New Proposed Feature Extraction Method to Enhance Speaker Recognition Rate with GMM

Masood Qarachorloo and Gholamreza Farahani
Institute of Electrical Engineering and Information Technology, Iranian Research Organization for Science and Technology (IROST), Tehran, Iran
Email: {m.gharachorlouei, farahani.gh}@irost.org

***Abstract*—In this paper, a novel speaker feature extraction with Gaussian Mixture Model (GMM) is proposed. In this new method Perceptual Linear Prediction (PLP) and Linear Predictive Cepstral Coefficient (LPCC) features have extracted and Gaussian Mixture Models (GMM) of speakers has built, then identification tests with clean and noisy TIMIT database have been carried out. With usage of TIMIT database, train and test samples of the speech ratio is 9 to 1. Implementation results with GMM have shown that GMM will model the structure of the vocal tract finely and minimize the distance between training and test feature vectors. Also experimental results show that LPCC feature coefficients will improve the results of speaker recognition rate. Thus in new proposed method with combination of PLP and LPCC features, the efficiency of the speaker recognition rate will increase 2.2% and speaker recognition efficiency will be 98.4%.**

***Index Terms*—speaker recognition, Gaussian mixture model, feature extraction, expectation maximization, TIMIT database**

I. Introduction

The human speech carries different types of information. The primary type is the meaning of words, which speaker tries to pass to the listener. But the other types that are also included in the speech are information about language being spoken, speaker emotions, gender and identity of the speaker. The goal of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity [1]. Speaker recognition is usually divided into two different branches, speaker verification and speaker identification. Speaker verification task is to verify the claimed identity of person from his voice [2], [3]. This process involves only binary decision about claimed identity. In speaker identification there is no identity claim and the system decides who the speaking person is [2].

The purpose of a speaker recognition system is identifying a set of sound samples of various sounds, which have the best matching with the characteristics of an unknown sample sound input [4]. Speaker recognition is a two-step process includes training and testing phases. In the training phase, speaker feature vectors X_m^{training}, dependent to *M* speakers, extracted from the speech signal training and model of each speaker (λ_s) for each feature vector is made. Generally, in the speaker identification systems, Mel Frequency Coefficients Cepstral (MFCC) [5] is used as feature vectors with dimensions $L \times 1$ and a Gaussian mixture model [6] to model speakers. In the testing phase, the speaker feature vectors X_m^{testing} related to *M* speakers (for anonymous speaker) extracted by calculating the likelihood, according to (1), a decision took on the identity S, which has maximum likelihood in comparison with all S speakers models.

$$\hat{s} = \arg max \sum_{\substack{m=1 \\ 1\le s\le S}}^{M'} \log p(X_m^{test}|\lambda_s) \quad (1)$$

In evaluating a speaker identification, recognition accuracy obtained by dividing the number of the true identification tests to the total number of tests. For many years it has been shown that systems based on the GMM have significant success in the speaker recognition in large populations [4], [5].

The purpose of this article is extracting and combining features from speech signals to enhance speaker recognition rate using the Matlab simulation software. The main characteristics of the matrix of coefficients are shown using feature extraction techniques. By the features extracted from speech signals and a statistical model, a unique identity for each person who is registered in the system, will be extracted. Laboratory assessments have been carried out on TIMIT English database consist of 630 audio speakers which is recorded by a good quality microphone. System uses a large amount of input speech of all speakers for universal model in training phase and a model is created for each speaker. For the testing phase, some other speech utterances which are different from training set are used.

At continuation of paper, features and vector space will describe. Then Gaussian mixture model and expected maximization algorithm which is used in new proposed method will discuss and finally description of database, results and conclusion will present.

II. Features and Vector Space

A. Frame Blocking

In speaker recognition, the first step is feature extraction which these features will obtain periodically. The time duration of signal which is considered for processing is called window and the data acquired in the

Manuscript received June 14, 2015; revised September 9, 2015.

doi: 10.18178/ijsps.4.4.276-281

window is named frame. Investigations show that speech signal characteristics will stay stationary in a sufficiently short period of time interval (quasi-stationary). For this reason, speech signals are processed in short time intervals. It is divided into frames with sizes generally between 20 and 30 milliseconds. Each frame

overlaps its previous frame by a predefined size. The overlap is usually selected between 10 to 15 milliseconds. The goal of the overlapping scheme is to smooth the transition from frame to frame [7].

