-mail: {smarajit,pamita}@isical.ac.in).

Anish Mukherjee is with the Department of Statistics, University of Missouri, Colombia, MO, USA (e-mail: anishmk9@gmail.com).

Debasmita Das is with Department of Statistics, University of Connecticut, Storrs, CT, USA (e-mail: debasmita88@yahoo.com).

doi: 10.18178/ijmlc.2017.7.5.635

MFCCs. For testing, to determine the likelihood values with respect to a given target speaker model, the MFCCs from the test utterance are transformed by the principal component transformation corresponding to that speaker.

2) *Using ensemble classifiers:* It is well known now that the use of an ensemble of classifiers instead of a single classifier can improve the accuracy to a great extent. In this paper we have used a clever idea to design an ensemble classifier which further improved the classification accuracy.

The paper is organized as follows. MFCCs are introduced in the following section, while Gaussian Mixture Models (GMMs) are briefly described in Section III, which also outlines how speaker recognition is carried out using MFCCs as features and GMMs as speaker models. The proposed approach is delineated in Section IV. Section V gives a brief description of the speech corpus used, namely, NTIMIT, and contains results obtained by applying the proposed approach on it, which clearly establish its effectiveness. Section VI contains concluding remarks.

II. Mel Frequency Cepstral Coefficients

The Mel Frequency Cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log-energy spectrum on a nonlinear mel scale of frequency. It exploits auditory principles, as well as the decorrelating property of the cepstrum, and is amenable to compensation for convolution distortion. As such, it has turned out to be one of the most effective feature representations in speech-related recognition tasks [7].

Mel-frequency cepstral coefficients (MFCCs) [8] are coefficients that collectively make up an MFC. A given speech signal is partitioned into overlapping segments or frames, and MFCCs are computed for each such frame. Based on a bank of K filters, a set of *M* MFCCs is computed from each frame as follows:

Let $x[m]$, $w[m]$ denote respectively the speech signal and a window function at a time point m within the frame. The speech waveform $x[m]$ is windowed with $w[m]$, and its short-time Fourier transform (STFT), $Y(n, \omega_k), n = 1,2, \dots, N,$ is computed as

$$Y(n, \omega_k) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\omega_k m},$$

where $\omega_k = \frac{2\pi}{N}k$, N being the length of the discrete Fourier transform. The magnitude of $Y(n, \omega_k)$ is then weighted by a series of filter frequency responses whose center frequencies and bandwidths roughly match those of the auditory critical band filters, that is, the so-called mel-scale filters, collectively referred to as a mel-scale filter bank (see below). If the frequency response of the *l*-th mel-scale filter is denoted by $V_l(\omega)$, then its energy at n is

$$E_{mel}(n, l) = \frac{1}{A_l} \sum_{k=L_l}^{U_l} |V_l(\omega_k) Y(n, \omega_k)|^2,$$

where L_l and U_l denote respectively the lower and upper frequency indices over which the *l*-th filter is non-zero, and

$$A_l = \sum_{k=L_l}^{U_l} |V_l(\omega_k)|^2.$$

Finally, the i-th MFCC computed from the frame is

$$MFCC_i = \sum_{k=1}^{K} \log(E_{mel}(i, k)) \ \cos\left[i\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right],$$
$$i = 1,2, \dots, M.$$

A. Mel-scale Filter Banks

A mel-scale filter bank (Fig. 1) is a set of filters spaced uniformly on the mel scale (described below), which has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval.

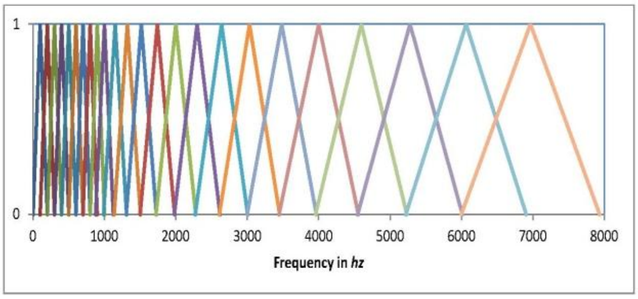

Fig. 1. A mel scale filter bank.

