Two-Microphone Binary Mask Speech Enhancement in Diffuse and Directional Noise Fields

Roohollah Abdipour, Ahmad Akbari, and Mohsen Rahmani

Two-microphone binary mask speech enhancement (2mBMSE) has been of particular interest in recent literature and has shown promising results. Current 2mBMSE systems rely on spatial cues of speech and noise sources. Although these cues are helpful for directional noise sources, they lose their efficiency in diffuse noise fields. We propose a new system that is effective in both directional and diffuse noise conditions. The system exploits two features. The first determines whether a given time–frequency (T–F) unit of the input spectrum is dominated by a diffuse or directional source. A diffuse signal is certainly a noise signal, but a directional signal could correspond to a noise or speech source. The second feature discriminates between T–F units dominated by speech or directional noise signals. Speech enhancement is performed using a binary mask, calculated based on the proposed features. In both directional and diffuse noise fields, the proposed system segregates speech T–F units with hit rates above 85%. It outperforms previous solutions in terms of signal-to-noise ratio and perceptual evaluation of speech quality improvement, especially in diffuse noise conditions.

Keywords: Two-microphone speech enhancement, source separation, binary mask, diffuse noise, directional noise.

Manuscript received Sept. 14, 2013; revised Mar. 29, 2014; accepted Apr. 9, 2014.

This work was supported by Iran Telecommunication Research Centre.

Roohollah Abdipour (r_abdipour@iust.ac.ir) and Ahmad Akbari (corresponding auther, akbari@iust.ac.ir) are with the School of Computer Engineering, Iran University of Science & Technology, Tehran, Iran.

Mohsen Rahmani (m-rahmani@araku.ac.ir) is with the Department of Computer Engineering Faculty of Engineering, Arak University, Arak, Iran.

I. Introduction

Speech enhancement systems remove the interfering noise signal from the input noisy signal(s) to improve speech quality or intelligibility. These systems are highly beneficial in voice-based applications, such as telecommunication, automatic speech recognition (ASR) and hearing aid devices lose performance in the presence of background noise.

Among existing speech enhancement approaches, binary mask (BM) methods have shown promising results [1]–[6]. These methods emulate the human ear's capability to mask a weaker signal by a stronger one [7]. This goal is achieved by eliminating spectral components in which the local energy of the speech signal is smaller than that of the noise. Such components do not contribute to the understanding of the underlying utterance and eliminating them will improve speech intelligibility for normal and hearing-impaired listeners ([3] and [8]), as well as the accuracy in ASR systems ([2], [6], and [9]).

BM solutions are broadly categorized into single- and two-microphone methods. Single-microphone methods rely on spectral cues for speech/noise discrimination. These cues include pitch continuity [5], harmonicity [6], a-priori SNR estimation ([1] and [10]), and long-term information about the spectral envelope ([4] and [11]). Due to the availability of only one signal, these methods cannot use spatial cues such as interaural time difference (*ITD*) and interaural level difference (*ILD*), which are highly useful in source separation ([5] and [12]–[16]).

On the other hand, two-microphone BM speech enhancement (2mBMSE) methods recruit localization cues along with spectral information to gain a better insight into

http://dx.doi.org/10.4218/etrij.14.0113.0917

acoustical situations. For example, [12], [13], and [16] find the location of peaks in a two-dimensional histogram of *ITD* and *ILD* features and associate each peak to a source. References [2] and [16] employ localization cues to train a classifier for separating sources with different directions of arrival (different *ITD*s). In [14], *ITD* is used to estimate the local signal-to-noise ratio (SNR) before exploiting it for speech segregation.

Most 2mBMSE methods rely on localization cues for speech segregation.[1] But, these cues are only useful when each sound source is located at a single point; hence, each signal will be arriving from a specific direction. Although this condition holds for speech and directional noise sources, in various environments the noise is diffuse and does not arrive from a specific direction (for example, consider restaurants). In these environments, traditional two-microphone BM methods lose their performance [17].

In this paper, we propose a 2mBMSE system with high performance in both directional and diffuse noise conditions. We employ two-channel features that discriminate between directional and diffuse noise environments, as well as separating speech and noise T–F units accordingly. The proposed system learns the rules of diffuse/directional source discriminations, as well as rules of speech/noise separation for each of these noise fields. The learned rules are then used to calculate a BM for denoising input signals..

In short, the contributions of this paper include: (a) incorporating new two-microphone features for BM calculation, (b) proposing a simple and effective algorithm for BM calculation based on the employed features, and (c) proposing a 2mBMSE system with acceptable performance in both directional and diffuse noise fields.

The detailed description of the proposed system is given in Section II. Then Section III details the experimental setup and the evaluation process that validates the performance of the system. Finally, the paper concludes with Section IV.

II. System Description

The proposed system is portrayed in Fig. 1. The input signal of microphone *i* can be written as

$$x_i(t) = s_i(t) + d_i(t) \quad \text{for} \quad i \in \{1,2\}, \tag{1}$$

where $s_i(t)$ and $d_i(t)$ denote, respectively, the speech and additive noise signals received at microphone *i*. By dividing this signal into overlapping frames, applying a window, and calculating its fast Fourier transform (FFT), the spectrum of this signal is obtained as

1) Other works employ supplementary cues (such as pitch period) in conjunction with localization cues; for example see [18] and [19].

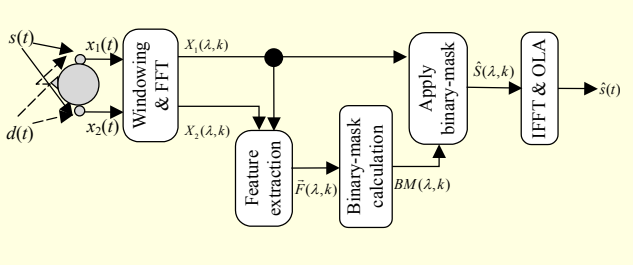

Fig. 1. Block diagram of proposed system.

$$X_i(\lambda,k) = S_i(\lambda,k) + D_i(\lambda,k) \quad \text{for} \quad i \in \{1,2\}, \tag{2}$$

where capital letters show the short-time Fourier transform (STFT) of their lowercase counterparts and λ and k represent frame and frequency bin indices, respectively. Based on the spectra of the input signals, the set of features $\vec{F}(\lambda,k)$ is extracted to calculate the binary mask as

$$BM(\lambda,k) = g(\vec{F}(\lambda,k)) = \begin{cases} 1 & \text{if } X_1(\lambda,k) \text{ is an SD T-F unit,} \\ 0 & \text{if } X_1(\lambda,k) \text{ is an ND T-F unit,} \end{cases} \tag{3}$$

where $g(.)$ is a function that assigns the values 1 and 0 to speech-dominated (SD) and noise-dominated (ND) units, respectively. By "SD units" we mean T–F units in which the power of speech is greater than that of the noise. In other words, the T–F unit $X_1(\lambda,k)$ is SD, if and only if $|S_1(\lambda,k)|^2 > |D_1(\lambda,k)|^2$. The ND units are defined similarly.

