IJICTR International Journal of Information & Communication Technology Research

Volume 6- Number 3-Summer 2014 (15-24)

Binaural Speech Separation Using Binary and Ratio Time-Frequency Masks

A. Mahmoodzadeh
Speech Processing Research Lab
Elec. and Comp. Eng. Dept.
Yazd University
Yazd, Iran
azar_mahmoodzadeh@yahoo.com

H. R. Abutalebi
Speech Processing Research Lab
Elec. and Comp. Eng. Dept.
Yazd University
Yazd, Iran
habutalebi@yazduni.ac.ir

H. Soltanian-Zadeh
Control and Intelligent Processing Center
of Excellence,
University of Tehran
Tehran, Iran
hzadeh@ut.ac.ir

H. Sheikhzadeh
Elec. Eng. Dept.
Amirkabir University of Technology
Tehran, Iran
hsheikhzadeh@aut.ac.ir

Received: November 7, 2013 - Accepted: June 5, 2014

Abstract— **In many speech applications, the target signal is corrupted by highly correlated noise sources. Separating desired speaker signals from the mixture is one of the most challenging research topics in speech signal processing. This paper proposes a binaural system combined with a monaural incoherent post processor for speech segregation. The proposed binaural system is based on spatial localization cues: Interaural Time Differences (ITD) and Interaural Intensity Differences (IID). A target speech is separated from interfering sounds by estimating time–frequency binary and ratio masks. The binary mask is estimated using the multi-level extension of the Otsu thresholding algorithm used in image segmentation. ITD and IID are important features for mask estimation in low and high frequencies, respectively. The ratio mask is estimated using the incoherent monaural speech separation system as the post processing stage. Systematic evaluations show that the proposed system can separate the target signal with acceptance quality.**

Keywords- Interaural intensity differences; interaural time differences; speech separation; time-frequency binary mask; ratio mask.

I. INTRODUCTION

Speech segregation under both monaural and binaural conditions is a challenging problem that has received considerable attention due to its potential application in hearing aid design, robust speech recognition or audio information retrieval. For us, filter out and comprehend a multitude of acoustic events that surround us in every moment. Imagine, for example, a cocktail party where we hear multiple voices, some background music and other environmental sounds at the same time. In this case, every acoustic source produces a vibration of the medium (i.e., air) and our hearing is determined by the

eardrums. As Helmholtz noted in 1863, the final waveform is "complicated beyond conception" [1]. Nonetheless, we are able to attend to and understand one particular talker in this situation. This perceptual ability is known as the "cocktail-party effect" – a term introduced by Cherry [2].

When target and intrusions are presented at different locations, the human auditory system is able to separate the sources from each other using two ears. This binaural advantage is because of the ability to utilize the interaural differences at the two ears. Monaural separation algorithms rely primarily on the pitch cue. On the other hand, the binaural algorithms use the source location cues, time and intensity differences between the ears.

Many Computational Auditory Scene Analysis (CASA) systems perform based on a Time-Frequency (T-F) mask. This mask selectively weights the T-F units in the acoustic mixture in order to enhance the desired signal. The weights can be binary or real [3]. The binary T-F masks are inspired by the masking system in human audition, in which a weaker signal is masked by a stronger one when they are presented in the same critical band [4].

In recent years, several binaural speech separations based on binary mask estimation using the ITD and IID features have been presented [5]-[9]. Roman has introduced a classification method based on supervised learning for binary mask estimation in order to separate the target signal from the interference [10]. The feature space of this classification is a two dimensional space based on IID and ITD local features. Srinivasan extended the Roman system by adding a ratio mask for modification of speech recognition [3].

In 2008, Cobos and López presented a method for binary mask estimation according to the Otsu thresholding algorithm [11], originally used in image processing for image segmentation. By applying this threshold to the IID feature and consequently binary mask estimation, some of the T-F units are multiplied by one while the rest are multiplied by zero. As a result, the target signal is separated from the interference signal.

It is generally accepted that for human audition, ITD is the main localization cue in low frequencies (< 1.5 kHz), whereas IID is used in the high-frequency range [12]. The resolution of the binaural cues affects both localization and recognition tasks.

Based on the above, the goal of this paper is to propose a binaural speech separation system that is able to extract a target speech signal from an acoustic mixture. By extending the binaural separation system proposed in [11], the proposed binaural speech separation system estimates the Time-Frequency (T-F) binary masks using binaural cues extracted from the responses of a KEMAR dummy head [13], which simulates the filtering process of the head, torso, and external ear. The novelty of the proposed binaural system is using the ITD cue for estimation of the T-F binary masks of low frequency subbands and the IID cue for high frequency subbands. The T-F binary mask is estimated based on the Degenerate Unmixing Estimation Technique (DUET), which is used for the separation of stereo anechoic mixtures [14]-[15]. The general approach in DUET based methods is to define a two-dimensional histogram constructed from the ratio of the T-F representations of the mixtures. A main assumption of this technique is that the signal T-F units of different sources do not overlap significantly in the T-F domain [16], that is called the W-disjoint orthogonality assumption. It should be mentioned that the proposed method has been partially presented in a preliminary form in [17].

In the proposed binaural system, first, the ITD and IID cues are calculated between the T-F units of the left and right channels and then, the positive and negative values of the ITD and IID are separated. The separated values are normalized and used for forming two weighted histograms, one for the positive values and other for the negative values. Next, based on the number of sources and using the Otsu thresholding algorithm [18], several thresholds are obtained from the weighted histograms. The range constrained by any two thresholds corresponds to a source and the T-F units corresponding to these ranges are used to estimate the T-F binary masks.

One of the limitations of the proposed binaural system is that most of the interference signals overlap the target speech signal, reducing the system performance. To overcome this limitation, we use a post processing stage to process the estimated target signal and improve its quality.

In this stage, a monaural speech separation system that we proposed in [19] is used. The estimated target speech signal obtained from the binaural speech separation system is entered to the monaural separation system as a new noisy signal. The simulation results using the PESQ evaluation index indicate that the proposed system extracts a majority of target speech without including much interference.