B. The Mel Scale

Psychophysical studies show that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. For each tone with an actual frequency, *f*, measured in *hertz*, a subjective pitch is measured on the so-called 'mel' scale. The mel scale is a scale of pitches judged by listeners to be equal in distance from one another. The word mel comes from the word melody to reflect this. This scale has linear frequency spacing below 1000 hz and logarithmic spacing above 1000 hz (Fig. 2).

A popular formula to convert *f* hertz into *m* mel is:

$$m = 2595 \ \log_{10}\left(1 + \frac{f}{700}\right).$$

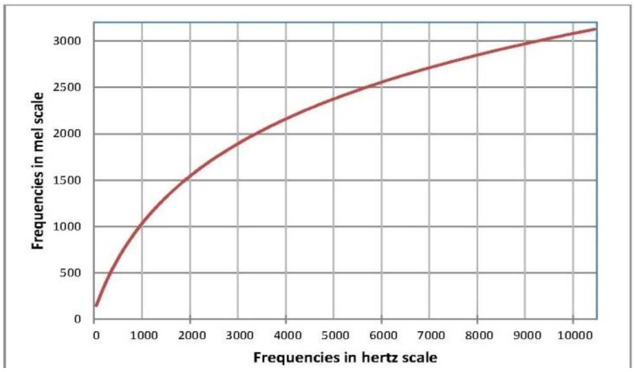

Fig. 2. The mel scale.

C. Computation of MFCCs

This involves the following steps:

1) Partitioning the speech signal into overlapping segments or frames
2) Taking the Fourier transform of signal from each frame.
3) Mapping the powers of the spectrum obtained above onto

the mel scale, using triangular overlapping windows.

4) Taking the logs of the powers at each of the mel frequencies.
5) Taking the discrete cosine transform of the list of mel log powers, as if it were a signal.

III. Perceptual Linear Predictive Coefficients

Perceptual Linear Prediction is a method of spectral estimation proposed by Hermansky [9]. In different psycho-acoustic experiments it was observed that human frequency resolution varies over different frequency ranges and low frequencies mask higher ones. Moreover, it has been found that hearing is most sensitive at mid-frequencies. While listening people generally integrate 1 bark of spectrum, whereas for discrimination purpose people seem to integrate about 3.5 barks of spectrum. These observations inspired the development of Perceptual Linear Predictive Coefficients (PLPC) which turned out to be superior in many ways to the Linear Predictive (LP) coefficients in the task of speaker identification.

In this technique of speech analysis, mainly three psycho-acoustic concepts are used to estimate the auditory spectrum which are critical-band spectral analysis, the equal loudness curve and the intensity power law. PLP algorithm can be described using the following steps -- first in the spectral analysis phase the speech signal is partitioned into overlapping segments and each segment is weighted by the Hamming window.

The short-term power spectrum $P(\omega)$ is computed for each of these segments. In the next stage, the spectrum $P(\omega)$ is warped along the frequency axis into the Bark Frequency which is then convolved with power spectrum of the simulated critical band masking curve that results in samples of the critical-band power spectrum. In this step, spectral resolution is significantly reduced. The sampled power spectrum is then pre-emphasized by an equal-loudness curve and a cubic-root amplitude compression is performed simulating the power law of hearing. Finally in the autoregressive modeling phase, the resulting spectrum is modeled by a 5th order model using the autocorrelation method of all-pole spectral modeling. The following block diagram shows the steps of PLP algorithm.

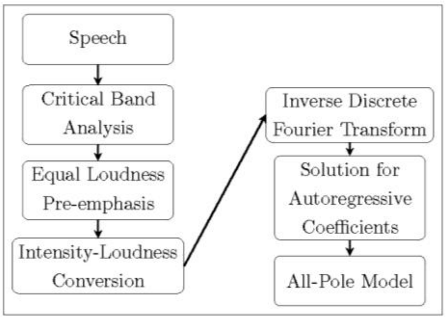
Speech
Critical Band Analysis
Equal Loudness Pre-emphasis
Intensity-Loudness Conversion
Inverse Discrete Fourier Transform
Solution for Autoregressive Coefficients
All-Pole Model

Fig. 3. PLP algorithm.