The BM is then applied to the spectrum of the reference signal (signal of microphone 1) to get the enhanced spectrum

$$\hat{S}(\lambda,k) = BM(\lambda,k) \times X_1(\lambda,k). \tag{4}$$

Finally, the enhanced signal is obtained using Inverse FFT (IFFT) and overlap-add (OLA) operations

$$\hat{s}(n) = \text{OLA}\{\text{IFFT}[\hat{S}(\lambda,k)]\}. \tag{5}$$

One of the challenges in 2mBMSE systems is which features to use. Existing 2mBMSE methods utilize localization cues such as *ITD* and *ILD* (for example, see [2], [5], [12]–[16], [20], and [21]). The assumption behind using these localization cues is that the speech and noise sources are positioned at fixed locations, and thus are emitted from specific directions of arrival. Although this assumption holds for environments with directional noise sources (such as car and street noise), it is not true in environments such as restaurants with diffuse noise signals. By "diffuse" we mean that the noise signal arrives from different directions with equal power. In these environments, the localization cues lose their meaning; hence, the performance of corresponding methods drops drastically. To have acceptable performance in both directional and diffuse

noise fields, we propose two new features to be used. These features and the motivations for using them are given in Section II-1.

Another challenge in 2mBMSE methods is to decide upon the filter calculation algorithm (the function $g(.)$). The filter calculation can be supervised or unsupervised. For example, [12]–[16], [20], and [21] work in an unsupervised manner by clustering T–F units based on their *ITD* and *ILD* values, and then assigning each cluster to a source. On the other hand, the methods of [2], [5], and [22] are supervised solutions that employ localization cues to train a classifier in advance. This is then utilized for mask calculation. In this paper, we adopt a supervised solution that learns a simple decision-making algorithm based on the proposed features. This algorithm is described in Section II-2.

1. Feature Extraction

We propose two features for BM calculation. These features are introduced in this section.

A. Coherence Feature

The "coherence" of the two spectra $X_1(\lambda,k)$ and $X_2(\lambda,k)$ is defined as [23]

$$COH(\lambda,k)=\frac{\left|P_{X_1X_2}(\lambda,k)\right|}{\sqrt{\left|P_{X_1}(\lambda,k)\right|\times\left|P_{X_2}(\lambda,k)\right|}}, \quad (6)$$

where $P_{X_i}(\lambda,k)$ is the smoothed spectrum of signal x_i, $i\in\{1,2\}$. This is calculated as

$$P_{X_i}(\lambda,k)=\alpha P_{X_i}(\lambda-1,k)+(1-\alpha)\left|X_i(\lambda,k)\right|^2. \quad (7)$$

The smoothed cross (power) spectral density (CPSD) of $X_1(\lambda,k)$ and $X_2(\lambda,k)$ is denoted by $P_{X_1X_2}(\lambda,k)$ and computed as

$$P_{X_1X_2}(\lambda,k)=\alpha P_{X_1X_2}(\lambda-1,k)+(1-\alpha)\,X_1(\lambda,k)\,X_2^*(\lambda,k). \quad (8)$$

In the above relations, α is the smoothing parameter (α=0.7 is used in our implementations) and $*$ denotes the conjugate transpose operation.

The coherence feature has been widely used for speech enhancement [23]–[27]. The coherence of two signals shows the level of correlation or similarity of two signals. For a directional source, the signals received at the two microphones are highly similar to each other (they only differ in their time of arrival and amplitude attenuation). So, their coherence is near one. But for a diffuse source, the received signals have lower similarity; hence, their coherence is noticeably smaller than one. This property is shown in Fig. 2. This figure depicts the coherence of two spectra for 256 sub-bands of a frame for directional and diffuse signals. The directional signal is a clean speech signal played at 30° angle. The diffuse signal is a two-microphone babble noise signal recorded in a crowded cafeteria [28]–[30]. The microphones were 180 mm away from each other. According to Figs. 3(a) and 3(b), it is observed that coherence takes different ranges of values for diffuse and directional sources. So, it is capable of determining whether a T–F unit is arriving from a directional or diffuse source.

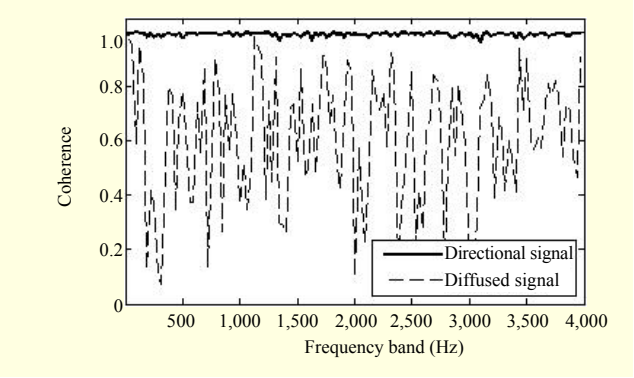

Fig. 2. Coherence values for 256 sub-bands of a frame for directional and diffuse signals.

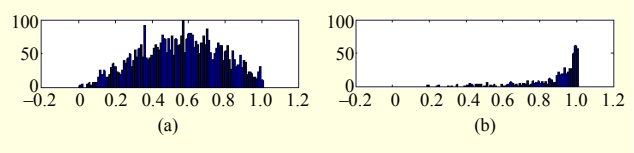

Fig. 3. Histogram of $COH(\lambda, k)$: (a) diffuse-dominated T–F units and (b) directional-dominated T–F units.

The above observation describes the behavior of the coherence feature when only a single source signal exists (that is, when each T–F unit of the spectrum comes from either the diffuse or directional source). We now consider situations where both diffuse and directional sources are active simultaneously. Examples of these situations are environments with diffuse noise and a single speaker (for example, someone in a restaurant talking on his mobile phone). In these situations, any T–F unit of the spectrum possibly contains components of both directional and diffuse signals. The coherence feature has the potential to determine whether a T–F unit is dominated by its diffuse or directional component. This property of the coherence feature, which has recently been pointed out in [31] and [32], can be observed in Fig. 3. Figures 3(a) and 3(b) depict, respectively, the histogram of the coherence feature for diffuse-dominated and directional-dominated T–F units in the sub-band centered at 2.5 kHz. The signals in this experiment are the same signals used in Fig. 2; however, the signals are played simultaneously. The two signals were mixed at 5 dB SNR level. Similar behavior of the coherence feature could be

observed for other sub-bands and SNR levels, and noise types. If a T–F unit is a diffuse-dominated, it is undoubtedly dominated by a noise source because anechoic speech signals cannot be diffuse (they always arrive from a single direction). So, if $COH(\lambda, k)$ is far from one, we can assign that T–F unit to a noise source. On the other hand, if $COH(\lambda, k)$ is near to one, the corresponding T–F unit is dominated by a directional source. This source could be a speech or directional noise source. To discriminate between these two directional sources, phase error (PE) is helpful.

