Text Independent Speaker Recognition Using Mixed MFCC and WOCOR Methods in Persian Language

Hassan Farsi[1] and Saber Amjadi[2]

***Abstract*—Voiced speech is usually used for speaker recognition. But in text-independent speaker recognition it would be better to use special voiced letters which are appeared in all words. In this paper, we have employed the certain letters for speaker recognition. As we know in Persian language, each consonant letter must be followed by a vowel letter. These are A - I - U – æ- e – ɔ:. Therefore it is enough for text-independent speaker recognition, to find and use these letters. For speaker recognition we employ both vocal source excitation signal and vocal tract system of these letters. Also we use the most prevalent feature parameters for speech/speaker recognition, that is the Mel-frequency cepstral coefficients (MFCC) and for speaker recognition by using vocal source excitation, we have employed Wavelet Octave Coefficients of Residuals (WOCOR). Since these methods are highly sensitive to noise, we use spectral subtraction method to cancel the noise.**

***Index Terms*—Source-tract features, Text independent speaker recognition, spectral subtraction, similarity coefficient, Mel-frequency cepstral coefficients.**

I. Introduction

Speaker recognition could be divided into verification and identification tasks. The verification task is to decide whether or not an unlabeled voice belongs to a claimed speaker. The identification task is to classify the unlabeled voice as belonging to one of the registered speakers. Speaker recognition can also be divided into text-dependent and text-independent recognitions. In text-dependent recognition, the system knows exactly the spoken text which could be either fixed phrase or prompted phrase. In text-independent recognition, the system does not know the text of the spoken utterance, which could be user selected keywords or conversational speech.

One of the most prevalent feature parameters for speech/speaker recognition is the MFCC parameters [7].The primary purpose of using cepstral features is to characterize the spectral envelope of a quasi-stationary speech segment. In the source-filter model of speech production, spectral envelope corresponds to the vocal tract filter, which determines the articulation of sounds [3]. It is believed that vocal source information, which is reflected in the details of short-time spectrum, e.g., harmonic peaks, also plays an important role in identifying and discriminating speakers. Residual signals of linear predictive (LP) analysis contain useful information about the excitation source and can be exploited for speaker recognition applications [4,5].In [6], vocal source related features were derived by time-frequency analysis of LP residual signals. Pitch-synchronous wavelet transform was then applied to capture the spectral-temporal characteristics of the excitation signal the resulting feature parameters, named wavelet octave coefficients of residues (WOCOR), were generated from wavelet coefficients of individual octave subband. It has been shown that WOCOR features provide complementary information to MFCC [6] [7].

This paper has been organized as following. In section 2 act of noise cancellation has been explained. In section 3 the method of obtaining the voiced letter mentioned in the abstract and then vocal source feature extraction have been explained. By using the extracted feature parameters in sections 3 and 4, the proposed speaker recognition method has been expressed in section 5. Finally in section 6, the obtained results of simulation by combining the methods MFCC and WOCOR have been shown.

II. Noise Cancellation

Assuming additive noise, i.e, $Y(e^{jW}) = S(e^{jW}) + D(e^{jW})$,the power spectral subtraction (PSS) algorithm is employed to estimate the clean speech spectrum as follows:

$$|\hat{S}(e^{jw})| = \begin{cases} |Y(e^{jw})| - |\bar{D}(e^{jw})| & if\ |Y(e^{jw})| > |\bar{D}(e^{jw})| \\ 0 & other\ wise \end{cases} \quad (1)$$

Along the timeline, during intervals where speech is absent for more than 300 msec, the noise power spectral density (PSD) $D(e^{jw})$ is estimated to update the noise statistics until the end of an utterance.

III. Representing the Voiced Letters A - I - U - U – æ- e – ɔ:

In this section, the methodology for finding a letter 'A' based on MFCCs method is explained. The same method for finding other letters is applied.

Each voiced segment is divided into non-overlapping frames of 30 msec. Then through the many experiments we find the best frame representing 'A' letter.