The rest of the paper is organized as follows: Section II describes the proposed system including the IID and ITD binaural cue extraction and T-F binary and ratio mask. Section III presents the evaluation results of the proposed system, and Section IV includes conclusion and discussion of the work.

II. Proposed Speech Segregation System

The main target of the proposed system is to produce the binary and ratio masks for separation of voiced and unvoiced parts of speech from interference. The proposed system consists of four stages: (1) auditory periphery, (2) binaural cues extraction, (3) estimation of a time-frequency binary mask and speech separation, and (4) post-processing for estimation of a ratio mask. Figure 1 shows the system architecture for the two sound sources.

In the proposed system, the inputs of the two microphones are two mixtures of target speech and interference signals measured at different, but fixed locations. As a standard method for binaural synthesis, the measurements of Head-Related Transfer Functions (HRTF) are employed. Here, we use a catalog of HRTF measurements, collected by Gardner and Martin from a KEMAR dummy head under anechoic

conditions [13]. The catalog consists of left/right KEMAR measurements for the source located at the distance of 1.4 m in the horizontal plane. Each measurement is actually a 128-point impulse response (at a sampling rate of 44.1 kHz). Binaural signals are achieved by convolving monaural signals with the impulse responses of HRTFs, with respect to the direction of incidence. The responses to multiple sources are added at each microphone. Due to the differential filtering effects between the two ears, HRTFs introduce a natural combination of ITD and IID into the signals to be extracted by subsequent stages of our system.

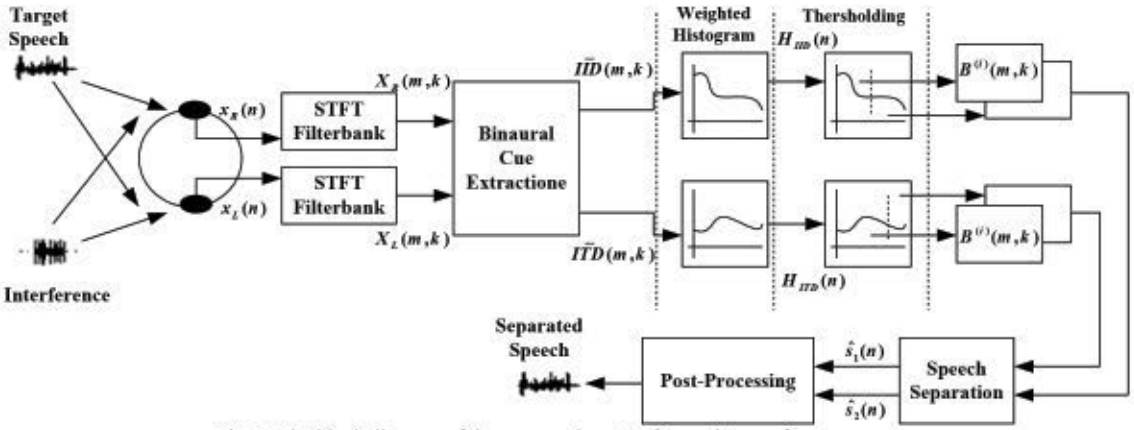

Fig. 1. The block diagram of the proposed system for a mixture of two sources.

A. *Auditory periphery*

Short Time Fourier Transform (STFT) is usually employed as a sparse representation of speech signals in T-F domain. In addition, it is used as a uniform filter-bank for decomposing a speech signal into narrowband subband signals. The location-dependent ITD and IID can be extracted independently in each T-F unit. These features are used for estimation of T-F binary masks; then these masks are applied to the mixtures for obtaining the separated sources in the T-F domain.

B. *Binaural cue extraction*

When speech and interference signals are orthogonal, the linear MMSE filter is the Wiener filter [20]. With a frame-based processing, the MMSE filter corresponds to the ratio of speech eigen values to the sum of eigen values of speech and noise [20]. Ephraim and Malah have shown that the optimal MMSE estimate of speech spectral amplitude in a local T-F unit is strongly related to the a priori SNR [21]. To estimate the speech in a local T-F unit, we approximate the frame-based filter with an ideal ratio mask defined using the a priori energy ratio $R(m,k)$:

$$R(m,k)=\left[\frac{|S(m,k)|^2}{|S(m,k)|^2+|N(m,k)|^2}\right] \quad (1)$$

where $S(m,k)$ and $N(m,k)$ are respectively the target and interference spectral values and m and k refers to time and frequency indices, respectively. $R(m,k)$ can be computed for each microphone signal before mixing the target and interference signals. According to Equation (1), the ideal binary mask is defined as [22]:

$$B(m,k)=\begin{cases}1 & \text{if } R(m,k)\geq\delta_3\\ 0 & \text{otherwise}\end{cases} \quad (2)$$

where δ_3 is set to be 0.5. Such masks can generate high-quality reconstruction for a variety of signals but, in practice, there is no access to target and interference signals. The objective of our pre-processing stage is to develop effective algorithm for estimating ideal binary mask.

Roman et al. [10] have shown that ITD and IID undergo systematic shifts as the energy ratio between the target and the interference changes. In a particular frequency channel, the ITD and IID corresponding to the target source exhibit location dependent characteristic values. As the SNR in this frequency channel decreases due to the presence of interference, the ITD and IID systematically shift away from the target values. The ITD and IID are computed independently in each T-F unit based on the spectral ratio at the left and right microphones:

$$IID(m,k)=20\log_{10}\left(\frac{|X_L(m,k)|}{|X_R(m,k)|}\right) \quad (3)$$

and

$$ITD(m,k)=-\frac{N}{2\pi k}A\left(\frac{X_L(m,k)}{X_R(m,k)}\right) \quad (4)$$

where $X_L(m,k)$ and $X_R(m,k)$ are the spectral values of the noisy speech signals respectively at the left and right ears, at frequency index k and time index m. N is the STFT length or equivalently the number of the filter bank channels. Also, $A(\cdot)$ is the phase angle function that returns the phase angle of a complex number in radians (i.e., $A(re^{j\phi})=\phi, -\pi<\phi\leq\pi$). The relative magnitude and consequently, the intensity difference between the spectral values of the left and right ears is calculated by IID. Also, ITD estimates the time difference between the signals of the left and the right ears via dividing the relative phase angle by the $2\pi k / N$.

