Bashirpour and Geravanchizadeh *EURASIP Journal on Audio, Speech, and Music Processing* (2018) 2018:9
https://doi.org/10.1186/s13636-018-0133-9

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Robust emotional speech recognition based on binaural model and emotional auditory mask in noisy environments

Meysam Bashirpour and Masoud Geravanchizadeh*

Abstract

The performance of automatic speech recognition systems degrades in the presence of emotional states and in adverse environments (e.g., noisy conditions). This greatly limits the deployment of speech recognition application in realistic environments. Previous studies in the emotion-affected speech recognition field focus on improving emotional speech recognition using clean speech data recorded in a quiet environment (i.e., controlled studio settings). The goal of this research is to increase the robustness of speech recognition systems for emotional speech in noisy conditions. The proposed binaural emotional speech recognition system is based on the analysis of binaural input signal and an estimated emotional auditory mask corresponding to the recognized emotion. Whereas the binaural signal analyzer has the task of segregating speech from noise and constructing speech mask in a noisy environment, the estimated emotional mask identifies and removes the most emotionally affected spectro-temporal regions of the segregated target speech. In other words, our proposed system combines the two estimated masks (binary mask and emotion-specific mask) of noise and emotion, as a way to decrease the word error rate for noisy emotional speech. The performance of the proposed binaural system is evaluated in clean neutral train/noisy emotional test scenarios for different noise types, signal-to-noise ratios, and spatial configurations of sources. Speech utterances of the Persian emotional speech database are used for the experimental purposes. Simulation results show that the proposed system achieves higher performance, as compared with automatic speech recognition systems chosen as baseline trained with neutral utterances.

Keywords: Emotional speech recognition, Binaural model, Emotional auditory mask, Classification of emotional states, Kaldi speech recognition system, Noise robustness

1 Introduction

Speech is the most convenient means of communication for humans. In the last years, scientific and technical improvements in speech technology have resulted in a more natural human-machine speech interaction and natural language processing systems.

Despite all the recent advances in speech technology, often these systems struggle with issues caused by speech variabilities. These variabilities can occur due to speaker-dependent characteristics (e.g., shape of the vocal tract, age, gender, and emotional states), environmental noise, channel distortion, speaking rate, and accent variabilities [1, 2].

The automatic speech recognition (ASR) systems have been employed in many applications (e.g., voice-controlled personal computers) in the last decades. However, noise and speech variabilities such as emotions degrade the performance of the ASR systems, and this greatly limits the deployment of the systems in realistic situations [3, 4].

Characterization of the effect of emotional expression on speech, together with related techniques to improve the performance of speech processing systems, is a major research topic [1, 2, 5]. The emotional states such as anger, happiness, fear, sadness, and disgust affect the speech production by introducing changes in speech loudness, muscle tension, breathing rate, etc., and these in turn modify the factors such as glottal waveform, intensity, speech quality, prosody, and timing. Although most of the research has been focused on the recognition of speech

* Correspondence: geravanchizadeh@tabrizu.ac.ir
Faculty of Electrical & Computer Engineering, University of Tabriz, Tabriz 51666-15813, Iran

emotions, a limited work has been performed in the area of emotion affected speech recognition (EASR) [6–8].

Generally, the performance degradation in the EASR systems arises mainly due to the statistical mismatch between neutral training and emotional testing conditions. The proposed solutions for the EASR systems can usually be classified into three main categories, namely, feature level, acoustic model (AM) level, and language model (LM) level.

The feature-level approaches aim to find more robust acoustic features or to compensate for the effects of emotional states during the recognition phase. As the main work in this category, Sun et al. [9] have improved the performance of the EASR system by increasing the resolution of important frequency bands in extracting Mel-frequency cepstral coefficients (MFCCs) and perceptual linear prediction (PLP) features. They utilized Fisher's *F*-ratio analysis method in statistics to analyze the significance of different frequency bands for EASR. In another work, Sheikhan et al. [8] have used the neutralized MFCCs in a hidden Markov model (HMM)-based ASR system trained by neutral speech to improve the speech recognition rate for emotional speech utterances. In their method, the frequency warping is performed in the second formant (F2) frequency range to obtain the neutralized MFCCs. Here, the warping factor is first calculated by employing a hybrid structure of dynamic time warping (DTW) and multi-layer perceptron (MLP) neural network. Then, the frequency warping is applied to the stages of Mel-filterbank and/or discrete cosine transform (DCT) in the MFCC feature extraction to obtain the neutralized MFCCs.

The goal of the acoustic model-based methods, on the other hand, is to tune the models in the training stage to make AMs more matched to emotional speech [10, 11]. Pan et al. [10] have used an adaptation technique to construct the emotion-dependent AMs with a small amount of emotional speech. In their work, a model selection approach based on emotion classification is proposed using the Gaussian mixture models (GMMs) to improve the performance of the EASR system. A rapid model adaptation technique has been developed by Ijima et al. [11] for EASR which utilizes the multiple regression HMM (MRHMM) framework. In this method, first, MRHMM is trained using a speaker-independent neutral style model with a small amount of target speaker's data. Then, the acoustic model for speech recognition is adapted to emotional input speech from the trained MRHMM.

In the category of LM techniques, some high-level knowledge, i.e., emotion-specific clues, are added to the model. Athanaselis et al. [3] improved the recognition rate for spontaneous emotionally colored speech by using an emotionally enhanced language model. In this approach, the emotionally enriched LM was derived by adapting an already existing corpus, the British National Corpus (BNC). Here, first, an emotional dictionary is used to identify emotional words in the BNC. Then, sentences containing these words are recombined with the BNC to form a corpus with a raised proportion of emotional material.

Another major factor that leads to the degradation in the performance of ASR systems is the presence of environmental noise. Many techniques have been developed that attempt to address this issue. These include (among others) model-based techniques which use noise models in the recognition procedure [12], noise-robust feature extraction approaches [13], and speech enhancement methods such as spectral subtraction [14].

In contrast to the performance of ASR systems, human speech perception is remarkably robust. Listeners can follow a conversation in the presence of background noise, even in cases where two or more speakers are simultaneously active [15, 16]. This robustness is, for the most part, due to the ability of the auditory system to analyze and decompose complex acoustic scenes into its constituent acoustic sources. The capability of the auditory system to segregate a target sound from an acoustic mixture is termed as auditory scene analysis (ASA) by Bregman [17]. Bregman's work has inspired interest in the development of computational auditory scene analysis (CASA) systems, which aim at modeling the human process of ASA. Typically, the techniques of CASA operate on a time-frequency (T-F) representation of the input and produce an output that can be viewed as a binary T-F mask. A reasonable objective of CASA is the ideal binary mask (IBM) [18], which assigns the values of 0 or 1 to each T-F unit. A value of 1 indicates that the corresponding T-F unit is grouped into the segregated target, and a value of 0 indicates that the unit is considered a part of interference, and hence, removed.

The other obvious advantage of human is the listening with two ears (binaural hearing). This advantage arises from the spatial separation of target and interfering sources, which causes differences between the time of arrival and the sound level of the two ears. These cues are referred to as interaural time difference (ITD) and interaural level difference (ILD). Given a complex acoustic scene, human listeners are able to separate and localize sounds in space by measuring the ITDs and ILDs. The task of understanding speech with two ears in the presence of other concurrent talkers was termed the "cocktail party problem" [19].