B. Windowing

The second step is to window all frames. This will carry out to eliminate discontinuities at the edges of the frames. If the windowing function is defined as *w(n)*, $0 \leq n \leq N$ where *N* is the number of samples in each frame, the resulting signal will be *y(n)= x(n).w(n)*. There are different types of windows which can be used:

- Rectangular window
- Bartlett window
- Hamming window.

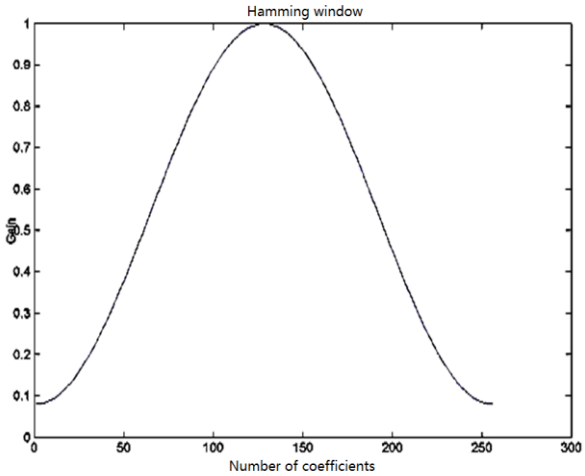


Figure 1. Hamming window.

The system uses hamming window as it introduces the least amount of distortion. Furthermore, the most widely used window is hamming window. Impulse response of the hamming window is a raised cosine impulse and is shown in Fig. 1. Transfer function of hamming window is such as (2) [7]. Then features are extracted from each of frame.

$$W(n) = 0.54 - 0.46\, Cos\frac{2\pi n}{N-1}, 0 \leq n \leq N \qquad (2)$$

C. Linear Predictive Cepstarl Coefficient (LPCC)

Linear Prediction is widely used in speech recognition and synthesis systems, as an efficient representation of a spectral envelope for speech signal. According to [8], it was first applied to speech analysis and synthesis by Atal and Schroeder Saito and Itakura [9].

There are two ways to compute the LP analysis, including autocorrelation and covariance methods. In this paper, LPC-related features are extracted using the autocorrelation method. Assume the nth sample of a given speech signal is predicted by the past M samples of the speech such as (3).

$$\hat{x}(n) = a_1 x(n-1) + a_2 x(n-2) + \cdots + a_M x(n-M) = \sum_{i=1}^{M} a_i x(n-i) \qquad (3)$$

To minimize the sum squared error between actual and predicted present sample, the derivative of *E* with respect to a_i is set to zero which is shown in (4).

$$\sum_n x(n-k)(x(n) - \sum_{i=1}^{M} a_i x(n-i)) = 0 \qquad (4)$$

If there are M samples in the sequence indexed from *0* to *M-1*, (4) can be expressed in the matrix form as (5) and (6).

$$\begin{bmatrix} r(0) & \cdots & r(M-1) \\ \vdots & \ddots & \vdots \\ r(M-1) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_{M-2} \end{bmatrix} = \begin{bmatrix} r(1) \\ r(M-2) \end{bmatrix} \qquad (5)$$

$$r(k) = \sum_{n=0}^{N-1-k} x(n)x(n+k) \qquad (6)$$

To solve the matrix (5) and (6), $O(M^3)$ multiplications is required. However, the number of multiplications can be reduced to $O(M^2)$ with the Levinson-Durbin algorithm which recursively compute the LPC coefficients. The recursive algorithm is described in (7).

