IJICT International Journal of Information & Communication Technology

ITRC

Volume 2- Number 2- July 2010 (pp.9 - 19)

A Robust Voice Activity Detection Based on Short Time Features of Audio Frames and Spectral Pattern of Vowel Sounds

Mohammad H. Moattar
Laboratory for Intelligent Signal and Speech Processing, Computer Engineering and IT Dept.
Amirkabir University of Technology (AUT)
Tehran, Iran
moattar@aut.ac.ir

Mohammad M. Homayounpour
Laboratory for Intelligent Signal and Speech Processing, Computer Engineering and IT Dept.
Amirkabir University of Technology (AUT)
Tehran, Iran
homayoun@aut.ac.ir

Received: December 14, 2009- Accepted: July 28, 2010

Abstract— **This paper presents a set of voice activity detection (VAD) methods, that are easy to implement, robust against noise, and appropriate for real-time applications. The common characteristic is the use of a voting paradigm in all the proposed methods. In these methods, the decision on the voice activity of a given frame is based on comparing the features obtained from that frame with some thresholds. In the first method, a set of three features, namely frame energy, spectral flatness, and the most dominant frequency component is applied. In the second approach however, the spectral pattern of the frames of vowel sounds is used. To use the strengths of each of the above methods, the combination of these two decision approaches is also put forth in this paper. The performance of the proposed approaches is evaluated on different speech datasets with different noise characteristics and SNR levels. The approaches are compared with some conventional VAD algorithm such as ITU G. 729, AMR and AFE from different points of view. The evaluations show considerable performance improvement of the proposed approaches.**

Keywords- voice activity detection, spectral flatness, vowel spectral pattern, noise robustness, vowel sounds

I. INTRODUCTION

Voice Activity Detection (VAD) is a critical task in many speech/audio processing applications. According to [1], the required characteristics for an ideal voice activity detector are: reliability, robustness, accuracy, adaptation, simplicity, real-time processing and no prior knowledge of the noise. The most challenging characteristics of an ideal VAD algorithm are robustness against noisy environments and its computational complexity, especially when a real-time application is targeted. The performance of all VAD algorithms degrades to a certain extent when the SNR decays. The ultimate goal of the newly proposed approaches is to lessen this degradation. Most of the previously proposed methods have partially overcome this problem at the price of higher computational complexity. Simplicity and robustness against noise are two essential characteristics for a voice activity detection procedure, and should be considered simultaneously in an applicable VAD algorithm.

Many different VAD approaches have been proposed, which have considered this task from

different points of view, but the main concern of these methods is the robustness of the approach. The difference between most of the previously proposed methods is the features they use. Various kinds of robust acoustic features, such as autocorrelation function based features [2], spectrum based features [3], the power in the band-limited regions [4, 5], Mel-frequency cepstral coefficients [2], delta line spectral frequencies [5], and features based on higher order statistics [6] have been proposed for VAD. Multiple features used in parallel provide more robustness against different noises. In some previous works, multiple features are applied in combination with some modelling and decision algorithms such as classification and regression tree (CART) [7] or artificial neural network (ANN) [8].

Some of the most widely proposed voice activity detection methods are based on statistical pattern classification [9]. These methods require the noise model to be trained, using a set of corresponding noisy speech data. This limits their use in unknown noisy environments. Also, most of these methods assume the noise to be stationary during a certain time period, making them sensitive to changes in the SNR of the observed signal or the nature of noises. To overcome this shortcoming, some previous works propose noise estimation and adaptation for improving VAD robustness [10], but these kinds of methods are computationally expensive.

In this paper a set of voice activity detection approaches are proposed which are relatively robust to noise and environmental varieties. The main goal of these approaches is to improve the robustness of the VAD algorithm while maintaining the processing speed - convenient for real-time applications. Section 2 presents the proposed VAD algorithms. Section 3 introduces the evaluation datasets and performance metrics. The experimental results are discussed in Section 4. Finally, the conclusions and future works are mentioned in Section 5.

II. Proposed Approaches

In the first proposed method, which is previously published by the authors [11], three different features per each frame are used. The first feature is the widely used short-term energy. Energy is the most common feature for speech/silence detection. However, this feature loses its efficiency in noisy conditions especially in lower SNRs.

The second feature is spectral flatness measure (SFM). Spectral flatness is a measure of the noisiness of spectrum and is a good feature in voiced/unvoiced/silence detection. In some resources, SFM is defined by the ratio of the geometric mean to the arithmetic mean of the power spectral density components in each critical band [12]. Alternatively, in some other resources, SFM is computed on the entire frame spectrum and no sub-band division is required [13, 14]. Naturally, geometric mean is higher than arithmetic mean. Therefore, SFM has the property of being bounded between zero and one. For an unvoiced frame, since the spectral magnitude has a similar amount of power in all spectral bands and the spectrum is relatively flat, the difference between arithmetic and geometric means is inconsiderable. Therefore, for an unvoiced frame, SFM is close to 1. For voiced frames, the spectral density is concentrated in a small number of frequencies. For such frames, the geometric mean of the power spectrum is much lower than the arithmetic mean. Therefore, SFM is relatively higher for unvoiced frames compared to voiced frames [15].

In this paper, since we intend to capture the tonal characteristics of a speech frame in a single value, we use the definition of SFM in [13, 14] and calculate SFM on the entire frame spectrum. Therefore, the spectral flatness measure of the ith frame is computed as follows:

$$SFM(i) = 10\log_{10}(G/A) \qquad (1)$$

where A and G are arithmetic and geometric means of the frame spectrum, respectively. In Eq. (1), the SFM value is expressed in decibels. Since the ratio of the geometric mean to the arithmetic mean of the frame spectrum is lower than or equal to 1, SFM in Eq. (1) is always negative. Therefore, the absolute value of SFM for voiced frames is typically higher than its absolute value for unvoiced speech. We use this property of SFM for threshold based VAD.

In addition to these two features, our experiments showed that the most dominant frequency component of speech frame spectrum can be very useful in discriminating between speech and silence frames. We found out that this feature is relatively higher for voiced speech than the silence parts. Also we observed that this feature is relatively robust against noise. The most dominant frequency component of an audio frame is denoted by F in the rest of the paper. Fig. 1 illustrates average spectral magnitude of voiced and silence frames for clean speech and White noise (i.e. 5 dB SNR). The fact that the average value of F feature (marked with *) is higher for voiced frames than silence and noisy frames is clearly observable in this figure.