A number of different computational methods were developed to segregate a target source from background noise or to perform speech recognition based on estimating an ideal binary mask [20, 21]. May et al. [22] proposed a novel binaural scene analyzer to robustly localize and

detect a known number of speech sources in the presence of spatially distributed interfering noise signals. The proposed system has two processing stages: In the first stage, a binaural unit analyzes the input acoustic mixture to detect the activities of relevant sound sources. Then, in the second stage, based on an estimated binary mask, a speech detection module is constructed which has the task of choosing most likely speech positions from a set of candidate source positions. This is achieved by a two-class Bayesian classifier that is trained to discriminate between speech and noise signals.

Previous studies in the field of EASR [8–11] have focused on improving emotional speech recognition using clean (noiseless) speech data recorded in a quiet environment (i.e., controlled studio settings). However, in real-world scenarios, emotional speech signals are usually disturbed with different types and levels of noises, which decrease the performance of the speech recognition systems. On the other hand, human beings are capable of perceiving emotional speech even in noisy scenarios. Therefore, it is reasonable to believe that the performance of a speech recognition system can be improved by adopting an approach that models the (known) mechanisms of auditory processing [23].

This paper proposes a binaural emotional speech recognition (BESR) system based on known principles of CASA. Our proposed method combines binaural processing with an emotion recognition module to improve the recognition rate of ASR systems in the presence of emotion and background noise. The binaural front-end employs the technique of binaural scene analysis proposed by May et al. [22]. This unit aims at segregating speech from noise using an estimated binary mask. The remaining parts of the proposed BESR system have the task of estimating a binary emotional mask based on the identification of emotional state. This mask is used to identify and retain the speech T-F units that are most likely neutral and remove those units that are mostly affected by emotional states.

The paper is organized as follows. Section 2 explains in detail the main components of the proposed BESR system, including binaural scene analysis, feature extraction, emotion recognition, and emotional mask estimation. In Section 3, the evaluation procedure and the experimental results are provided. The concluding remarks together with a general discussion are given in Section 4.

2 Binaural emotional speech recognition

The block diagram of the proposed BESR system is illustrated in Fig. 1, which is composed of train and test processing phases. In the training phase, based on the extracted features from the input utterance, the Gaussian mixture models (GMMs) are obtained for each emotional state.

The test phase includes three main processing stages, namely, binaural processing, emotion recognition, and emotional masking and recognition. In the binaural processing stage, the target speech is segregated from the noise using the method proposed by May et al. [22]. In the emotion recognition stage, the emotional content of input speech is detected using some pre-trained models of emotional states. In the final stage, based on the recognized emotion, an emotional mask is estimated and applied to the target speech to obtain the most neutral-like segments of speech. Then, the resynthesized noise- and emotion-free signal is fed into a neutrally trained ASR system.

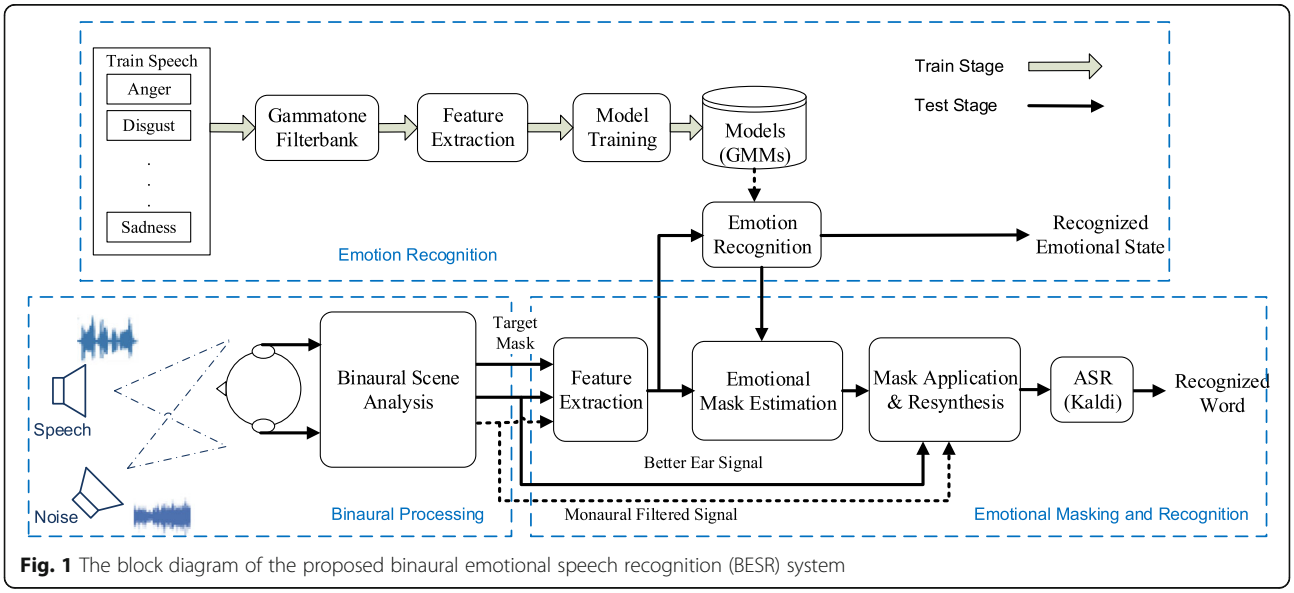

Fig. 1 The block diagram of the proposed binaural emotional speech recognition (BESR) system

As shown in Fig. 1, the proposed BESR system enables us to obtain both the emotional state (i.e., paralinguistic information) and the recognized words (i.e., linguistic information) contained in a speech signal at the same time. The different computational stages of the proposed system are described in detail below.

2.1 Binaural processing

The acoustic input to the proposed BESR model is a binaural mixture signal consisting of speech and noise sources that are positioned at pre-specified spatial locations. The processing unit, called the binaural scene analysis, takes the binaural signal and returns the associated mask of the segregated target speech together with the better ear (BE) signal at its output. The procedure is based on the approach provided by May et al. [22], where the localization and detection of the sources are performed in sequence. The building blocks of the method are shown in Fig. 2.

In the case of monaural input, the block diagram consists of a monaural path (dashed line) which processes the signal only via the gammatone filterbank without the processing stages required for the binaural input signal. The monaural path signal is fed to the emotional masking and recognition unit (see Fig. 1). This structure, called monaural emotional speech recognition (MESR), makes possible the application of the emotional speech recognition in the monaural scenario. In our experiments, the MESR structure serves as a comparison baseline system to assess the efficiency of the proposed BESR system.

The building blocks of the binaural scene analysis are described in detail below.

2.1.1 Gammatone filterbank

In the first stage of the binaural scene analysis, the binaural signal is decomposed by a bank of gammatone filters consisting of $Q = 32$ filters [24], with the center frequencies equally distributed on the equivalent rectangular bandwidth (ERB) rate scale from 80 to 5000 Hz. The impulse response of the gammatone filter is given as:

$$g_{f_c}(t) = t^{N-1} \exp[-2\pi b(f_c)] \cos(2\pi f_c t + \phi)\, u(t), \tag{1}$$

where t refers to time, $N = 4$ is the order of the filter, $b(f_c)$ is the equivalent rectangular bandwidth, f_c is the center frequency of the filter channel c, ϕ is the phase, and $u(t)$ is the step function.

After decomposing the signal with the gammatone filterbank, the signal in each frequency channel is processed by units of half-wave rectification and square-root compression which model the auditory inner hair-cell behavior.

2.1.2 Binaural cue extraction

The binaural cues of the interaural time difference (ITD) and interaural level difference (ILD) are computed independently for each channel by the cross-correlation analysis and the energy comparisons between the right and left ear signals, respectively. The extraction of ITDs and ILDs at each auditory channel is performed by overlapping the frames of 20 ms with a 10-ms shift.

