

Binaural Classification-Based Speech Segregation and Robust Speaker Recognition System

R. Venkatesan¹ · A. Balaji Ganesh¹ 

Received: 5 June 2017 / Revised: 2 November 2017 / Accepted: 4 November 2017 /
Published online: 23 November 2017
© Springer Science+Business Media, LLC, part of Springer Nature 2017

Abstract The paper presents an auditory scene analyser that comprises of two joint simultaneous modules, namely binaural speech segregation and speaker recognition. The binaural speech segregation is realized by incorporating interaural time and level differences, interaural phase difference and interaural coherence along with direct-to-reverberant ratio into deep recurrent neural network. The performance of deep recurrent network-based speech segregation is validated in terms of source to interference ratio, source to distortion ratio and source to artifacts ratio and compared with existing architectures including deep neural network. It is observed that performance of conventional deep recurrent neural network can be improved further by involving discriminative objectives along with soft time–frequency masking as a layer in the network structure. The system also proposes a spectro-temporal extractor which is referred as Gabor–Hilbert envelope coefficients (GHEC). The proposed monaural feature is responsible for extracting discriminative acoustic information from segregated speech sources. The performance of GHEC is validated under various noisy and reverberant environments and the results are compared with existing monaural features. The results of binaural speech segregation have shown better signal-to-noise ratio at an average of 0.7 dB even in the presence of higher reverberation time, 0.89 s over other baseline algorithms.

✉ A. Balaji Ganesh
abganesh@velammal.edu.in
R. Venkatesan
venky88an@gmail.com

¹ Electronic System Design Laboratory, Department of Electrical and Electronics Engineering, Velammal Engineering College, Chennai 600066, India

Keywords Binaural cues · Computational auditory scene analysis · Automatic speaker recognition · Gabor Hilbert envelope features · Deep recurrent neural network · Soft time–frequency masking

1 Introduction

Human auditory system has the fascinating ability to segregate acoustic signals from a complex mixture of input speech signals amidst reverberant and noisy environment. It can also locate and eventually estimate the distance of acoustic signals, even in the absence of visual information [6, 7, 34, 40, 50]. Significant numbers of computational techniques referred as computational auditory scene analysis (CASA) have been devised to analyse the complex acoustic mixture mainly because of inspiration received from human auditory system [59]. The speech segregation process involves separation of an interested single speech source from multiple sound mixtures [59, 60]. The implementation of an automatic speaker recognition system in a real-world application is considered as a difficult task due to the presence of additive noises as well as room reverberations. Several spectral subtraction filtering methods have been investigated for improving the robustness in automatic speaker recognition [59, 60]. Currently, researchers have focused on designing reliable speech processing system by combining speech separation and robust speaker recognition module [60]. The combined architecture contains speech segregation module followed by robust speaker recognizer. Further, the joint architecture is found to be improving the speech intelligibility and efficiently degrading the combined effects of reverberation and noise [59, 60].

In the literature, most of the speech segregation studies have concentrated on monaural speech signals and related features [8, 51]. Various algorithms, such as spectral subtraction, inverse and Wiener filtering techniques are suggested for the problems in the segregation of monaural speech signals [24, 48]. Woodruff et al. [54] describe a robust joint localization as well as segregation of voiced speech sources by means of combined Pitch and azimuth cues in reverberant environments. Also, Woodruff et al. [53] propose an azimuth-dependent classifier-based localization method in which segregation process is carried out by using monaural feature to improve the estimation of azimuth cues from binaural input. The comparative analyses are largely carried out by using various classifiers, including support vector machine (SVM) and Gaussian mixture model (GMM) [12]. Recently, many researchers use neural network-based classifiers due to its improved robustness and performance. A multi-layer perceptron (MLP)-based monaural feature evaluation framework is demonstrated in a speech segregation application [20]. Several studies have considered pitch and azimuth cues as features for the segregation of speech sources. Wrigley et al. [55] propose recurrent timing neural network-based speech segregation of two acoustic sources in which pitch and location features are considered. Alinaghi et al. [3] suggest binaural speech segregation algorithm on the basis of weighted combination of binaural cues, such as IPD, ILD, IC and the mixing vector models. Weiss et al. [52] have proposed a binaural source separation technique that combines spatial models with a priori trained source models and derived the expectation maximization (EM) algorithm for the determination of maximum likelihood parameters. Abdipour et al. [1] have suggested a novel

system for the segregation of multiple moving sources from stereo signals, and it is based on the statistical model where maximum likelihood estimation is realized by using expectation-maximization (EM) technique.

Many researchers have successfully used direct-to-reverberant ratio (DRR) for the estimation of distance [30,31]. The direct-to-reverberant ratio depends upon various factors, including room volume, directivity, source to receiver distance and also reverberation time [13,26,30,31,56]. Lu et al. suggest a binaural equalization–cancellation technique that estimate direct energy ratio by locating acoustic source in a delay-line structure [30,31]. A direct and reverberant sound spatial correlation matrix model is suggested for the estimation of absolute distance between acoustic source and microphone array [30,31]. Hioka et al. [13] suggest a spatial correlation matrix model for the segregation of direct and reverberant components. The estimated DRR from the above method is restricted to smaller distances but the method has shown better improvement in the speech segregation process.

Automatic speaker recognition (ASR) is considered very important, especially in applications, such as speech and speaker indexing, document content structuring, call routing, data entry and dictation and speaker attributed speech to text transcription [5,29]. The performance of an ASR is adversely affected by two forms of reverberation, namely self-masking and overlap-masking [42]. Self-masking occurs due to early reflections and diffractions, whereas high impact overlap-masking is due from late reverberation [42]. The binary time–frequency mask is considered as a core of computational auditory scene analysis which is used to segregate the desired target from multiple acoustic mixtures [29,47].

ASR is well supported by various techniques, such as Gaussian mixture models, pattern matching, support vector machine (SVM), hidden Markov models (HMM) and neural networks [5,29]. Sadjadi et al. [43] suggest the mean Hilbert cepstral coefficients (MHEC) method as a replacement to traditional Mel-frequency cepstral coefficients (MFCC) within I -vector-based speaker acoustic model under noisy reverberant environments. Recently, Gabor filter banks have been efficiently used to construct monaural features for various applications, such as facial emotion recognition, robust speaker recognition and automatic speech recognition [25,27,45,46]. A joint optimization of spectro-temporal features (Gabor filters) along with neural net acoustic model is demonstrated and proposed as an improved ASR [25]. Kanagasundaram et al. [22] have shown an improvement in the I -vector-based speaker verification by involving channel compensation method. Further, research findings are also available on the basis of simultaneous localization and recognition of target speaker to suppress the combined effects of noise and reverberation [24,34]. May et al. [34] have proposed a noise robust binaural scene analyser for the localization and also recognition of speakers in the presence of competing sound sources, instantaneously. More specifically, the effects of reverberation and noises in speaker verification are addressed in recent research works [2,36–38]. Al-Ali et al. [2] have introduced a forensic speaker verification system that investigates the combined features of MFCC and DWT–MFCC of input speech signal under different noisy reverberant conditions. Naik et al. have proposed a novel method on the basis of evaluation of super- and sub-Gaussian signals which are computed by using different objective measures of speech qualities to improve the quality of separated audio sources [37].