Initial values:

$$E_0 = r(0) \qquad (7)$$

With $m \geq 1$, the recursion formulas (8) to (12) are performed.

$$q_m = r(m) - \sum_{i=1}^{m-1} a_{i(m-1)}\, r(m-i) \qquad (8)$$

$$k_m = \frac{q_m}{E_{(m-1)}} \qquad (9)$$

$$a_{mm} = k_m \qquad (10)$$

$$a_{im} = a_{i(m-1)} - k_m a_{(m-1)(m-1)} \quad for \;\; i = 1,2,\ldots,m-1 \qquad (11)$$

$$E_m = E_{m-1}[1 - {k_m}^2] \qquad (12)$$

where k_m is the reflection coefficient and the prediction error E_m decreases as m increases. Thus, LPC coefficients are generally transformed into other representations, including LPC reflection coefficients and LPC Cepstral coefficients. LPC Cepstral coefficients are important LPC-related features which are employed in speech recognition research commonly. They will compute directly from the LPC coefficients a_i with using the recursion formulas (13) to (15).

$$c_0 = r(0) \quad initial \qquad (13)$$

$$c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}, 1 < m < M \qquad (14)$$

$$c_m = \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}, m > M \qquad (15)$$

Based on the recursive formulas (13) to (15), an infinite number of Cepstral coefficients can be extracted from a finite number of LPC coefficients. However, typically the first 12-20 Cepstrum coefficients are employed depending on the sampling rate.

D. Perceptual Linear Prediction (PLP)

Perceptual Linear Prediction (PLP) was introduced by Hermansky [10] to work in the warping of the frequency and spectral magnitude, based on auditory perception

tests, into pitch and loudness to be used mainly as a preprocessor for the Linear Prediction method. The goal of this model is to perceptually approximate the human hearing structure in the feature extraction process. In this technique, several hearing properties such as frequency banks, equal-loudness curve and intensity-loudness power law are simulated by mathematic approximations. The output spectrum of the speech signal is described by an all-pole autoregressive model. Fig. 2 shows the block diagram of the PLP method which was first introduced in [10].

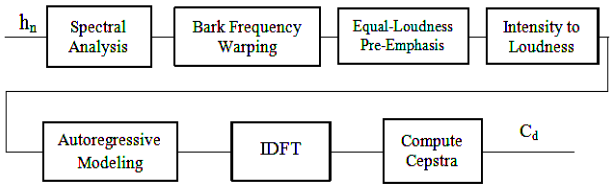


Figure 2. Perceptual linear predictive method diagram.

The extraction process of conventional PLP [11] is described below.

1) Spectral analysis

Each of the speech frames is weighted by hamming window. The windowed speech samples *s(n)* are transformed into the frequency domain *P(ω)* with use of Fast Fourier Transform (FFT). For a 10 kHz sampling frequency, a 256- point FFT is needed. Transforming 200 speech samples from the 20 ms window, will pad by 56 zero-valued samples. The real and imaginary components of the short-term speech spectrum are squared and added to get the short-term power spectrum.

$$P(\omega) = \mathrm{Re}[\mathrm{S}(\omega)]^2 + Im[\mathrm{S}(\omega)]^2 \qquad (16)$$

2) Bark frequency warping

The spectrum *P(ω)* is warped along its frequency axis ω into the Bark frequency *Ω* according to (17). The new spectral magnitude in the bark scale will be given by the following conversion formula due to Schroeder [12] which converts the linear version of the angular frequency (*ω*) to the Bark frequency.

$$\Omega(\omega) = 6\, ln\left(\frac{\omega}{1200\pi} + \sqrt{\left(\frac{\omega}{1200\pi}\right)^2 + 1}\right) \qquad (17)$$

The convolution of *Ω(ω)* and *P(ω)* yields the critical-band power spectrum (*Θ (ω)*).