We are considering the frame related 'A' letter as input and passing from Hanning window

$$x_w(n) = x(n)\, W(n) \quad (2)$$

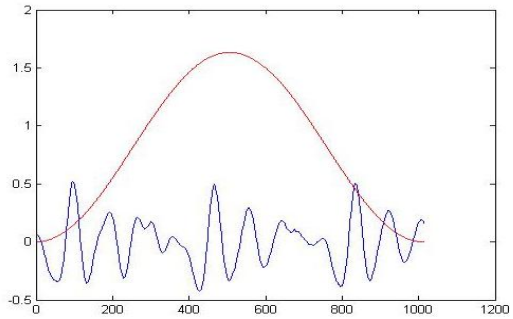
And define Hanning window as mentioned in the bellow,

$$W(n) = \beta_w \left(.5 - .5\cos\left(\frac{2\pi n}{N_w - 1}\right)\right) \quad (3)$$

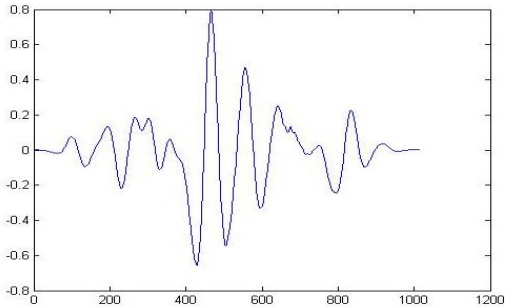
$$0 \le n \le N_w$$

Manuscript received February 1, 2011; revised May 12, 2011.

where N_w, corresponding to 30 ms, is the dimension of the Hanning window .β_w is a normalization factor defined so that the root mean square value of the window is unity.



(a)

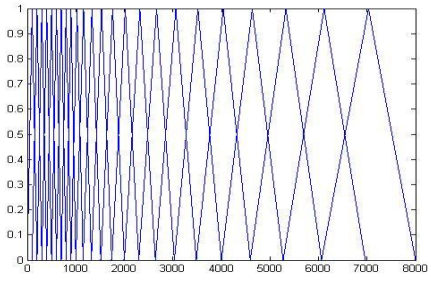


(b)

Figure 1. (a) One frame of Letter 'A' signal and window (b) windowed frame of Letter 'A' signal

The frequency band of 1 to 8 KHz is divided by 24 filters as illustrated in figure 2, with the following constraint:

$$\sum_{k=0}^{k-1} \varphi_j(k) = 1 \tag{6}$$



1-200	100-300	200-400	300-500	400-600	500-700	600-800	700-900
800-1000	900-1149	1000-1320	1149-1516	1320-1741	1516-2000	1741-2297	2000-2639
2297-3031	2639-3482	3031-4000	3482-4595	4000-5272	4595-6063	5272-6964	6063-8000

Figure 2. Filter allocation in the frequency domain before normalization with frequency bands.

The energy in each frequency band is calculated as:

$$E_j = \sum_{k=0}^{k-1} \varphi_j(k)\, X(k) \tag{7}$$

MFCC parameter is then calculated as:

$$c_m = \beta_c \sum_{j=0}^{J-1} \cos\left(m \frac{\pi}{J}(j + .5)\right) \log(E_j) \tag{8}$$

The last equation can also be shown as a scalar product between the log spectral energy vector and a vector of weighting factors given by

$$V_m = \left\{ \left(\cos\left(m \frac{\pi}{J}(j + .5)\right) \;\middle|\; 0 \le j \le J \right) \right\} \tag{9}$$

which leads to the following equation:

$$c_m = \beta_c \sum_{j=0}^{J} V_{m,j} \log(E_j) \tag{10}$$

The amplification factor, β_c, which accommodates the dynamic range of the coefficients c_m , depends on the value of the normalization factor β_w. In our implementation β_c is taken equal to 200 and its value does not change. Generally, only the first 15 values of c_m are retained.