To show the relationship between IID/ITD and the energy ratio R, the scatter plot and histogram of R in terms of IID/ITD have been shown in Figures 2 and 3, for a typical case where the target is in the middle of the horizontal plane and the interference is located on the right side at 30°. Here, we assume that the target and interference signals are separately accessible; thus, the value of R can be computed for each mixture signal. The scatter plot in Figure 2(a) shows the distribution of R with respect to IID for a frequency channel at 1kHz. Figure 2(b) depicts the histogram of IID samples. Figure 3(a) describes the variation of ITD and R for a frequency channel at 3.5 kHz and Figure 3(b) shows the histogram of ITD samples.

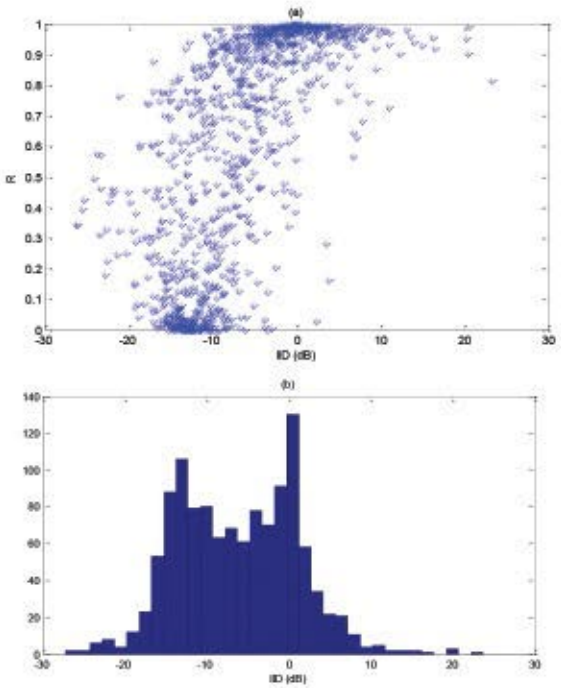

Fig. 2. Relationship between IID and the energy ratio R when the target is in the middle of the horizontal plane and the interference is on the right side at 30 degrees. (a) Scatter plot of the distribution of R over IID for a frequency channel at 3.5 kHz. (b) Histogram of IID values.

As seen in Figures 2 and 3, for the T-F units dominated by the target ($R \approx 1$), the binaural cues are clustered around the target IID/ITD values. Likewise, for the T-F units dominated by the interference ($R \approx 0$), the binaural cues are clustered around the interference IID/ITD values. Moreover, each peak in the histogram corresponds to a distinct active source.

C. *Time-frequency binary mask*

The aim of this stage is to estimate time-frequency binary masks for speech separation based on IID and ITD cues. It is well-known that IID is a salient cue at high frequencies while ITD becomes a more prominent cue at low frequencies [12]. This can be justified as follows.

Based on the duplex theory, Rayleigh [23] explained the human ability of localizing sounds by ITD and IID between the sounds reaching each ear. Woodworth [24] performed some experiments to test the duplex theory. He used a solid sphere in order to model the shape of the head and then, measured the ITDs as a function of azimuth for different frequencies. In the employed model, the distance between the two ears was approximately 22–23 cm. It was inferred from the initial measurements that there is a maximum time delay of approximately 660 μs when the source of the sound is located at directly 90° azimuth to one ear. This time delay is correlated with the wavelength of a sound source with a frequency of 1.5 kHz. Thus, Woodworth achieved to the important result that when a played sound had a frequency less than 1.5 kHz, the respective wavelength is greater than this maximum time delay between the ears (times the sound velocity). Consequently, a phase difference exists between the sound waves entering the ears, providing an appropriate acoustic localization cue. Hence, a sound input with a frequency closer to 1.5 kHz has the wavelength of the sound wave similar to the natural time delay. Thus, the size of the head and the distance between the ears results in a reduced phase difference; so localizations errors may be occurred. For a sound input with a frequency greater than 1.5 kHz, the wavelength is shorter than the distance between the two ears; hence, a head shadow is produced and the ITD will be no longer a reliable localization cue in such frequencies.

On the other side, it is well-known that when the sound is propagated in the air, it faces with an attenuation which is directly proportional to the sound frequency. So, because of higher attenuations in higher frequencies, IID is a more reliable cue at high frequencies.

Based on these principles, and by extending the binary mask estimation algorithm proposed by Cobos and López [11], we propose a method of mask estimation for high frequency channels considering the IID cue and for low frequency channels based on the ITD cue. The same procedure may be followed for low frequency channels based on the ITD cue.

Since the procedure of the proposed algorithms for IID and ITD features and for high and low frequency channels are similar, in the following, we only explain the algorithm for the IID (at high frequencies) and do not repeat the explanation for the ITD (at low frequencies).