3) Equal-Loudness pre-emphasis

The sampled *Θ(ω)* is pre-emphasized by the simulated equal-loudness curve with use of (18).

$$\Psi(\omega) = F_{p.e}(\omega).\Theta(\omega) \qquad (18)$$

where $F_{p.e}(\omega)$ is the approximation to the non-equal sensitivity of human hearing at different frequencies [13]. This simulates hearing sensitivity at 40dB level. This step is designed to carry out some pre-emphasis in the spirit of combining the concept of equal loudness curves and the concept of pre-emphasis which changes the weights of the spectral magnitudes. Reference [10] uses an approximation to the equal loudness curves, due to Makhoul and Cosell [14] to compute the pre-emphasis factor, $F_{p.e}$, as a function of the angular frequency (*ω*) for band-limited signals with an upper cut-off frequency of 5kHz according to (19).

$$F_{p.e}(\omega) = \frac{(\omega^2+5.68*10^7)\omega^4}{(\omega^2+6.3*10^6)^2(\omega^2+3.8*10^8)} \qquad (19)$$

This pre-emphasis filter causes a 12dB/octave drop in the signal strength for frequencies up to 400Hz. However, here, the drop is only 6dB/octave for frequencies between 1200Hz and 3100Hz and 0 for all other frequencies up to the Nyquist critical frequency of 5kHz.

For signals with a higher frequency content, an additional term is utilized which adds a sharp drop of 18dB/octave in the power for frequencies higher than 5kHz as (20).

$$F_{p.e}(\omega) = \frac{(\omega^2+5.68*10^7)\omega^4}{(\omega^2+6.3*10^6)^2(\omega^2+3.8*10^8)(\omega^6+9.58*10^{26})} \qquad (20)$$

4) Intensity-Loudness power law

To approximate the power law of human hearing, which has a nonlinear relation between the intensity of sound and the perceived loudness, the emphasized *Ψ(ω)* is compressed by cubic-root amplitude given by (21).

$$\Phi(\omega) = \Psi(\omega)^{\frac{1}{3}} \qquad (21)$$

5) Autoregressive modeling

In the last stage of PLP analysis, *Φ(ω)* which calculated with (21) will approximated by an all-pole spectral modeling through autocorrelation LP analysis [15]. The first *M+1* autocorrelation values are used to solve the Yule-Walker equations for the autoregressive coefficients of the *M* order all-pole model.

III. Gaussian Mixture Model

Gaussian mixture model clustering is a measure of the probability distribution used to create clusters. Each cluster actually looks that it has a Gaussian distribution. Gaussian mixture models are one of the best known and most widely used methods to identify the speaker. Gaussian mixture models are based on the division of sounds into different classes and these classes are compared with the input speech. In this model, the segmentation of phonemes to classes is implicitly based on a division of unsupervised clustering, therefore tag will not use for classes (identify the exact phoneme). On the other hand, Gaussian mixture model tries to model the probability density function of the speaker. This modeling is performed with a linear combination of some Gaussian functions, which is the reason that it has called Gaussian mixture model [16].

Gaussian mixture model is similar to the single-state Hidden Markov Model (HMM) and is a probability density function of the state, with many normal mixtures. The probability of test vector *x* belongs to a Gaussian mixtures model with *M* mixtures, will calculate in the form of (22).

$$P(x|GMM) = \sum_{t=1}^{M} c_t.N(\mu_t, \Sigma_t) \qquad (22)$$

where c_t is weight of mixtures, and μ_t and Σ_t are the normal distribution mean vector and covariance matrix respectively. Covariance matrix of GMM, usually considered diagonal, although there is the possibility of

using full matrix as well. Equation (22) can be also stated using normal probability density function as expressed in (23).

$$P(x|GMM) = \sum_{i=1}^{M} c_i . \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} . exp\left\{-\frac{1}{2}(\vec{x} - \overrightarrow{\mu_i})' \sum_i^{-1} (\vec{x} - \overrightarrow{\mu_i})\right\} \quad (19)$$

where d is an input space dimension. To obtain GMM parameters, including Gaussian distributions mean, covariance and weight, EM algorithm is used. It should be noted that the number of Gaussian mixtures have a direct relationship with the existing training models and GMM models cannot be trained with an excessive number of the mixtures with poor data collection. In the formation and training of GMMs, like all other models, consideration of the complexity of the model and training samples is necessary [4].

IV. Expected Maximization Algorithm

The Expectation Maximization (EM) algorithm as an example of the Baum - Welch algorithm, is used in the training of GMMs. In EM algorithm, a method of testing, it is possible to get the maximum or minimum, or it may be getting into the trap in the local maximum or minimum. EM method is a general method to find the parameters with estimating the Maximum Likelihood (ML). Certainly in the each iteration, likelihood logarithm will increase. The EM algorithm guarantees convergence to a local maximum of likelihood function, in both phases of the expected value and likelihood.

