New Features to Improve Speaker Recognition Efficiency with Using LPCC and SSC Features

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, the new method to improve speaker recognition efficiency has been proposed. In this new method firstly, silence of the sentences has removed and secondly different train and test data with -5, 0, 5 and 10 dB signal to noise ratio has made artificially. Spectral Subband Centroid (SSC) and Linear Predictive Cepstral Coefficient (LPCC) features were extracted and then Gaussian Mixture Models (GMM) of speakers has built and identification tests with clean and noisy TIMIT database have been used. In the used TIMIT database, train and test samples of the speech at a ratio of 9 to 1 are used. Implementation results with comparison between different characteristics for speaker recognition, shows that SSC feature coefficients versus the other feature in different SNRs, has the better results. Because of some weakness of LPCC method, the new proposed method, with using SSC feature could cover these weaknesses, therefore combination of SSC and LPCC features will increases the efficiency of the speaker recognition 2.9 percent, and speaker recognition efficiency will be 99.1%.**

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

I. Introduction

Feature extraction is the key part of the front-end process in speaker identification systems. The performance of Speaker Identification (SI) system is highly related on the quality of the selected speech features. Most of the current proposed SI systems use Mel frequency Cepstral Coefficients (MFCC) and Linear Predictive Cepstral Coefficients (LPCC) as feature vectors [1].

Currently, researches are focusing on improving these two cepstral features [2] or appending new features on them [3]. In feature extraction techniques, Cepstral features will be shown with a matrix of coefficients. By the features extracted from speech signals and a statistical model, a unique identity for each person who is registered in the system build, will be extracted.

The first step in this paper, Gaussian Mixture Models (GMM) as a statistical model for speaker independent recognition is used. Secondly, from a global model, the context will be used for speaker recognition [4]. It has been proven that GMMs are very effective to model the speaker's identity, Because of the Gaussian model shows spectral forms of speaker properly. In addition, universal model improves Gaussian mixture model calculations for speaker decision-making. Also, Expectation Maximization (EM) algorithm which is an effective technique for discovering answers of maximum likelihood of model, used in the universal model training phase [5].

This paper briefly describes advanced techniques to improve the accuracy of speaker recognition. Laboratory assessments have been carried out on TIMIT English database consisting of 630 audio speakers which is recorded by a good quality microphone. System uses a large amount of input speech of all speakers for the universal model training phase and a model is created for each speaker. For the testing phase, some other speech utterances different from training set is used. At continuation of this paper, feature extraction method and results and conclusions will present.

II. Feature Extraction

A. Framing

Variations of voice in a large area are not stationary; therefore short time consideration of a voice will use in calculations which is steady state. In other words, due to the characteristics changing of the speech signal over time and in a nonstationary way (namely the statistical properties of the signal varies over time), feature extraction, does not provide reliable information from a relatively large area.

Audio signals are generally produced with stable position in 80-200ms from vocal tracts. That is why speech is divided into frames of 20 to 30ms, and then features are extracted from each frame, in which case the audio signal can be assumed stationary. Frames are usually selected so that they overlap each other. The overlap is usually selected 10 to 15ms [5].

B. Pre-Processing

To eliminate the effects of sudden changes in continuous time signal, the signal must pass through a filter first, called pre-emphasis filter. One of the reasons for the use of pre-emphasis filter is that filter effectively removes the spectral effects of the larynx (the poles) and lips (a zero). The filter also eliminates sudden signal

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

doi: 10.18178/ijsps.4.4.295-299

changes in the environment caused by the severe noise and it will be the same as the signal [6].

C. Windowing

At this phase, each frame separately multiplied by a window signal to reduce the effect discontinuity at the beginning and end of each frame. Select the window is very important because the margins of a frame are effective in decreasing and increasing of error signal. For this reason, you should use a window narrow uniformly margins of the frame. The best window for use has a narrow main lobe and side lobe levels slightly in frequency response. If the window displayed with $w(n)$, applying the window will be according to (1).

$$\bar{X}_k(n) = X_k \cdot W(n)\,, 0 \le n \le N-1 \qquad (1)$$

where N is the number of samples in a frame, and k is the frame number.