At the first step, the IID values are divided into two parts corresponding to the sources located at left and right. For this purpose, two binary masks, one for the positive and another for the negative values of the IID are respectively defined as:

$$U_{IID}^{(1)}(m,k) = \begin{cases} 1 & if\ IID(m,k) \geq 0 \\ 0 & if\ IID(m,k) < 0 \end{cases} \quad (5)$$

$$U_{IID}^{(2)}(m,k) = \begin{cases} 1 & if\ IID(m,k) \leq 0 \\ 0 & if\ IID(m,k) > 0 \end{cases} \quad (6)$$

By multiplying by the above binary masks, the IID values are split into two parts:

$$IID^{(i)}(m,k) = IID(m,k) U_{IID}^{(i)}(m,k), \quad i \in \{1,2\} \quad (7)$$

The second step involves an estimation of the perceptually weighted histogram using the absolute value of $IID^{(i)}(m,k)$. To this end, first, the values of

$\left|IID^{(1)}(m,k)\right|$ and $\left|IID^{(2)}(m,k)\right|$ are normalized; this results in $\overline{IID}^{(1)}(m,k)$ and $\overline{IID}^{(2)}(m,k)$, respectively. Then, to count the points in the frequency range $[k_{\min},k_{\max}]$, two histograms of L uniform bins in the range of [0, 1] are formed for $\overline{IID}^{(i)}(m,k)$. The center of each bin is computed as:

$$z_n=\frac{1}{2L}(2n+1),\quad n=0,\ldots,L-1 \tag{8}$$

where n is the number of uniform bins. As a matter of fact, most of the speech energy is concentrated in the range of [100 Hz, 4 kHz]. Therefore, to facilitate

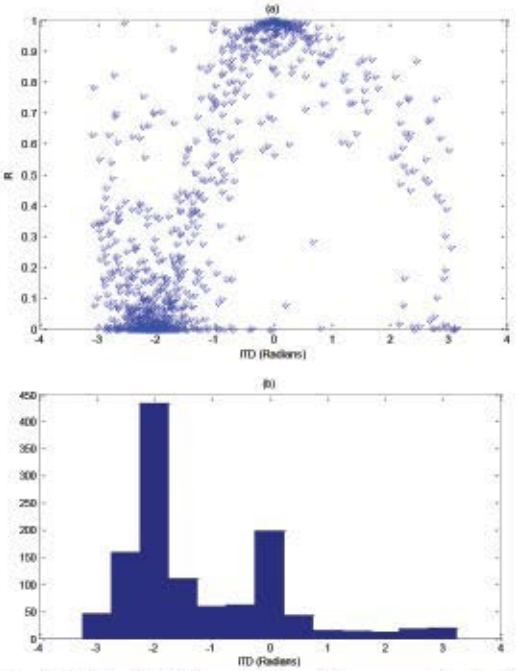

Fig. 3. Relationship between ITD and the energy ratio R when the target is in the middle of the horizontal plane and the interference is on the right side at 30 degrees. (a) Scatter plot of the distribution of R over ITD for a frequency channel at 1 kHz. (b) Histogram of ITD values.

the search for an optimal threshold for a perceptually weighted histogram, we take $k_{\min}$ and $k_{\max}$ as the index of the closest frequency to 100 Hz and 1.5 kHz for the IID cue histograms and 1.5 kHz and 4 kHz for the ITD cue histograms. Despite STFT (with linearly spaced frequency axis), the human auditory system has a logarithmic-like behavior on the frequency axis. Also, the energy of the sound (music, speech, and audio in general) is concentrated in low frequencies. Based on this observation and considering the human auditory system, we use the following weighting function proposed in Cobos and López system [11]:

$$g(k)=\frac{\log(100)}{\log(100+k-k_{\min})} \tag{9}$$

This function gives a greater weight to the points with lower frequencies. The weighted histogram is calculated as a summation of the weights of the T-F points that lie in each of the bins.

$$H^{(i)}(n)=\sum g(k_l),\quad i=1,2, \tag{10}$$

where $g(k_l)$ is the weight of a point (k_l,m_l), whose value is contained in the value range of bin n.

At the third stage, we need a threshold to assign each T-F point to a source using the weighted histogram. For this purpose, we use a thresholding algorithm, proposed by Otsu [18] for selecting a range of values in the histogram that correspond to a source. In the Otsu algorithm, the sources are extracted by maximizing their inter-class variance. Thresholding is used to segment targets from their backgrounds based on the distribution of the pixel intensities in the image. In our separation system, image segmentation and source separation are considered from the same point of view. The thresholds are found to maximize the inter-class variance of the distribution of mixing ratios in the T-F transform domain (see [11] for more details). The threshold values are those in the middle of bins $n=t_i^*$:

$$Th_i=z_n\Big|_{n=t_i^*} \tag{11}$$

In the next stage, the T-F binary masks corresponding to each class are defined using the optimal thresholds calculated for $\overline{IID}^{(i)}(m,k)$ in the previous stage. $Th^{(1)}$ and $Th^{(2)}$ are referred to as the optimal thresholds for the sources tend to the left and right ears, respectively. Note that we can search for an arbitrary number of classes in each channel, even if the number of sources is less than the number of classes. In such a case, a reassignment step for clustering several classes to the same source should be carried out. This will be discussed further in the next step. The binary masks for the sources that tend to the left are given by:

$$B_i^{(1)}(m,k)=\begin{cases}U^{(1)}(m,k) & \text{if } Th_{i-1}^{(1)}<\overline{IID}^{(1)}(m,k)\le Th_i^{(1)},\\ 0 & \text{elsewhere}\end{cases} \tag{12}$$

where $i=1,\ldots,M_1$, M_1 is the number of classes to be estimated in the histogram of the left microphone, $Th_0^{(1)}=0$, and $Th_{M_1}^{(1)}=1$. Similarly, for the right ear:

$$B_i^{(2)}(m,k)=\begin{cases}U^{(2)}(m,k) & \text{if } Th_{i-1}^{(2)}<\overline{IID}^{(2)}(m,k)\le Th_i^{(2)},\\ 0 & \text{elsewhere}\end{cases} \tag{13}$$

where $i=1,\ldots,M_2$, M_2 is the number of classes to be estimated in the histogram of the right microphone, $Th_0^{(2)}=0$ and $Th_{M_2}^{(2)}=1$.

As already stated, there is no restriction in the multilevel thresholding process for defining the number of classes M_i in $\overline{IID}^{(i)}$. This means that if the number of sources located at the left (or right) is less than the number of classes defined in the thresholding step, more than one mask may correspond to the same source. Thus, a reassignment process for the grouping of the class masks corresponding to the same source should be carried out. Regardless of the number of classes, when a source is panned to the center, there will be always two masks corresponding to that source: $B_1^{(1)}$ and $B_1^{(2)}$.