The EM algorithm, with using hidden variables λ is formed where the maximum likelihood is achieved by using the training set X as shown in (24).

$$p(X|\lambda) = \prod_{t=1}^{T} p(X_t|\lambda) \quad (24)$$

To maximize the likelihood between the Gaussian distribution and the samples based on these relationships, model parameters changes frequently. EM algorithm consists of two steps:

1) The expected value: In expectation value, GMM parameters are obtained for each sample of d dimensional data $x \in \{X\}_{t=1,\dots,T}$ using inductive probability and for i^{th} component using (25).

$$P(i|x_t, \lambda) = \frac{w_i g(x_t|\mu_i, \Sigma_i)}{\sum_{k=1}^{n} w_k g(x_t|\mu_k, \Sigma_k)} \quad (25)$$

where $g(x_t|\mu_k, \Sigma_k)$ is introduced according to (26).

$$g(x_t|\mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d \times \Sigma_i}} \exp\{\frac{-1}{2}(x_t - \mu_i)' \sum_i^{-1} (x_t - \mu_i)\} \quad (26)$$

2) The maximization: At maximum, the parameters are calculated in accordance with inductive probability estimated in the previous step. GMM parameters updated as well according to (27) to (29).

$$\bar{w}_i = \frac{1}{T} \sum_{t=1} P(i|x_t, \lambda) \quad (27)$$

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} P(i|x_t, \lambda) x_t}{\sum_{t=1}^{T} P(i|x_t, \lambda)} \quad (28)$$

$$\bar{\sigma}_i{}^2 = \frac{\sum_{t=1}^{T} P(i|x_t, \lambda) x_t{}^2}{\sum_{t=1}^{T} P(i|x_t, \lambda)} - \bar{\mu}_i{}^2 \quad (29)$$

Algorithm steps are repeated until the boundary of the convergence is achieved. The EM algorithm, guarantee converging to a local maximum likelihood, in both expected and likelihood phases [16].

V. Database, Results and Conclusion

A. TIMIT Speech Database

TIMIT database is an English connected speech prepared by company of TI and university of MIT and US which bureau of standards (NIST) has approved it. TIMIT database contains the 6300 speech, which were uttered by 630 speakers and 8 common North American accents. TIMIT database 70% male and 30% female speakers is included. Each speaker has uttered 10 sentences that the 2 sentences of them have been uttered by other speakers. Totally there are 2432 distinct sentences in TIMIT which includes two common sentences among all speakers, 450 common sentences among groups of seven people of speakers and 1890 sentences including a single speaker. All words and phonemes in TIMIT sentences have the time tags. TIMIT database is free of noise and generally is used to assess the rate of recognition of phonemes in continuous speech recognition and speaker recognition types. However, despite the time tags for words and phonemes, it could be used separately to assess word recognition rate. To use this database for evaluation of speech recognition in noise, noise must be artificially added to the database [17].

B. Tests and Results

In this paper for implementation, 6300 utterances in TIMIT database is used which 5670 of them is utilized for the training system, and 630 utterances were used for the test. Eligibility criteria to be considered in models, like the likelihood ratio logarithm. Since the data are used consistently, therefore after initialization parameters of the models, EM algorithm is used for re-estimating of parameters. Finally, models of recognition will achieve. Then, the recognition models will adapt with each model. The calculation accuracy based on the relationship is like as (30).

$$CIR\ (Correct\ Identification\ Rate) = \frac{No.of\ Successful\ Identification}{Total\ No.of\ Attempts} \times 100 \quad (30)$$

In this paper, 12 coefficients of LPCC and 9 of the PLP coefficients were used. Another characteristic which are used is as below.

Pre-emphasized factor is 0.975, the length of a window is 25 milliseconds, and the step in the window or overlap of windows will occur every 10 milliseconds. The 26 filter banks are selected. In calculation of features, hamming window is used.