PLP has an advantage of approximating the speaker independent effective second formant. It reduces the disparity between voiced and unvoiced speech. It has been shown in different experiments that there exists a strong correlation between the perceptually estimated second formant and that estimated by the PLP method.

IV. Speaker Recognition with MFCC-Based GMM Speaker Models

A. Gaussian Mixture Models (GMMs)

If $\boldsymbol{x}$ is a d-dimensional feature vector, then for a K-speaker problem, the probability distribution of the MFCCs obtained from speaker i, i=1, 2,…, K is modeled as a mixture of N component probability densities as follows:

$$p(\mathbf{x}|\lambda_i) = \sum_{j=1}^{N} p_{ij}\, f_j\big(x|\boldsymbol{\theta}_{ij}\big), \qquad \sum_{j=1}^{N} p_{ij} = 1.$$

where, for the i-th speaker, p_{ij} is the prior probability for the j-th component of the mixture, $\lambda_i = \{p_{ij},\ \boldsymbol{\theta}_{ij}, j = 1,2,\cdots,N$ is the collection of unknown parameters, and $f(\mathbf{x}|\boldsymbol{\theta}_{ij})$ is the probability density of $\boldsymbol{x}$ in the j-th component, assumed to be Gaussian in this case. That is, for a GMM,

$$p(\mathbf{x}|\lambda_\mathrm{i}) = \sum_{j=1}^{N} p_{ij}\ \frac{1}{(2\pi)^{\frac{d}{2}}\,|\boldsymbol{\Sigma}_{ij}|}\ e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{ij})^T \boldsymbol{\Sigma}_{ij}{}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{ij})},$$

and $\boldsymbol{\theta}_{ij} = \{\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}\},\ i = 1,2,\ldots,K,\ j = 1,2,\ldots,N.$

GMM models for all speakers are trained by the Expectation-Maximization algorithm [10]. An unknown speech sample is split into a number of overlapping segments, with MFCCs computed from each segment. The likelihood function for the sample is computed, based on all MFCC vectors obtained from it, and it (the unknown sample) is classified by the principle of maximum likelihood, described below.

B. Speaker Recognition by Maximum Likelihood

Consider a speaker database consisting of K speakers where the i-th speaker is being represented by a GMM $p(x|\lambda_i)$ as defined above. If a speech utterance of unknown origin is presented, and it is known that is speaker is represented in the speaker database, the objective of speaker recognition is to identify which of the K speakers could have uttered it.

Suppose the unknown utterance is split into P overlapping frames using the same procedure as for the training samples, and MFCCs are computed from each segment. If $\mathbf{x}_p$ denotes the MFCC vector computed from the p-th segment, then the overall likelihood of the unknown utterance under the i-th speaker model is

$$L(\lambda_i) = \prod_{p=1}^{P} p\big(\mathbf{x}_p|\lambda_i\big),$$

assuming the MFCC vectors from different segments to be mutually independent statistically.

Speaker number k is identified as the speaker of the unknown speech utterance if

$$L(\lambda_k) = \max_{1\le i\le K} L(\lambda_i).$$

Since the logarithm function is monotonically increasing in its argument, maximizing the likelihood function $L(\lambda_i)$ is equivalent to maximizing the log-likelihood

$$\ell(\lambda_i) = \log L(\lambda_i) = \sum_{p=1}^{P} \log p\big(\mathbf{x}_p|\lambda_i\big).$$

Thus speaker number k is identified as the speaker of the unknown speech utterance if

$$\ell(\lambda_k) = \max_{1 \leq i \leq K} \ell(\lambda_i).$$

V. The Proposed Approach

The objective of the proposed approach is to enhance significantly the classification accuracy of the basic MFCC-GMM speaker recognition system, by a combination of the following:

1) *Combining the MFCCs and the PLPCs*: Results were obtained with both these feature sets. Further investigations revealed that the classifiers built based on the two feature sets could identify different speakers accurately. It was then natural to see whether a more powerful classifier can be built with an enhanced feature set by combining both the feature sets.
2) *Using ensemble of classifiers:* Since there were quite a few parameters in the MFCC-GMM model, one could build many classifiers by choosing different combination of values for the parameters. An ensemble classifier based on 3-4 such classifiers was employed for the final classification.