IV. Time-Frequency Feature Extraction from the LP Residual Signal

In the linear predictive modeling of speech, a speech sample s(n) is approximated as the weighted sum of a limited number of past samples, i.e.,

$$\hat{S}(n) = \sum_{k=1}^{p} \alpha_k\, S(n-k) \tag{11}$$

where p is the order of prediction, and[ak ,k =1;2;.,.,p] are the predictor coefficients. In this study, a 12th-order LP filter is used. The prediction error, also referred to as residual signal, is given by,

$$e_n = S(n) - \sum_{k=1}^{p} \alpha_k\, S(n-k) \tag{12}$$

In the frequency domain, the residual signal can be derived as

$$H(z) = \frac{S(e^{jw})}{U(e^{jw})} \tag{13}$$

where$S(e^{jw})$ is the speech spectrum and

$$H(e^{jw}) = \frac{1}{1 - \sum_{j=1}^{p} \alpha_j e^{jw}} \tag{14}$$

Therefore the speech signal s(n) is the convolution output of the residual signal e(n) and the impulse response of the vocal tract filter $H(e^{jw})$;

So, we are calculating excitation signal for 'A' sounds as:

$$e(n) = S(n) - \sum_{j=1}^{p} \alpha_k\, S(n-k) \tag{15}$$

This has been typically shown in Fig. 3. A Hamming window of two pitch periods long is applied to pitch pulse of e(n), as shown in Fig. 4. Let $e_w(n)$ be the windowed residual signal.

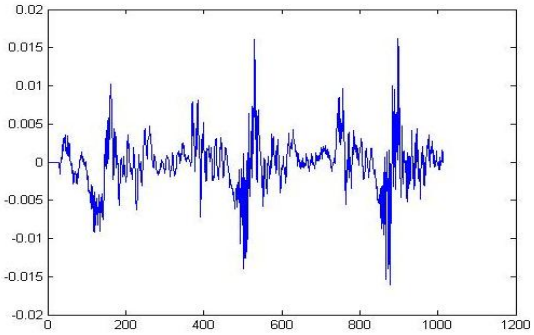
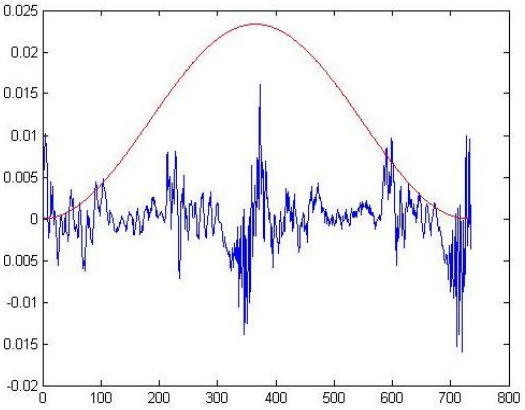
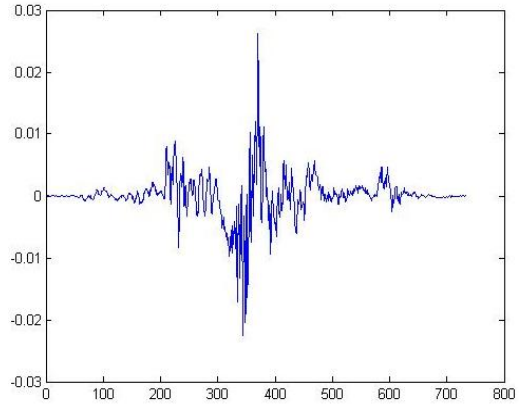


Figure 3. Residual signal of letter 'A'.



(a)



(b)

Figure 4. (a) The Residual signal of Letter 'A' (b) windowed frame of Residual signal of Letter 'A'.

The wavelet transform of $e_w(n)$ is computed as;

$$W(a,b) = \frac{1}{\sqrt{a}} \sum_{n=1}^{N} e_w(n) \varphi^* \left(\frac{n-b}{a}\right) \quad (16)$$

$\varphi^*(n)$ is the conjugate of the fourth-order Daubechies wavelet basis function (n), "a" is the scaling parameter and "b" is the translation parameter. The LP residual signal is analyzed in K octave subbands. In a specific subband, the temporal characteristics within the analysis window are measured as "b" changes.