B. PE

The PE of $X_1(\lambda,k)$ and $X_2(\lambda,k)$ is defined as [33]

$$PE(\lambda,k) = \Delta\varphi(\lambda,k) - 2\pi k \times ITD, \quad (9)$$

where $\Delta\varphi(\lambda,k) = \angle X_1(\lambda,k) - \angle X_2(\lambda,k)$ and ITD is the time-delay-of-arrival of signals $x_1(t)$ and $x_2(t)$. The $PE(\lambda,k)$ values are constrained to the interval $(-\pi, \pi]$.

This feature is used in several papers for speech enhancement (for example, see [29] and [33]). It is shown [33] that PE is near zero for a clean speech signal and its absolute value increases as SNR is decreased. This behavior is restricted to directional noise conditions because ITD makes no sense in diffuse environments; as a result, the PE estimation will be unreliable in these environments. The SNR-like behavior of the PE feature makes it possible to separate SD and ND T–F units in directional noise conditions. $PE(\lambda, k)$ is centered around zero for SD T–F units, and is far from zero (around $\pm\pi$) for ND T–F units. This property is shown in Fig. 4. In this figure, the histogram of $PE(\lambda, k)$ is drawn for SD and ND samples at a frequency band centered at 1 kHz. The noise and speech signals were played at +30° and –30° direction of arrivals, respectively. We used street noise in this experiment with overall 0 dB SNR. It is seen that the PE feature takes different values for SD and ND samples.

Finally, we include the frequency band index, k, to the feature set, because we expect the system to learn BM calculation rules for each sub-band separately. So, the final proposed feature set is as follows:

$$\vec{F}(\lambda,k) = \langle k,\ COH(\lambda,k),\ PE(\lambda,k) \rangle. \quad (10)$$

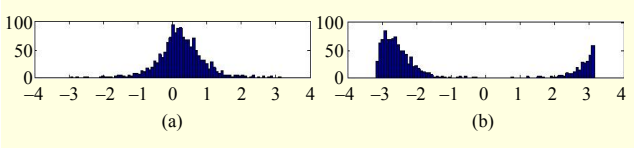

Fig. 4. Histogram of $PE(\lambda, k)$ at frequency band centered at 1 kHz: (a) speech-dominated T–F units and (b) noise-dominated T–F units.

2. BM Calculation

According to the characteristics of the coherence and PE features, a simple solution for BM calculation, which works in both diffuse and directional noise conditions, could be similar to the following algorithm:

```
if |COH(λ, k)| < δ(k)
  BM(λ, k) = 0;
else
    if |PE(λ, k)| < ε(k)
      BM(λ, k) = 1;
    else
      BM(λ, k) = 0;
```

where $0<\delta(k)<1$ is a threshold value on coherence for discriminating diffuse and directional sources in the kth sub-band and $0<\varepsilon(k)<\pi$ is a threshold value on PE for separating SD and ND T–F units in the kth sub-band in directional source conditions. If the coherence is noticeably smaller than one at the given T–F unit, that T–F unit is dominated by its diffuse component. So, the algorithm considers that T–F unit as ND and sets the corresponding BM cell to zero. But, if the coherence is near to one, that T–F unit is dominated by a directional component that could be either speech or noise. To distinguish between these two cases, the algorithm checks the value of $|PE(\lambda, k)|$. If this value is near zero, that T–F unit is considered as SD and the corresponding BM cell is set to one. Otherwise, that T–F unit is classified as an ND unit, and the related BM cell is set to zero.

Although the above algorithm seems to be simple, one should determine the threshold values $\delta(k)$ and $\varepsilon(k)$ for each sub-band. To avoid the exhaustive process of threshold tuning, we take a supervised approach. We train a classifier that learns the BM calculation rules from a train set. The train set contains samples of both SD and ND classes in directional and diffuse noise fields. This classifier learns the above algorithm for SD/ND separation. The classifier receives the feature set $\vec{F}(\lambda,k)$ as an input and generates outputs of zero and one for ND and SD classes, respectively. The performance of this classifier is reported in Section III-2 for different classifier types.

III. Evaluation and Comparison

To evaluate the proposed system, at first, we synthesized the train and test sets of SD and ND samples. Then these sets were used for training and testing the classifier. The trained classifier was subsequently utilized for BM calculation. The enhanced files were evaluated using objective measures. The details of the evaluation process and the corresponding results are described in the following subsections.

1. Dataset Description

We selected 120 clean files (60 male and 60 female) from the TIMIT database [34]. The files were downsampled from 16 kHz to 8 kHz. To make the two-microphone signals, we recruited the "image" method [35] with reverberation coefficient equal to zero. The speech source was placed in directions 30°, 75°, 120°, 165°, 210°, 255°, 300°, and 345° with respect to the perpendicular bisector of the connecting line of the two microphones. For each direction, the two signals received at the microphones were saved as the corpus of clean speech files. Similarly, to make the corpus of directional noise files, we placed a source of white noise in directions 10°, 55°, 100°, 145°, 190°, 235°, 280°, and 325° and saved the received signals. In addition, to make the corpus of diffuse noise files, we placed eight noise sources simultaneously at the above-mentioned directions and recorded the signals received at the two microphones. The signal of each source was randomly selected from a large noise file. Finally, to synthesize the corpuses of noisy files in directional and diffuse noise conditions, we mixed the utterances of the clean speech corpus and the files of directional and diffuse noise corpuses with –10 dB, –5 dB, 0 dB, 5 dB, 10 dB, and 15 dB SNR levels.[2] For each recording, we also saved the clean and noise components of mixture received at the reference microphone (that is, microphone 1).

Each pair of mixed noisy files $x_1(t)$ and $x_2(t)$ were divided into frames of 32 ms duration with 50% overlap. A Hanning window was applied to each frame, and its spectrum was calculated using 256-point FFT. Then the coherence and PE of each frequency bin were calculated. The *ITD* in (9) was estimated using the well-known GCC-PHAT method [36]. In addition, having the true noise and speech signals received at the reference microphone, the true local SNR of each T–F unit was determined as

$$SNR(\lambda,k) = 10\log_{10}\left(\frac{|S_1(\lambda,k)|^2}{|D_1(\lambda,k)|^2}\right). \quad (11)$$

Finally, T–F units with true local SNRs greater than and less than the threshold $Thr = 0$ dB were considered as SD and ND data samples, respectively.