There are many different ways to combine the decisions of different classifiers in an ensemble classifier. Majority voting is quite popular. However we used aggregation of likelihood values of different classifiers and maximized the aggregated likelihood values.

Thus if there are c number of classifiers in an ensemble, then the aggregated likelihood $L'(\lambda_i)$ is

$$L'(\lambda_i) = \sum_{j=1}^{c} L(\lambda_j).$$

3) *Incorporation of the individual correlation structures of the feature sets for each speaker into the corresponding speaker model*: This is achieved through the use of the Principal Component Transformation (PCT) [6], which is described below. The basic MFCC-GMM system ignores this totally by assuming the MFCCs to be independent.

Since correlation structures differ from speaker to speaker, these transformations are also different for different speakers. The GMM for a particular speaker is fitted on the MFCCs transformed by the principal component transformations for that speaker, rather than the original MFCCs. As far as testing is concerned, to determine the likelihood values with respect to a given target speaker model, the MFCCs from the test utterance are transformed by the principal component transformation corresponding to that speaker.

4) *Implementation of robust estimation procedures like the trimmed mean to eliminate the effect of outliers*: Outliers are observations that are too different from the majority of observations, and may be due to the inherent variability in the data set or to measurement error.

This is motivated by the observation that the log-likelihood function $\ell(\lambda_i)$ can be interpreted as being equal to $P\rho(\lambda_i)$, where

$$\rho(\lambda_i) = \frac{1}{P}\sum_{p=1}^{P} \log p(\boldsymbol{x}_p|\lambda_i),$$

which is nothing but the average or arithmetic mean of the P $\log p(\boldsymbol{x}_p|\lambda_i)$ values. Also, maximizing $\ell(\lambda_i)$ over i is equivalent to maximizing $\rho(\lambda_i)$ over i. To obtain a more robust estimate of the quantity $\rho(\lambda_i)$, we make use of the well-known trimmed mean procedure [11], which is described below.

A. Principal Component Transformation (PCT)

This is a widely-used linear orthogonal transformation for converting a set of observations on possibly correlated variables into a set of observations on linearly uncorrelated variables called principal components [6]. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components. Principal components are guaranteed to be independent only if the data set is jointly normally distributed. PCT is sensitive to the relative scaling of the original variables. Depending on the field of application, it is also called the Karhunen–Lo ève transform (KLT), and so on.

Let **X** be a $m \times n$ data matrix each of whose n columns represents an observation on an m-variate random variable $\boldsymbol{U}$. It is assumed that the columns have zero empirical mean (that is, the arithmetic mean of the n observations has been subtracted from each of them). If the $m \times m$ matrix $\boldsymbol{\Sigma}$ is the dispersion matrix of the observations, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, the corresponding eigenvectors being $\boldsymbol{P}_1, \boldsymbol{P}_2, \cdots, \boldsymbol{P}_m$, then the principal component transformation of **X** that preserves dimensionality (that is, gives the same number of principal components as the original variables) is given by

$$\mathbf{Y} = \mathbf{PX},$$

where **P** is a $m \times m$ orthogonal matrix having,$\boldsymbol{P}_1, \boldsymbol{P}_2, \cdots, \boldsymbol{P}_m$ as its columns, and the columns of **Y** are the transformed versions of the original m-variate observations on $\boldsymbol{U}$ forming the columns of **X**.

B. The Trimmed Mean

Let $x_1, x_2, \cdots, x_n$ be n univariate observations, and let the corresponding ordered observations be $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$. Then, for some $\alpha \in (0,0.5)$, where $\alpha = \alpha_1 + \alpha_2$, for some $\alpha_1, \alpha_2 \in [0,0.5)$, the α-trimmed mean of the n observations is defined as

$$\bar{x}_\alpha = \frac{1}{n - A_1 - A_2} \sum_{i=A_1+1}^{n-A_2} x_{(i)},$$

where $A_j = \lfloor n\alpha_j/2 \rfloor, j = 1,2, \lfloor\cdot\rfloor$ being the floor function. It is nothing but the average of observations excluding the A_1 smallest and A_2 largest, so that a proportion α of the observations (that are supposedly extreme) are excluded.