2 Related Works

Recently, deep learning-based binaural speech segregation shows better results than monaural speech segregation for the applications where reverberations are significantly considered [19,58]. Zhang et al. [58] have proposed a novel deep learning-based binaural speech segregation by employing a fixed beam-former before extracting spectral features and have successfully validated in various reverberant environments. Jiang et al. [19] have proposed a binaural classification by using deep neural network (DNN) for stereo signals to handle complex auditory scenes, effectively. The DNN-based binaural classification is found to be providing good performance for speech segregation in a multi-source environment. The performance of automatic speech recognition is further improved when a combined architecture of deep neural network and recurrent neural network are used [32,55]. Maas et al. [32] have introduced a noise reduction technique by applying a deep recurrent auto encoder neural network to ensure robustness in automatic speech recognition system. Recently, Yu et al. [57] have proposed localization-based stereo speech segregation process in which, the generated soft time–frequency mask by using deep neural network is compared and proved as a better model than GMM/EM for the segregation process. Huang et al. have proposed a solution for monaural speech separation problem by jointly optimizing soft masking layer with deep recurrent neural network [15,16]. Zhao et al. [60] have introduced a combined approach that consists of deep neural network-based speech segregation followed by robust speaker identification module which is tested under various noisy and reverberant conditions. It is observed that the combined perceptual architecture helps to improve the speaker identification performance. Also, issues related to reverberation time and signal-to-noise ratio are efficiently addressed. Mowlae et al. [35], have proposed a joint system by combining speech separation modules and speaker identification to enhance intelligibility of automatic speaker recognition. Trowitzsch et al. [49] have suggested a systematic approach to improve the robustness of the classifier through multi-conditional training and also by super-imposing general environmental sounds.

The present study proposes two major contributions. At first, binaural classification-based speech segregation is carried out in which a total number of 83-dimensional features, such as 32-D interaural time difference, 32-D interaural phase difference, 16-D interaural level difference, 2-D interaural coherence and 1-D direct to reverberant ratio are considered. The concatenated above discussed resultant features are incorporated into deep recurrent neural network (DRNN)-based joint discriminative training classifier for the segregation of speech signals. The present work considers various performance evaluation metrics, such as source to interference ratio, source to distortion ratio and source to artifacts ratio for the validation of proposed model. Eventually, the obtained results are compared with the existing architectures, including deep neural network and observed better performance. Secondly, a spectro-temporal pattern extractor referred as Gabor–Hilbert envelope coefficients (GHEC) is proposed. The performance of GHEC is compared with existing monaural features using acoustic speaker models, such as GMM–UBM and *I*-vector. The results found that the joint architecture consists of binaural speech segregation followed by robust speaker rec-

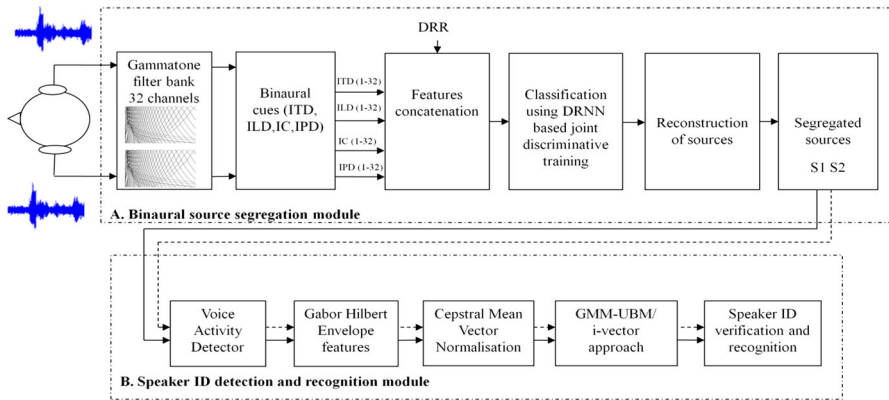


Fig. 1 Block diagram representation of binaural speech segregation along with automatic speaker recognition

ognizer that helps to improve the speech intelligibility, even in the presence of both noise and reverberation.

3 Model Architecture

Figure 1 shows the block diagram representation of proposed joint automatic speech signal segregation and recognition system.

The following section describes the various functionalities of each component presented in both binaural signal segregation and recognition modules.

3.1 Binaural Source Segregation Module

The predominant component of binaural speech segregation module is deep recurrent neural network classifier. The binaural cues, such as interaural time and level difference (ITD/ILD), interaural phase difference (IPD), and interaural coherence (IC), are first extracted from binaural auditory front-end. As an important contribution in this study, the DRR is also estimated from binaural signals through equalization–cancellation techniques [31] and combined with binaural cues. It is understood that these binaural cues are dependent on the various factors, including reverberation time, quality and energy of acoustic source, noises, obstacles and distance especially in an enclosed space. The resultant features are then incorporated into DRNN-based joint discriminative training model in order to generate soft mask.

3.1.1 Binaural Cues Extraction

The *basilar* membrane in *cochlea* is found to be responsible to segregate the acoustic signals on the basis of its frequencies in the human auditory system [34]. Gammatone filters are modelled as an inspiration of frequency selectivity and other functional prop-

erties of human *cochlea*. The speech signals arrived at two ears are decomposed into auditory channels ($N = 32$) by using fourth-order Gammatone filter bank followed by inner hair cell processing. Further, these phase-compensated filter banks are used to adjust the binaural features at common time intervals. The centre frequencies of the filter banks are equally spaced on the Equivalent Rectangular Bandwidth (ERB) scale between 80 Hz and 5 kHz [53,54].

The output of Gammatone filter bank is processed by computing half-wave rectification and square root compression for the transduction process in the inner hair cells. The auditory binaural features are processed by using a rectangular window of 20 ms at a sampling frequency of 44.1 KHz with an overlap of 50% between the successive frames at frame shift of 10 ms [53,54]. The estimation of interaural time and level differences are achieved by using normalized cross-correlation analysis in time domain and by calculating energy per frame, respectively. In general, interaural time difference (ITD) or interaural phase difference (IPD) deals with discrepancy in arrival times and phases at each ear at low level frequencies. They are sensitive to source distance, whereas ILD is considered as more robust at higher frequencies (higher than 1600 Hz) [50]. The estimated peak position for ITD across the time interval I between two ears is defined as,

$$\tau_i(t) = \arg \max C_i(t, \gamma) \quad (1)$$

where t is frame number, γ is time lag and normalized cross-correlation function [34] of channel C_i is given by,

$$C_i(t, \gamma) = \frac{\sum_{n=0}^{I-1} (s_i(t \frac{I}{2} - n) - \bar{s}_i) (k_i(t \frac{I}{2} - n - \gamma) - \bar{k}_i)}{\sqrt{\sum_{n=0}^{I-1} (s_i(t \frac{I}{2} - n) - \bar{s}_i)^2} \sqrt{\sum_{n=0}^{I-1} (k_i(t \frac{I}{2} - n - \gamma) - \bar{k}_i)^2}} \quad (2)$$

where \bar{k}_i and \bar{s}_i are the mean values of left and right ear signals, respectively. The comparison of energy arrived at two ears are used to derive interaural level differences, especially in a reverberant environment. The ILD estimation [34] across time interval, I between two ears are given by,

$$\text{ild}_i(t) = 20 \log_{10} \left(\frac{\sum_{n=0}^{I-1} k_i(t \frac{I}{2} - n)^2}{\sum_{n=0}^{I-1} s_i(t \frac{I}{2} - n)^2} \right) \quad (3)$$

Interaural coherence (IC) [3,50] is considered as a more salient feature for analysing similarity and strength of correlation between two ear canals.

$$\text{IC}(t) = \frac{\varnothing_{1,r}(\omega, t)}{\sqrt{\varnothing_{1,l}(\omega, t) \cdot \varnothing_{r,r}(\omega, t)}} \quad (4)$$

where $\varnothing_{r,r}(\omega, t)$, $\varnothing_{1,l}(\omega, t)$ represent the auto-power spectral densities (APSD) of the left and right ears, respectively.