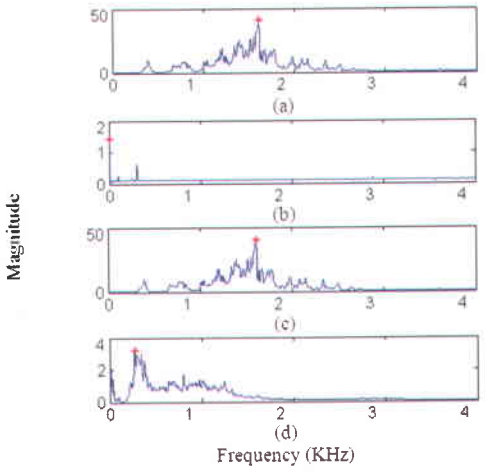

Fig 1. Average spectral magnitude of (a) voiced frames and (b) silence frames with their corresponding F value (marked with *) in clean speech. Figures (c) and (d) respectively show the average spectrum of the voiced frames and silence frames in White noise at 5 dB SNR. White noise is extracted from NOISEX-92 corpus.

Also, Fig. 2 shows the spectrums of a voiced frame (Fig. 2(a)) and a silence frame (Fig. 2(d)) and their

corresponding most dominant frequency component. Also the spectrum of these two frames in White and Babble noise is illustrated in Fig. 2. As you can see, the F value for voiced frame (marked with *) is relatively higher than the F value for silence and noisy frames. This figure shows that even when the silence frame spectrum gets flatter by the noise effect (Fig. 2(e)), the maximum spectral peak is still in lower frequencies, compared to the voiced frame spectrum.

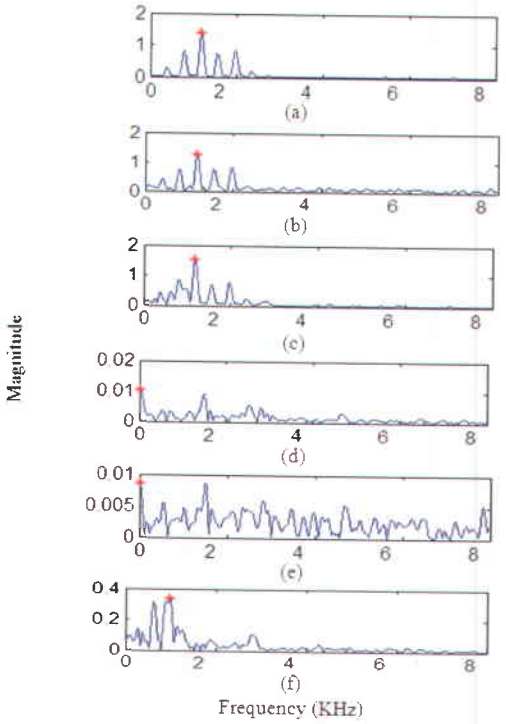

Fig 2. Spectral magnitude of (a) a voiced frame and (d) a silence frame with their corresponding F value (marked with red *). Figures (b) and (c) show the spectrum of the voiced frame in White and Babble noises respectively. Also the spectrums of the silence frame in White and Babble noise are illustrated in (e) and (f), respectively. Signal SNR is 5 dB. Noises are extracted from NOISEX-92 corpus.

However, the above property is not generalizable to all silence/speech frames and all noise types. For example, if the speech signal is affected by the channel noise, or the signal SNR is relatively low, the maxima of the silence frame spectrum may move to higher frequencies and exceeds the F value of speech frames. This can be observed in Figs 3 and 4, in which the overlap between the distributions of F for speech and non-speech frames changes in different noises. Fig. 3 and Fig. 4 illustrate the relative distribution of this feature for speech and silence classes in clean speech condition and for different noise conditions, respectively. For better demonstration, the feature values are scaled to the range of [0, 1].

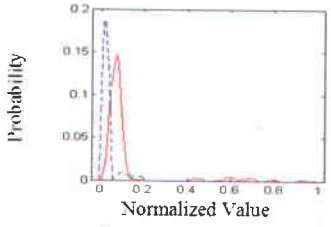

Fig. 3. Distribution of the F feature for speech (strict red contour) and silence (dashed blue contour) classes in clean speech

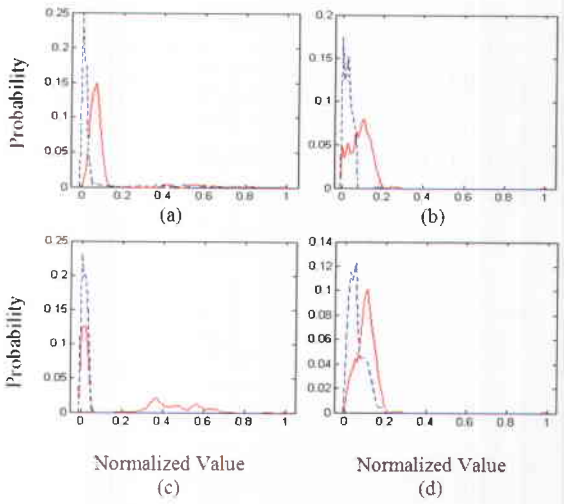

Fig. 4. Distribution of F feature for speech (strict red contour) and non-speech (dashed blue contour) classes in (a) White, (b) Babble,(c) Factory, and (d) Volvo noises at 5 dB SNR. Noises are extracted from NOISEX-92 corpus.

The discriminative power of this feature is observable for clean speech and White, Babble and Volvo noises. Fig. 4(c) depicts that the mentioned feature is less discriminative in Factory noise.

The most dominant frequency component is simply computed by finding the frequency corresponding to the maximum value of spectrum magnitude, $|S(k)|$. Figs 5, 6 and 7 represent the effectiveness of these three features (i.e. energy, SFM and F) when speech is clean or is corrupted by White and Babble noises.

The proposed algorithm starts with framing the audio signal. In our implementations, no windowing is done on the frames. The first N frames are used for threshold initialization. For each incoming speech frame, the mentioned features are computed. The audio frame is marked as a speech frame if more than one of the feature values exceeds the pre-computed threshold.