2.1.3 Spatial log-likelihood map computation

After the extraction of the aforementioned binaural cues vector, $\mathrm{x}_{t,f} = (\mathrm{ITD}_{t,f}, \mathrm{ILD}_{t,f})$, a GMM classifier, which has been trained with the azimuth-dependent distribution of feature vectors$\mathbf{x}_{t,f}$, is used to determine the log-likelihood of each source location. Here, the training of the GMM classifier is performed at $K = 37$ different sound source positions in the steps of 5° within the range of $[-90°, 90°]$. The likelihood is a three-dimensional map that represents the probability that the kth source direction is active at frame t and frequency f:

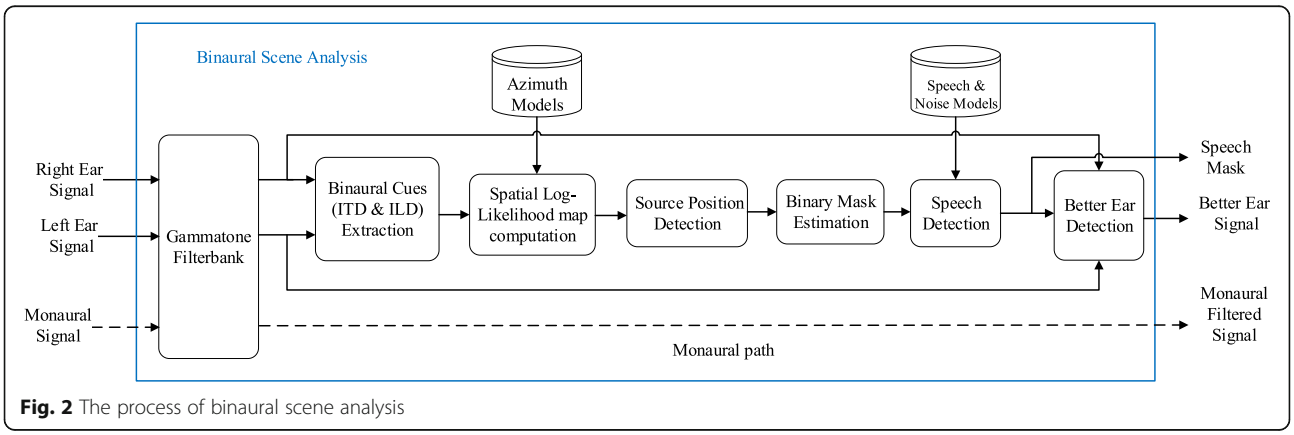

Fig. 2 The process of binaural scene analysis

$$\mathfrak{L}(t,f,k) = \log p\left(\mathbf{x}_{t,f}|\lambda_{f,\phi_k}\right), \tag{2}$$

where $p(\mathbf{x}_{t,f}|\lambda_{f,\phi_k})$ is a Gaussian mixture density with 15 components, and ϕ_k represents the kth source direction.

2.1.4 Source position detection

In this stage, at each time frame t, the likelihood of a source location is summed across all frequency channels to determine the most probable sound source position:

$$\hat{P}(t) = \arg\max_k \sum_{f=1}^{Q} \mathfrak{L}(t,f,k). \tag{3}$$

Then, an azimuth histogram is computed based on the estimated $\hat{P}(t)$. This histogram is used to determine all active sound sources. Here, it is assumed that throughout the time intervals the azimuth histogram is computed, the source positions do not change. The peaks in this azimuth histogram correspond to a set of speech source candidate positions $L = \{l_1, \ldots, l_A\}$.

2.1.5 Mask estimation

In order to determine and isolate the T-F units of each individual sound source, the spatial log-likelihood map $\mathfrak{L}(t,f,k)$ is employed to create a binary mask $M_m(t,f)$ for each candidate position$m = \{1, \ldots, A\}$:

$$M_m(t,f) = \begin{cases} 1, & \text{if } m = \arg\max_{k\in L} \mathfrak{L}(t,f,k) \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

2.1.6 Speech detection

For each estimated mask, the type of corresponding source (i.e., speech or noise) is determined by employing a simple log-likelihood classifier [22]:

$$\log\left(\frac{p(F|\lambda_{\text{Speech}})}{p(F|\lambda_{\text{Noise}})}\right) \underset{\text{Noise}}{\overset{\text{Speech}}{\gtrless}} 0. \tag{5}$$

Here, the feature vector F represents the mean absolute deviation of the smoothed envelope which is extracted separately for noise and speech and used to train the corresponding speech and noise GMMs, λ_{Speech}, and λ_{Noise}, with 32 components and diagonal covariance. In this work, the GMM models of speech and noise are obtained by using the Persian ESD [25] and the NOISEX [26] databases, respectively.

After the detection of the speech source, its corresponding mask, called speech mask, is given to the next processing units, as shown in Fig. 1. Unlike the approach taken by T. May [22], in which the detected speech mask is used for missing data -based speaker recognition task, in our present work, the estimated mask is employed to compute eventually an emotional mask from the target speech in the proposed system.

2.1.7 Better ear selection

Generally, the binaural input signals of the left and right ears differ in their signal-to-noise ratio (SNR) values. This motivates us to select the signal for the next processing stages with the highest SNR value, named as the better ear (BE) signal. The underlying effect is referred to as the better ear effect [27]. After detecting the target speech from a set of speech source candidates, the corresponding azimuth of the target is used to select the BE signal. This is achieved by choosing the closest ear signal to the estimated azimuthal position of the corresponding target speech source. The output of BE selection procedure is one of the left or right mixture signals with the highest SNR, which is used in the "emotional masking and recognition" unit.

2.2 Emotion recognition

In the proposed BESR system, employing the appropriate model in the emotional mask estimation process requires the recognition of the underlying emotion from the test utterance. To this aim, probabilistic models are trained for the emotional states during the train phase. The details of the emotion recognition process are described below.

2.2.1 Feature extraction

Here, since a binary decision is to be made within each T-F unit in the stage of the mask estimation, an appropriate representation should be found for each T-F unit. So, the acoustic features are derived at the frame level for both the train and test stages. To extract the features for the T-F unit $u_{b,q}$ in channel b and frame q, the gammatone filterbank output of the channel b (i.e., the signal $x_b(t)$) is divided into 20 ms time frames with 10 ms overlapping. Then, feature extraction techniques are employed at the frame level to calculate the feature vector for $u_{b,q}$.

In this work, power normalized cepstral coefficients (PNCC) [28] are extracted and employed in the train and test stages of the proposed BESR system. This feature has been shown to provide better recognition accuracies compared to other features [29, 30]. The results of our previous work [30] conducted for EMO DB [31] and Persian ESD [25] databases confirm the robustness and effectiveness of the PNCC for speech emotion recognition in both clean and noisy conditions.

PNCC employs medium-time processing to alleviate the noise corruption and uses power-law compression instead of a log compression [29]. This feature is extracted from the signal in each T-F unit. First, the short-time power spectrum of the input signal is computed using the gammatone frequency summation procedure where the center

frequencies of a 40-channel gammatone filterbank are linearly spaced in equivalent rectangular bandwidth (ERB) scale between 200 and 8000 Hz. Then, based on the medium-duration temporal analysis, an asymmetric filtering and temporal masking are carried out to subtract the background noise. Finally, a power-law nonlinearity and DCT are applied to obtain a 31-D feature vector. The final 93-D feature vector is constructed by employing the first and second derivatives.