$\varnothing_{l,r}(\omega, t)$ represents cross power spectral density (CPSD) of the two time-aligned input channels.

The IPD model [24, 50] provides high temporal resolution of robust binaural information. The interaural transfer function (ITF) is computed by using left–right pair of complex Gammatone filter outputs, such as $g_l(t)$ and $g_r(t)$. Computed ITF contains complex terms along with amplitude and phase information and it is given by,

$$\text{ITF}(t) = g_l(t) \cdot \overline{g_r(t)} = A_l(t) \cdot A_r(t) \cdot e^{j(\varnothing_l(t) - \varnothing_r(t))} \quad (5)$$

where $A_l(t)$ and $A_r(t)$ represent the amplitude information whereas $\varnothing_l(t)$, $\varnothing_r(t)$ represent phase information of ITF for left and right channels, respectively. The temporally smoothed IPD is obtained by using low-pass filtered ITF [24, 50] and it is given as,

$$\text{IPD}(t) = \arg([\text{ITF}(t)]_{\text{lp}}) \quad (6)$$

3.1.2 Direct-to-Reverberant Ratio (DRR)

The distance estimation of a sound source is closely associated with the energy ratio between direct and reverberant sound signals. It is understood that direct-to-reverberant ratio is dependent upon two factors, namely acoustic properties of the room and source to receiver configuration (i.e. distance and orientation). The DRR is considered as one of the most widely analysed parameters by many researchers for the estimation of distance between source and destination [30, 31]. Further, it is observed that DRR decreases with increasing distance between source and target and it is also affected by certain properties of the room, such as the reverberation time and room volume. It is described in terms of dB and defined as,

$$\text{DRR} = 10 \log_{10} \frac{\sum_{k=0}^{S_d} |h[K]|^2}{\sum_{k=S_d+1}^{\infty} |h[K]|^2} \quad (7)$$

where S_d represents the sample length of the direct sound arrival, and $h[k]$ represents the room impulse response.

3.1.3 DRNN-Based Joint Discriminative Training Classifier

The level of complexity in recognizing a target speaker in a reverberant environment depends on two important factors, namely number of target speakers and nature of noise sources [16, 60]. The computational goal of adopting deep learning model in this study is mainly to separate target speech source from input acoustic mixture. The concatenated features, such as binaural cues and direct-to-reverberant ratio are given as input features. The deep recurrent structure is characterized by temporal connections of recurrent neural network. The deep learning model is successfully processed to reform the magnitude spectra of output targets and predictions. The optimized deep learning structure with different layers is shown in Fig. 2.

The network parameters are updated by involving back-propagation through time (BPTT) method. The limited memory Broyden–Fletcher Goldfarb Shannon (L-BFGS)

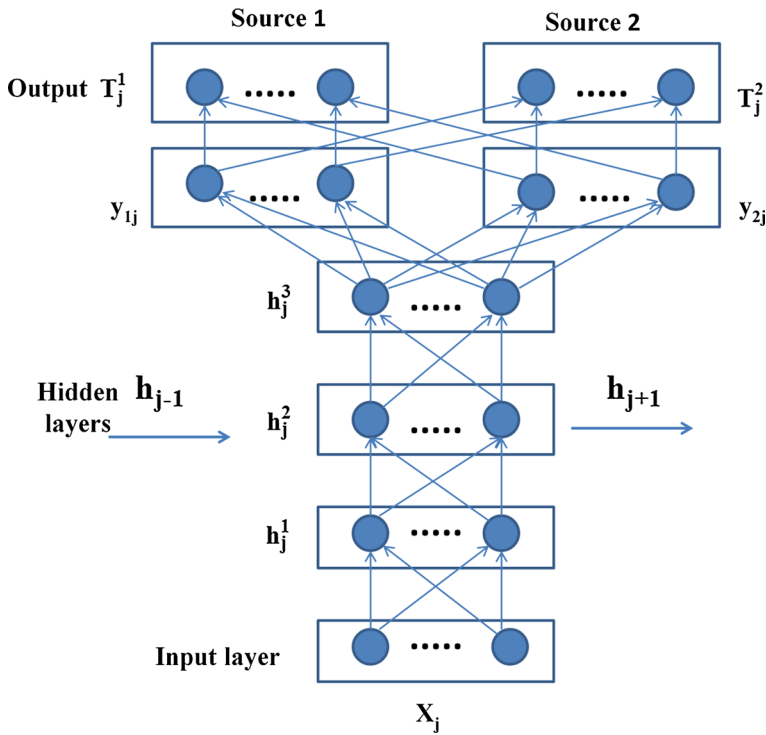


Fig. 2 Deep recurrent neural network structure with different layers

algorithm [15, 16] is processed to train the models at the time of training phase. The estimation of DRNN model parameters is carried out by using the error back-propagation algorithm with stochastic gradient learning in order to achieve state-of-the-art performance in neural network structure. The time–frequency masking function is integrated as one of the layers in neural network structure that reduced computational complexities. It is assumed an M intermediate layer with recurrent connection is presented at the k th layer. The hidden activation at this layer [16, 55] is calculated on the basis of current input at time, j by using Eq. (8).

$$\begin{aligned}
 h^k(x_j) &= f_h((x_j) \cdot h_{j-1}^k)) \\
 &= \gamma_k(U^k h_{j-1}^k + W^k \gamma_{k-1}(W^{k-1}(\dots \gamma_1(W^1 x_j))))
 \end{aligned}
 \tag{8}$$

where f_h represents a state transition function and x_j is given as input to the neural network at time j . U^k and W^k are the two weight matrices for the k th layer and recurrent connection at that layer, respectively. $\gamma_k(\cdot)$ is the element-wise nonlinear function in k th layer. The output function is given by,

$$y = f_o(h_i^k)
 \tag{9}$$

where f_o represents an output function.

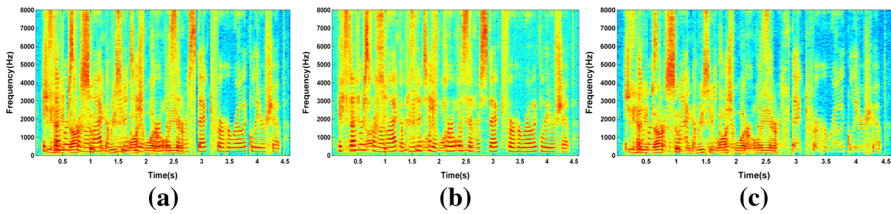


Fig. 3 Spectrogram of segregated speech sources using DRNN-based joint discriminate training. **a** Mixture of two sources; **b, c** segregated source signals

The soft time–frequency masking phase is carried out immediately after the training phase in order to improve intelligence in speech signal. The recent reviews reported that soft masking can be used to reduce the artifacts and also to improve smoothness in the predicted results. It is applied on the predicted magnitude spectrogram for the reconstruction which is followed by inverse short-term Fourier transform.