The threshold value *Thr* affects the performance of the system. In [1], the effect of this value on the intelligibility of the enhanced signal is studied, and best intelligibility scores are achieved when the ideal binary mask (IdBM) is constructed with $-12 \text{ dB} \le Thr \le +12$ dB. So, the authors of [1] have proposed to use $Thr = -6$ dB for intelligibility improvement. This threshold value is also proposed in [8]. It is reported in [8] that an IdBM with $Thr = -6$ dB improves human speech recognition. Several other studies have also shown that a threshold value lower than 0 dB is suitable for both intelligibility and speech recognition (for example, see [37]–[39]), especially when the input SNR is as low as –5 dB. While the above works focus on intelligibility improvement purposes, our experiments on different values of *Thr* showed that, for the purpose of speech quality improvement, threshold values smaller than 0 dB are not promising and will result in a noticeable amount of annoying residual noise. On the other hand, an IdBM with $Thr = 0$ dB removes the interfering noise to a large extent, without introducing noticeable speech distortion and results in an enhanced signal of higher quality. It is also confirmed in [40] that $Thr = 0$ dB is suitable for SNR-gain purposes. For these reasons, we choose this threshold value in this contribution.

The above process was performed for both diffuse and directional noisy files, and the samples were saved separately as diffuse and directional datasets.

In addition, to study the performance of the system for different inter-microphone distances (IMDs), we performed the above process for IMDs of 180 mm, 66 mm, and 20 mm and saved the corresponding datasets separately. These IMDs correspond to the distance between pairs of microphones in a headset that we utilized for audio recording in real situations (more details are given in Section III-3). The 180 mm IMD corresponds to the average distance between a person's ears and is related to applications such as binaural hearing aids. The smaller IMDs (that is, 66 mm and 20 mm) are desired in applications like two-microphone mobile phones.

2. Classifier Training and Evaluation

The performance of the 2mBMSE system depends on the accuracy of the SD/ND classifier. If an ND T–F unit is misclassified as an SD, its noise component will remain in the enhanced signal and will be heard as annoying audio artifacts. On the other hand, misclassifying an SD T–F unit as an ND, causes that T–F unit to be removed from the enhanced spectrum, which means speech distortion will occur. To quantify these two classification errors, we measure the *hit* and false alarm (*FA*) rates of the classifier. The *hit* rate criterion measures the percentage of SD samples that are classified correctly. Higher *hit* rates mean that lower speech distortion will occur. The *FA* rate shows the percentage of ND samples that are misclassified as SD. The lower the value of *FA*, the lower the residual background noise.

We evaluated the classifier performance through four-fold

2) It is worth pointing out that the overall SNR of the input files of the train set does not have a high impact on the performance of the system (thus, there is no need to consider all possible overall SNR levels in the train set). This is because the system works at the T–F level, and even in a file with a specific overall SNR, there are different local SNRs at the T–F level. So, the classifier will see different possible local SNR levels.

ETRI Journal, Volume 36, Number 5, October 2014
http://dx.doi.org/10.4218/etrij.14.0113.0917

cross validation. In other words, we randomly divided the noisy files into four subsets. Each time, three subsets were jointly used to train the classifier. The remaining subset was saved as a test set and used to measure the *hit* and *FA* rates of the classifier. The process of classifier training was performed separately for each IMD. Then each classifier was evaluated utilizing either diffuse or directional samples. The average of the evaluation criteria is shown in Table 1 for the four classifier types — namely, neural networks (NN) (with two hidden layers with 10 neurons each), decision tree (DT) with C4.5 learning algorithm [41], Gaussian mixture model (GMM) with 16 mixtures, and support vector machine (SVM). We report the experimental results of the different classifier types to show that the achieved performance does not depend on the utilized classifier; rather, it is due to the proposed set of features.

According to Table 1, all the classifiers have consistently high *hit* rates for all IMDs. These results are comparable to other works, such as [3]. This behavior is observed for both diffuse and directional noise types. So, the noise reduction process will result in negligible speech distortion. It is also seen that the *FA* rate is small. Therefore, speech enhancement will be performed with a low amount of residual noise. The authors of [37] have argued that *FA* rates lower than 20% are needed for intelligibility improvement purposes. According to Table 1, this condition holds true for nearly all classifiers and IMDs.

Among the studied classifier types, the DT classifier obtains the highest *hit* rates. So, we only consider this classifier in the following evaluations. Moreover, for the sake of brevity, we only consider the 180 mm IMD in the following evaluations.

Table 1. Mean *hit* and *FA* rates in diffuse and directional conditions (%).

Classifier	IMD	Directional test set		Diffuse test set	
		Hit	*FA*	*Hit*	*FA*
NN	180 mm	86.42	19.35	85.91	18.94
	66 mm	85.81	20.61	85.41	19.74
	20 mm	84.94	20.74	84.07	20.49
DT	180 mm	85.68	18.10	84.68	17.44
	66 mm	85.29	18.23	84.53	18.35
	20 mm	85.37	18.80	84.36	19.01
GMM	180 mm	83.76	17.89	84.24	19.57
	66 mm	82.49	18.59	82.98	19.73
	20 mm	82.60	18.86	82.36	20.66
SVM	180 mm	84.34	17.18	84.58	18.21
	66 mm	84.82	18.03	83.91	19.94
	20 mm	84.09	18.30	83.39	19.72

Table 2. Average *hit* and *FA* rates for each input SNR level.

SNR	Directional test set		Diffuse test set	
	Hit	*FA*	*Hit*	*FA*
–8 dB	87.31	18.90	85.90	18.18
–3 dB	85.80	18.71	85.14	17.42
7 dB	84.95	17.86	83.91	16.94
12 dB	84.48	17.32	83.64	16.79

Table 3. Average *hit* and *FA* rates for different angles between speech and noise sources.

Angle	*Hit* (%)	*FA* (%)
0°	85.48	18.90
45°	84.70	16.87
90°	86.52	17.90
135°	83.79	17.72
180°	84.47	18.61

The results are consistent for other IMDs and classifier types.

We also evaluated the SD/ND classifier in each SNR level separately. The *hit* and *FA* rates of the classifier for different input SNR levels are shown in Table 2. We used the same clean and noise files, as well as the same experimental setup, as described above. We considered –8 dB, –3 dB, 2 dB, 7 dB, and 12 dB SNR levels in these experiments, which are not used in the training of the classifier. It is seen that the classifier performance does not depend on SNR level. The small differences between *hit* rates in Table 2 are consistent with results in [42].

We also evaluated the classification performance for different angles between speech and noise sources. We fixed the speech source at 10° and put the noise source at 10°, 55°, 100°, 145°, and 190° (that results in angles of 0°, 45°, 90°, 135°, and 180° between speech and noise). The overall SNR level was set to 0 dB. The classification performance for each angle is shown in Table 3. It is seen that the results do not depend on the angle between speech and noise sources. This is because, unlike many 2mBMSE methods, we do not employ localization cues in our system.