VI. Results

A. The Benchmark Telephone Speech Corpus NTIMIT

The database NTIMIT [12], [13], like TIMIT [14], [15] is an acoustic-phonetic speech corpus in English, belonging to the Linguistic Data Consortium (LDC) of the University of Pennsylvania. TIMIT consists of clean microphone recordings of 10 different read sentences (2 *sa*, 3 *si* and 5 *sx* sentences, some of which have rich phonetic variability), uttered by 630 speakers (438 males and 192 females) from eight major dialect regions of the USA. It is characterized by 8-kHz bandwidth and lack of intersession variability, acoustic noise, and microphone variability or distortion. These features make TIMIT a benchmark of choice for researchers in several areas of speech processing.

B. Baseline Performance of the MFCC-GMM Method

Using a number of competing MFCC-GMM classifiers, the overall classification accuracy obtained by us with all 630 speakers in NTIMIT was 34.96% when 6 out of the 10 recordings per speakers were used for training and the remaining 4 were used for testing (referred to as the 6:4 dataset). This improved to 42.14% when 8 out of the 10 recordings per speakers were used for training and the remaining 2 were used for testing (referred to as the 8:2 dataset)

The competing classifiers referred to above were obtained by varying certain tuning parameters of the generic MFCC-GMM model. The number of MFCCs as well as the number of mel-scale filters was 38 for all these classifiers, and a 32-component GMM was used in each case.

C. Improvement after Principal Component Transformation

Using the same set of competing MFCC-GMM classifiers as above, the greatest recognition accuracy obtained, after transformation of MFCCs by PCT, for all 630 NTIMIT speakers was 42.26% in the 6:4 case and 52.30% in the 8:2 case.

D. Improvement after Incorporating Robust Approach

To compute the quantity $\rho(\lambda_i)$, the trimmed mean procedure was applied for different combinations of α_1, α_2, and it was found that in the best scenario, there was a maximum improvement of a modest 3% over the baseline MFCC-GMM system, with $\alpha_1 = \alpha_2 = 0.1$. However, when this was applied in conjunction with the PCT-transformation, the improvement was even more substantial, 48.29% and 58.97% for the 6:4 and 8:2 cases respectively.

E. Improvement after Combining Feature Sets

We have used a combined feature set with 13 MFCCs, 13 delta features based on them and 13 PLPCs. This combined set performed much better in improving the classification accuracy which climbed up to 54.8% and 64.76% respectively.

F. Improvement Using Ensemble Classifiers

In Table I, we show the results of using the MFCC based classifier and the combined feature based classifier for two experiments where the window time for generating the MFCCs were set at two different values. When we employed an ensemble classifier using these four classifiers by aggregated likelihood method, the results improved to 60.24% in the 6:4 case and 70.48% in the 8:4 case.

Thus the overall improvement over the baseline GMM-MFCC classifier was nearly 70% (60% compared to 35% in 6:4 and 70% compared to 42% in 8:2) in both cases.

These results are summarized in Table I.

TABLE I: RESULTS ON NTIMIT DATABASE

Type of features used	6:4		8:2	
	MFCC	**MFCC-PLPC**	**MFCC**	**MFCC-PLPC**
Window time (0.02)	48.29	51.27	58.97	59.92
Window time (0.03)	47.94	54.82	59.60	64.76
Combined	**60.24**		**70.48**	
Baseline (ordinary GMM-MFCC)	34.96		42.14	
% Increase over baseline performance	**72.3**		**67.3**	

VII. CONCLUSION

From the results presented in the previous section, it is quite evident that the proposed approach, using an ensemble classifier with a combined feature set in conjunction with the principal component transformation, can significantly improve the performance of the Gaussian Mixture Model-based speaker identification system. Extensive experimentation with the benchmark NTIMIT data empirically establishes that it has tremendous potential for improving the degraded performance of the MFCC-GMM model particularly in the case of noisy speech data. In this work, we have experimented solely with the maximum likelihood approach. As a future direction, one could perhaps compare an estimated GMM model based on the test utterance with those based on different speakers. For this comparison one could possibly try minimizing an appropriate divergence measures between the estimated densities by the GMM models.