The soft time–frequency masking, $S_j(f)$ is given by

$$S_j(f) = \frac{|y_{1j}(f)|}{|y_{1j}(f)| + |y_{2j}(f)|} \quad (10)$$

where y_{1j} and y_{2j} represent the obtained output predictions. The soft masking is applied to the magnitude spectra, $T_j(f)$ of the original mixture signal in order to obtain segregated spectra, s'_1 and s'_2 which are given by,

$$\begin{aligned} s'_1(f) &= S_j(f) T_j(f) \\ s'_2(f) &= (1 - S_j(f)) T_j(f) \end{aligned} \quad (11)$$

The time-domain signal can be obtained by applying the inverse short-time Fourier transform. Further, the signal-to-interference ratio (SIR) can be improved by applying discriminative training criterions, such as mean squared error (MSE) and Kullback–Leibler (KL) divergence [15, 16]. DRNN-based speech segregation is an appropriate technique to facilitate dynamic temporal behaviour. The spectrograms of input mixture and segregated speech signals are shown in Fig. 3a–c, respectively.

3.2 Speaker Identification Module

The speaker identification process is initiated by using resultant input signals from the speech segregation module. The identification module includes three stages, namely feature extraction, speaker modelling and pattern classification-based decision making [23]. The present study proposes the Gabor filter banks for the extraction of monaural features. Further, it uses Gaussian mixture model-universal background model (GMM–UBM) and I -vector methods for the recognition of a speaker [44]. The training phase confines the distribution of extracted features by involving one or more types of statistical models. The unidentified utterances are then classified in the recognition phase on the basis of its similarities with the corresponding speaker model [23].

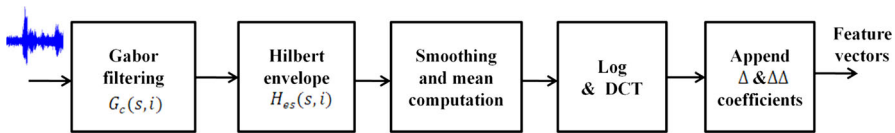


Fig. 4 The structure of proposed envelope coefficients extraction

3.2.1 Energy-Based Voice Activity Detector (VAD)

All speech samples are pre-processed by involving down sampling to 8 kHz, pre-emphasis and removing silence speech regions [44]. The energy of all input speech frames are calculated for a given speech utterance and then the empirical threshold is chosen from frame energies. The accurate discrimination of speech and non-speech regions is achieved by an energy-based detector that rejects frames if their energy decreases below a threshold value.

3.2.2 Gabor Hilbert Envelope Coefficients (GHEC)

The present study proposes a Gabor Hilbert Envelope Coefficients (GHEC) method in which Gabor filters are convolved with Hilbert Envelope and it is illustrated in Fig. 4. The characteristics of spectral, temporal and spectro-temporal components are extracted by using set of 41-Gabor filters. The feature extraction process uses local patches of log-Mel scaled spectrogram of 26-channels. Log-Mel spectrogram considers the basic qualities of human auditory system which includes resolution across entire frequencies and logarithmic intensity perception [45]. The extracted feature components are dependent upon the output of Gabor filters and its convolution with Hilbert envelope. The Hilbert envelopes perform the exact envelope of the auditory nerve response at particular centre frequencies. The output of Hilbert transform, $H_t(s, i)$ contains both real and transformed part which are used to obtain envelope, $H_e(s, i)$.

$$H_t(s, i) = G_c(s, i) + iG'_c(s, i) \quad (12)$$

where $G_c(s, i)$, $G'_c(s, i)$ are the real and Hilbert transformed signal, respectively, and i , is the imaginary unit. The Hilbert envelope [44], $H_e(s, i)$ is obtained by using Eq. (13).

$$H_e(s, i) = G_c(s, i)^2 + G'_c(s, i)^2 \quad (13)$$

The Hilbert envelope is smoothed by using low-pass filter with cut-off frequency of 20 Hz in order to remove redundant undesired higher frequencies. The smoothed envelope, $H_{es}(s, i)$ is grouped into 25 ms duration with a skip rate of 10 ms. Further, discontinuities at the edges of each frame are minimized by using a Hamming window. The sample means are estimated as,

$$M(t, i) = \frac{1}{N} \sum_{s=0}^{N-1} w(s) H_{es}(s, i) \quad (14)$$

where $w(s)$ is a Hamming window. The natural logarithm is applied on the estimated resultant parameter, $M(t, i)$ which is used here as a channel normalization factor in order to bring human perception of loudness as well as to compress the dynamic range [43]. In the final step, discrete cosine transform (DCT) is used to perform two functions, namely conversion of spectral features into cepstrum and also to de-correlate various over-lapped feature dimensions [44]. The first and second cepstral derivatives are calculated and appended to the features in order to capture various 57-dimensional dynamic patterns.

3.2.3 GMM–UBM Model-Based Speaker Verification and Identification

The speaker recognition is used to identify an individual person by analysing the spectral contents of his/her speech signal [23,41]. Generally, performance of this process is degraded when reverberation time increases. The source speech signal reaches the target after experiencing series of reflections and diffractions in a reverberant room environment. The training phase is involved with coefficients and the trained data is processed by using Gaussian mixture model with universal background model (GMM–UBM) [23,41] with 57-dimensional Gabor Hilbert envelope coefficient (GHEC) features. In the literature, the Mel-frequency Cepstral Coefficients (MFCCs) as well as Gabor Filter Bank Features (GBFB) [45,46] have been extensively applied in many speech signal processing applications, including speech, emotion and language recognition as well as for the speaker recognition. The performance of proposed Gabor Hilbert envelope coefficient (GHEC) is compared with various known existing methods, including Mel-frequency cepstral coefficient (MFCC) and Gammatone frequency cepstral coefficients (GFCC). The cepstral mean and variance normalization (CMVN) are applied on monaural features for adopting feature normalization in order to reduce the effect of channel influence and to increase the robustness of automatic speaker recognition systems. The Gaussian mixture model (GMM) [23,41] is considered as a stochastic model which comprised of weighted sum of M multivariate component Gaussian densities. For a D -dimensional feature vector, x , the Gaussian mixture model is referred by its probability density function and it is given by,

$$p(x|\lambda) = \sum_{i=1}^M w_i p_i(x) \quad (15)$$

where w_i and $p_i(x)$ denotes the mixture weights and component densities, respectively

$$p_i(x) = \frac{1}{(2\pi)^{\frac{D}{2}} |\sum_i|^{-\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (x - \mu_i)' \sum_i^{-1} (x - \mu_i) \right\}$$

The uni-modal Gaussian densities depend on the mean $D \times 1$ vector, μ_i and a $D \times D$ covariance matrix, \sum_i . The parameters of the density model are defined as $\lambda =$

$\{w_i, \mu_i, \sum_i\}$ and the mixture weights are required to satisfy the constraint $\sum_{i=1}^M w_i = 1$. In this study, 64 mixture component of universal background model (UBM) is trained with expectation-maximization (EM) algorithm. The mean vectors of universal background model are adapted by utilizing the relevance factor of 19. The speaker model is frequently trained by using maximum a Posteriori (MAP) method in order to promote the consistency. The mean vectors of universal background model (UBM) are concatenated into super vector for each speaker in the enrolment set and eventually a target speaker model is constructed [23,41].