In implementation, uniformly, 32 GMM mixtures are used. Also 90% of the database, including 5670 sentences in the training phase and 10% database includes, 630 sentences will use for test phase. The average length of each sentence is 3 seconds, therefore 27 seconds of speech for training and 3 seconds for the test is used. In the training phase, for each speaker model, likelihood

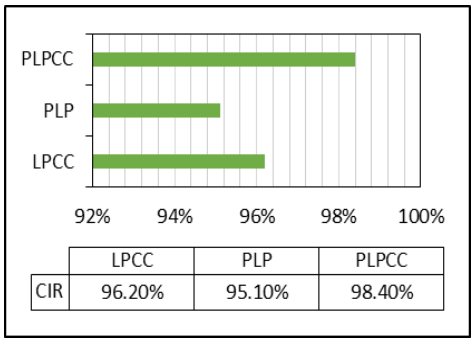
score of input sequence from the input feature vectors is calculated with (31).

$$L(X, G_s) = \sum_{i=1}^{M} P(\overrightarrow{x_t}|G_s) \quad (31)$$

where L stands the likelihood, and it is in concept of derived vectors from the model G_s, such that $X = \{\overrightarrow{x_1}, \overrightarrow{x_2}, \dots, \overrightarrow{x_M}\}$ is speaker feature vectors sequence, and M is the total number of feature vectors. The highest score $L(X, G_s)$ of the generated GMM is selected as most similar to the original speaker.

GMM provides a robust basic model to compute likelihoods between a test speaker and a given model. This method has proven its effectiveness on small populations, with few noise components and intersession variability. Computation of the likelihood ratio makes it interesting for speaker verification making the match score range of different speakers comparable. LPC analysis is an effective method to estimate the main parameters of speech signals. The conclusion extracted was that an all-pole filter, is a proper approximation to estimate the speech signals. In this way, from the filter parameters, the speech samples could be synthesized by a difference equation. Thus, the speech signals resulting can be seen as linear combination of the previous samples.

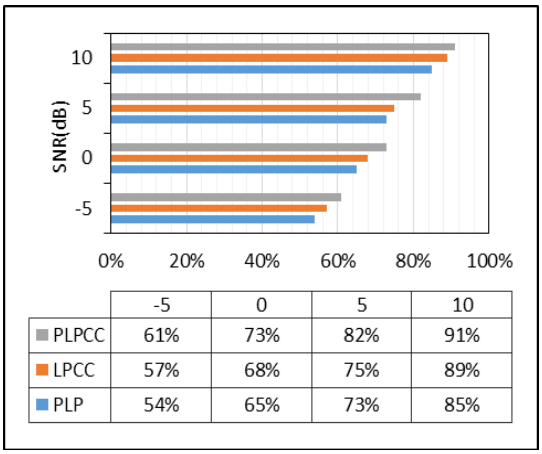
The PLP speech analysis method is more adapted to human hearing, in comparison to the classic Linear Prediction Coding (LPC). The main difference between PLP and LPC analysis techniques is that the LP model assumes the all-pole transfer function of the vocal tract with a specified number of resonances within the analysis band. The LP all-pole model approximates power distribution equally well at all frequencies of the analysis band. This assumption is inconsistent with human hearing, because beyond 800 Hz, the spectral resolution of hearing decreases with frequency and hearing is also more sensitive in the middle frequency range of the audible spectrum.



	LPCC	PLP	PLPCC
CIR	96.20%	95.10%	98.40%

Figure 3. Speaker recognition rate using TIMIT database.

All implementation such as feature extraction and combination of speech signals has been carried out in Matlab software. As is shown in Fig. 3, the proposed new method named PLPCC (Perceptual Linear Prediction Cepstral Coefficient) uses 21 coefficients for speaker recognition which it has speaker recognition rate of 98.4% versus 96.2% for LPCC and 95.1% for PLP. Results of recognition rate in noisy condition with 630 speakers are shown in Fig. 4. The results clearly demonstrate the improvements of new proposed method PLPCC over PLP and LPCC. When PLPCC features are used as complementary features, the efficiency of speaker recognition will improve. This will be shown that PLP features have the additional information, which is not in the Cepstral coefficients. Therefore it is logical to get better results with PLPCC features in comparison with PLP and LPCC methods.