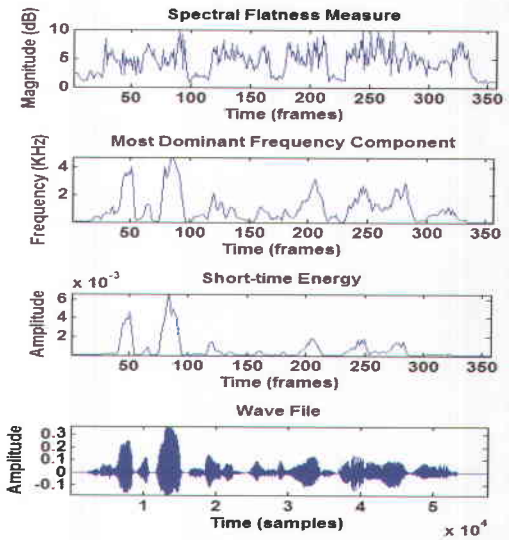

Fig. 5. Feature values for a clean speech signal

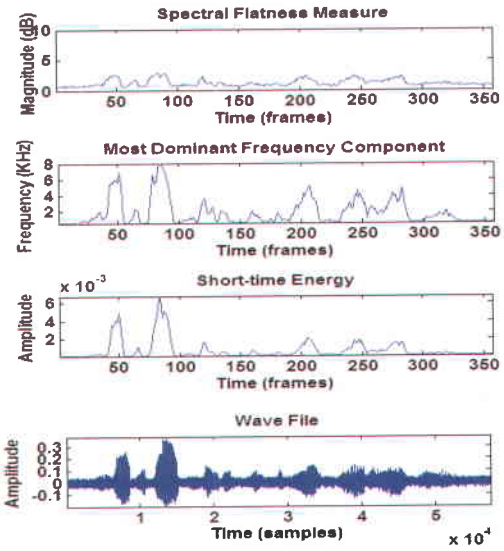

Fig. 6. Feature values for a speech signal corrupted with White noise at 5 dB SNR. Noises are extracted from NOISEX-92 corpus.

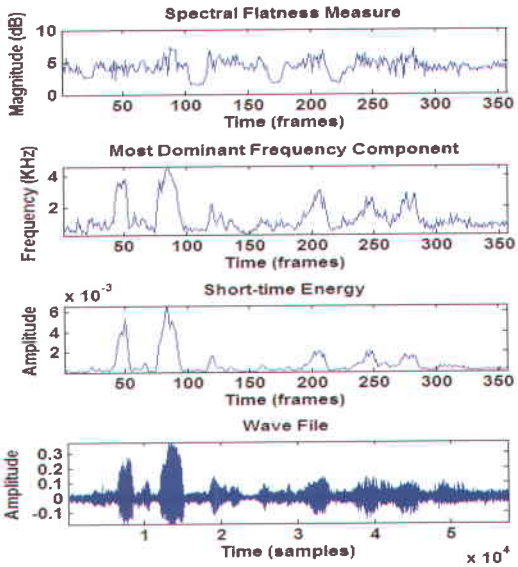

Fig. 7. Feature values for a speech signal corrupted with Babble noise at 5 dB SNR. Noises are extracted from NOISEX-92 corpus.

The complete procedure of the proposed method is described below:

Algorithm 1. VAD algorithm with the voting approach using energy, spectral flatness and most dominant frequency component

1) Set *Frame_Size* = 10ms and compute number of frames (*nbFrames*)(no overlap is required)
2) For *i* from 1 to *nbFrames*
   a) Compute frame energy, $E(i)$.
   b) Apply FFT on each speech frame and compute the frame spectrum, $|S(i)|$.
      I) Find $F(i) = \arg\max_k(|S(k)|)$ as the most dominant frequency component.
      II) Compute the absolute value of spectral flatness measure, $SFM(i)$.
   c) Supposing that the first N=20 frames are non-speech, find the minimum value for E (Min_E), F (Min_F) and SFM (Min_SFM).
   d) Set decision thresholds for E, F and SFM according to Eq. (3).
   e) Set $Counter = 0$.
      I) If $E(i) > Thresh_E$ then $Counter++$,
      II) If $F(i) > Thresh_F$ then $Counter++$,
      III) If $SFM(i) > Thresh_{SFM}$ then $Counter++$,
   f) If $Counter > 1$ then mark current frame as speech else mark it as silence
   g) If current frame is marked as silence, update the energy minimum value:

$$Min\_E = \frac{(Silence\_Count * Min\_E) + E(i)}{Silence\_Count + 1} \quad (2)$$

   h) $Thresh_E = Param_E * \log10(Min\_E)$
3) Ignore silence-run less than 5 successive frames.
4) Ignore speech-run less than 5 successive frames.

The decision thresholds in the proposed approach are defined as follows:

$$Thresh_{Feature} = Min\_Feature + Param_{Feature} \quad (3)$$

where $Feature \in \{E, F, SFM\}$. Using the $Min\_Feature$ term in the threshold computation and updating formula helps the VAD approach adapt to different noises and SNR levels.

The optimal values of $Param_{Feature}$ is found on a sufficiently big set of clean speech data, called the development dataset, so that the total performance of the algorithm on this set is maximized. Different search strategies can be applied to span the space of possible parameter values, including linear search, simulated annealing and genetic algorithm. In this paper, we have applied a linear search strategy to find the proper values for parameters, $Param_{Feature}$. We have used the test set of the TIMIT speech database as the development dataset, the details of which is mentioned in the next section

In [16] it is suggested that the positions of spectral peaks are important factors in discriminating vowel sounds from others. Also it is claimed that spectral peaks of vowel sounds are robust against the noisy environments even in severe noisy conditions. This important feature can be applied in voice activity detection. Using this approach, the problem of voice activity detection will change to the problem of detecting vowels. In speech/silence discrimination problem, the vowel-like sounds also appear in silence parts corrupted by Babble noise, and may influence the VAD performance in such environments. In such cases, other discriminating features can help to improve the VAD accuracy. Otherwise, the occurrence of vowel-like patterns is almost improbable in noise corrupted silence signal. Fig. 8 illustrates the spectrum of a vowel sound in three different SNR and noise

conditions. As can be seen in this figure, the positions of the peaks in the frame spectrum are relatively unchanged for all three cases.