In the recognition phase, log-likelihood ratio (LLR) score [23,41] for the given test feature vectors, X is estimated from two models, namely target (γ_{tar}) speaker model and universal background model (γ_{impost}) and the ratio is derived as,

$$\text{LLR}_{\text{score}} = \log (X|\gamma_{\text{tar}}) - \log (X|\gamma_{\text{impost}}) \quad (16)$$

where γ_{tar} is the utterances related to target speaker and γ_{impost} is the utterances that are not related to target speaker. The speaker verification is carried out to confirm whether a speech source signal can be accepted or rejected and it is mainly based on the decision threshold θ ,

$$\text{LLR}_{\text{score}} = \begin{cases} \geq \theta & \text{accepted} \\ < \theta & \text{rejected} \end{cases} \quad (17)$$

The normalization value is resulted from universal background model (UBM) which is done by shifting the log-likelihood scores obtained from various feature vectors. The score normalization is applied to reduce the score variability across different speakers and sessions. It improves the accuracy and also provides a common (speaker-independent) decision threshold value. In addition, the Z -normalization [23,41] is also carried out for the enhancement and it is given by,

$$\text{LLR}_{\text{norm}} = \frac{\text{LLR}_{\text{raw}} - \mu}{\sigma} \quad (18)$$

where μ and σ denotes mean and variance of imposter score of a speaker, respectively.

3.2.4 *I-Vector-Based Speaker Recognition System*

The experiment uses 57-dimensional Gabor Hilbert envelope features with the appended delta coefficients that are extracted as acoustic features from speech material for i -vector-based speaker verification methods. The low-dimensional representation of Gaussian mixture model (GMM) super vectors is referred as I -vector which was introduced very recently as a major refinement in existing speaker recognition system [9,10]. I -vector extraction along with Gaussian probabilistic linear discriminant analysis (GPLDA) has been experimentally proved as an enhanced and computationally efficient technique in comparison with conventional joint factor analysis (JFA) and support vector machine (SVM) [21]. In general, channel and session variability is referred as a mismatch between trained and test utterances which is induced by

various factors, including noise sources, variations in voice of the speaker and environmental conditions. The same can be compensated by involving various methods, such as within class covariance normalization (WCCN), linear discriminative analysis (LDA) and source-normalized weighted linear discriminant analysis (SN-WLDA) [22]. The joint factor analysis (JFA) [9, 10] is mainly based on decomposition of speaker-dependent Gaussian mixture super vector, k that consists of separate speaker- and channel-dependent components, S and C , respectively, and are given as,

$$k = S + C \quad (19)$$

where $S = m + Vy + Dz$; $C = Ux$; where, m is a session- and speaker-independent super vector extracted by using universal background model (UBM); x , y and z are the speaker- and session-dependent factors in their respective subspace. V and D specify the speaker subspace, whereas U represents session subspace.

I -vector model has shown significantly better performance, especially for short utterances (< 10 s). The total variability space simultaneously represents speaker and channel variability [21]. The speaker- and channel-dependent Gaussian mixture super vector in an I -vector-based speaker recognition, k is computed as,

$$k = m + Tw \quad (20)$$

where m is the session- and speaker-independent universal background model (UBM) super vector, T is a low-rank rectangular matrix representing the primary directions of variability across all development data and w denotes the independent normal distributed random vector with parameter $N(0, 1)$.

3.2.4.1 Within-Class Covariance Normalization Along with Linear Discriminant Analysis Within-class covariance normalization (WCCN) is used to compensate dimensions of the high within-class variance. The within-class covariance normalization (WCCN) additionally removes the dimensions of between-class variance while reducing the dimensions of within-class variability which is considered as a major demerit. This can be overcome by combining within-class covariance normalization (WCCN) along with linear discriminant analysis (LDA). The combined compensation of WCCN + LDA [21, 22] minimizes within-class variance as well as maximizes between-class variance and it is derived by following Eigen-value decomposition which is denoted as,

$$V_b v = \lambda V_w v \quad (21)$$

Linear discriminative analysis (LDA) [21] is computed by the usage of between-class variance (V_b) and within class variance (V_w), respectively, and it is given as,

$$\begin{aligned}
 V_b &= \sum_{s=1}^S n_s (\bar{\mu}_s - \bar{w})(\bar{\mu}_s - \bar{w})^T, \\
 V_w &= \sum_{s=1}^S \sum_{i=1}^{n_s} (w_i^s - \bar{\mu}_s)(w_i^s - \bar{\mu}_s)^T,
 \end{aligned} \tag{22}$$

where S is the total number of speakers, w_i^s denotes the i -vector representation of i session of speaker s and n_s is the number of utterances of speakers. The mean I -vectors, μ_s for each speaker and w is the global mean across which all speaker are specified as,

$$\begin{aligned}
 \bar{\mu}_s &= \frac{1}{n_s} \sum_{i=1}^{n_s} w_i^s \\
 \bar{w} &= \frac{1}{N} \sum_{s=1}^S \sum_{i=1}^{n_s} w_i^s
 \end{aligned} \tag{23}$$

where N is the total number of sessions. As described, linear discriminative analysis (LDA) is responsible for producing reduced set of axes A through Eigen-value decomposition whereas WCCN transformation matrix (B) is derived by using Cholesky decomposition, $BB^T = W^{-1}$ where W is computed by using,

$$W = \frac{1}{S} \sum_{s=1}^S \sum_{i=1}^{n_s} (A^T (w_i^s - \bar{w}_s))(A^T (w_i^s - \bar{w}_s))^T \tag{24}$$

The resultant WCCN [LDA] is obtained by computing,

$$\text{WCCN [LDA]} = B^T A^T W \tag{25}$$

3.2.4.2 Gaussian Probabilistic Linear Discriminant Analysis (GPLDA) Classifier In the literature, significant numbers of work have been presented the probabilistic linear discriminant analysis (PLDA)-based I -vector speaker recognition system by creating session and speaker variability within I -vector space, effectively. Recently, length-normalized Gaussian probabilistic linear discriminant analysis (GPLDA) approach is introduced that converts I -vector feature behaviour from heavy-tailed to Gaussian [21,22]. The Gaussian probabilistic linear discriminant analysis-based I -vector speaker recognition technique involves extraction of I -vector, session variability compensation, likelihood ratio scoring. The tested results of proposed technique with the baseline methods for TIMIT dataset [11] are given in Table 1. I -vectors are extracted for Gabor Hilbert envelope features by using front-end factor analysis. Gaussian probabilistic linear discriminant analysis (GPLDA) classifier is applied on channel compensated I -vector features [21,22]. The speaker- and channel-dependent length-normalized I -vector w can be defined as,

Table 1 Comparative analysis of joint factor analysis and i -vector-based techniques for different truncation of testing and training utterance under reverberant condition (RT = 0.38 s) are given below

Utterance size (training–testing)	JFA system		WCCN + LDA		GPLDA	
	EER%	DCF	EER%	DCF	EER%	DCF
Full(20 s)–2 s	20.08	0.064	18.53	0.067	19.73	0.068
Full(20 s)–5 s	15.95	0.054	13.70	0.056	14.00	0.060
Full(20 s)–10 s	13.36	0.047	11.43	0.045	11.46	0.049
Full(20 s)–12 s	10.02	0.042	10.54	0.040	9.43	0.043
Full–Full	5.37	0.032	5.56	0.033	5.26	0.034

$$w = \bar{w} + U_1x_1 + U_2x_2 + \gamma_r \quad (26)$$

where γ_r is speaker residuals with mean zero, U_1 and U_2 are the Eigen voice matrix and Eigen channel matrix, respectively.