2.2.2 Model training

The recognition of the underlying emotions during the test phase requires some pre-trained class (i.e., emotional) models. To this aim, first, for each emotional state, $\mathscr{E}$, and frequency sub-band b, feature vectors are extracted from the training utterances. Then, the extracted features are used to obtain a Gaussian mixture model (GMM), $\lambda_b^{\mathscr{E}}$, with 32 Gaussian components and diagonal covariance matrices. The computed GMMs for all emotional states (i.e., anger, disgust, fear, happiness, sadness, and neutral) are utilized in the testing phase to recognize the underlying emotions from the input signal and subsequently in the estimation of the emotional mask.

2.2.3 Emotion recognition

In our recent work, the performance of the emotion recognition system was evaluated using different acoustical features in a real environment [30]. Here, in a similar approach, the simple maximum-likelihood estimation is used as the emotion classification method. During the test phase, the pre-trained GMMs are employed to determine the underlying emotion using simple likelihood estimation. Let $x_{b,q}$ represents the speech feature vector obtained from the T-F unit, $u_{b,q}$, in frequency sub-band b, and time frame q. The recognized emotion of the signal is the one that maximizes the likelihood function over all frequency sub-bands and time frames:

$$\hat{\varepsilon} = \arg\max_{\mathscr{E}} \sum_{b\in B}\sum_{q\in Q} p\left(x_{b,q}|\lambda_b^{\mathscr{E}}\right), \tag{6}$$

where B and Q are the number of bands in the gammatone filterbank and time frames, respectively. After the recognition of a specific emotion, its corresponding GMM is used in the emotional mask estimation procedure.

2.3 Emotional masking and recognition

Using the selected model for the underlying emotion and the information obtained from the binaural processing unit (i.e., the detected speech mask and the BE signal), an emotional mask is estimated based on the likelihood ratio of the emotional and neutral states for each T-F unit. Then, the output signal is resynthesized after the application of the emotional mask to the BE signal. Finally, the reconstructed signal is fed to the ASR system to achieve the final recognition process. More details are given as follows.

2.3.1 Feature extraction

The BE signal and the target mask obtained from the binaural processing unit are used to extract appropriate features from the target-dominated T-F units. Note that in the case of monaural scenario, only the monaural filtered signal (refer to Fig. 2) is used for the feature extraction process. The computed auditory features are used in the next stage for the estimation of the emotional mask. Here, the PNCC algorithm is employed for feature extraction as described in Section 2.2.1.

2.3.2 Emotional mask estimation

Our proposed EASR system is based on employing a binary mask to select portions of a speech which are less affected by emotions. The motivation behind using such an auditory mask is the following: emotional states of humans do not affect all parts of speech in the same way and to the same amount. Using the idea of the auditory mask, a speech recognition system can be equipped with a pre-processing unit to eliminate the parts of speech which are more affected with emotions. It is expected that this will remove the regions of the emotional speech which degrades the performance of the ASR system. The following binary mask is proposed and estimated in the test stage of the system shown in Fig. 1:

$$M_{\mathscr{E}}(b,q) = \begin{cases} 1, & p\left(x_{b,q}|\lambda_b^{\mathcal{N}}\right)/p\left(x_{b,q}|\lambda_b^{\mathscr{E}}\right) > \theta_b \\ \alpha, & \text{otherwise,} \end{cases} \tag{7}$$

where $M_{\mathscr{E}}(b,q)$ is the mask computed for the sub-band b and frame q, $p(x_{b,q}|\lambda_b^{\mathscr{E}})$ and $p(x_{b,q}|\lambda_b^{\mathcal{N}})$ are the likelihood functions of emotional (denoted as $\mathscr{E}$) and neutral (denoted as $\mathcal{N}$) states, respectively. The threshold θ_b in Eq. (7) is used as an adjustment parameter in sub-band b and determines the extent of the spectro-temporal regions to be retained as the most likely neutral in the specified sub-band. The parameter α is a weighting parameter which controls the amount of removal of most emotionally affected spectro-temporal units from the final speech. This is achieved by removing the affected unit partially (e.g., $\alpha = 0.5$) or completely (i.e., $\alpha = 0$).

The motivation behind the definition of the parameters of α and θ_b is to preserve the linguistic content of speech as much as possible while improving the recognition rate.

2.3.3 Mask application and resynthesis

Using the estimated binary emotional mask, it is straightforward to resynthesize the speech signal from the output of the gammatone filterbank (the BE signal or monaural).

This can be achieved by employing a method introduced by Weintraub [32]. In this approach, after some pre-processing stages, the energy in each T-F unit is weighted by the corresponding T-F mask value obtained from the estimated binary mask. Then, the weighted responses are summed across all frequency channels to yield a speech waveform which is mostly neutral.

2.3.4 ASR

A typical ASR system comprises two stages: feature extraction and decoding. In the first stage, the input speech signal is processed by the feature extraction unit to provide a stream of acoustic feature vectors or observations. In the second stage, the extracted observation sequence is fed into a decoder to recognize the most likely word sequence. Three main knowledge sources, i.e., lexicon, language model, and acoustic model, are used in this stage [33]. In the statistical framework, the Bayesian decision rule is employed to find the most probable word sequence $\hat{W}$ given the observation sequence $O = (o_1, o_2, \ldots, o_n)$:

$$\hat{W} = \arg\max_{W} P(W|O), \tag{8}$$

Using the Bayes' rule, we obtain:

$$\begin{aligned} \hat{W} &= \arg\max_{W} \frac{P(W)P(O|W)}{P(O)} \\ &= \arg\max_{W} P(W)P(O|W), \end{aligned} \tag{9}$$

where the prior probability $P(W)$ and $P(O|\,W)$ specify the language model and acoustic model, respectively. The probability $P(O)$ is discarded in the second equation since it does not alter the search for the best hypothesis.

In order to recognize the underlying words for the input target speech, the resynthesized signal is given to an ASR system. This system which was trained by neutral and noise-free (i.e., clean) speech utterances is used to determine the recognition accuracy of speech from noisy emotional test signals.

3 Experiments and evaluations

3.1 Experimental setup

In a conventional speech recognition system, neutral speech is used for both train and test stages. This establishes a neutral train/neutral test scenario. However, in real conditions (i.e., emotional and noisy speech), the neutrally trained ASR system results in poor performance when employed directly in the emotional noisy speech recognition condition. In this paper, clean-neutrally train/noisy-emotionally test situation is considered in the recognition experiments. Here, to verify the effectiveness of the proposed BESR system, its performance is evaluated and compared with those of the two baseline systems: a neutrally trained Kaldi ASR and the monaural ESR (MESR) (refer to Section 2.1) in noisy emotional test conditions.

The analysis and evaluation results are presented in two different experimental conditions. In the first experiment, the speech recognition task is performed using clean emotional speech utterances taken from the Persian ESD [25]. In the second experiment, the recognition task is conducted with the noisy emotional speech in which four different types of noise, including babble, white, factory, and speech-shaped noise (SSN), and, taken from the Noisex-92 database [26] are artificially added to each utterance at various SNRs. The effect of noise addition is investigated for different SNR values ranging from the − 5 dB SNR to 30 dB SNR.

The binaural target and noise signals in the experiments are created by convolving the signals with their corresponding head-related impulse responses (HRIRs) of the KEMAR artificial head [34]. The binaural acoustic mixture signal is generated by combining a binaural target speech located at 0° of azimuth with a binaural noise positioned at azimuths of 0°, ± 30°, ± 60°, and ± 90°. Here, it is assumed that the acoustic mixture is created in an anechoic room.