We also evaluated the system in echoic conditions. To do so, we employed the image method [35] to simulate a 10 m × 8 m × 3 m room with different reverberation coefficients. We used the same direction of arrivals for speech and noise sources, as described above. The speech and directional noise sources were 1 m and 3 m away from the microphones, respectively. The classification accuracy is shown in Table 4 for different

Table 4. Mean *hit* and *FA* rates for directional noise with reverberation.

Reverberation coefficient	Directional test set (SNR = 0 dB)	
	Hit (%)	*FA* (%)
0	86.52	17.90
0.2	79.11	18.29
0.4	73.92	18.38
0.6	67.08	18.36
0.8	60.33	18.43

Table 5. Average *hit* and *FA* rates for different diffuse-to-directional noise level ratios.

Diff./dir. ratio	*Hit* (%)	*FA* (%)
–10 dB	86.64	19.27
–5 dB	84.01	18.13
0 dB	83.92	17.64
5 dB	85.07	18.26
10 dB	85.93	18.90

reverberation coefficients (r). It is seen that *hit* rate decreases with r. This means that in highly reverberant situations, more speech segments are misclassified as noise.

Finally, we considered the situation where a mixture of diffuse and directional noises is present. We considered the same configuration as described in Section III-1 for the generation of the test set. We considered the 0 dB SNR level with no reverberation. The babble and car noises were employed as diffuse and directional noises, respectively. These noise signals were selected from our recordings in real situations ([28]–[30]). We considered different diffuse-to-directional level ratios and evaluated the *hit* and *FA* rates separately for each condition. The results are shown in Table 5. It is seen that the results do not change with diffuse-to-directional level ratio. This behavior is assigned to the simultaneous employment of coherence and PE features, which are useful in diffuse and directional noise conditions, respectively.

3. Speech Quality Evaluation

To evaluate the quality of the enhanced signals, we utilized the DT classifier trained in the previous section for 180 mm IMD. But, the input noisy files are selected from the dataset recorded by our lab members in real situations ([28]–[30]). This dataset was recorded using four omnidirectional microphones installed on a headset on a dummy head. Half of

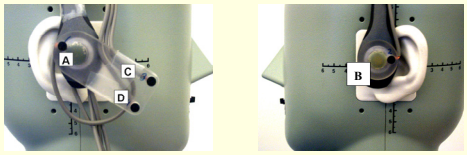

Fig. 5. Configuration of microphones (A to D) [27].

the recorded clean speech files were uttered by human speakers wearing the headset. Different pairs of microphones had 180 mm, 66 mm, and 20 mm distance between them. The configuration of the microphones is shown in Fig. 5. In our experiments, we used the signals recorded using microphones with 180 mm distance (that is, the microphones on the ears). The clean speech signal was played from a loudspeaker installed on the mouth of the dummy head. Speech and noise signals were recorded separately using the same configuration. Speech files were recorded in a quiet room. Car noise files were recorded in a Peugeot 405 with the speed around 80 km/h. Babble noise signals were recorded in a cafeteria. To make the noisy signal with a desired SNR level, the noise signal of each microphone was scaled and added to speech signals received at that microphone. In these experiments, we considered –8 dB, –3 dB, 2 dB, 7 dB, and 12 dB input SNR levels, which are not used in the training of the classifiers. More than 30 minutes of noisy signals were prepared for each SNR level and each IMD.

We used two objective evaluation criteria — namely, SNR improvement (SNRI) [43] and Perceptual Evaluation of Speech Quality (PESQ) measures [44]. SNRI determines the level of improvement of SNR in speech regions during a speech processing operation. The SNRI is computed by subtracting the SNR of the input signal from that of the output signal. The PESQ measure is a psychoacoustics-based measure that is correlated with subjective evaluation measures with correlation values around 0.8 [44]. The PESQ values range from –0.5 (for the worst case) to 4.5 (for the best case) [44]. The details of SNRI and PESQ calculation can be found in [43] and [44], respectively.

We compare our proposed method with a two-channel Wiener filter (2CWF), Rickard and others [22], Roman and others [5], MESSL [12], and Roman+Wiener methods. To implement the 2CWF method, the smoothed spectrum and CPSD of input signals were computed using (7) and (8), respectively. We employed the minimum-statistics method [45] to estimate the noise power of each input signal, which was used to calculate the CPSD of a noise signal similar to that in (8). The Roman+Wiener baseline is the serial application of Roman and others [5] and single-microphone Wiener methods. Such a serial system is considered as a baseline for removing directional noises (using the Roman and others method) as well

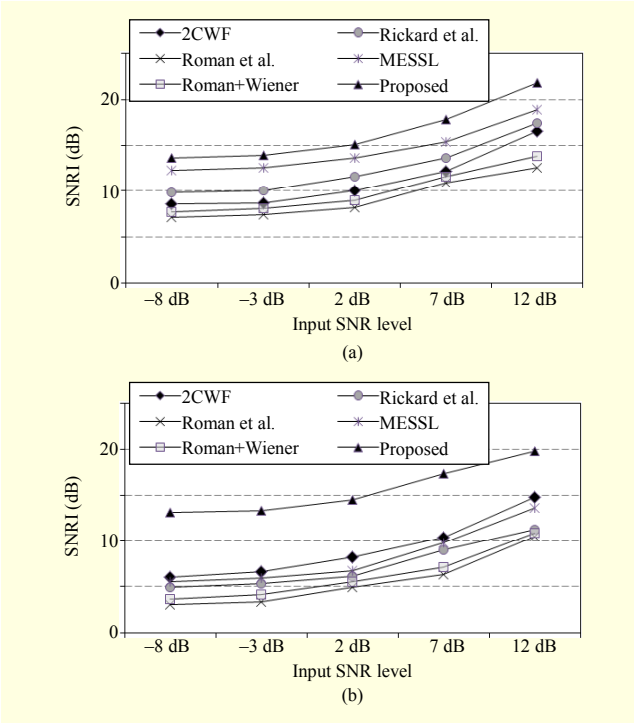

Fig. 6. SNRI results: (a) directional car noise and (b) diffuse babble noise.

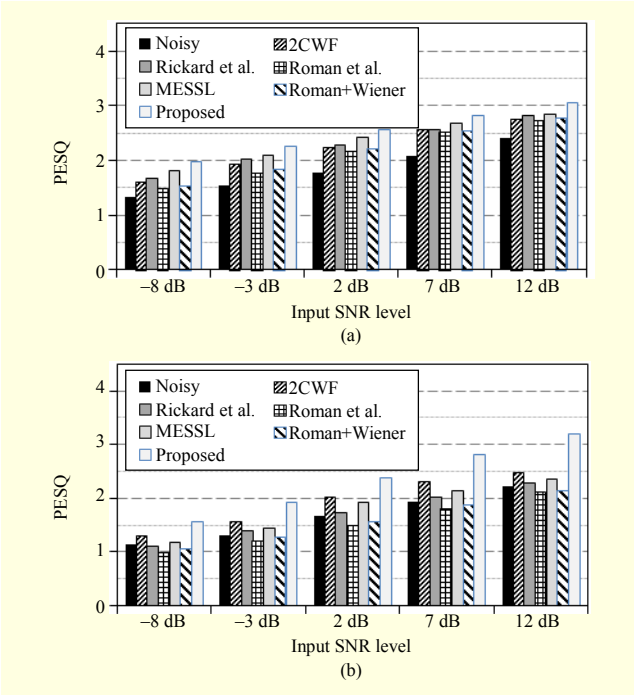

Fig. 7. PESQ results: (a) directional car noise and (b) diffuse babble noise.

as diffuse noises (using the Wiener filter). In the implementation of the Wiener filter, the noise power was estimated using the minimum-statistics method [45]. Roman and others' and Rickard and others' methods are selected for comparison because, similar to the proposed system, they are supervised 2mBMSE systems that rely on classification algorithms for BM calculation.