The Gaussian probabilistic linear discriminant analysis (GPLDA) scoring [21] is computed using batch likelihood ratio which provides a ratio between two I -vectors of target and test speakers. It is calculated as,

$$\ln \frac{P(w_{\text{targ}}, w_{\text{test}} | H_1)}{P(w_{\text{targ}} | H_0)P(w_{\text{test}} | H_0)} \quad (27)$$

where H_1 the speakers are same, H_0 the speakers are different.

The equal error rate (EER) and detection cost function (DCF) are used as performance evaluation metrics [44]. The EER is obtained where false acceptance rate (FAR) and false rejection rate (FRR) are found to be equal. The Detection Cost Function is investigated by using weighted sum of the two error probabilities and it is defined as

$$\text{DCF} = C_{\text{miss}}E_{\text{miss}}P_{\text{target}} + C_{\text{FA}}E_{\text{FA}}(1 - P_{\text{target}}) \quad (28)$$

where $C_{\text{miss}} = 10$ and $C_{\text{FA}} = 1$ represents cost factors, $P_{\text{target}} = 0.01$ gives the probability of target and E_{miss} , E_{FA} denotes probability of miss and false alarm, respectively.

From Table 1, it is observed minimum change in the computed detection cost function for joint factor analysis and I -vector-based techniques. The equal error rate (EER) by following Within-Class Covariance Normalization (WCCN) along with LDA is found to be relatively better than other methods. The equal error rate (EER) of Gaussian probabilistic linear discriminant analysis produces better results than joint factor analysis. It should be noted that the purpose of involving compensation techniques is to promote efficiency in speaker discrimination and attenuate channel effects/variability. It is observed that the equal error rate increases as length of test utterance decreases.

4 Results and Discussions

The speech signals are convolved with impulse responses (BRIR) and are obtained from Aachen Impulse Response (AIR) database for different rooms [18] and from [24]. The study also uses impulse responses obtained from the University of Surrey [17, 19] in four different reverberation rooms (*A*, *B*, *C*, and *D*) for azimuths between -90° and 90° spaced by 5° at a distance of 1.5 m. The TIMIT database [11] is composed of high-quality, read speech collected from a total of 630 speakers (comprises of 192 female and 438 male). Each speaker supplies 10 short utterances, phonetically rich English language sentences and average quantity of speech available per speaker is 30 s. In this study, 9 short utterances are used for training and the remaining one utterance is used as a test sample. Similarly, 530 speakers are used for background model training, and the rest of 100 speakers are used as test samples. The NOIZEUS dataset [14] contains various noises which are utilized for the experiments. The binaural signals and deep learning algorithms are computed by using a workstation (ThinkStation P300) with an Intel Xeon (E3-1271) 3.6 GHz processor, 32 GB of RAM and dedicated NVIDIA (Quadro K620) graphics card. The software includes, MATLAB (R2015a) installed in Windows 7 operating system.

4.1 Module 1: Feature Extraction and Classification-Based Speech Segregation

In this module, the features of binaural and direct-to-reverberant ratio cues are chosen to generate soft time-frequency mask and also to handle issues during binaural source separation process. The concatenated mixture of these features is given as an input to the deep recurrent classifier, DRNN. Each layer in the typical deep neural network is further enhanced with temporal feedback loops in order to make existing network structure as deep recurrent neural network. In this study, the deep recurrent structure is implemented as three hidden layers of 1000 hidden units integrated with joint discriminative training criterion.

It is understood that the number of neural network parameters, such as weights and bias increase when the number of input feature dimensions increase. These network parameters are updated through back-propagation through time (BPTT) and the *epoch* is adjusted to 500. Each layer in the network is added with temporal context information; thus each network in the recurrent structure is updated with new information and travels up ensuring a hierarchical architecture. Each layer in the hierarchy is characterized by recurrent neural network. The performance analysis metric, mean square error (MSE) is computed for each feature vector of the network structure in order to produce better signal-to-noise ratio and also to optimize network parameters. The limited memory Broyden–Fletcher Goldfarb Shanon (L-BFGS) algorithm [15, 16] is considered during optimization stage to train the networks from random initialization. Further, long short-term memory (LSTM) optimizer is explored in the recurrent structure that creates the possibility to store and callback temporal information over time to handle vanishing gradient problem.

It should be noted that better combination of concatenated features are selected by estimating output signal-to-noise ratio and HIT-FA (success-false alarm rate) [11].

The correctly identified speech-dominant time–frequency (T–F) units represent HIT rate and the wrongly classified noise-dominant T–F units represent false rate (FA). The binaural cues obtained from binaural auditory front-end are combined along with direct-to-reverberant ratio. The various combinations of resultant concatenated auditory features are validated by estimating output signal-to-noise ratio as well as HIT-FA rate. It is presumed that better combination of binaural features can further improve the performance of deep recurrent neural network-based speech segregation process.

All the segregation-related experiments are carried out by analysing various impulse responses associated with four different reverberation rooms, i.e. *A*, *B*, *C* and *D* [22, 34]. The babble noise is considered as a noise source as it is known for its higher efficiency, especially in speech masking-related applications. The babble noise is spread across the speech spectrum in azimuth between -90° and 90° spaced by 5° at a distance of 1.5 m and used to train deep recurrent neural network. An untrained interference angle of 15° is considered for the testing in all the experiments. The results from classification-based speech segregation analysis are shown in Tables from 2 to 4 which performed under reverberation time (T_{60}) at 0.32 s.

The estimation of signal-to-noise ratio is considered as one of the popular evaluation metrics that expresses the performance of source segregation system [60] and it is given by,

$$\text{SNR} = 10 \log_{10} \left[\frac{\sum_m x^2(m)}{\sum_m [\hat{x}(m) - x(m)]^2} \right] \quad (29)$$

where $\hat{x}(m)$ is the estimated target signal, $x(m)$ is the target signal. The estimation of HIT-FA is not only considered as a best evaluation criteria and also it is widely used to correlate with human speech intelligibility.

The performance of concatenated binaural cues and direct-to-reverberant ratio in the segregation process are validated by estimating output signal-to-noise ratio and the results are shown in Table 2. It is observed that the classifier-based segregation process produces better results as when the dimensionalities of interaural level difference increases. The effect of reverberation of rooms and noise play the major role than the output signal-to-noise ratio in the performance of segregation process. It is found that the computational complexities further increased when the dimensionality of interaural level difference increases than 16 dimensions. The study observes lower HIT-FA rate when 16-dimensional interaural level differences are used. A non-linear behaviour is observed between interaural time difference and better output signal-to-noise ratio. The better output HIT-FA rate and output signal-to-noise ratio are observed for the combination of 32-dimensional interaural time and phase differences and 16-dimensional level differences along with direct-to-reverberant ratio. The computational time increases when dimensions of interaural level difference increases above 16-D. It is observed that minimum change in output SNR when the dimension of interaural coherence is increased and it is chosen as 2-D interaural coherence (IC) as one of the concatenated features.

The next step involves concatenation of total number of 83-dimensional features that include four binaural and direct-to-reverberant ratio cues. These combined feature cues

Table 2 Performance of combined binaural features along with direct-to-reverberant ratio-based source segregation under babble noise at -5 dB

Features	HIT (%)	FA (%)	HIT-FA (%)	Output SNR (dB)
32D-ITD + 32D-IPD + 16D-ILD + 2D-IC + 1D-DRR	88	2	86	10.67
32D-ITD + 16D-IPD + 8D-ILD + 2D-IC + 1D-DRR	86	3.5	84.5	9.64
32D-ITD + 16D-IPD + 4D-ILD + 2D-IC + 1D-DRR	86	4.5	82.5	9.57

are then incorporated into deep recurrent neural network-based joint discriminative training model.