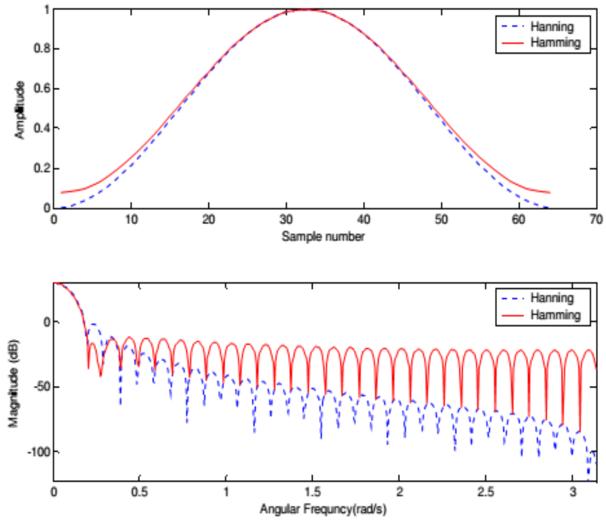


Figure 1. Hanning and Hamming windows.

Hanning and Hamming windows usually use in such applications and their mathematically relationships are in the form of (2) and (3) [7]. Also Fig. 1 shows the shape of hanning and hamming windows.

$$Hamming: W(n) = 0.54 - 0.4\,6Cos\frac{2\pi\, n}{N-1}, 0 \le n \le N \quad (2)$$

$$Hanning: W(n) = \frac{1}{2}\left(1 - Cos\,\frac{2\pi\, n}{N-1}\right), 0 \le n \le N-1 \quad (3)$$

D. Spectral Subband Centroid (SSC)

Spectral Subband Centroids [8]-[11] (SSC) are an alternative for Cepstral coefficients. SSCs are computed as the centroid frequencies of subband spectra and they give the locations of the local maxima of the power spectrum. SSCs have been used for speech recognition [8], [9], audio fingerprinting [10] and speaker recognition [11].

If the frequency band [0, Fs/2] would be divided into M subbands, Fs is the sampling frequency. For the m^{th} subband, let its lower and higher edges be l_m and h_m respectively. Furthermore, let the filter shape be $\omega_m(f)$ and $P^{\gamma}(f)$ be the power spectrum at location f raised to the power of γ. The m^{th} subband centroid, according to [9], is defined as (4).

$$C_m = \frac{\int_{l_m}^{h_m} f\,\omega_m(f)P^{\gamma}(f)df}{\int_{l_m}^{h_m} \omega_m(f)P^{\gamma}(f)df} \qquad (4)$$

Note that the term ω_m(f). P^{γ}(f) can be viewed as a bias which influences the location of subband centroid. A peak in this term leads to a higher weight in the corresponding f. Typically, ω_m(f) takes on the shape of either a square window (ones over the m^{th} subband and zeros everywhere else) or a triangular window (which gives a maximum response around its center and decreases towards both of it edges). In the case of MFCCs, ω_m is a triangular window. This same window is used here. The use of parameter in this function is rather a design parameter and is not motivated by any psychological aspect of hearing. This parameter has been used elsewhere in the literature [12] as part of feature extraction (which is called a two-dimensional root spectrum) for speech recognition. According to [12], this design parameter can be optimized according to the given data set and task. Hence, the introduction of γ is only for practical reasons from engineering point of view. In this paper, γ is set to 1.

Firstly, when there is no speech, SSCs in a given frequency subband tend to be the center of the band. On the other hand, with the presence of speech, SSCs show some regular trends: the trajectory of SSCs in a given subband actually locates the peaks of the power spectrum limited in that given subband [8]. Secondly, the medium to long-term time-trajectory of SSCs can be an interesting feature set as well, as demonstrated in [13]. Thirdly, if there are not enough centroids, then SSCs will not cover enough information. On the other hand, if there are too many centroids, additional centroids will only add to the unnecessary dimensionality of the data, without adding any more information.

E. Linear Predictive Cepstarl Coefficient (LPCC)

Linear Prediction is widely used in speech recognition and synthesis systems, as an efficient representation of a speech signal's spectral envelope. According to [14], it was first applied to speech analysis and synthesis by Saito and Itakura [15] and Atal and Schroeder [14].

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 n^{th} sample of a given speech signal is predicted by the past M samples of the speech such as (5).

$$\hat{x}(n) = a_1x(n-1) + a_2x(n-2) + \cdots + a_Mx(n-M)$$
$$= \sum_{i=1}^{M} a_i x(n-i) \qquad (5)$$

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 (6).

$$\sum_n x(n-k)\left(x(n) - \sum_{i=1}^{M} a_i x(n-i)\right) = 0 \qquad (6)$$

If there are M samples in the sequence indexed from 0 to M-1, the (6) can be expressed in the matrix form as (7) and (8).

$$\begin{bmatrix} r(0) & \cdots & r(M-1) \\ \vdots & \ddots & \vdots \\ r(M-1) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_{M-2} \end{bmatrix} = \begin{bmatrix} r(1) \\ r(M-2) \end{bmatrix} \quad (7)$$

$$r(k) = \sum_{n=0}^{N-1-k} x(n)x(n+k) \quad (8)$$

To solve the matrix (7) and (8), $O(M^3)$ 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 (9).