ACKNOWLEDGMENT

The authors gratefully acknowledge the funding provided for the initial exploratory work in this area by Department of Information Technology, Ministry of Communication and Information Technology, Government of India. They are also indebted to Prof. Hema Murthy of the Indian Institute of Technology Madras for her invaluable assistance in the form of useful discussions and for loan of software related to the MFCC-GMM speaker recognition system.

REFERENCES

[1] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Communication*, vol. 52, pp. 12–40, 2010.

[2] J. P. Campbell, “Speaker recognition: a tutorial,” *Proceedings of the IEEE*, vol. 85, pp. 1437-1462, 1997.

[3] D. A. Reynolds, “Large population speaker identification using clean and telephone speech,” *IEEE Signal Processing Letters*, vol. 2, pp. 46-48, 1995.

[4] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture models,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72-83, 1995.

[5] A. Pal., S. Bose, G. K. Basak, and A. Mukhopadhyay, “Improved speaker identification by aggregating Gaussian mixture models (GMMs),” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 28, no. 4, 2014.

[6] C. R. Rao, *Linear Statistical Inference and Its Applications*, 2nd (reprint) edition, John Wiley & Sons, New York, 2001.

[7] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Pearson Education, Inc, 2008.

[8] S. B. Davies and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously

spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, pp. 357-366, 1980.

[9] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoustical Society of America*, vol. 87, no. 4, pp. 1738-1752, 1990.

[10] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1-38, 1977.

[11] P. J. Huber, *Robust Statistics*, Springer Berlin Heidelberg, 2011.

[12] W. M. Fisher, G. R. Doddington, K. M. Goudie-Marshall, C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT," *Linguistic Data Consortium*, Philadelphia, 1993.

[13] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 1990.

[14] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, 1986, pp. 93-99.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, Philadelphia, 1993.

Smarajit Bose received B. Stat. and M. Stat. degrees from the Indian Statistical Institute, Calcutta in 1984 and 1986 respectively, and a Ph. D. degree from the University of California, Berkeley in 1992, all in statistics. Between 1991 and 1992, he was a visiting scientist in IBM Almaden Research Center, San Jose where he worked with the Advance Process Monitoring Group. He visited the Statistics Departments of the University of Washington, Seattle during 1992–1993 and the Ohio State University, Columbus during 1993–1996 where he worked on classification and clustering problems and their applications in medical imaging and ergonomics. In 1996, he joined the Indian Statistical Institute, Calcutta and now is a professor in the Applied Statistics Division of the Institute. He has also visited the University of California, Santa Barbara in 2001, and also in 2002–2003. His current research interests are pattern recognition and its applications in image processing and speaker identification.

Amita Pal (Nee Pathak) obtained a B.Sc. degree with Honours and a M.Sc. degree in statistics from the University of Calcutta in 1979 and 1981 respectively, and a Ph.D. degree in statistics from the Indian Statistical Institute, Kolkata in 1991. She joined the Indian Statistical Institute as a lecturer in 1994, and is currently working as an associate professor in the Interdisciplinary Statistical Research Unit of the same institute. She visited the Imperial College of Science, Technology and Medicine, London in 1994 on a six-month UNDP fellowship. Her research interests included pattern recognition, image processing and statistical machine learning

Anish Mukherjee received a bachelor of science (B.Sc.) degree in statistics from St. Xavier's College, Kolkata in 2011, followed by a B.Tech. degree in computer science and engineering from the University of Calcutta in 2014. He has been involved in research on robust speaker identification since 2014 in the Interdisciplinary Statistical Research Unit of the Indian Statistical Institute, Kolkata. His areas of interest are pattern recognition, computer vision, high-dimensional data analysis and bayesian nonparametrics. He is currently pursuing a Ph.D. degree in statistics in University of Missouri, Columbia, USA.

Debasmita Das received a bachelor of science (B.Sc.) degree in statistics from Ashutosh College, Kolkata in 2012 and a master of science (M.Sc.) degree in statistics from University of Calcutta in 2014. She has been involved in research on Robust Speaker Identification since 2014 in the Interdisciplinary Statistical Research Unit of the Indian Statistical Institute, Kolkata. Her areas of interest are pattern recognition, computer vision, high-dimensional data analysis and biostatistics. She is currently pursuing a Ph. D. degree in statistics in University of Connecticut, Storrs, USA.