The training of deep recurrent neural network is done by using randomly chosen 100 speakers from TIMIT database which has concatenated 9 sentences for each speaker. Eventually, testing is carried out with an advent of unused 1 sentence. The efficiency of deep recurrent neural network is evaluated by estimating three performance metrics, namely source to interference ratio (SIR), source to artifacts ratio (SAR) and source to distortion ratio (SDR).

Source to distortion ratio (SDR): It is referred as the ratio between target source and the difference between estimated and target source signals. The higher SDR denotes the better performance [15], [16], [57].

$$\text{Source to distortion ratio (SDR)} = 10 \log_{10} \frac{\|S_{\text{tar}}\|^2}{\|e_{\text{intf}} + e_{\text{noise}} + e_{\text{artif}}\|^2} \quad (30)$$

where S_{tar} denotes target source signal, e_{intf} denotes interferences from other sources, e_{noise} denotes deformation caused by the noise, e_{artif} denotes artifacts established by the separation algorithm.

Source to interference ratio (SIR): It carries the information about errors caused by failures during the interfering signal elimination process [38], [16], [57] and it is specified as,

$$\text{Source to interference ratio (SIR)} = 10 \log_{10} \frac{\|S_{\text{tar}}\|^2}{\|e_{\text{intf}}\|^2} \quad (31)$$

Source to artifacts ratio (SAR): It informs about errors due to extraneous artifacts that established at the time of source segregation process [38], [16], [57] and it is defined as,

$$\text{Source to artifacts ratio (SAR)} = 10 \log_{10} \frac{\|S_{\text{tar}} + e_{\text{intf}} + e_{\text{noise}}\|^2}{\|e_{\text{artif}}\|^2} \quad (32)$$

In this study, both masking and discriminative training model are convolved with deep recurrent neural network in order to improve robustness and the results are shown in Table 3. The analysis is carried out by using 83-dimensional (32D ITD + 32D IPD + 16D ILD + 2D IC + 1D DRR) cues at a reverberation time of 0.32 s in the presence of babble noise at 20 dB. It is observed that the addition of soft masking within a layered architecture of network produces better performance and it is evaluated in terms of SDR, SIR and SAR. The deep recurrent network is also validated with and without discriminative training model. From the results, it is observed that the model with discriminative training outperforms other structures. The performance metrics have observed higher values for all other structures than deep neural network (DNN) model. The temporal hierarchy of deep recurrent network plays a major role in producing better performance. Further, the improved performance is mainly due to back-propagation of gradients with respect to training objectives that provide optimized structure for the model. The present study attempts to further improve existing deep recurrent network

Table 3 Performance of DRNN with and without masking and discriminative training model

Models	Male versus female			Female versus female		
	SDR (dB)	SIR (dB)	SAR (dB)	SDR (dB)	SIR (dB)	SAR (dB)
DRNN with joint masking	8.12	11.04	9.93	5.77	8.02	7.53
DRNN without joint masking	7.73	9.58	9.02	4.36	6.02	8.03
DRNN with discriminative training	8.73	12.15	10.32	5.83	8.05	6.16
DRNN without discriminative training	8.54	11.46	8.57	4.74	8.57	7.34
DRNN with joint discriminative training	11.07	16.35	13.90	8.04	9.42	8.15
DNN	7.52	8.93	8.72	5.02	7.53	7.02

Table 4 Performance of classification-based source separation for the different input SNR values

Input SNR (dB)	HIT (%)	FA (%)	HIT-FA (%)	Output SNR (dB)
– 15	76	0.6	75.4	2.3
– 10	83.2	0.8	82.4	6.0
0	87.7	3	84.7	11.7
10	90.2	5.2	85	13.9
20	92	6	86	16.3

model by integrating both masking as well as discriminative model criterion within layered structure. The results obtained observe better performance metrics than any other existing models. Obviously, the addition of soft time–frequency (T–F) mask as an internal layer along with exploration of mean square error (MSE) is believed as the main reason for the better results.

The source segregation process is carried out by incorporating the features, total number of 83-dimensional features (32D ITD + 32D IPD + 16D ILD + 2D IC + 1D DRR) into deep recurrent neural network-based joint discriminative model. The performance is validated by estimating input signal-to-noise ratio in the presence of babble noise and the results are shown in Table 4. It is experimentally observed that the higher-dimensional interaural level difference above 16 dimensions not merely increases the complexity but also the stability of the system by consuming more time. Also, output signal-to-noise ratio decreases when dimensionality of interaural time difference/interaural phase difference decreases from 32 dimensions. For the obvious reason, the system produces better output signal-to-noise ratio when input SNR increases.

The source segregation process is performed in four different reverberant rooms in the presence of babble noise at 0 dB and the results are shown in Table 5. The system produces better signal-to-noise ratio and HIT-FA rate for the rooms which has lower reverberant time period. However, the performance of classifier-based source segre-

Table 5 Performance analysis of binaural-DRR combined source segregation process under babble noise at 0 dB

Rooms	HIT (%)	FA (%)	HIT-FA (%)	Output SNR (dB)
A (0.32 s)	87.5	7.4	80	10.8
B (0.47 s)	76.4	5.1	71.3	7.9
C (0.68 s)	77.9	5.5	72.4	9.3
D (0.89 s)	81.27	11.4	69.87	8.4

gation system is not significantly affected by increasing values of reverberation time periods. In other words, the system is observed with significantly good performance metrics for all the rooms with various reverberant time periods even at 0 dB.

4.1.1 Comparison Analysis with Baseline Method

In this study, the developed speech separation framework is compared with results of [4, 33, 54]. Woodruff and Wang [54] have proposed binaural detection, localization and segregation of speech source that mainly depend on pitch and azimuth cues. Also, the system uses hidden Markov model for the estimation of number of active sources across time. Mandel et al. [33] have suggested model-based EM source separation and localization (MESSL) system. Here, a mask is generated and successfully utilized for the separation of a desired sound source from stereo signals. Every source in a mixture has been illustrated by a probabilistic model of interaural parameters. Probabilistic model of interaural parameters is evaluated at each spectrogram points, independently. Alinaghi et al. [4] have investigated the strength and weakness of mixing vector estimation along with other techniques, such as interaural level and phase differences (ILD and IPD) for the separation of stereo speech signals.

The comparative analysis is carried out by considering various rooms at different reverberant time periods in the presence of babble noise at -5 dB. It is understood that the incorporation of combined binaural cues along with direct to reverberant features into deep recurrent-based joint discriminative model show a better performance and validated in terms of output signal-to-noise ratio than Woodruff–Wang and MESSL models which is shown in Table 6. The results are found to be higher in all four rooms, invariably. It is observed that the proposed model produces 2.57 times better results than Woodruff–Wang model in Room A which has reverberation time period of 0.32 s. The proposed system shows improved signal-to-noise ratio (SNR) at an average of 0.58 dB over Woodruff and Wang [54] and Alinaghi et al. [4] models in Room D.