Our proposed method for improving the EASR system is based on applying a binary acoustic mask which is estimated using the pre-trained emotional GMM models. The GMMs used for emotional models of the proposed EASR system are composed of 32 components with diagonal covariance matrices. As an acoustic feature, the PNCC is extracted [29] and employed in the train and test stages of the proposed EASR system. This feature is used together with its first and second derivatives to construct the final feature vector.

In the emotional mask estimation procedure, the threshold values, θ_b, in each frequency sub-band are set to retain 95% of the regions most related to the neutral state. This means that in each sub-band, 5% of the spectro-temporal regions are suppressed by the masking procedure.

3.2 Database

Speech utterances of Persian ESD [25] are used for the experimental purposes. Persian ESD is a comprehensive emotional speech database for colloquial Persian. The database was produced in a professional recording studio in Berlin, Germany, under the supervision of an expert linguist and an acoustician. It contains a set of 90 validated novel sentences uttered by two native Persian speakers (one male and one female) in different emotional states. As shown in Table 1, the Persian ESD comprises 472 speech utterances, each with a duration of 5 s on average, which are classified into five basic emotional groups of anger, disgust, fear, happiness, and sadness, as well as

Table 1 Number of utterances per state for Persian ESD database used in the experiments

Emotions	Number of Utterances
Anger	62
Disgust	58
Fear	58
Happiness	58
Sadness	56
Neutral	180
Total	472

the neutral state. The database was articulated in three situations: (1) congruent (emotional lexical content spoken in a congruent emotional voice), (2) incongruent (neutral sentences spoken in an emotional voice), and (3) baseline (all emotional and neutral sentences spoken in neutral voice). The validity of the database was assessed by a group of 34 native speakers in a perception test. Utterances having a recognition rate of 71.4% or better were regarded as valid descriptions of the target emotions. The recordings are available at a sampling rate of 44.1 kHz and mono channel.

3.3 Evaluation criterion

The performance of an ASR system for a particular task is often measured by comparing the hypothesized and reference transcriptions. In this context, word error rate (WER) is the most widely used metric which is used to assess the quality of our proposed BESR system. After the alignment of the two-word sequences (i.e., hypothesis and reference), the number of substitutions (S), insertions (I), and deletions (D) is obtained in the Kaldi ASR. These are all considered as errors, and the WER is calculated by the rate of the number of errors to the total number of words (N) in the reference:

$$\mathrm{WER} = \frac{S + I + D}{N} \times 100\%. \tag{10}$$

3.4 Kaldi ASR system

The effectiveness of the proposed emotional mask in removing most emotional areas is assessed by an ASR system that has been trained by neutral speech utterances. This paper uses the ASR system implemented by the Kaldi toolkit, which is based on finite-state transducers (FST) [35]. Kaldi uses the weighted finite-state transducers (WFSTs) for training and decoding algorithms. The FST framework provides graph operations, which can be effectively used for decoding. Using the FST, the speech decoding task is expressed as a beam search in a graph, which is a well-studied problem [36].

Here, a GMM-HMM model is trained on neutral utterances of Persian ESD corpus using the Kaldi recipe. In the process of feature extraction, first, 13 MFCCs are extracted. Then, cepstral mean-variance normalization (CMVN) and delta and delta-delta operations are applied to compute the final 39 acoustic features. For decoding, the trained acoustic models are employed within which the phonemes are modeled by three-state single-mixture left-to-right monophone HMMs. For this purpose, 30 phonemes are used including silence and pause. The corresponding HMMs of the Kaldi system are trained using the neutral utterances of the Persian ESD which comprises 180 utterances (2 speakers, 90 utterances per speaker). The statistical language model used in the ASR system is the word-pair bigram language model. Because of the lack of a standard lexicon in the Persian language for speech recognition tasks, a lexicon is created from the Persian ESD dataset.

3.5 Results and discussions

3.5.1 Kaldi baseline

We first examine the effects of noise and emotion on the performance of the speech recognition system. To this aim, the trained ASR system with clean and neutral utterances is first tested with clean emotional utterances and then with noisy ones. In this experiment, the test utterances from different emotional states are mixed with four different noise types and in six different SNR levels, ranging from − 5 to 30 dB. Tables 2 and 3 show the results of speech recognition in terms of WER obtained in clean and noisy conditions, respectively.

The recognition result for the neutral case is obtained through a distinct experiment in which 80% of the total neutral utterances are used for the training of the acoustic model, and the rest is reserved for the testing.

The results in Table 2 indicate that the neutrally trained ASR system can achieve a high performance on neutral speech while performing poorly on emotional input. This confirms the fact that emotion has a negative impact on speech recognition accuracy. The average WERs for emotional speech obtained is 12.47%, which are much higher than the results achieved by neutral speech (1%).

As the results of Table 3 show, the performance of the ASR system degrades in the presence of noise. Specifically, it is observed that by increasing the SNR value, the recognition performance is enhanced. However, it is seen that even at high SNR values (e.g., SNRs of 20, 30 dB),

Table 2 The performance of speech recognition for clean emotional speech using Kaldi ASR trained with neutral speech

Emotional states	Anger	Disgust	Fear	Happiness	Sadness	Neutral
WER (%)	11.31	13.92	13	12.52	11.60	1

Table 3 The performance of Kaldi system in terms of WER (%) for noisy emotional speech trained with neutral speech

Noise types	Emotional states	SNR					
		− 5	0	5	10	20	30
Babble	An.	94.43	91.64	86.23	65.08	35.74	26.56
	Dis.	93.04	87.91	75.09	62.27	35.9	17.95
	Fe.	95.05	91.76	87.91	75.64	38.46	20.33
	Ha.	93.74	90.88	81.4	68.34	34.88	19.32
	Sad.	95.06	93.92	89.92	76.81	33.27	18.25
	Mean	94.26	91.22	84.11	69.63	35.65	20.48
White	An.	97.38	93.77	83.44	70.82	40	25.74
	Dis.	94.69	91.39	86.63	76.92	56.96	29.12
	Fe.	98.35	94.51	90.66	85.53	53.48	28.75
	Ha.	97.14	90.52	83.54	70.13	43.83	26.65
	Sad.	96.58	94.3	88.97	83.65	64.83	38.21
	Mean	96.83	92.90	86.65	77.41	51.82	29.69
SSN	An.	98.36	94.75	88.03	71.8	39.02	24.75
	Dis.	96.34	89.74	84.62	69.05	36.63	16.12
	Fe.	99.27	97.07	91.21	80.4	43.41	23.81
	Ha.	99.46	95.53	86.4	75.13	36.49	20.75
	Sad.	99.05	95.82	92.4	83.84	44.11	19.2
	Mean	98.50	94.58	88.53	76.04	39.93	20.93
Factory	An.	96.56	91.31	84.1	66.07	36.72	25.74
	Dis.	94.14	91.58	83.7	68.86	38.83	19.78
	Fe.	98.53	95.79	91.94	81.32	44.87	23.44
	Ha.	96.78	93.56	84.97	70.3	37.57	16.99
	Sad.	97.34	92.59	88.02	78.33	40.11	20.53
	Mean	96.67	92.97	86.55	72.98	39.62	21.30

the WER ranges from 35 to 52 and 20 to 30, respectively, signifying the detrimental effect of noise.

3.5.2 MESR baseline

In the second experiment, the performance of the monaural ESR system is evaluated in the emotional clean and noisy scenarios. In comparison with the ASR system of the first experiment (Kaldi baseline), this system utilize a pre-processing to decrease the effect of emotion. It is expected, therefore, that this system performs better than the Kaldi baseline in the clean emotional conditions. The results of this experiment are shown in Tables 4 and 5 for clean and noisy emotional utterances, respectively.