To generate the feature parameters for pattern recognition, the wavelet coefficients with a specific scaling parameter are grouped as

$$W_k = \{w(2^k, b) | b = 1,2,\dots,N\} \quad (17)$$

where N is the window length. Each W_k is called an octave group. Then the WOCOR parameters can be derived as.

$$WOCOR_1 = \{\|W_k\| \; | k = 1,2,\dots,6\} \quad (18)$$

WOCOR1 has 6 elements and contains only spectral information. It includes no temporal characteristics within the analysis window. To retain the temporal details, each octave group can be equally divided into M sub-groups

$$W_k = \left\{w(2^k, b) \middle| b \in (m-1:m] \times ROUND\left(\frac{N}{M}\right)\right\} \quad (19)$$

m=1,2,…,M

Finally, a feature vector with 6M parameters can be generated as

$$WOCOR_M = \{\|W_k(m)\| | m = 1,2,..,M, K = 1,2,\dots,K\} \quad (20)$$

With multi-level wavelet transform, the pitch-related low frequency properties and the high frequency information associated with pitch epochs can be captured with different time-frequency resolutions. Pitch synchronization and dividing each octave group into several sub-groups enable the measuring of temporal variations of spectral components within a pitch period and that over consecutive periods. Therefore, WOCORM is capable of capturing the spectral-temporal characteristics of the LP residual signal. In this study, we let K = 6 to deal with speech data in the frequency range of 0 to 4000 Hz.

This leads to six frequency subbands at different octave levels: 2000-4000Hz (W1), 1000-2000Hz (W2), 500-1000Hz (W3), 250-500Hz (W4), 125-250Hz (W5), and 62.5-125Hz (W6). This could be observed in Fig. 5.

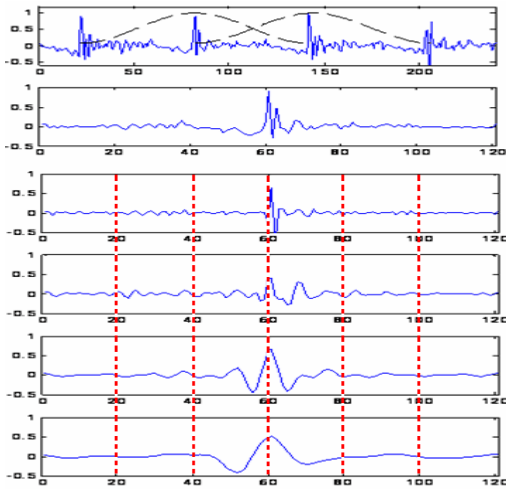


Figure 5. Up to down: respectively residual signal, a windowed frame of residual signal, Wavelet coefficients of residual signals in 4 octave subbands. Each octave group have been equally divided into M=6 sub-groups

The parameter M controls the temporal resolution attained by the WOCOR parameters. If all subbands have the same value of M, the dimension of WOCOR feature vectors will be 6M.

V. Speaker Recognition: The Proposed Method

First, test speech similarity coefficients with one of the registered speakers are calculated. We call this speaker, for instance, M.ESMITH. We can calculate test speech similarity coefficients with other registered speakers in the

same way.

For speaker recognition we need to first find the frame related to the mentioned letters in the test speech. We explain this method for finding letter 'A'. We use the same method to find the other letters.

First each voiced segment is divided into overlapping frames of 30 msec. overlapping is due to find the best candidate representing letter 'A'. In this paper, the overlapping time of 8ms has been considered. Figure 6 shows it more clearly.

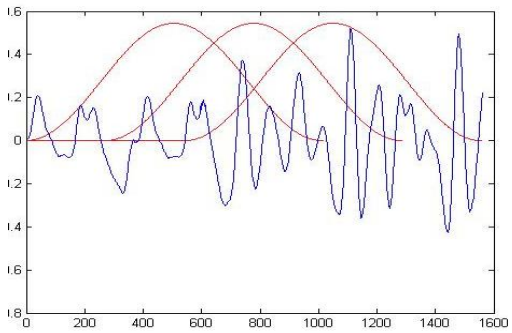


Figure 6. Overlapping frames with 8ms overlap length.