The noisy files were enhanced using the proposed method as well as other studied methods. The SNRI and PESQ values were calculated for each enhanced file. The average of these values was calculated for each enhancement method and SNR level. The results are shown in Figs. 6 and 7 for directional and diffuse noise types. According to Figs. 6 and 7, although

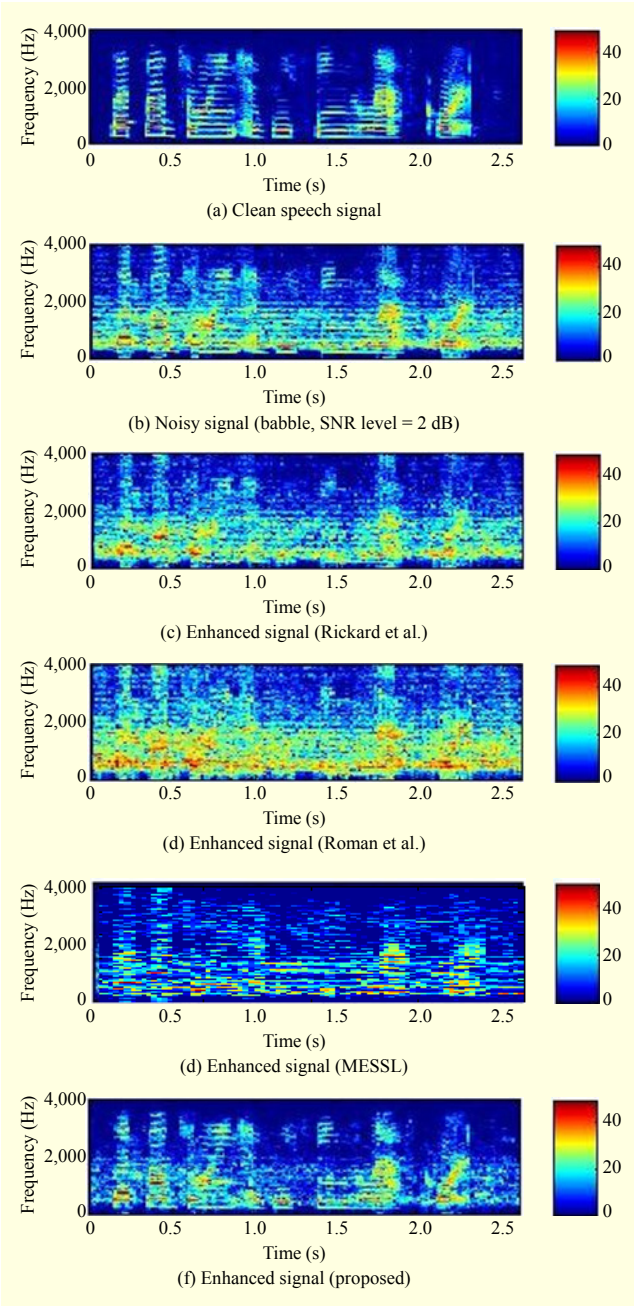

Fig. 8. Spectrograms comparison in diffuse noise condition.

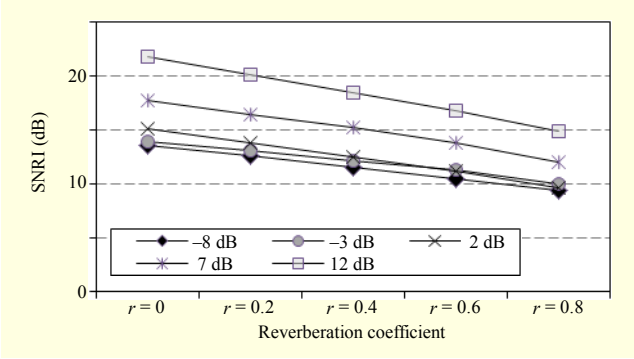

Fig. 9. SNRI results for directional noise in echoic conditions.

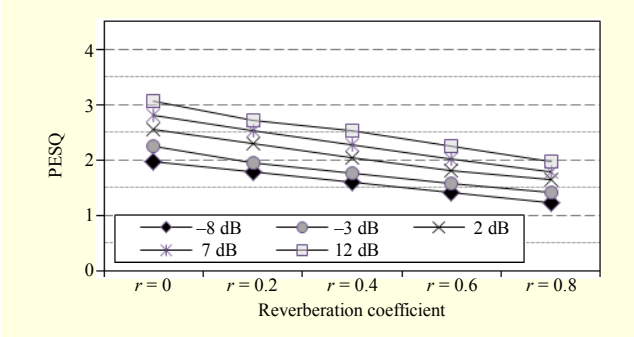

Fig. 10. PESQ scores for directional noise in echoic conditions.

competing methods show acceptable performance in the case of directional noise, their performance drops dramatically in diffuse noise fields. This fact is due to the usage of localization cues, which are not meaningful in diffuse noise conditions. But the proposed method results in acceptable qualities in both diffuse and directional noise conditions. This behavior is assigned to the proposed features for BM calculation.

To further compare our method with existing 2mBMSE methods in diffuse noise conditions, we compare the spectrograms of a file enhanced using the proposed method with that of one enhanced using the methods of Rickard and others, Roman and others, and MESSL (see Fig. 8). The clean file is selected from the NOIZEUS database [46]. The babble noise is selected from the corpus recorded in real conditions [28]–[30]. The clean and noise files are mixed at the 2 dB SNR level. Comparing the enhanced and noisy spectra, it is clearly seen that the proposed method outperforms other studied methods in noise removal as well as speech restoration in diffuse noise fields. To investigate the performance of the system in reverberant conditions, we conducted an experiment with the same setup as described in Section III-2. We set the reverberation coefficient (r) of the walls to 0, 0.2, 0.4, 0.6, and 0.8 in the image method [35] and evaluated the SNRI and PESQ scores of the system for different input SNR levels. The results are shown in Figs. 9 and 10. It is seen that the

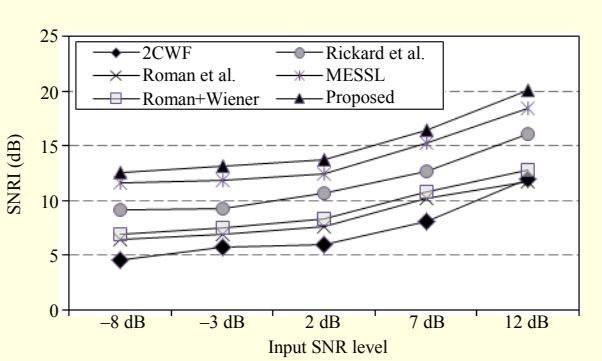

Fig. 11. SNRI results of studied methods in echoic conditions ($r = 0.2$).