For constructing a unique mask for each source, the obtained masks form a set as: $B=\{B_1,B_2,B_3,B_4\}=\{U_2^{(1)},U_1^{(1)},U_1^{(2)},U_2^{(2)}\}$. Then, for comparing the binary masks, a $Z\times W$ grid is taken for each mask B_i and the number of non-zero points, m_n, is computed in each cell. This way, a vector for each mask $\mathbf{m}_i=[m_1 m_2 \ldots m_{Z\times W}]^T$ is formed for $(i=1,2,\ldots,M_1+M_2-1)$. Then, we calculate the mean distance between all the adjacent vectors $\mathbf{m}_i$ and $\mathbf{m}_{i+1}$:

$$d_{i,i+1}=\frac{1}{Z\times W}\sum_{n=1}^{Z\times W}\left|\mathbf{m}_i(n)-\mathbf{m}_{i+1}(n)\right| \qquad (14)$$

where $i=1,2,\ldots,M_1+M_2-1$. If $d_{i,i+1}$ is a local or absolute minimum of the whole distance sequence, then their corresponding masks are added: $B_i'=B_i\cup B_{i+1}$. After this reassignment step, a set of J' different masks are available for retrieving the original sources ($J'<M_1+M_2$).

Finally, the separated target signal in each channel is estimated by applying the calculated masks to amplitude and phase of noisy signal in T-F domain, $\hat{S}_{ij}(k,m)$. Therefore, the estimated signal is reconstructed in time domain as: $\hat{s}_j=STFT^{-1}\{S_{1j}+S_{2j}\}$, for $j=1,\ldots,J'$ and $i=1,2$.

D. Post-processing

As mentioned before, one of the main advantages of the proposed binaural separation system is the ability to separate the target speech signal from the interference in both voiced and unvoiced portions. Since two microphones record target and interference signals, another advantage of this system is the access to two mixture signals with different SNR's. This means that if two sources are not located at the same side, then one of the microphones records the interference with more energy whereas the other one has a better recording from target signals. On the contrary, in a monaural system, only one mixed signal is accessible.

The main limitation of the binaural method is that for the signals without W-disjoint orthogonality (i.e., those with major overlapping parts in the STFT domain), the system performance is considerably reduced. This limitation is most pronounced for the noisy mixture of speech signals with large overlaps and more than two sources.

To overcome this limitation, we use a post processing stage using the incoherent monaural speech separation system proposed in [19]. The incoherent speech separation system produces a ratio mask for single channel speech separation in the modulation spectrogram domain. Producing the ratio mask for speech separation needs determining the pitch range of target and interference speech signals. In each subband, the value of this mask depends on the obtained pitch range of target and interference in that subband. The pitch ranges of target and interference speakers are determined based on the modulation spectrogram of the speech signal and, then, a proper mask is calculated for speech separation.

An important feature for determining the pitch range is finding the distribution of modulation spectrogram energy in the modulation frequency domain. For this purpose, we use an onset and offset detection algorithm [25] for segmentation of modulation spectrogram energy (see Section 3.2.2 in [19] for details). The resulting segments are grouped in order to estimate the pitch range of each speaker. A detailed description of incoherent monaural speech separation system is described in [19].

III. Experimental Results

To evaluate the performance of the proposed system for binaural speech separation, the signals are taken from the TIMIT and Cooke's databases [26]-[27]. The interference signals are: N0) 1 kHz pure tone; N1) white noise; N2) noise bursts; N3) babble noise; N4) rock music; N5) siren; N6) trill telephone; N7) female speech; N8) male speech; and N9) female speech (taken from [27]). As shown in Table I, these interferences are classified into three categories: 1) those with no periodicity; 2) those with quasi-periodicity; and 3) speech utterances.

TABLE I
CATEGORY OF INTERFERENCE SIGNALS

Category 1	white noise, noise bursts
Category 2	1 kHz pure tone, babble noise, rock music, siren, trill telephone
Category 3	female and male speech signals

To evaluate the proposed system, we use Perceptual Evaluation of Speech Quality (PESQ), Weighted Spectral Slope (WSS) distance, and Log Likelihood Ratio (LLR) as objective measures that correlate well with subjective Mean Opinion Score (MOS) evaluations. The input signal is sampled at 16 kHz. The filterbank has 256 subbands with a prototype Hanning filter of 32 ms duration and a frame rate of 8 ms. The proposed binaural system is the extension of Cobos and López system [11] using the ITD and IID features for low and high frequency subbands and the post processing stage. Therefore, the performance of the proposed method is compared with that of the Cobos and López system.

Table II presents the performance of the three speech separation systems for the separated signal in terms of objective measures: PESQ; WSS; and LLR before and after enhancement. The results are averaged for the target signal separated from the mixture of a target speaker with interference signals N0-N9. In this part of evaluation, the number of sources is two (one target, one interference) which are respectively located at directions $+30^\circ$ and -30°.

The first and the second rows of Table II show PESQ, WSS, and LLR of noisy speech signal in the left and right microphones before the separation. The third row indicates performance of the binary masking ("binaural system"). The fourth row presents the

evaluation results of the proposed binaural system with post processing using the incoherent speech separation system ("proposed system"). Finally, the last row shows the performance of the Cobos and López system.

The results show that the use of ITD feature for low frequency subbands and the IID feature for high frequency subbands improves the performance of the proposed binaural separation system compared to the Cobos and López system, which only uses the IID feature for the separation of the target signal.

TABLE II
SPEECH SEPARATION RESULTS FOR DIFFERENT METHODS IN TERMS OF OBJECTIVE MEASURES LLR, WSS, AND PESQ FOR EACH INTRUSION (N0-N9) FOR TWO SOURCES, TARGET AT $+30^\circ$ AND INTERFERENCE AT -30°.