Table 4 represents the recognition results achieved by the MESR system based on two different mask estimation methods (i.e., PNCC-mask and MFCC-mask) for two different parameter values of α (0 and 0.5) used in the mask estimation process compared with the results obtained with the Kaldi baseline. As the table shows, in general, the monaural ESR system with different masks

Table 4 The performances of the Kaldi baseline and the MESR systems in terms of WER (%) based on three mask estimation methods and different values of α in clean condition

Emotional states	Kaldi baseline	MESR baseline			
		PNCC-mask		MFCC-mask	
		$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0$	$\alpha = 0.5$
Anger	11.31	11.97	4.10	19.34	5.08
Disgust	13.92	13.92	13.19	21.25	15.02
Fear	13.00	14.10	8.97	27.11	11.36
Happiness	12.52	12.34	9.48	29.86	8.94
Sadness	11.60	8.94	9.70	17.30	9.70
Average	12.47	12.25	9.08	22.97	10.02

Table 5 The performances of MESR system in terms of WER (%) based on PNCC-mask for noisy emotional speech. For comparison, the mean values of the recognition rates for the Kaldi system have been included (see Table 3)

Noise types	Emotional states	SNR					
		− 5	0	5	10	20	30
Babble	An.	93	89	73	50	13	7
	Dis.	92	86	75	59	20	11
	Fe.	92	88	72	46	14	8
	Ha.	94	90	79	58	26	11
	Sad.	93	89	78	52	18	11
	Mean	92.8	88.4	75.4	53	18.2	9.6
	Kaldi mean	94.26	91.22	84.11	69.63	35.65	20.48
White	An.	92	92	83	62	21	9
	Dis.	94	92	90	73	35	16
	Fe.	92	89	82	70	27	12
	Ha.	93	91	84	69	31	15
	Sad.	94	91	86	74	32	14
	Mean	93	91	85	69.6	29.2	13.2
	Kaldi mean	96.83	92.90	86.65	77.41	51.82	29.69
SSN	An.	93	91	78	57	16	8
	Dis.	93	90	86	69	24	13
	Fe.	92	90	82	58	16	10
	Ha.	94	92	83	66	27	13
	Sad.	95	90	85	71	22	13
	Mean	93.4	90.6	82.8	64.2	21	11.4
	Kaldi mean	98.50	94.58	88.53	76.04	39.93	20.93
Factory	An.	93	90	83	65	21	9
	Dis.	91	91	88	76	32	15
	Fe.	95	94	89	68	25	11
	Ha.	95	93	87	72	33	15
	Sad.	92	91	89	79	36	16
	Mean	93.2	91.8	87.2	72	29.4	13.2
	Kaldi mean	96.67	92.97	86.55	72.98	39.62	21.30

attains lower WERs as compared with the Kaldi system, which confirms the effectiveness of the system. Here, it is observed that the mask estimation based on PNCC achieves the lowest WER value. On the average, the error rate obtained by the PNCC-mask is 9.08% for $\alpha = 0.5$, which is almost 1% lower than that achieved with the MFCC-Mask for the same value of α. Moreover, for this value of the parameter (i.e., $\alpha = 0.5$) the performance of the PNCC-mask is 3.4% better than that obtained from the Kaldi baseline. As the results show, among different mask-based implementations of MESR system, the lowest WER is obtained for the case when $\alpha = 0.5$. This improvement in the performance can be interpreted by the fact that using this value of the weighting parameter (i.e., $\alpha = 0.5$) in the mask calculation removes partially the most emotionally affected regions, and this in turn improves the recognition rate in the sense of reducing WER.

Experimental results demonstrate the effectiveness of the PNCC for emotional speech recognition. This is due to incorporating different processing stages in the implementation of the PNCC, including the use of a power-law nonlinearity, employing a noise suppression algorithm based on asymmetric filtering, and using a module that accomplishes temporal masking [29]. In all the experiments that follow, we will use the PNCC-mask with $\alpha = 0.5$ due to its high performance.

The performance of MESR system based on the PNCC-mask is depicted in Table 5 in terms of WER for distinct values of SNRs and noise types. For the purpose of comparison, the mean recognition values of the Kaldi system have also been included. Comparing these results with those of the Kaldi baseline shows that the MESR system has a better performance in noisy situations. Specifically, the amount of improvement increases generally by increasing the SNR value.

3.5.3 Proposed BESR

In the third experiment, the performance of the BESR system is evaluated in emotional noisy scenarios. Here, for each azimuth position of the interferer, including azimuths of 0, ± 30, ± 60, and ± 90°, 120 distinct tests (five emotional states, four noise types, and six SNR values) are performed comprising a total of 840 different test conditions. The results of the experiment are shown in Fig. 3 for noisy emotional utterances. To show the results in a compact way, the outcomes of speech recognition are averaged for symmetric azimuths (i.e., ± 30, ± 60, ± 90) and emotional states. For comparison, the mean values of WER for the Kaldi baseline and monaural ESR (MESR) systems are also included.

As shown in Fig. 3, in general, the proposed BESR system improves the emotional speech recognition rate in noisy conditions. Among various azimuths, the BESR recognition rates yield the worst results at the azimuth of 0°, as expected. This observation relies on the fact that in this case, both of target and noise sources are located at the same position (collocated scenario), where the binaural scene analysis is not able to segregate the sources. This situation is equivalent to the monaural (i.e., MESR) case as the results show. For other azimuths, the results of BESR are close to each other. With the exception of the babble noise scenario, the best result is achieved for the case when the noise source is located at the azimuth of 90°. In this case, the distance between the target and the noise is maximized, and therefore, the segregation is performed well as compared with other spatial configurations.

The amount of improvement obtained by the BESR system in comparison with other systems varies with noise types and SNR values. The maximum improvement is attained in the case of babble and SSN noises which is probably due to the gaps occurring in time-frequency representations of these noise types.

At the SNR of − 5 dB, the highest improvement is achieved for the babble noise and the least improvement is obtained with the white and factory noises. Generally, as the level of SNR increases, the amount of improvement of the BESR increases gradually in contrast to the other systems and reaches its maximum in the SNR range of 5–10 dB, after which it starts to decrease. Specifically, for the babble and SSN, the highest improvement is reached to approximately 40 to 50%, compared to the Kaldi and MESR baselines. However, this amount reaches 30% for white and factory noises.

The results of the figure also show that the performance difference between MESR and BESR systems is observed below the SNR values of 20 dB, and as expected, the two systems have the same performance at higher SNR levels (e.g., 20 and 30 dBs).

3.5.4 Emotion recognition

It is expected that enhancing the recognition of emotion results in improving emotional speech recognition. Once the type of emotion is recognized, the appropriate emotional model can be used in the emotional mask estimation procedure. This expectation is verified in an experiment in which the performance of the proposed binaural system (i.e., BESR) for the emotion recognition task (refer to Fig. 1) is evaluated in both clean and noisy conditions against that of the monaural system (i.e., MESR) for different noise types and SNR values. The results of this experiment are shown in Fig. 4.