Initial values:

$$E_0 = r(0) \quad (9)$$

With m≥1, the recursion formulas (10) to (14) are performed.

$$q_m = r(m) - \sum_{i=1}^{m-1} a_{i(m-1)}\, r(m-i) \quad (10)$$

$$k_m = \frac{q_m}{E_{(m-1)}} \quad (11)$$

$$a_{mm} = k_m \quad (12)$$

$$a_{im} = a_{i(m-1)} - k_m a_{(m-1)(m-1)} \;\; for\; i = 1,2,\dots,m-1 \quad (13)$$

$$E_m = E_{m-1}[1 - {k_m}^2] \quad (14)$$

If m<M then m will increase or stop.

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 frequently. They will compute directly from the LPC coefficients a_i using the recursion formulas (15) to (17).

$$c_0 = r(0) \;\; initial \quad (15)$$

$$c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}, 1 < m < M \quad (16)$$

$$c_m = \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}, m > M \quad (17)$$

Based on the recursive formulas (15) to (17), 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.

III. Gaussian Mixture Model

Gaussian mixture model clustering is a measure of the probability distribution used to create clusters. Each cluster actually looks 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, this 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 (18).

$$P(x|GMM) = \sum_{t=1}^{M} c_t . N(\mu_t, \Sigma_t) \quad (18)$$

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 (18) can be also stated using normal probability density function as expressed in (19).

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

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 [1].

IV. Expected Maximization Algorithm

The expectation maximization, or 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 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 (20).

$$p(X|\lambda) = \prod_{t=1}^{T} p(X_t|\lambda) \quad (20)$$

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 ith component using (21).

$$P(i|x_t, \lambda) = \frac{w_i g(x_t|\mu_i, \Sigma_i)}{\sum_{k=1}^{n} w_k g(x_t|\mu_k, \Sigma_k)} \quad (21)$$

where $g(x_t|\mu_k, \Sigma_k)$ is introduced according to (22).

$$g(x_t|\mu_k,\Sigma_k)=\frac{1}{\sqrt{(2\pi)^d\times\Sigma_i}}\exp\{\frac{-1}{2}(x_t-\mu_i)'\textstyle\sum_i^{-1}(x_t-\mu_i)\} \quad (22)$$

2) The maximization: At maximum, the parameters are calculated in accordance with inductive probability estimated in the previous step. GMM parameters updated as well as the relations (23) to (25):

$$\bar{w}_i=\frac{1}{T}\sum_{t=1}P(i|x_t,\lambda) \quad (23)$$

$$\bar{\mu}_i=\frac{\sum_{t=1}^{T}P(i|x_t,\lambda)x_t}{\sum_{t=1}^{T}P(i|x_t,\lambda)} \quad (24)$$

$$\bar{\sigma}_i{}^2=\frac{\sum_{t=1}^{T}P(i|x_t,\lambda)x_t{}^2}{\sum_{t=1}^{T}P(i|x_t,\lambda)}-\bar{\mu}_i{}^2 \quad (25)$$

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 a database with 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. 70% male and 30% are women speakers. Each speaker has uttered 10 sentences that the 2 sentences of the 10, by other speakers have been uttered. In total 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*

From 6300 utterances in TIMIT database, 5670 of them is used 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. The final models of recognition will achieve. Then, the recognition models will adapt with each model. The calculation accuracy based on the relationship is like as (26).

$$CIR\ (Correct\ Identification\ Rate)=\frac{No.of\ Successful\ Identification}{Total\ No.of\ Attempts}\times 100 \quad (26)$$

In three steps, the system will end its work. The first step is feature extraction. In this paper, 12 coefficients of LPCC and 26 of the SSC 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. Inc. calculation of features, hamming window is used.

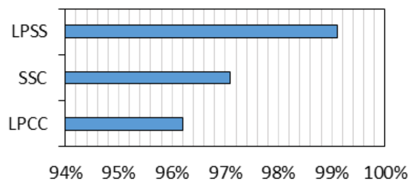
In the second phase, from specified files, models for each of the English sentences have been built. This requires the use of matrices of features which is obtained in the previous phase. Uniformly, 32 GMM mixtures are used. Finally, comparison of output files with models will carry out.