Interference signal		N0	N1	N2	N3	N4	N5	N6	N7	N8	N9
Left mixture	PESQ	2.09	0.89	0.43	0.76	1.26	0.31	0.70	1.01	0.96	1.05
	WSS	87.79	67.92	98.22	96.45	88.79	119.4	95.66	107.6	54.53	107.6
	LLR	0.38	5.07	0.94	1.24	2.47	1.11	2.15	1.69	1.93	1.03
Right mixture	PESQ	2.37	1.66	1.46	1.97	1.58	1.58	2.38	2.09	1.79	2.61
	WSS	40.04	29.67	5.19	35.29	38.74	59.91	43.13	43.62	23.61	17.33
	LLR	0.14	2.75	0.50	0.49	1.07	0.32	0.75	0.57	0.59	0.25
Binaural system	PESQ	3.75	3.01	2.43	3.14	3.29	3.49	3.87	3.11	2.67	3.30
	WSS	2.98	29.35	5.47	27.36	23.38	8.55	10.41	14.38	13.28	9.84
	LLR	0.04	1.05	0.11	0.19	0.25	0.28	0.12	0.13	0.31	0.12
Proposed system	PESQ	4.13	3.79	4.03	3.60	3.70	3.96	4.23	3.66	3.29	3.71
	WSS	1.44	13.38	3.14	18.40	12.15	0.37	6.12	7.56	9.91	6.44
	LLR	0.05	0.37	0.08	0.06	0.09	0.19	0.09	0.06	0.23	0.09
Cobos and López system	PESQ	3.64	2.89	2.95	2.64	2.79	3.31	3.64	2.77	2.15	3.15
	WSS	3.39	37.13	7.02	38.56	34.05	20.48	17.18	20.97	19.54	15.45
	LLR	0.04	1.37	0.15	0.33	0.41	0.48	0.18	0.26	0.46	0.24

The results also show that the use of monaural incoherent speech separation system (as the post processing stage) improves performance of the proposed system compared to the binaural and Cobos and López systems. In fact, the binaural separation system cannot separate those parts of target and interference signals that overlap in the T-F domain.

The results also show that the use of monaural incoherent speech separation system (as the post processing stage) improves performance of the proposed system compared to the binaural and Cobos and López systems. In fact, the binaural separation system cannot separate those parts of target and interference signals that overlap in the T-F domain.

In the next experiment, we evaluate the performance of the system in terms of the distance between target and interference sources (or the resolution of the separation system). To this end, the target and interference sources are placed in the fixed and symmetric locations with respect to the reference axis (at 0°) angle. The target signal is an utterance from a male speaker which is placed at the distance of 1.4 meters (according to KEMAR measurements [13]) in directions $+40^\circ$, $+30^\circ$, $+20^\circ$, $+10^\circ$, $+5^\circ$, $+4^\circ$, $+3^\circ$, $+2^\circ$ and $+1^\circ$. The interference signals are selected from each category listed in Table I. The interference source is assumed to be placed in directions -40°, -30°, -20°, -10°, -5°, -4°, -3°, -2° and -1°. Therefore, the resolution of the proposed system is evaluated at the angular differences of 80°, 60°, 40°, 20°, 10°, 8°, 6°, 4° and 2° for the target and interference sources.

Figures 4, 5, and 6 show the performance of the binaural and proposed speech separation systems in terms of PESQ, for different angular distances. For this evaluation, we have selected sample interference from each category listed in Table I. The selected interference signals are white noise, cocktail party, and male speaker, respectively.

According to the results shown in Figures 4 and 5, we can conclude that the performance of the binaural and proposed systems degrade for angular distance differences smaller than 6° and 10°, respectively. In addition, by comparing the results of Figures 4 and 5, it is seen that the proposed system is able to improve the performance of the binaural system for the cocktail party noise.

Figure 6 shows that in the case of male speaker interference, for angular distance differences smaller than 10° the performance of the binaural separation system is poor and PESQ values before and after applying the binaural separation system are almost the same. However, the proposed system has improved the quality of the separated speech signal even when the local distance difference is small.

As mentioned before, the performance of the binaural separation system is degraded when the number of sources is more than two. As the number of sources increases, the overlaps of the T-F units in the spectrum of noisy signals (in the left and right microphones) increase. Consequently, the performance of this system in separating the target signal from the interference drastically degrades.

In the next experiment, we evaluate the performance of the separation systems in the case of three sources (one target and two interferences) which are respectively located at the fixed directions 0°, $+30^\circ$ and -30°. Table III presents the performance of the three speech separation systems in terms of objective measures: PESQ; WSS; and LLR before and after enhancement. The results are averaged for the target signal separated from mixtures of a target speaker with interference signals N0-N9. The interferences located at the directions $+30^\circ$ and -30° are from the same type.

The results of applying the proposed binaural speech separation system show acceptable performance in all except for the N2 and N3 interferences. The results listed in Tables II and III show that increasing the number of interference sources degrades the performance of the binaural system.

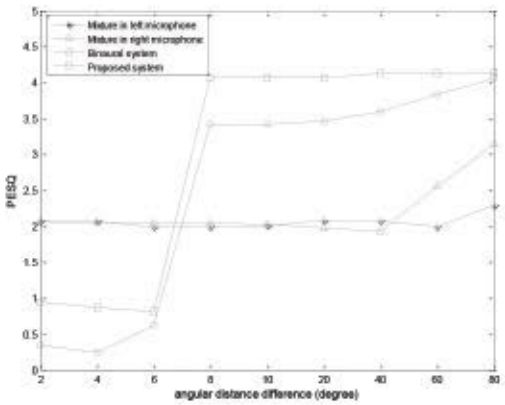

Fig. 4. Speech separation results in terms of PESQ versus angular distance differences for a mixture of the male target speaker and white noise.