The figure illustrates the average values of the emotion recognition rates for the BESR and MESR systems obtained among different noise types (babble, white, SSN, and factory) and different emotional states (anger, disgust, fear, happiness, neutral, and sadness) of the Persian ESD. As it can be seen from the figure, both systems have the

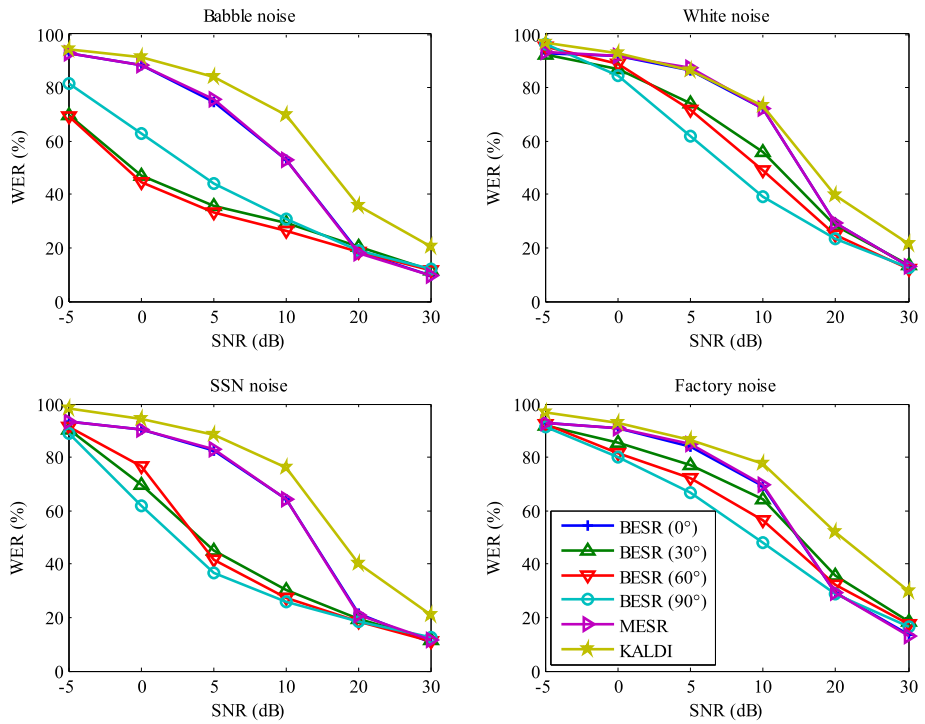

Fig. 3 Comparing the performance of the BESR system with MESR and Kaldi baseline systems in motional noisy scenarios

same performance in the clean condition. In this condition, the average accuracy obtained by the MESR and BESR systems is about 97%. Obviously, as the SNR is decreased, the recognition rates of the systems also decrease. However, the difference between the performances of the systems increases, in a way that the BESR maintains always its higher performance against the MESR. Specifically, a large gap in the performances of both systems is observed at the SNR of – 5 dB where the recognition rates for the MESR and BESR systems reaches to 25% and 55%, respectively. The improvement in emotion recognition of the BESR system is due to binaural processing in suppressing the effect of the noise when the target and noise source have different azimuthal positions.

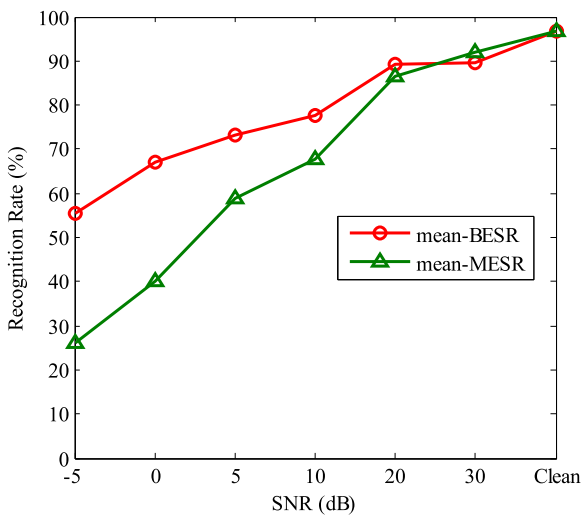

Fig. 4 Emotion recognition rates of the BESR and MESR systems based on the PNCC feature in both clean and noisy conditions for different SNR values obtained as mean values among different noise types and emotional states

The above discussion confirms the notion that enhancing the emotion recognition enables the system to select the appropriate emotional model which improves the emotional mask estimation. This in turn increases the recognition rate of the ASR in the final stage of the proposed binaural emotional speech recognition.

4 Conclusions

In this paper, we consider the problem of emotional speech recognition in noisy conditions and propose a new approach to improve the recognition rate of the EASR systems. The proposed binaural emotional speech recognition (BESR) system is based on the binaural processing of the input signal and estimation of an emotional auditory mask corresponding to the recognized emotion. The proposed binaural system employs a preprocessing stage in which the target is first segregated from the noise and then the most emotionally effected spectro-temporal regions are removed. This is achieved by using the idea of a binary mask and the recognition of underlying emotions.

The performance of the proposed binaural system is evaluated against two baseline systems, namely monaural ESR (MESR) and Kaldi, in neutral train/noisy emotional test conditions for different noise types, SNRs, and spatial configurations of sources. Speech utterances of Persian ESD are used for experimental purposes.

In the experiments, first, we show the effect of noise and emotion on the performance of speech recognition. Then, the performance of MESR is evaluated in clean and noisy conditions. This stage of experiments shows an improvement in emotional speech recognition as compared with the Kaldi ASR system, which can be justified as a result of incorporating a preprocessing stage to remove the emotionally affected regions. Finally, the speech recognition performance of the proposed binaural system is compared with those of the baselines, which shows further improvements in the sense of WER measures.

The results of assessment for the proposed BESR system in different spatial configurations show the best WER score when the noise source is located at the azimuth of 90° whereas the worst score is attained at 0° where the target and noise signals stem from the same spatial location. In the experiments involving different noise types, high amounts of improvement (up to 50%) are obtained for babble and SSN compared with the performances of baselines, while this improvement reaches to a maximum of 30% for factory and white noises.

Another contribution of the proposed binaural system concerns its capability in the recognition of different emotions. As to this, the performances of the proposed binaural system (i.e., BESR) and the monaural system (i.e., MESR) are evaluated in the framework of emotion recognition task in different noisy conditions. Here, the proposed system shows again satisfactory results in terms of recognition rates.

The experimental results show a higher performance of the proposed system as compared with the baseline systems, namely, Kaldi and monaural ESR. This is mainly due to the use of binaural processing and emotional masks in the removal of emotionally affected areas.

As future work, the authors plan to evaluate the proposed system in other environmental conditions, such as reverberant rooms, to study more on the effects of other auditory masking methods on the recognition accuracies, to incorporate other peripheral analysis techniques based on more physiologically justified auditory models, and also to consider the feasibility of missing data handling methods in the framework of emotional speech recognition.

Abbreviations

AM: Acoustic model; ASA: Auditory scene analysis; ASR: Automatic speech recognition; BE: Better ear; BESR: Binaural emotional speech recognition; CASA: Computational auditory scene analysis; CMVN: Cepstral mean-variance normalization; DCT: Discrete cosine transform; DTW: Dynamic time warping; EASR: Emotion affected speech recognition; ERB: Equivalent rectangular bandwidth; ESD: Emotional speech database; ESR: Emotional speech recognition; FST: Finite-state transducers; GMM: Gaussian mixture model; HMM: Hidden Markov model; HRIR: Head-related impulse responses; IBM: Ideal binary mask; ILD: Interaural level difference; ITD: Interaural time difference; LM: Language model; MESR: Monaural emotional speech recognition; MFCC: Mel-frequency cepstral coefficient; MLP: Multi-layer perceptron; MRHMM: Multiple regression HMM; PLP: Perceptual linear prediction; PNCC: Power normalized cepstral coefficient; SNR: Signal-to-noise ratio; SSN: Speech-shaped noise; WER: Word error rate

Availability of data and materials

The Persian ESD is not publicly available but can be obtained upon request (refer to reference [25]). Other used and publicly available datasets are, respectively, NOISEX-92 [26] and KEMAR [34].