	-5	0	5	10
PLPCC	61%	73%	82%	91%
LPCC	57%	68%	75%	89%
PLP	54%	65%	73%	85%

Figure 4. Speaker recognition rate in TIMIT database with white Gaussian noise.

C. Conclusion

In this paper, the importance of feature extraction to improve the speaker recognition rate is mentioned. This paper tried to evaluate the effect of the combination of some features in automatic speaker recognition systems.

Speaker recognition has encountered lots of advancements within the past few years. Emerging new technologies will improve robustness, more particularly in speaker identification. The new PLPCC proposed method based on GMM has speaker recognition rate of 98.4% using 21 coefficients, which is better results in comparison with 96.2% for LPCC and 95.1% for PLP.

Study of researchers will concentrate on the new techniques as well as improvement in the existing methods. Field of studies become larger, from physiological to behavioral with the use of high level features. The research in the field of speaker recognition contributes substantially to a better management in security for various uses, although the behavioral aspects are only emerging recently. Future researches are in the training GMMs using the median, which may provide improvement when the median is used in the evaluation stage.

REFERENCES

[1] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. ICASSP*, 2002, pp. 4072-4075.
[2] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. of the IEEE*, vol. 85, no. 9, pp. 1437-1462, Sept 1997.
[3] J. M. Naik, "Speaker verification: A tutorial," *IEEE Communications Magazine*, pp. 42-48, January 1990.
[4] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Signal Process.*, vol. 3, no. 1, pp. 72-83, Jan. 1995.
[5] T. Kinnunen, E. Karpov, and P. Franti, "Real-Time speaker identification and verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 277-288, Jan. 2006.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process*., vol. 10, pp. 19-41, 2000.
[7] M. Woszczyna, "JANUS 93: Towards spontaneous speech translation," in *Proc. IEEE Conference on Neural Networks*, 1994.
[8] M. R. Schroeder and B. S. Atal, "Predictive coding of speech signals," in *Proc. Conf. Commun. and Process*., 1967.
[9] S. Saito and F. Itakura, "The theoretical consideration of statistically optimum methods for speech spectral density," Report No. 3107, Electrical Communication Laboratory, N.T.T., Tokyo, 1966.
[10] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578-589, 1994.
[11] H. Hermansky, "Perceptual Linear Predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738-1752, April 1990..
[12] M. R. Schroeder, "Recognition of complex acoustic signals," in *Life Sciences Research Report*, T. H. Bullock, Ed., Abacon Verbag, Berlin, 1977, pp. 324.
[13] D. W. Robinson and R. S. Dadson, "A redetermination of the equal-loudness relations for pure tones," *Br. J. Appl. Phys.*, vol. 7, pp. 166-181, 1956.
[14] J. Makhoul, L. Cosell, "LPCW: An LPC vocoder with linear predictive spectral warping," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 1976, pp. 466-469.
[15] J. Makhoul, "Spectral linear prediction: Properties and applications," *IEEE Trans. ASSP*, vol. 23, pp. 283-296, 1975.
[16] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1-38, 1977.
[17] V. Zue, S. Seneff, and J. Glass, "Speech database developmentat MIT: TIMIT and beyond," *Speech Comm.*, vol. 9, pp. 351-356, September 1990.

Masood Qarachorloo received his BSc degree in electrical engineering from Iran University of Science and Technology (IUST), he is MSc student in electrical engineering from Iranian Research Organization for Science and Technology (IROST). His research interest is speech processing.

Gholamreza Farahani received his BSc degree in electrical engineering from Sharif University of Technology, Iran, in 1998 and the MSc and PhD degrees in electrical engineering from Amirkabir University of Technology (Tehran Polytechnic) in 2000 and 2006, respectively. Currently, he is an assistant professor with the Institute of Electrical Engineering and Information Technology, Iranian Research Organization for Science and Technology (IROST), Iran. His research interest is speech processing.