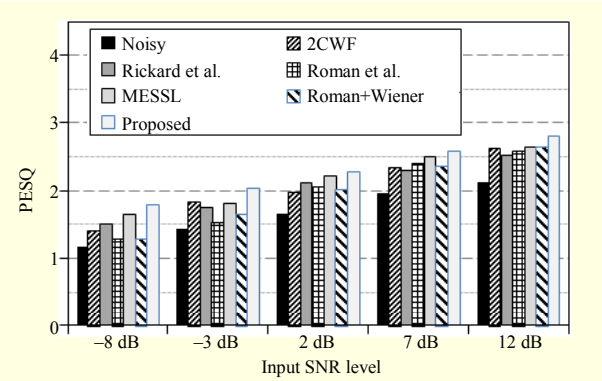

Fig. 12. PESQ scores of studied methods in echoic conditions ($r = 0.2$).

performance decreases as r increases. This is because in a highly reverberant environment, echoed signals make a semi-diffuse condition, which is considered as noise in the proposed algorithm. We also compared the performance of the proposed system with that of studied methods in conditions with moderate reverberation (r = 0.2). The results are shown in Figs. 11 and 12. Comparing with Figs. 6 and 7, it is observed that even though the performance of the proposed method is decreased, it is still comparable to competing methods.

IV. Summary and Conclusion

We proposed a 2mBMSE system that works effectively in both directional and diffuse noise fields. The proposed system was compared with existing 2mBMSE systems, and its superiority was confirmed in terms of SNR improvement and PESQ scores. The system owes its high performance to the two features it employs. We showed that the coherence feature has the potential to determine whether a T–F unit is dominated by a diffuse noise or directional signal. We also showed that the PE feature is capable of discriminating between SD and ND T–F units in directional noise situations. Using these features, the

system was able to build an effective binary mask for separating SD and ND units in both directional and diffuse noise fields.

It was shown that the performance of the system does not vary with the angle between speech and noise due to the usage of non-spatial cues. In highly reverberant conditions, SNR-gain decreased by 5 dB to 7 dB (analogous to one-level decrease of PESQ score). But, in moderate reverberation conditions, the PESQ decrease was only 0.2 and the proposed system outperformed the competing methods.

References

[1] D.S. Brungart et al. "Isolating the Energetic Component of Speech-on-Speech Masking with Ideal Time-Frequency Segregation," *J. Acoust. Soc. America*, vol. 120, no. 6, 2006, pp. 4007–4018.

[2] S. Harding, J. Barker, and G.J. Brown, "Mask Estimation for Missing Data Speech Recognition Based on Statistics of Binaural Interaction," *IEEE Trans. Audio Speech Language Proc.*, vol. 14, no. 1, Jan. 2006, pp. 58–67.

[3] G. Kim and P.C. Loizou, "Improving Speech Intelligibility in Noise Using a Binary Mask that is Based on Magnitude Spectrum Constraints," *IEEE Signal Proc. Lett.*, vol. 17, no. 12, Dec. 2010, pp. 1010–1013.

[4] G. Kim and P.C. Loizou, "Improving Speech Intelligibility in Noise Using Environment-Optimized Algorithms," *IEEE Trans. Audio Speech Language Proc.*, vol. 18, no. 8, Nov. 2010, pp. 2080–2090.

[5] N. Roman, D. Wang, and G.J. Brown, "A Classification-Based Cocktail Party Processor," *Neural Inf. Proc. Syst.*, 2003, pp. 1425–1432.

[6] M.L. Seltzer, B. Raj, and R.M. Stern, "A Bayesian Classifier for Spectrographic Mask Estimation for Missing Feature Speech Recognition," *Speech Commun.*, vol. 43, no. 4, Sept. 2004, pp. 379–393.

[7] B. Moore, *An Introduction to the Psychology of Hearing*, 5th ed., San Diego, CA, USA: Emerald Group Publishing Ltd, 2003, pp. 83–105.

[8] D. Wang et al., "Speech Intelligibility in Background Noise with Ideal Binary Time-Frequency Masking," *J. Acoust. Soc. America*, vol. 125, no. 4, 2009, pp. 2336–2347.

[9] S. Srinivasan, N. Roman, and D. Wang, "Binary and Ratio Time-Frequency Masks for Robust Speech Recognition," *Speech Commun.*, vol. 48, no. 11, Nov. 2006, pp. 1486–1501.

[10] Y. Hu and P.C. Loizou, "Techniques for Estimating the Ideal Binary Mask," *Int. Workshop Acoust. Echo Noise Contr.*, Seattle, WA, USA, 2008.

[11] Y. Hu and P.C. Loizou, "Environment-Specific Noise Suppression for Improved Speech Intelligibility by Cochlear Implant Users," *J. Acoust. Soc. America*, vol. 127, no. 6, 2010, pp. 3689–3695.

[12] M.I. Mandel, R.J. Weiss, and D. Ellis, "Model-Based Expectation-Maximization Source Separation and Localization," *IEEE Trans. Audio Speech Language Proc.*, vol. 18, no. 2, Feb. 2010, pp. 382–394.

[13] J. Nix and V. Hohmann, "Sound Source Localization in Real Sound Fields Based on Empirical Statistics of Interaural Parameters," *J. Acoust. Soc. America*, vol. 119, no. 1, 2006, pp. 463–479.

[14] E. Tessier and F. Berthommier, "Speech Enhancement and Segregation Based on the Localization Cue for Cocktail-Party Processing," *CRAC Workshop*, Alborg, Denmark, 2001.

[15] R.J. Weiss, M.I. Mandel, and D.P. Ellis, "Combining Localization Cues and Source Model Constraints for Binaural Source Separation," *Speech Commun.*, vol. 53, no. 5, 2011, pp. 606–621.

[16] O. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *IEEE Trans. Signal Proc.*, vol. 52, no. 7, July 2004, pp. 1830–1847.

[17] T. Lotter, C. Benien, and P. Vary, "Multichannel Direction-Independent Speech Enhancement Using Spectral Amplitude Estimation," *EURASIP J. Appl. Signal Proc.*, vol. 2003, no. 1, Jan. 2003, pp. 1147–1156.