Authors' contributions

MB and MG have equally contributed in proposing the ideas, discussing the results, and writing and proofreading the manuscript. MB carried out the implementation and experiments. Both authors read and approved the final manuscript.

Authors' information

Meysam Bashirpour was born in Tabriz, Iran, in 1984. He received the B.Sc. Degree in Electronic Engineering from the Sharif University of Technology in 2007 and M.Sc. Degree in Communication Engineering from the University of Tehran in 2010. He is currently pursuing Ph.D. Degree in the Faculty of Computer Engineering at the University of Tabriz, Tabriz, Iran. His current research interests are focused on speech recognition, binaural signal processing, emotion recognition from speech, and pattern classification. Masoud Geravanchizadeh received the B.Sc. Degree in Electronics Engineering from the University of Tabriz, Tabriz, Iran, in 1986 and the M.Sc. and Ph.D. Degrees in Signal Processing from the Ruhr-University Bochum, Bochum, Germany, in 1995 and 2001, respectively. Since 2005, he has been with the Faculty of Electrical and Computer Engineering at the University of Tabriz, Tabriz, where he is currently an associate professor. His research interests include binaural signal processing, auditory-based emotional speech recognition, improvement of speech quality and intelligibility for normal hearing and hearing-impaired listeners, sound source localization and separation, pattern classification, and stochastic signal processing.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 11 April 2018 Accepted: 5 August 2018
Published online: 28 August 2018

References

1. Benzeghiba, M, et al. (2007). Automatic speech recognition and speech variability: a review. *Speech Comm.*, ***49***(10), 763–786.
2. Cowie, R, et al. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Proc. Mag.*, ***18***(1), 32–80.
3. Athanaselis, T, Bakamidis, S, Dologlou, I, Cowie, R, Douglas-Cowie, E, Cox, C. (2005). ASR for emotional speech: clarifying the issues and enhancing performance. *Neural Netw.*, ***18***(4), 437–444.
4. Xiong, X (2009). Robust speech features and acoustic models for speech recognition, PhD thesis, Nanyang Technological University, School of Computer Egineering, 2009.
5. Bosch, L. (2003). Emotions speech and ASR framework. *Speech Comm.*, ***40***(1), 213–225.
6. Ayadi, M, Kamel, MS, Karray, F. (2011). Survey on speech emotion recognition: features, classification schemes, and databases. *Patt. Recognit.*, ***44***(3), 572–587.
7. Ververidis, D, & Kotropoulos, C. (2006). Emotional speech recognition: resources, features, and methods. *Speech Comm.*, ***48***(9), 1162–1181.
8. Sheikhan, M, Gharavian, D, Ashoftedel, F. (2012). Using DTW neural–based MFCC warping to improve emotional speech recognition. *Neural Comput. Appl.*, ***21***(7), 1765–1773.

9. Sun, Y, Zhou, Y, Zhao, Q, Yan, Y (2009). In Proc. International Conference on Information Engineering and Computer Science, 2009. In *Acoustic feature optimization for emotion affected speech recognition*, (pp. 1–4).
10. Pan, Y, Xu, M, Liu, L, Jia, P (2006). In Proc. Multiconference on Computational Engineering in Systems Applications. In *Emotion-detecting based model selection for emotional speech recognition*, (pp. 2169–2172).
11. Ijima, Y, Tachibana, M, Nose, T, Kobayashi, T (2009). in Proc. ICASSP 2009. In *Emotional speech recognition based on style estimation and adaptation with multiple-regression HMM*, (pp. 4157–4160).
12. Gales, MJ, & Young, SJ. (1996). Robust continuous speech recognition using parallel model combination. *IEEE Trans. Audio Speech Lang. Process*, ***4***(5), 352–359.
13. Hermansky, H, Morgan, N, Hirsch, HG (1993). In Proc. ICASSP-93. In *Recognition of speech in additive and convolutional noise based on RASTA spectral processing*, (pp. 83–86).
14. Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoustics Speech Signal Process.*, ***27***(2), 113–120.
15. Yost, WA (1997). *Binaural and spatial hearing in real and virtual environments*, ed. by RH Gilky, TRAnderson (Psychology press, New York, 2014), p. 329.
16. Hawley, ML, Litovsky, RY, Colburn, HS. (1999). Speech intelligibility and localization in a multi-source environment. *J. Acoust. Soc. Amer*, ***105***(6), 3436–3448.
17. Bregman, AS (1990). *Auditory scene analysis: the perceptual organization of sound* (MIT Press, Cambridge, 1994)
18. Wang, D (2005). *Speech separation by humans and machines*, (pp. 181–197). Boston: Springer.
19. Cherry, EC. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Amer*, ***25***(5), 975–979.
20. Wang, D, Kjems, U, Pedersen, MS, Boldt, JB, Lunner, T. (2009). Speech intelligibility in background noise with ideal binary time-frequency masking. *J. Acoust. Soc. Amer*, ***125***(4), 2336–2347.
21. Wang, D, & Brown, GJ (2006). *Computational auditory scene analysis: principles, algorithms, and applications* (Wiley-IEEE Press, New Jersey, 2006)
22. May, T, van de Par, S, Kohlrausch, A. (2012). A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation. *IEEE Trans. Audio Speech Lang. Process.*, ***20***(7), 2016–2030.
23. Hermansky, H. (1998). Should recognizers have ears? *Speech Comm.*, ***25***(1), 3–27.
24. Holdsworth, J, Nimmo-Smith, I, Patterson, R, Rice, P (1988). *Implementing a gammatone filter bank, Annex C of the SVOS final report: part A: the auditory filterbank*, (vol. **1**, pp. 1–5).
25. Keshtiari, N, Kuhlmann, M, Eslami, M, Klann-Delius, G. (2015). Recognizing emotional speech in Persian: a validated database of Persian emotional speech (Persian ESD). *Behav. Res. Methods*, ***47***(1), 275–294.
26. Varga, A, & Steeneken, HJ. (1993). Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Comm*, ***12***(3), 247–251.
27. Shinn-Cunningham, BG, Schickler, J, Kopčo, N, Litovsky, R. (2001). Spatial unmasking of nearby speech sources in a simulated anechoic environment. *J. Acoust. Soc. Amer*, ***110***(2), 1118–1129.
28. Kim, C, & Stern, RM (2009). in Proc. Interspeech. In *Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction*, (pp. 28–31).
29. Kim, C, & Stern, RM. (2016). Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.*, ***24***(7), 1315–1329.
30. Bashirpour, M, & Geravanchizadeh, M. (2016). Speech emotion recognition based on power normalized cepstral coefficients in noisy conditions. *Iranian J. Electr. Electron. Eng.*, ***12***(3), 197–205.
31. Burkhardt, F, Paeschke, A, Rolfes, M, Sendlmeier, WF, Weiss, B (2005). in Proc. Interspeech. In *A database of german emotional speech*, (pp. 1517–1520).
32. Weintraub, M. A theory and computational model of auditory monaural sound separation, PhD thesis, Stanford University (1985)
33. Jelinek, F. *Statistical methods for speech recognition*. (MIT press, Cambridge, 1997)
34. B Gardner K Martin, HRTF measurements of a KEMAR dummy-head microphone, MIT Media Lab Perceptual Computing Technical Report, (1994)
35. Povey, D, et al. (2011). in IEEE 2011 workshop on automatic speech recognition and understanding. In *The Kaldi speech recognition toolkit*.
36. Mohri, M, Pereira, F, Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Comput. Speech Lang.*, ***16***(1), 69–88.