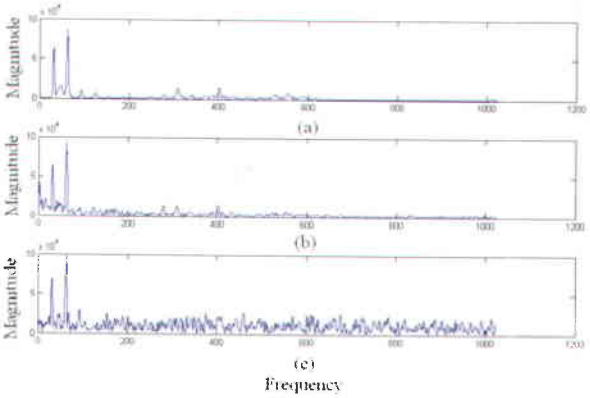

Fig. 8. Spectrum of a vowel sound (a) Clean (b) Corrupted by Pink noise (SNR= 5 dB) (c) Corrupted by White noise (SNR= -5 dB)

By exploiting this property, an approach to improve the performance of VAD algorithm in presence of various noises and SNR conditions is proposed. In this approach, a huge set of patterns of the locations of spectral peaks of various vowel sounds is extracted from a set of training data. In the test stage, for every incoming audio frame s, the relevance of the frame to the vowel family V is calculated with the following measure which is called the spectral relevance (SR) [16]:

$$SR_s = \max_{X \in V}(SR(S, X)) \quad (4)$$

$$SR(S,X) = \frac{\sum_{i=0}^{N-1}(X[i]S[i])}{\sum_{i=0}^{N-1} S[i]} - \frac{\sum_{i=0}^{N-1}(X[i](1-S[i]))}{\sum_{i=0}^{N-1}((1-S[i]))} \quad (5)$$

where S is the spectrum of the input frame and X is a sample pattern of the vowel spectrum. After computing SR_s for the incoming frame, the frame is marked as a vowel if SR_s exceeds a decision threshold. There is a difference between the above measure and the approach proposed in [16]. In the proposed approach, no peak detection is done and the unchanged spectrums of vowel sounds are determined as the vowel patterns. On the other hand, in [16] the peaks of vowel spectrums are detected and marked as 1 while the rest of the spectrum bands are marked as zeros. This binary vector is determined as the spectrum pattern of the vowel. Since the original spectrum includes the magnitude of the frequency components, it will function as weighted coefficients for the spectral peaks, while the binary patterns weigh all the peaks identically. In addition, peak detection is not simple and we may have concerns about the amplitude of the selected peaks. In our experiments, we observed that the proposed idea is more robust and computationally more efficient than the approach proposed in [16].

The second proposed approach includes two stages. In the first step (training step), the spectral signatures of vowel sounds (V) are extracted from a set of training data. The training phase of this approach is summarized in Algorithm 2.

Algorithm 2. Training phase of the second approach

1- A set of vowel segments is extracted from labeled training data.
2- Each segment is divided into 30 ms length frames at 10ms frame rate.
3- The Fourier transform is applied on every frame s.
4- The average spectrum of each sound is calculated.
5- The average spectra of vowel sounds are grouped using k-means clustering algorithm.
6- The centroids of the resulted clusters are extracted and stored as the spectrum signature of vowel for future references.

This VAD approach is similar to the previous one. The difference is that the decision on the existence of speech in a frame is done using a single thresholding on the SR_s value of the analysis frame. This approach is explained in more details in Algorithm 3.

Algorithm 3. The second VAD algorithm using spectral relevance of vowel sounds

1) Divide the input signal to 30 ms duration frames with 20 ms overlap. Compute the number of frames, *nbFrames*.
2) For i from 1 to *nbFrames*
   a) Apply FFT on each speech frame and compute the frame spectrum, $|S(i)|$.
   b) Compute the SR_s of $|S(i)|$, $SR_S(i)$.
   c) Supposing that the first N=20 frames are non-speech, find the minimum value of SR_s (Min_SR_S).
   d) Set decision threshold for SR_s according to Eq. (3).
   e) If $SR_S(i) > Thresh_{SR}$ then mark current frame as speech else mark it as silence.
   f) If the current frame is a speech (vowel-like) frame, mark 5 neighboring frames before and after the current frame as speech too.
   g) If current frame is marked as silence, update the minimum value of SR_s (Min_SR_S).
3) Ignore silence-run less than 5 successive frames.
4) Ignore speech-run less than 5 successive frames.

As denoted in the above algorithm, the frame length of the analysis window is increased to 30ms instead of 10ms in the first approach (Algorithm 1). This analysis window length is used to increase the frequency resolution of the analysis and the resolution of peak locations. Longer analysis window can also be used but the window length is limited to maintain the processing time of the VAD algorithm.

To use the strengths of each of the above approaches, the decision rules are combined in a voting scheme to see if the final performance will

improve. Algorithm 4 shows the combinational approach in details.

Algorithm 4. The combinational approach for voice activity detection

1) Divide the input signal to 30 ms duration frames with 20 ms overlap. Compute the number of frames, *nbFrames*.
2) For i from 1 to *nbFrames*
   a) Compute frame energy, $E(i)$.
   b) Apply FFT on each speech frame and compute the frame spectrum, $|S(i)|$.
      I) Find $F(i) = \arg\max_k(|S(k)|)$ as the most dominant frequency component.
      II) Compute the SR_s of $|S(i)|$, $SR_S(i)$.
      III) Compute the absolute value of spectral flatness measure, $SFM(i)$.
   c) Supposing that the first N=20 frames are non-speech, find the minimum value for E (Min_E), F (Min_F), SFM (Min_SFM) and SR_S (Min_SR_S).
   d) Set decision threshold for E, F and SFM and SR_S according to Eq. (3).
   e) Set $Counter = 0$.
      I) If $E(i) > Thresh_E$ then $Counter++$,
      II) If $F(i) > Thresh_F$ then $Counter++$,
      III) If $SFM(i) > Thresh_{SFM}$ then $Counter++$,
      IV) If $SR_S(i) > Thresh_{SR}$ then $Counter++$,
   f) If $Counter > 1$ then mark the current frame as speech else mark it as silence
   g) If current frame is marked as silence, update the energy minimum value (Eq. 2).
   h) $Thresh_E = Param_E * \log10(Min\_E)$.
3) Ignore silence-run less than 5 successive frames.
4) Ignore speech-run less than 5 successive frames.

III. Evaluations

To evaluate the proposed method, three different speech corpora are used. The first one is TIMIT Acoustic-Phonetic Continuous Speech Corpus [17] which is regularly used for the evaluation of speech recognition systems and contains only clean speech data. The training set of the TIMIT database is used for vowel sounds extraction to determine the spectral signature of the vowel frames. From this set, 2474 utterances are used for extracting 5000 vowel segments. These 5000 segments are the only samples that are used as the reference patterns in all experiments. Also we used the test data of this corpus as the development dataset for determination of the parameter values. It should be noted that the parameter estimation is performed on a noise-free dataset.