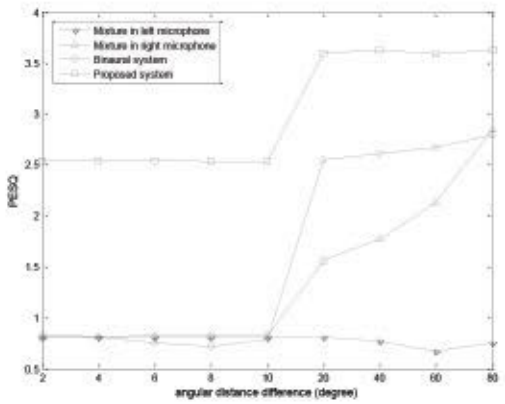

Fig. 5. Speech separation results in terms of PESQ versus angular distance differences for a mixture of the male target speaker and cocktail party.

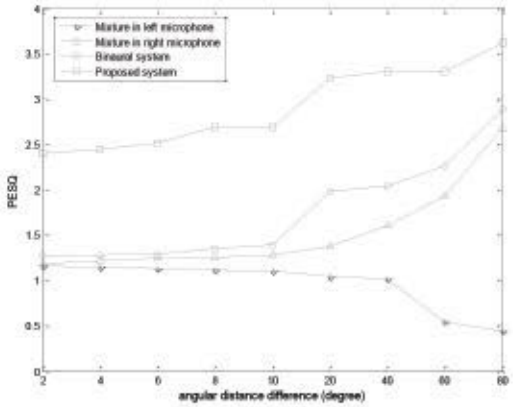

Fig. 6. Speech separation results in terms of PESQ versus angular distance differences for a mixture of the male target speaker and male speaker.

TABLE III

SPEECH SEPARATION RESULTS FOR DIFFERENT METHODS IN TERMS OF OBJECTIVE MEASURES LLR, WSS, AND PESQ FOR EACH INTRUSION (N0-N9) FOR THREE SOURCES, TARGET AT 0° AND INTERFERENCES AT $+30^{\circ}$ AND -30°.

Interference signal		N0	N1	N2	N3	N4	N5	N6	N7	N8	N9
Mixture signal	PESQ	1.85	1.32	0.53	0.74	1.27	0.69	1.28	1.20	0.97	1.34
	WSS	78.61	48.60	9.36	81.58	76.16	92.52	78.59	90.00	45.33	78.92
	LLR	1.15	4.48	0.89	1.10	2.12	2.23	1.95	1.32	1.45	0.76
Binaural system	PESQ	1.68	2.12	0.12	1.29	2.71	2.34	2.04	2.32	1.61	2.37
	WSS	61.40	62.12	71.04	84.98	31.09	26.47	37.26	48.33	51.03	49.95
	LLR	1.85	0.68	1.83	0.75	0.86	0.61	0.53	0.49	0.72	0.53
Proposed system	PESQ	2.31	3.31	0.53	1.86	3.52	3.38	3.29	2.95	2.46	2.90
	WSS	42.88	26.16	63.97	65.95	16.75	13.08	18.96	25.52	30.70	24.95
	LLR	0.11	0.07	0.99	0.04	0.31	0.06	0.10	0.21	0.49	0.25
Cobos and López system	PESQ	1.25	1.62	-0.31	0.89	2.16	1.60	2.00	2.07	0.82	2.13
	WSS	101.6	63.25	78.90	94.88	52.12	37.05	51.06	61.44	68.98	63.34
	LLR	3.52	1.97	3.50	1.20	1.10	1.05	0.71	0.69	1.05	0.66

According to the third row of Table III (the results of the proposed system in the case of one target and two interference sources), it is possible to conclude that the monaural separation system (the post processing stage) improves the performance of the binaural system. However, comparing these results with those in Table II reveals that this improvement is not significant for some interference signals.

IV. Discussions And Conclusions

The perceptual ability to detect, discriminate, and recognize one utterance in a background of acoustic interferences has been studied extensively under both monaural and binaural conditions. Our model is motivated by physiological and psychoacoustical results concerning the extraction of spatial features.

In this paper, we have proposed a binaural speech separation system based on T-F binary and ratio masks. The binaural speech separation system uses a DUET-like method that accurately estimates the T-F binary mask. The target speech signal is separated from interfering sounds using spatial localization cues IID and ITD. These cues estimate the binary mask for T-F points with low and high frequencies, respectively.

A major limitation of the binaural systems is the estimation of the binary mask for the T-F units of the sources that overlap in this domain. The overlap decreases the performance of the binaural system considerably. Therefore, we proposed a post processing stage to improve the performance of the proposed system in such cases. In this stage, a ratio mask is estimated using the monaural incoherent speech separation system. This ratio mask is applied on the signal estimated by the binaural system for separating the target signal from the residual interference signal. The ratio mask is constructed based on the pitch ranges of the target and interference signals in the monaural speech separation system. Using the onset and offset algorithm, the pitch ranges are estimated based on the distribution of the speaker energy in the modulation spectrogram domain.

The proposed system is able to separate both voiced and unvoiced parts of the target signal from the interference signal. Our extensive evaluations show that the perceived speech quality at the output of the proposed system is superior to those of the binaural and Cobos and López separation systems.

REFERENCES

[1] H. Helmholtz, On the Sensation of Tone (Ellis, A. J., Trans.), 2nd English ed., New York: Dover Publishers, 1863.

[2] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am., vol. 25, pp. 975-979, 1953.

[3] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," Speech Communication, vol. 48, pp. 1486-1501, 2006.

[4] B. C. J. Moore, An introduction to the Psychology of Hearing. 5th ed., San Diego, CA: Academic, 2003.

[5] R. M. Stern, DeL.Wang, and G. J. Brown, "Binaural sound localization," in Computational Auditory Scene Analysis: Principle, Algorithms and Applications, DeL. Wang and G. J. Brown, Eds. Wiley-IEEE Press, 2006.

[6] H. Park, and R. M. Stern, "Spatial separation of speech signals using amplitude estimation based on interaural comparisons of zero crossings," Speech Comm., vol. 51, no. 1, pp. 15–25, Jan. 2009.

[7] P. Arabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," IEEE Tran. Systems, Man, and Cybernetics-Part B:, vol. 34, no. 4, pp. 1763–1773, Aug. 2004.