The MFCC feature parameters for each of these frames are then calculated. Next these parameters are compared to MFCC feature vectors of M.SMIT's letter 'A' . For comparison the following equation is used:

$$p_i = \frac{[\sum C_{mi} C_{ma}]^2}{\sum C_{mi} \sum C_{ma}} \quad (21)$$

Pi is the similarity coefficient of i th frame to M.SMIT's letter 'A'. Cmi is feature vector of i th frame. Cma is frame vector of M.SMIT's letter A. Since the MFCC feature parameters of letter A in different speaker available in database have similarity coefficient between 0.6 until 0.7, the frames that their similarity coefficient whit Mr.SMITH 's letter A is more than 0.75 are considered as Mr.SMITH's letter A . Three frames that have the most similarity coefficient , are selected. We consider the average of these coefficient , as the first speaker similarity coefficient with test speech.

$$SS(A) = \frac{p_1 + p_2 + p_3}{3} \quad (22)$$

$$p_1 > p_2 > p_3 > p_i$$

The WOCOR feature parameter of the frame with highest SS is extracted. then the similarity coefficient of this feature parameter to WOCOR feature parameters of Mr.smith's letter A is considered as the second speaker similarity coefficient.

$$ES(A) = \frac{[\sum WOCOR_{sm} WOCOR_a]^2}{\sum WOCOR_{sm}^2 \sum WOCOR_a^2} \quad (23)$$

WOCOR_{sm} is the WOCOR feature parameter of the frame with highest SS and WOCOR_a is WOCOR feature parameter of the Mr.smith's letter 'A'. Finally the similarity coefficient to new test signal in 'A' letter is equal with;

$$\mathrm{sim}(A) = .7\,\mathrm{SS}(A) + .3\mathrm{ES}(A) \quad (24)$$

For all voiced letter mentioned before, the similarity coefficient of M.SMITH to test a speaker are obtained and finally the similarity coefficient is derived as:

$$\mathrm{SIM} = \frac{\mathrm{sim}(A)+\mathrm{sim}(I)+\mathrm{sim}(U)+\mathrm{sim}(æ)+\mathrm{sim}(e)+\mathrm{sim}(ɔ:)}{6} \quad (25)$$

For the all speaker, the similarity coefficients are obtained and the speaker with highest similarity coefficient is the speaker of test speech.

VI. Speaker Recognition Experiments

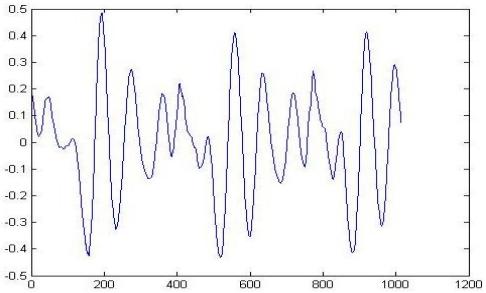
The commonly used performance evaluation metric for speaker identification is the identification error rate (IDER)

$$\mathrm{IDER} = \frac{\text{number of misidentified trials}}{\text{total number of identification trials}} \times 100\% \quad (26)$$

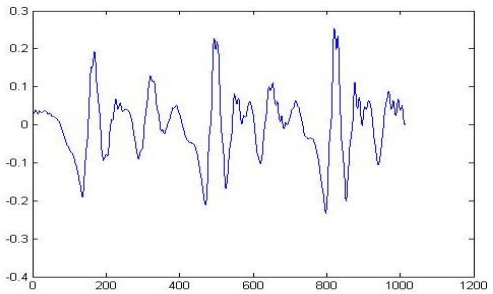
For speaker verification, the decision threshold should be selected to make trade-off between the false acceptance (FA) and false rejection (FR) errors. Generally, the threshold is selected such that the false acceptance rate equals to the false rejection rate, usually referred as the equal error rate (EER).