The second corpus is a Farsi telephony speech corpus named TPersianDat collected at AUT's Laboratory of Intelligent Signal and Speech Processing. This corpus is recorded for telephony speech and speaker recognition. This corpus is gathered in real world conditions and the speech files include background noise. Using this corpus is helpful, since the VAD algorithm will be evaluated for a different recording condition, namely telephony speech. Also, the language of the evaluation dataset will be different from the language from which the vowel spectral patterns are extracted (Farsi against English). This helps evaluate the performance of the algorithms in different environmental conditions which will be encountered in real applications.

The third dataset, commonly used for evaluating VAD algorithms, is the Aurora2 Speech Corpus. The Aurora2 Speech corpus includes clean speech data as well as noisy speech. Aurora2 database contains utterances with 8 different noises and 6 different SNRs (i.e. -5dB to 20dB with 5dB increment), from which we selected Subway, Babble and Car noises.

To show the robustness of the proposed methods against noisy environments, five different noises (i.e. White, Babble, Pink, Factory and Volvo) at four different SNRs (i.e. -5, 5, 15 and 25dB) were added to the clean speech signals in the first two corpora. No additional noise is added to the Aurora2 corpus.

To obtain a better viewpoint of the performance of the proposed methods, they were compared with four other VAD algorithms. The first one which is proposed in [10] finds an estimation of noise using minimum mean-squared error (MMSE) and is proposed for VAD in conditions with vehicular noise. The other methods, that are mostly used as a reference method for evaluation of VAD algorithms, are the ITU G.729 Annex B standard [4], the European Telecommunications Standards Institute (ETSI) Adaptive Multi-Rate (AMR) [18] with VAD option 1 (AMR1) and option 2 (AMR2), and the VAD module of the ETSI Advanced front-end feature extraction (AFE) [19].

Two common metrics known as silence hit rate (HR0) and speech hit rate (HR1) are used for evaluating the VAD performance. It is necessary to mention that there is often a trade-off between these two metrics and increasing one may lead to decreasing the other. To have a better metric for comparing two different VAD methods, we define a total performance metric (T) as the average of HR0 and HR1.

IV. Expriments

At first, the proposed methods are evaluated on TIMIT and TPersianDat datasets. For these evaluations, White, Babble, Pink, Factory and Volvo noises at 25, 15, 5 and -5 dB SNRs are added to the test utterances. Table 1, shows the results achieved from these evaluations in term of HR0, HR1 and T.

As seen in the above table, the average performance of the proposed voting scheme is 75.41%. Also Table 1 shows equal performance in detecting both silence and non-silence parts. The worst performance of this approach is under the condition

when the speech signal is corrupted by Volvo noise.

Table 1. Avearge accuracy of the first approach on TIMIT and TPersianDat

Noise	SNR (dB)	Accuracy %		
		HR0	HR1	T
None	---	85.77	92.94	89.36
White	25	92.33	87.21	89.77
	15	92.33	77.55	84.94
	5	80.54	65.68	73.12
	-5	52.08	64.51	58.3
Babble	25	87.01	83.48	85.17
	15	79.36	82.75	81.06
	5	65.35	79.51	72.43
	-5	28.52	84.49	56.51
Pink	25	92.11	86.24	89.18
	15	92.55	74.61	83.60
	5	95.24	58.84	77.04
	-5	99.71	34.795	67.255
Factory	25	91.30	85.84	88.57
	15	87.49	78.84	83.16
	5	80.85	68.645	74.75
	-5	71.76	54.67	63.22
Volvo	25	85.72	83.89	84.80
	15	63.95	78.74	71.35
	5	36.36	75.90	56.13
	-5	20.86	87.24	54.05
Average		75.29	75.54	75.41

The next evaluations concern the performance of the second method which uses the spectra of the vowel sounds as a reference for deciding on the existence of speech in audio signal. The results of these evaluations on the same databases are illustrated in Table 2.

Table 2. Avearge accuracy of the second approach on TIMIT and TPersianDat databases

Noise	SNR (dB)	Accuracy %		
		HR0	HR1	T
None	---	87.13	92.17	89.65
White	25	91.91	87.86	89.89
	15	92.35	78.335	85.345
	5	95.38	64.68	80.01
	-5	99.71	37.20	68.45
Babble	25	94.00	83.08	88.54
	15	77.16	85.33	81.24
	5	63.77	79.26	71.51
	-5	56.27	77.78	67.02
Pink	25	92.6	81.86	87.23
	15	96.35	73.27	84.81
	5	83.63	65.98	74.80
	-5	75.55	57.85	66.70
Factory	25	92.49	81.05	86.77
	15	95.07	75.82	85.45
	5	66.86	69.54	68.20
	-5	77.52	68.96	73.24
Volvo	25	87.26	82.76	85.01
	15	45.61	84.14	64.87
	5	22.94	88.31	55.62
	-5	17.96	94.46	56.21
Average		76.74	76.65	76.69

The result in the above table is almost the same as the previous one except that the average performance of this method is slightly higher. Again, in this approach the HR0 and HR1 measures are balanced and the worst case happens for the speech signal corrupted by Volvo noise. The third experiment concerns the evaluations of the combinational approach in voice activity detection. The results of these evaluations are listed in Table 3. The average performance of the combinational approach is higher than the performance of the previous approaches. This algorithm has inherited the main characteristics of the two original approaches and its main deficiency is its lower performance in the presence of Volvo noise.

Table 3. Avearge accuracy of the combinational approach on TIMIT and TPersianDat

Noise	SNR (dB)	Accuracy %		
		HR0	HR1	T
None	---	91.59	92.72	92.15
White	25	90.67	86.74	88.71
	15	92.93	73.88	83.40
	5	89.44	62.30	75.87
	-5	69.39	60.70	65.04
Babble	25	91.24	84.23	87.73
	15	72.22	87.21	79.71
	5	68.51	80.45	74.48
	-5	49.34	81.31	65.33
Pink	25	91.41	85.19	88.3
	15	92.66	72.67	82.66
	5	86.90	64.54	75.72
	-5	83.62	50.36	66.99
Factory	25	84.45	84.08	84.26
	15	85.87	77.13	81.5
	5	67.02	71.18	69.105
	-5	73.44	63.75	68.595
Volvo	25	77.14	83.68	80.41
	15	57.86	84.34	71.10
	5	50.41	89.22	69.81
	-5	42.98	96.24	69.61
Average		76.62	77.71	77.16

For better comparison, the experimental results that are mentioned in tables 1 to 3 are summarized in Fig. 9 and 10. Fig. 9 illustrates the silence hit rate (HR0) versus the speech miss detection rate in 5 different SNR levels (i.e. -5, 5, 15, 25 dBs as well as clean speech) on TIMIT and TPersianDat databases (averaged over noise type).