[8] J. Woodruff and D.L. Wang, "Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues," IEEE Tran. Audio, Speech, Lang. Process., vol. 21, pp. 806-815, 2013.

[9] A. Alinaghi, W. Wang and Jackson, , "Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013.

[10] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," J. Acoust. Soc. Am., vol. 114, no. 4, pp. 2236-2252, 2003.

[11] M. Cobos, and J. J. López, "Stereo audio source separation based on time–frequency masking and multilevel thresholding," Digital Signal Processing, vol. 18, pp. 960-976, 2008.

[12] J. Blauert, "Spatial Hearing—The Psychophysics of Human Sound Localization," MIT Press, Cambridge, MA, 1997.

[13] W.G. Gardner, K. D. Martin, ''HRTF measurements of a KEMAR dummy-head microphone,'' Technical Report #280, MIT Media Lab Perceptual Computing Group, 1994.

[14] A. Jourjine, S. Richard, O. Yilmaz, ''Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures,'' in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '00, vol. 5, pp. 2985–2988, Turkey, 2000.

[15] O. Yilmaz, and S. Rickard, ''Blind separation of speech mixtures via time–frequency masking, '' IEEE Trans. Signal Process., vol. 52 (7), pp. 1830–1847, 2004.

[16] P. O'Grady, B. Pearlmutter, S. Rickard, ''Survey of sparse and non-sparse methods in source separation, '' Int. J. Imaging Systems Technol, vol. 15, no. 1, pp. 18-33, 2005.

[17] A. Mahmoodzadeh, H. R. Abutalebi, H. Soltanian-Zadeh, and H. Sheikhzadeh, "Binaural Speech Separation Based on the Time-Frequency Binary Mask," in Proc. IST, 2012, pp. 329-332.

[18] N. Otsu, ''A threshold selection method from gray-level histogram, '' IEEE Trans. System Man Cybernet. SMC-9 (1), pp. 62–66, 1979.

[19] A. Mahmoodzadeh, H. R. Abutalebi, H. Soltanian-Zadeh, and H. Sheikhzadeh, "Single channel speech separation in modulation frequency domain based on a novel pitch range estimation method," EURASIP J. on Advances in Signal Process., vol. 67, 2012, doi:10.1186/1687-6180-2012-67.

[20] H. L. van Trees, Detection, Estimation, and Modulation Theory, Part I. Wiley, New York, NY, 1968.

[21] Y. Ephraim, and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans. Acoust. Speech, and Signal Process., vol. 32, pp. 1109-1121, 1984.

[22] Y. Li, and D. L. Wang, "On the optimality of ideal binary time–frequency masks," Speech Comm., vol. 51, pp. 230–239, 2009.

[23] L. Rayleigh, "On the dynamical theory of gratings," in Proc. R. Soc. Lond. A. 1907, pp. 399-416, doi:10.1098/rspa.1907.0051.

[24] R. S. Woodworth, Experimental Psychology. New York: Holt, Rinehart, Winston, 1938.

[25] G. Hu, and D. L. Wang, "Auditory segmentation based on onset and offset analysis," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 2, pp. 396–405, 2007.

[26] J. S. Garofolo, L. F. Lamel,W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, 1993, DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus. [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[27] M. P. Cooke, Modeling auditory processing and organization. Cambridge, U.K: Cambridge Univ. Press, 1993.

Azar Mahmoodzadeh received the B.Sc, M.Sc. and Ph.D. degrees in electrical engineering from Shiraz University (2004), Shahed University (2007) and Yazd University (2012), respectively. She is now a faculty member in the Electrical Engineering Dept., Fars Science and Research Branch, Islamic Azad University. Her research interests include Speech and Image Processing.

Hamid Reza Abutalebi received the B.Sc. and M.Sc. degrees from Sharif University of Technology, Tehran, Iran in 1996 and 1998, respectively and the Ph.D. degree from Amirkabir University of Technology, Tehran, Iran in 2003, all in electrical engineering (signal processing). Also, he was with University of Waterloo, Ontario, Canada as a visiting scholar during Mar. 2002- Feb. 2003. Since 2003, Dr. Abutalebi has been with Electrical and Computer Engineering Department of Yazd University, Yazd, Iran, where he is now an associate professor. During fall 2010-summer 2011, he was on sabbatical at Idiap Research Institute, Martigny, Switzerland. His research interests are microphone arrays, speech enhancement, sound source localization, and time-frequency analysis.

Hamid Soltanian-Zadeh received the B.Sc. and M.Sc. degrees in electrical engineering(electronics) from the University of Tehran, Tehran, Iran in 1986 and M.Sc. and Ph.D. degrees in electrical engineering(systems and bio-electrical sciences) from the University of Michigan, Ann Arbor, Michigan, USA, in 1992. He is currently a full Professor and a founder of Control & Intelligent Processing Center of Excellence (CIPCE) in the Department of Electrical and Computer Engineering at the University of Tehran, Tehran, Iran. Prof. Solatnian-zadehhas coauthored more than 700 publications in journals and conference records. His research interests include signal and image processing, medical imaging and analysis, pattern recognition, and neural networks. He has served on program committees of several scientific conferences and review boards of more than 40 peer-reviewed journals.

Hamid Sheikhzadeh received the B.Sc. and M.Sc. degrees in electrical engineering from Amirkabir University of Technology, Tehran, Iran, in 1986 and 1989, respectively, and the Ph.D. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 1994. He was a faculty member in the Electrical Engineering Department, Amirkabir University of Technology, until September 2000. From 2000 to 2008, he was a Principle Researcher with ON semiconductor, Waterloo, ON, Canada. During this period, he developed signal processing algorithms for ultra-low-power and implantable devices leading to many international patents. Currently, he is a faculty member in the Electrical Engineering Department of Amirkabir University of Technology. His research interests include signal processing, machine learning, biomedical signal processing and speech processing, with particular emphasis on speech recognition, speech enhancement, auditory modeling, adaptive signal processing, subband-based approaches, and algorithms for low-power DSP and implantable devices.