The database was concluded 40 speakers, each speaker utterances 5 sentences. Three sentences are used as training set and the spare sentences are used test set. The additive noise files were obtained by adding the white Gaussian noise to clean speech so as to maintain the segmental SNR round the desired value. The length of training speech is 10 second. Then we extract a frame of all of voiced letters from each speaker sentence. We depict a frame of 'A' and 'I' in figure 7. Each frame is equal with 25 ms. The dimension of the Each frame is equal to:

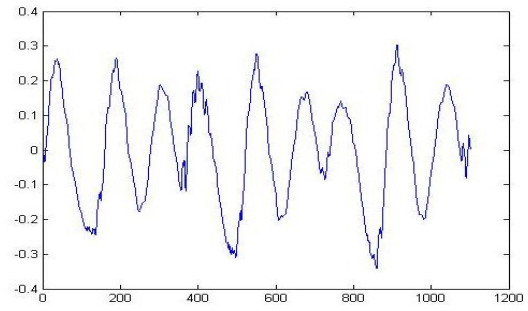
$$N_w = Fs \times 23m = 44100 * 25ms = 1014$$



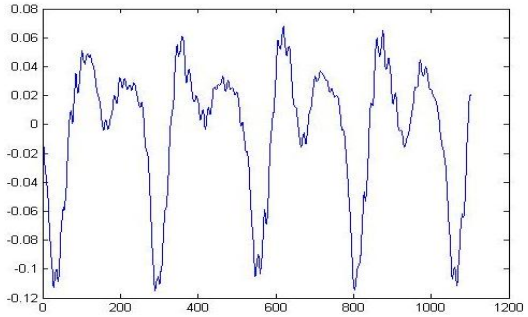
(a)



(b)



(c)



(d)

Figure 7. D1 and D2 are tow speakers. (a) one frame of Letter ‘A’ of speaker D1 (b) one frame of Letter ’A’ of speaker D2 (c)one frame of Letter ‘I’ of speaker D1 (d) one frame of Letter ‘I’ of speaker D2

MFCC and WOCOR feature parameters of these frames are extracted. In table 1 and 2 MFCC and WOCOR feature sets IDER (in %) for each letter is shown. It is observed that for all of these letters the efficiency of speaker recognition by using MFCC is better than WOCOR method. But speaker recognition by combining both of them is more exactly.

As a result in table 3 the performance of the MFCC in comparison with the proposed method has been shown.

TABLE1. MFCC FEATURE SETS USING VOICED LETTERS: IDER (in %).

IDER (MFCC)	0db	5db	10db	15db	20db
Sim(A)	10.5	5.2	1.6	0	0
Sim(I)	25	15.1	7.3	4.5	1.2
Sim(U)	7.7	3.5	0	0	0
Sim(æ)	13.2	5.3	1.4	0	0
Sim(e)	17	11.7	5.2	3	0.5
Sim(ɔ:)	4.9	3	0.6	0	0
SIM	3.6	1.6	0	0	0

TABLE2. WOCOR4 FEATURE SETS BY USING VOICED LETTERS: IDER (in %).

SNR / IDER (WOCOR)	0db	5db	10db	15db	20db
Sim(A)	35	25	14	8.8	2.5
Sim(I)	43.8	31	17.3	9.5	1.8
Sim(U)	16	12.6	8	4.7	0.5
Sim(æ)	33	27	18	7.4	3.3
Sim(e)	56.3	42.1	22.5	13	4.2
Sim(ɔ :)	24	11.3	6.8	3.4	0.6
SIM	6	3.8	2	1.1	0.4

TABLE 3. THE PERFORMANCE OF MFCC IN COMPARISON WITH THE PROPOSED METHOD

SNR / IDER	0db	5db	10db	15db	20db
MFCC	0.36	0.25	0.16	0.1	0.06
Proposed method	0.25	0.1	0.05	0.03	0.01

VII. CONCLUSIONS

Using all of voiced segment of speech in text-dependent speaker recognition has high efficiency because the number of voiced letters is fixed in each fixed sentence. However using all of voiced speech in text-independent speaker recognition has poor efficiency because: first the number of voiced letter is not specified and second different voiced letters have significant difference. Because in Persian language the number of vowel letters are a few (only 6 letters), and also every consonant letter is followed by a vowel letter, in this paper we have used the vowel letters for speaker recognition. Since in noisy environment the MFCC has poor performance due to sudden changes of speech signal and, the positions of pitch epoch and epoch pulses remain as the same places for clean speech, we employed WOCOR as a complimentary method for MFCC. Obtaining results show that the proposed method provides higher performance in comparison with the MFCC alone.