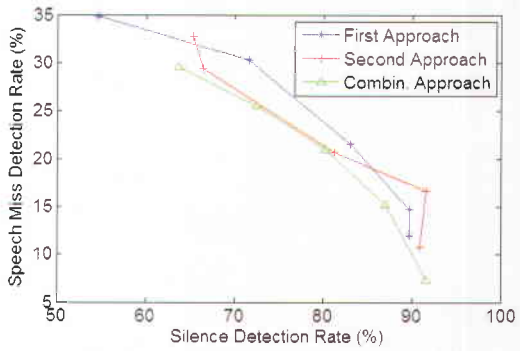

Fig. 9. Silence hit rate versus speech miss detection rate for the proposed approaches in 5 different SNR levels (i.e. -5, 5, 15, 25dBs as well as clean speech) on TIMIT and TPersianDat databases (averaged over noise type)

In Fig. 9, the left-most points show the performance measures for the lowest SNR levels. The points with the highest HR0 correspond to clean speech evaluations. Fig. 9 shows that the combinational approach is relatively more accurate than the other two methods. The performance measure of the proposed approaches is denoted by the average T measure for every SNR level for all noise types in Fig. 10. This figure shows the average accuracy of the described approaches in term of SNR levels. As can be seen, the relative accuracy of the combinational approach improves when SNR value decreases. This shows more robustness of this approach compared to other two approaches. Also, Fig. 10 shows the higher average VAD accuracy of the combinational method.

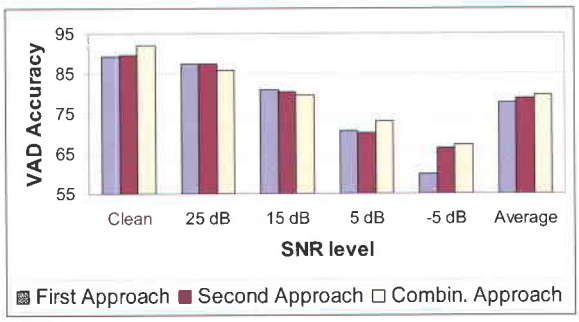

Fig. 10. Average accuracy of the proposed approaches on TIMIT and TPersianDat databases (averaged over noise types)

To demonstrate the efficiency of the proposed approaches especially the last one, the same evaluations are done for the MMSE approach proposed in [10] and the VAD algorithms of some common standard codecs including G.729, AMR (options 1 and 2) and AFE. Figs 11 and 12 illustrate the average accuracy of these approaches on TIMIT and TPersianDat databases for different SNR levels and noise types, respectively.

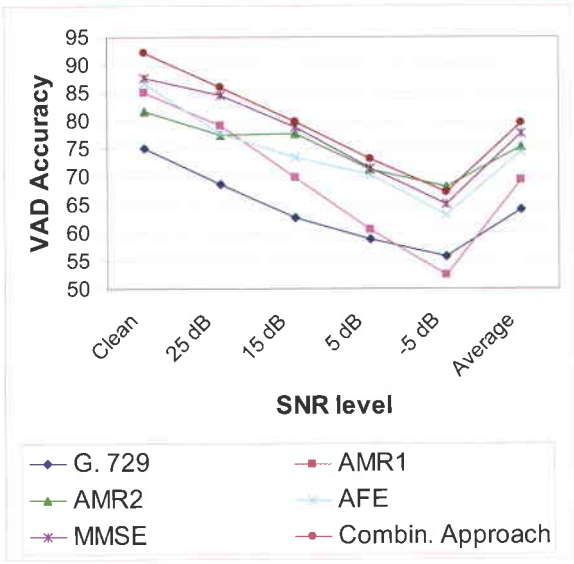

Fig 11. Average accuracy of the combinational approach and some other VAD approaches on TIMIT and TPersianDat databases (averaged over noise types)

Fig. 11 illustrates significant performance improvement of the proposed combinational approach compared to the VAD of G.729 and AMR option 1 codec. In addition, the approach relatively outperforms the other approaches almost in all SNR levels. Only in -5 dB SNR the average VAD accuracy of AMR option 2 is higher than the accuracy of the proposed approach. A similar performance comparison between the VAD approaches is illustrated in Fig 12 for various noise types.

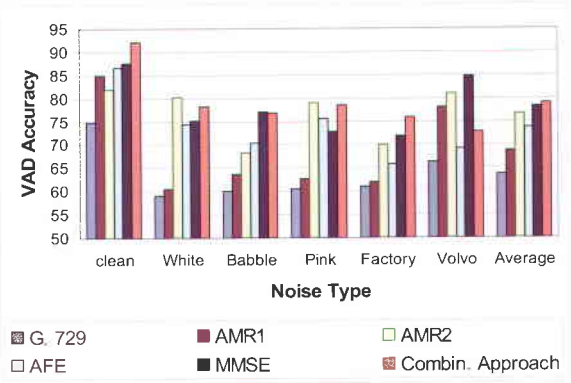

Fig 12. Average accuracy of the combinational approach and some other VAD approaches on TIMIT and TPersianDat databases (averaged over SNR levels)

Fig. 12 demonstrates better average performance of the proposed approach while some explanations on the results seem to be necessary. The MMSE approach has lower average performance compared to the proposed combinational method but has the best performance among all the methods in the presence of Volvo noise. Also, Fig. 12 shows that the average performance of the G.729 VAD is significantly low even in clean speech conditions. Experimental results show that the G.729 VAD standard is highly biased towards the HR1 measure which seems to be necessary for the VAD algorithms in speech codecs and therefore the average performance of this method degrades.

Since Aurora2 database has been widely used for evaluating VAD algorithms, the performance of the mentioned approaches on this database is summarized in Table 4 to obtain a better viewpoint on the performance of the proposed approaches.