[18] H. Christensen et al., "Integrating Pitch and Localization Cues at a Speech Fragment Level," *INTERSPEECH*, Antwerp, Belgium, Aug. 27–31, 2007.

[19] J. Woodruff and D.L. Wang, "Binaural Detection, Localization, and Segregation in Reverberant Environments Based on Joint Pitch and Azimuth Cues," *IEEE Trans. Audio Speech Language Proc.*, vol. 21, no. 4, Apr. 2013, pp. 806–815.

[20] S. Rennie et al., "Robust Variational Speech Separation Using Fewer Microphones than Speakers," *IEEE Int. Conf. Acoust. Speech Signal Proc.*, Hong Kong, China, vol. 1, 2003, pp. 88–91.

[21] K. Wilson, "Speech Source Separation by Combining Localization Cues with Mixture Models of Speech Spectra," *IEEE Int. Conf. Acoust. Speech Signal Proc.*, Honolulu, Hawaii, USA, vol. 1, Apr. 15–21, 2007, pp. 33–36.

[22] S. Rickard, R. Balan, and J. Rosca, "Real-Time Time-Frequency Based Blind Source Separation," *ICA*, San Diego, CA, USA, 2001.

[23] R. Le Bouquin and G. Faucon, "Using the Coherence Function for Noise Reduction," *IEE Proc. Commun. Speech Vis.*, vol. 139, no. 3, June 1992, pp. 276–280.

[24] D. Mahmoudi and A. Drygajlo, "Wavelet Transform Based Coherence Function for Multi-channel Speech Enhancement," *Euro. Signal Proc. Conf.*, Island of Rhodes, Greece, 1998.

[25] Q.H. Pham and P. Sovka, "A Family of Coherence-Based Multi-microphone Speech Enhancement Systems," *Radio Eng.*, vol. 12, no. 2, 2003, pp. 23–29.

[26] N. Yousefian and P.C. Loizou, "A Dual-Microphone Speech

Enhancement Algorithm Based on the Coherence Function," *IEEE Trans. Audio Speech Language Proc*., vol. 20, no. 2, Feb. 2012, pp. 599–609.

[27] B. Zamani, M. Rahmani, and A. Akbari, "Residual Noise Control for Coherence Based Dual Microphone Speech Enhancement," *Int. Conf. Comp. Elect. Eng*., Phuket, Thailand, Dec. 20–22, 2008, pp. 601–605.

[28] M. Rahmani, A. Akbari, and B. Ayad, "An Iterative Noise Cross-PSD Estimation for Two-Microphone Speech Enhancement," *Appl. Acoust*., vol. 70, no. 3, Mar. 2009, pp. 514–521.

[29] M. Rahmani et al., "Noise Cross PSD Estimation Using Phase Information in Diffuse Noise Field," *Signal Proc*., vol. 89, no. 5, May 2009, pp. 703–709.

[30] N. Yousefian, M. Rahmani, and A. Akbari, "Power Level Difference as a Criterion for Speech Enhancement," *ICASSP*, Taipei, Taiwan, Apr. 19–24, 2009, pp. 4653–4656.

[31] M. Jeub et al., "Blind Estimation of the Coherent-to-Diffuse Energy Ratio from Noisy Speech Signals," *EUSIPCO*, Barcelona, Spain, 2011.

[32] O. Thiergart, G. Del Galdo, and E.A. Habets, "On the Spatial Coherence in Mixed Sound Fields and its Application to Signal-to-Diffuse Ratio Estimation," *J. Acoust. Soc. America*, vol. 132, no. 4, 2012, pp. 2337–2346.

[33] P. Aarabi and S. Guangji, "Phase-Based Dual-Microphone Robust Speech Enhancement," *IEEE Trans. Syst. Man Cybern. Part B: Cybern*., vol. 34, no. 4, Aug. 2004, pp. 1763–1773.

[34] J.S. Garofolo et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *Linguistic Data Consortium*, 1993.

[35] J.B. Allen and D.A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. America*, vol. 65, no. 4, 1979, pp. 943–950.

[36] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. Acoust. Speech Signal Proc*., vol. 24, no. 4, Aug. 1976, pp. 320–327.

[37] N. Li and P.C. Loizou, "Factors Influencing Intelligibility of Ideal Binary-Masked Speech: Implications for Noise Reduction," *J. Acoust. Soc. America*, vol. 123, no. 3, 2008, pp. 1673–1682.

[38] U. Kjems et al., "Role of Mask Pattern in Intelligibility of Ideal Binary-Masked Noisy Speech," *J. Acoust. Soc. America*, vol. 126, no. 3, 2009, pp. 1415–1426.

[39] M.V. Segbroeck and H. Van Hamme, "Advances in Missing Feature Techniques for Robust Large-Vocabulary Continuous Speech Recognition," *IEEE Trans. Audio Speech Language Proc*., vol. 19, no. 1, Jan. 2011, pp. 123–137.

[40] Y. Li and D.L. Wang, "On the Optimality of Ideal Binary Time-Frequency Masks," *Speech Commun*., vol. 51, no. 3, Mar. 2009, pp. 230–239.

[41] J.R. Quinlan, *C4.5: Programs for Machine Learning*, 1st ed., San Francisco, CA, USA: Morgan Kaufmann, 1993.

[42] G. Kim et al., "An Algorithm that Improves Speech Intelligibility in Noise for Normal-Hearing Listeners," *J. Acoust. Soc. America*, vol. 126, no. 3, 2009, pp. 1486–1494.

[43] E. Paajanen and V.V. Mattila, "Improved Objective Measures for Characterization of Noise Suppression Algorithms," *IEEE Workshop Speech Coding*, Tsukuba, Japan, Oct. 2002, pp. 77–79.

[44] ITU-T Recommendation P.862, *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, 2001.

[45] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech Audio Proc*., vol. 9, no. 5, July 2001, pp. 504–512.

[46] Y. Hu and P.C. Loizou, "Subjective Comparison and Evaluation of Speech Enhancement Algorithms," *Speech Commun*., vol. 49, no. 7–8, 2007, pp. 588–601.

Roohollah Abdipour received his BSc and MSc degrees in computer engineering in 2002 and 2004, respectinely from the School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran and he now pursues his PhD degree. His research interests include audio and speech processing, especially speech enhancement.

Ahmad Akbari received PhD degrees in signal processing and telecommunications from the University of Rennes 1, Rennes, France, in 1995. In 1996, he joined the Computer Engineering Department, Iran University of Science and Technology, where he now works as an associate professor. His research interests include speech processing and network security.

Mohsen Rahmani received his PhD degrees in computer engineering from the Iran University of Science and Technology, Tehran, Iran, in 2008. In 2008, he joined the Engineering Department at Arak University, Arak, Iran, where he works as an assistant professor. His research interests include signal processing, especially speech enhancement.