REFERENCES

[1] Ning Wang , P. C. Ching , Nengheng Zheng, andTan Lee, “*Robust Speaker Recognition Using Denoised Vocal Source and Vocal Tract Features speaker verification*,” IEEE Transaction on Audio Speech and Language processing, Vol. 1, No. 2, pp. 25-35 , 2010.

[2] N. Zheng, P. C. Ching, and T. Lee, “*Time frequency analysis of vocal source signal for speaker recognition*,” in Proc. ICSLP, 2004, pp. 23.33.

[3] L. R. Rabiner and B. H. Juang, “*Fundamentals of Speech Recognition”*. Englewood, NJ: Prentice Hall, 1993.

[4] P. Thevenaz and H. Hugli, “*Usefulness of the LPC-residue in text independent speaker verification*,” Speech Commun., vol. 17, no. 1-2, pp. 145–157, 1995.

[5] M. Chetouani, M. Faundez-Zanuy, B. Gas, and J. L. Zarader, "*Investigation on LP-residual representations for speaker identification,*" Pattern Recognition letter, vol. 42, pp. 487–494, 2009
[6] N. Zheng, P. C. Ching, and T. Lee, "*Time frequency analysis of vocal source signal for speaker recognition,*" in Proc. ICSLP, 2004, pp. 2333– 2336.
[7] N. Zheng, T. Lee, and P. C. Ching, "*Integration of complementary acoustic features for speaker recognition,"* IEEE Signal Proc. Lett., vol. 14, no. 3, pp. 181–184, Mar. 2007.
[8] W. N. Chan, N. Zheng, and T. Lee, "*Discrimination power of vocal source and vocal tract related features for speaker segmentation,*" IEEE Trans. Audio Speech and Lang. Process., vol. 15, no. 6, pp. 1884–1892, Aug. 2007.
[9] D. O'Shaughnessy, "*Speaker recognition,*" IEEE ASSP Mag., vol. 3,no. 4, pp. 4–17, Oct. 1986.
[10] P. S. Aleksic and A. K. Katsaggelos, "*Audio-visual biometrics*," in IEEE Trans. Audio Speech and Lang. Process, vol. 94, no. 11, pp. 2025–2044, Nov. 2006.
[11] K. Chen, "*Personalize mobile access by speaker authentication* in *Biometric solutions: For authentication in an E-world*," Ed. Springer, 2002.
[12] J. P. Campbell, "*Speaker recognition: A tutorial,*" in IEEE Trans. Audio Speech and Lang. Process, vol. 85,no. 9, pp. 1437–1462, Sep. 1997.
[13] D. A. Reynolds, "*Speaker identification and verification using Gaussian mixture speaker models,*" Speech Gommun., vol. 17, pp. 91–108, 1995.

Hassan Farsi received the B.Sc. and M.Sc. degrees from Sharif University of Technology, Tehran, Iran, in 1992 and 1995, respectively. From 1992 to 1993, he worked in research communications centre, Tehran, Iran on waveform speech coders. In 1995, he was employed in University of Birjand , Birjand , Iran as a lecturer in the department of Electronics and communications engineering. Since 2000, he started his Ph.D in the Centre of Communications Systems Research (CCSR), University of Surrey, Guildford, UK, in field of speech quality improvement provided by low bit rate speech coders and received the Ph.D degree in 2004. His research has also led to improving speech quality coded by 2.4 kb/s standard MELP coder. Besides speech processing, he is also interested in image and video processing and he is doing his research in this area as well. Now, he works as an academic staff in department of Electronics and Communications eng., university of Birjand, Birjand, Iran. His email address is: hfarsi@birjand.ac.ir and also hs_farsi@yahoo.co.uk.

Saber Amjadi was born in Savadkuh, Iran, in 1985. He received the B.S and M.Sc degree in Electrical engineering {communication} from University of Tabriz, Tabriz, Iran, in 2007 and University of Birjand, Iran in 2011 respectively. His research interests include image processing and intelligent computing. E-mail: S.Amjadi@Birjand.ac.ir and also amjadisa@yahoo.com.