Table 4. Summarized evaluation results for different approaches on Aurora2 database

Approach	Subway noise			Babble Noise			Car Noise		
	HR0	HR1	T	HR0	HR1	T	HR0	HR1	T
G.729B	16.7	92.6	58.1	20.61	98.2	59.42	30.27	98.4	64.3
AMR1	62.9	95.3	79.1	64.4	74.3	65.3	64.18	79.5	71.8
AMR2	60.4	96.1	78.2	46.5	93.5	70.1	75.04	85.07	80.0
AFE	57.4	94.3	75.9	58.9	92.8	75.8	67.27	93.76	80.5
MMSE Approach	97.3	47.4	75.4	89.54	61.2	75.4	98.22	44.25	71.2
First Approach	90.1	70.6	80.4	79.96	77.6	78.82	90.24	74.81	82.5
Second Approach	84.5	78.2	81.4	74.85	70.3	72.6	88.29	78.29	83.2
Combin. Approach	91.4	73.4	82.4	80.68	79.8	80.25	91.47	76.9	84.1

The above table shows considerable performance improvement for the combinational method compared to other evaluated methods in all noise conditions. The results above are also illustrated in Fig. 13. The integrated results show that all three proposed approaches have satisfying performance compared to the other previously proposed methods, but the performance of the combinational approach is slightly higher than the other two approaches and the parallelism scheme has compensated for the weaknesses of each of the methods used separately.

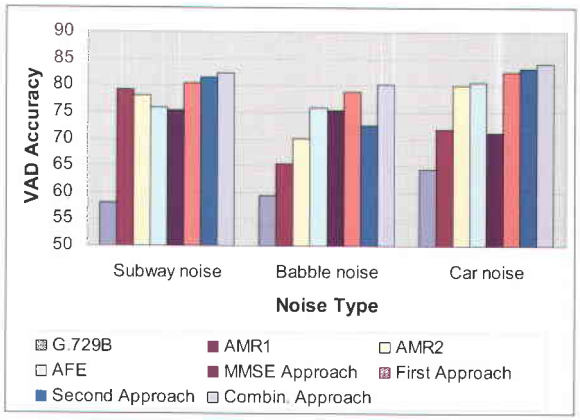

Fig 13. Average performance of different VAD algorithms when Aurora2 database is used

To compare the proposed approaches with the other VAD methods from the computational time point of view, the average run time of the mentioned VAD procedures for 1 second of input speech is measured on the evaluation databases. Table 5 shows the run time for three proposed methods and 5 reference approaches. These run times are achieved on an AMD Athlon Dual-Core processor with 2.7 GHz clock rate and 3 GB of RAM.

Table 5. Average run time of different VAD approaches for 1 second of input speech

VAD Approach	Average run time (seconds)
G. 729	0.030
AMR1	0.017
AMR2	0.036
AFE	0.030
MMSE	0.197
First Approach	0.012
Second Approach	0.027
Combin. Approach	0.031

Table 5 shows that the first approach is computationally the most efficient one, while the MMSE method has, as expected, the highest run time. It also demonstrates that the approaches that use the spectral patterns of vowel sounds for silence/non-silence discrimination (i.e., the second and combinational approaches), need a relatively high computing time. However the average run time of these approaches are still comparable to other reference methods. Therefore, we can conclude that the run time of the proposed methods, even the combinational approach, are near to that of reference methods such as AMR and AFE, while their accuracy is relatively higher.

In order to observe the strengths of the proposed methods in silence/non-silence detection, which is crucial in various tasks especially in audio indexing and retrieval, the proposed methods are also evaluated on a speech-music dataset, produced artificially from a set of speech and music files. To generate the evaluation dataset, the speech utterances of the TIMIT corpus were randomly concatenated with a set of music data extracted from a domestic music dataset with durations ranging from 5 to 30 seconds.

The experimental results of these evaluations with the same parameters as the previous experiments are illustrated in Fig. 14. These evaluations are only performed for White and Babble noises. To avoid multiple figures, only the music miss detection rate (MDR) is depicted in the following figure.

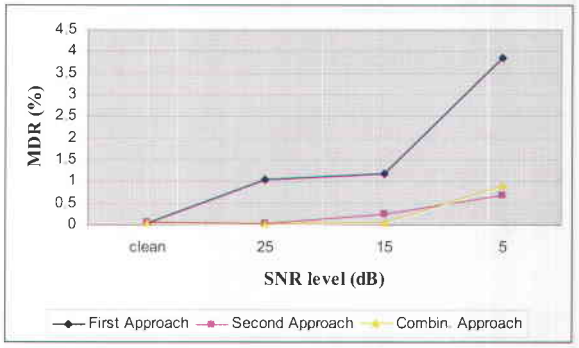

Fig 14. Music MDR of the proposed VAD approaches for music/silence discrimination in noisy environments

As can be seen in the above figure, the music MDR of the proposed approaches is inconsiderable and hence we can conclude that the proposed VAD approaches are relatively good in discriminating silence parts from music parts in different noise conditions. This performance is due to the fact that the proposed methods use the harmonic pattern of speech frames to discriminate silence from speech and the music parts of an audio signal are highly harmonic.

It is an inevitable fact that all the VAD algorithms are weak in detecting unvoiced parts of speech signals which is an inconvenience in applications such as speech recognition. Following the procedure of the proposed algorithms, especially those that use the spectral pattern of the vowel sounds, this idea may emerge that the proposed algorithms may be efficient in detecting voiced parts while they could be extremely weak in discriminating unvoiced parts from background noise. To answer the above question, we also evaluated the precision of the proposed approaches in discriminating silence from voiced and unvoiced speech. Fig. 15 depicts the results of these evaluations.

However, there is a trade-off between silence hit rate and speech hit rate, and the voiced/unvoiced detection performance is proportional to the speech hit rate. For example, if we simply apply the miss detection rate of voiced or unvoiced speech frames,

when no VAD detection is performed on the signal and all the speech frames are present, the highest performance will be achieved, but is not acceptable. To avoid this incoherence, we applied two performance measures including modified voiced hit rate (MVHR) and modified unvoiced hit rate (MUHR) which are the ratio of voiced/unvoiced hit rate to the speech hit rate and are formulated as follows:

$$MVHR = Voiced\ Hit\ Rate/HR1 \quad (6)$$

$$MUHR = Unvoiced\ Hit\ Rate/HR1 \quad (7)$$

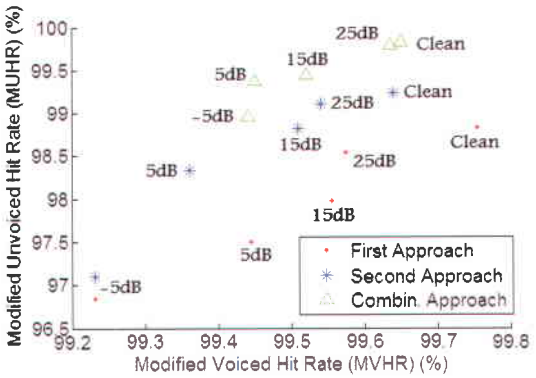

Fig 15. Performance of the proposed VAD approaches in detecting voiced and unvoiced parts of speech signal.