In implementation, 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 (27).

$$L(X,G_s)=\sum_{i=1}^{M}P(\overrightarrow{x_t}|G_s) \quad (27)$$

where *L* stands the likelihood, and it is in concept of derived vectors from the model G_s. So 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 the most similar to the original speaker.

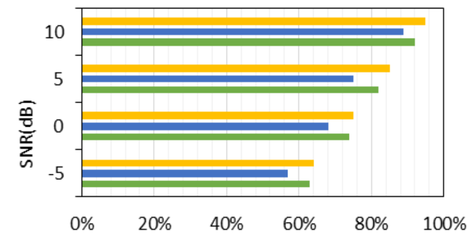
All implementation such as feature extraction and combination of speech signals has been carried out in Matlab software.



	LPCC	SSC	LPSS
CIR	96.20%	97.10%	99.10%

Figure 2. Speaker recognition results using a database TIMIT.

As is shown in Fig. 2, the proposed new method named LPSS (Linear Predictive Spectral Subband) uses 38 coefficients for speaker recognition which it has recognition rate of 99.1% versus recognition rate of 96.2% for LPCC and 97.1% for SSC. Results of recognition rate in noisy condition with 630 speakers are shown in Fig. 3. Because the human voice is naturally non-linear, methods such as LPC which has linear calculations are not suitable. In practice, LPC coefficients themselves are often not a good feature since polynomial coefficients are sensitive to numerical precision. The advantage of the new proposed method (LPSS feature) over the baseline SSC feature is that the subband boundaries are adapted for each frame. The partitions of the scalar quantizer themselves has no overlap. These forces the centroid frequencies to be monotonically increase, thereby limiting their dynamic range. Features obtained from SSC have similar to formant frequencies and are completely resistant to the noise. When these features are used as complementary features, efficiency of speaker recognition will improve. This will be shown that SSC features have the additional information, which is not in the Cepstral coefficients. Therefore it is logical to get better results with LPSS features in comparison with SSC and LPCC methods.



	-5	0	5	10
🟨 LPSS	64%	75%	85%	95%
🟦 LPCC	57%	68%	75%	89%
🟩 SSC	63%	74%	82%	92%

Figure 3. Speaker recognition results in white noise and database TIMIT.

C. Conclusion

In this paper, the importance of feature extraction for 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.

A new LPSS method based on GMM in proposed. The speaker recognition rate has reached to 99.1% using 38 coefficients, which has better results in comparisoSn with 96.2% for LPCC and 97.1% for SSC.

Future research will focus on two main issues, the evaluation algorithms such as Hidden Markov Model (HMM), GMM and Neural Networks (NN), and assess the impact of channel compensation techniques to achieve better speaker recognition results in the noisy condition with short length.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] M. R. Hasan, M. Jamil, M. G. Rabbani, and M. S. Rahman, "Speaker identification using Mel frequency Cepstral coefficients," in *Proc. 3rd International Conference on Electrical & Computer Engineering*, December 28-30, 2004.

[5] S. Furui, "Speaker-Independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 1, 1986.

[6] L. R Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, N.J.: Prentice-Hall, 1978.

[7] J. P. Campell, "Speaker recognition: A tutorial," *Proceeding of the IEEE*, vol. 85, pp. 1437-1462, 1997.

[8] B. Gajic and K. Paliwal, "Robust speech recognition in noisy environments based on subband spectral centroid histograms," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 600-608, 2006.

[9] K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Seattle, USA, 1998, pp. 617-620.

[10] J. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C. Yoo, "Audio fingerprinting based on normalized spectral subband centroids," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, 2005, pp. 213-216.

[11] N. Thian, C. Sanderson, and S. Bengio, "Spectral subband centroids as complementary features for speaker authentication," in *Proc. First Int. Conf. Biometric Authentication*, Hong Kong, China, 2004, pp. 631-639.

[12] F. Bimbot, J. F. Bonastre, *et al.*, "A tutorial on text-independent speaker verifcation," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 4, pp. 430-451, 2004.

[13] E. Piatkowska, A. N. Belbachir, S. Schraml, and M. Gelautz, "Spatiotemporal multiple persons tracking using dynamic vision sensor," in *Proc. Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 16-21, 2012.

[14] M. R. Schroeder and B. S. Atal, "Predictive coding of speech signals," in *Proc. Conf. Commun. and Process.*, 1967.

[15] 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.

[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 associate 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.