The above figure shows the improvement of the precision of the VAD algorithm when the combinational approach is applied. In addition, the above figure shows that the combinational approach has the best performance among the proposed methods in detecting unvoiced parts, while the detection accuracy for the voiced speech is maintained and the general performance of the VAD operation is improved.

V. CONCLUSIONS

In this paper we presented three structurally similar VAD approaches. Their similarity is in the way they use the discriminating features of the speech/non-speech content of a single frame and its neighboring frames. All three methods use a voting paradigm for decision on voice activity detection; while the feature set they use is different. Evaluations show that the proposed approaches have a satisfactory performance both in the sense of speech and non speech hit rates and improve the total performance of the operation. Also, the proposed approaches have a relatively low response time compared to the reference approaches. This makes them appropriate for real-time voice activity detection.

The most crucial factor on the prosperity of the proposed approaches is the appropriate selection of the decision thresholds. A well-selected set of thresholds will improve the performance of the VAD in almost all of the environmental conditions. In our implementations, the threshold values were selected automatically on a finite set of clean development speech signals to obtain the best performance. More intelligent threshold selection or update will also be helpful for improving the performance of the algorithms.

ACKNOWLEDGMENT

The authors would like to thank Iran Telecommunication Research Center (ITRC) for supporting this work under contract No. T/500/14939.

REFERENCES

[1] M. H. Savoji, "A robust algorithm for accurate end pointing of speech," Speech Communication, pp. 45–60, 1989.

[2] T. Kristjansson, S. Deligne and P. Olsen, "Voicing features for robust speech detection," Proc. Interspeech, pp. 369-372, 2005.

[3] R. E. Yantorno, K. L. Krishnamachari and J. M. Lovekin, "The spectral autocorrelation peak valley ratio (SAPVR) – A usable speech measure employed as a co-channel detection system," Proc. IEEE Int. Workshop Intell. Signal Process. 2001.

[4] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin and J. P. Petit, "ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," IEEE Communications Magazine, vol. 35, pp. 64-73, 1997.

[5] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," IEEE Trans. Speech Audio Process, 10, pp. 109-118, 2002.

[6] K. Li, N. S. Swamy and M. O. Ahmad, "An improved voice activity detection using higher order statistics," IEEE Trans. Speech Audio Process., vol. 13, pp. 965-974, 2005.

[7] W. H. Shin, "Speech/non-speech classification using multiple features for robust endpoint detection," In Proceeding of ICASSP, 2000.

[8] G. D. Wuand and C. T. Lin, "Word boundary detection with mel scale frequency bank in noisy environment," IEEE Trans. Speechand Audio Processing, 2000.

[9] A. Davis, S. Nordholm, and R. Togneri, "Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 2, pp. 412-424, 2006.

[10] B. Lee and M. Hasegawa-Johnson, "Minimum Mean Squared Error A Posteriori Estimation of High Variance Vehicular Noise," in Proc. Biennial on DSP for In-Vehicle and Mobile Systems, Istanbul, Turkey, June 2007.

[11] M. H. Moattar and M. M. Homayounpour, "A Simple but Efficient Real-Time Voice Activity Detection Algorithm," Eusipco 2009, Glasgow, Scotland, pp. 2549-2553, 2009.

[12] O. Izmirli, "Using a Spectral Flatness Based Feature for Audio Segmentation and Retrieval," (Abstract) Proc. of the International Symposium on Music Information Retrieval (ISMIR2000), Plymouth, Massachusetts, USA, October 23-25, 2000.

[13] A. B. Aicha and S. B. Jebara, "Perceptual musical noise reduction using critical bands tonality coefficients and masking thresholds", *INTERSPEECH*, pp. 822-825, 2007.

[14] J. H. L. Hansen, V. Radhakrishnan and K. Hoberg Arehart, "Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system," IEEE Trans. Audio, Speech, Language Process., vol. 14, pp. 2049–2063, Nov. 2006.

[15] G. Peeters, "A Large Set of Audio Features for Sound Description (Similarity and Classification) in The Cuidado Project," Cuidado Project Rep. Ircam, 2004.

[16] I.C. Yoo and D. Yook, "Robust Voice Activity Detection Using the Spectral Peaks of Vowel Sounds,"ETRI Journal, vol. 31, Number 4, pp. 451-453, August 2009.

[17] J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett and N. L. Dahlgren, The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM, Linguistic Data Consortium, 1993.

[18] ETSI EN 301 708, 1999. Digital cellular telecommunications systems (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels; General description (GSM 06.94 version 7.1.1 Release 1998). V7.1.1.

[19] ETSI ES 202 050, 2001. Speech processing, transmission and quality aspects (STQ). Distributed speech recognition. Advanced front-end feature extraction algorithm. Compression algorithms. V1.1.1.

Mohammad Hossein Moattar received the BSc degree in computer engineering from Azad University of Mashhad, Mashhad, Iran, in 2003, and the MSc degree in computer engineering from Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2005. He is now a PhD candidate with the Department of computer engineering and information technology, Amirkabir University of Technology, Tehran, Iran. His current research interests include speech and signal processing and audio indexing and retrieval.

Mohammad Mehdi Homayounpour was born in Shiraz, Iran, in 1960. He received the BSc degree in Electronics engineering from Amirkabir University of technology in 1986, and the MSc degree in Telecommunications from Khajeh Nasireddin Toosi University, Tehran, Iran, in 1989. He received his PhD in Electrical Engineering from University of Paris 11 (Orsay), Paris, France, in 1995. He is actually an Associate Professor of Computer Engineering and Information Technology Department of Amirkabir University of Technology, Iran, since 1995. His current research interests are natural language processing, speech and signal processing, and audio indexing. He is a member of Computer, Information and Telecommunication, and Cryptology Societies of Iran.

His email address is homayoun@aut.ac.ir and his link is http://www.aut.ac.ir/homayoun