4.2 Module 2: Speaker Identification and Recognition Module

In this study, total of 630 speakers (192 female, 438 male) from TIMIT database are chosen in which 9 short utterances are used for training and the remaining one utterance is used as a test sample. Nearly, 530 speakers are used for background model training and the rest of 100 speakers are used as test samples. The performance of proposed

Table 6 Comparative analysis with baseline methods

Model	Output SNR in dB at Room A (0.32 s)	Output SNR in dB at Room B (0.47 s)	Output SNR in dB at Room C (0.68 s)	Output SNR in dB at Room D (0.89 s)
This work	4.50	2.25	1.27	1.05
Alinaghi et al. [4]	1.93	1.64	1.02	0.68
Woodruff and Wang [54]	1.75	1.50	0.95	0.65
Mandel et al. [33]	1.52	0.90	0.60	0.55

Gabor Hilbert envelope coefficient (GHEC) is compared with other standard feature extraction techniques, namely Gammatone Frequency Cepstral Coefficients (GFCC) [60], Mel-Frequency Cepstral Coefficients (RASTA-MFCC), Mean Hilbert Envelope Coefficients (MHEC) [43] and Gabor Filter Bank (GBFB) [45,46] at various SNR values with different noise sources and reverberant room environments. The Gammatone Frequency Cepstral Coefficients (GFCC) feature extraction [59,60] is done by using a total number of 64 channel Gammatone filter banks with central frequencies ranges from 50 Hz to 8 KHz. The outputs from rectified filter are decimated into 100 Hz that yields time frames of 10 milli-seconds. The magnitudes of the decimated outputs are then compressed by a cubic root operation in order to minimize the loudness. The resultant matrix represents time–frequency (T–F) decomposition of the input and referred as GF (Gammatone Frequency) components which are correlated with each other. The discrete cosine transform (DCT) is applied on GF components to reduce dimensionality and also to de-correlate the components.

The Mel coefficients are computed by segregating input signal into 20 ms frames with 10 ms frame shift. Each frame is applied with hamming window and also short-time Fourier transform is utilized to derive power spectrum. Then, the derived power spectrum is converted into Mel scale. Finally, 39-dimensional coefficients are obtained by applying log compression and also discrete cosine transforms (DCT). The discrete cosine transform is used to perform two functions, such as conversion of spectral features into *cepstrum* and also de-correlation of various over-lapped feature dimensions. Appending RASTA filtering after DCT yields 39-dimensional RASTA-MFCC coefficients. The mean Hilbert envelope coefficients (MHEC) feature extraction [43] is performed by using 24 channels Gammatone filter banks with centre frequencies spaced on Equivalent Rectangular Bandwidth (ERB) scale between 300 and 3400 Hz which are utilized in order to decompose the speech signal into 24 bands.

The Hilbert envelope [43] is computed in addition to mean computation and smoothing. Then, first and second derivatives are computed and appended to the features in order to construct final 36-dimensional mean Hilbert envelope coefficient (MHEC) feature patterns. In this study, accuracy measures of different features are tabulated by using *I*-vector technique, which is used as an acoustic speaker model. The performance overview of various feature extraction techniques under different noisy environments in the presence of various signal-noise ratio (SNR) values are shown in Table 7.

Table 7 Accuracy results of speaker recognition system in various noisy conditions

	– 5 dBA	0 dBA	5 dBA	10 dBA	20 dBA	Clean	Average
<i>Babble noise</i>							
GHEC (proposed)	27.25	74.63	87.02	90.14	96.78	98.00	78.97
RASTA–MFCC	23.54	66.38	68.54	88.03	94.37	96.20	72.84
GFCC	27.00	70.12	80.32	92.65	95.50	98.16	77.29
MHEC	27.53	75.04	86.53	92.34	96.38	99.00	79.47
GBFB	26.50	65.67	73.43	90.37	97.00	98.32	75.21
<i>Street noise</i>							
GHEC (proposed)	33.73	70.52	75.29	86.93	90.10	97.27	75.64
RASTA–MFCC	25.07	60.72	81.75	85.53	87.95	95.76	72.79
GFCC	32.26	64.35	83.03	87.62	89.73	96.72	75.62
MHEC	35.37	71.55	73.47	90.87	92.19	98.57	77.00
GBFB	32.97	63.04	72.08	88.56	87.80	97.60	73.67
<i>Car noise</i>							
GHEC (proposed)	40.74	71.03	84.79	88.42	92.44	96.02	78.90
RASTA–MFCC	35.07	67.09	74.45	79.32	81.39	95.13	72.07
GFCC	35.53	68.15	80.03	89.07	92.14	96.13	76.84
MHEC	36.15	70.34	79.35	88.64	91.24	94.47	76.69
GBFB	39.53	69.27	80.35	86.06	92.32	94.06	76.93
<i>Airport noise</i>							
GHEC (proposed)	45.26	61.42	73.57	87.90	94.92	97.28	76.72
RASTA–MFCC	36.92	50.12	71.56	80.14	90.33	96.34	70.90
GFCC	42.53	68.52	74.05	82.16	92.15	98.34	76.29
MHEC	45.84	70.46	81.07	84.25	95.56	96.90	79.01
GBFB	41.53	71.56	75.63	85.09	94.12	97.94	77.64
<i>Exhibition noise</i>							
GHEC (proposed)	39.27	62.04	70.47	76.46	90.30	95.83	72.39
RASTA–MFCC	27.24	62.43	71.50	72.33	91.17	95.35	70.00
GFCC	27.18	60.15	69.14	74.12	85.46	96.92	68.82
MHEC	30.54	61.36	75.63	80.03	94.76	97.02	73.22
GBFB	25.72	59.62	65.69	73.48	88.15	92.48	67.52

The accuracy of speaker identification (SID) has been evaluated for the various feature extraction techniques, including the proposed Gabor–Hilbert envelope coefficients (GHEC) which is computed by convolving the Gabor filtered components with Hilbert envelope. The Gabor Filter is well documented for its spectro-temporal patterns that find similarity with certain brain-cortex neurons. The performance of Gabor–Hilbert Envelope Coefficients (GHEC) is compared with other well-known techniques, including RASTA-MFCC, GFCC, MHEC and GBFB and the results are shown in Table 7. It is observed that the performance of speaker identification by involving GHEC (57-dimensional features) outperforms RASTA-MFCC (36-dimensional fea-

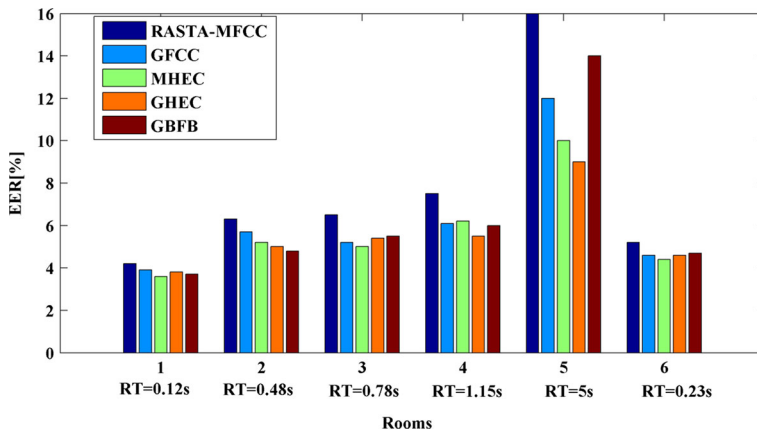


Fig. 5 Equal error rate (EER) for different features paradigm in the presence of babble noise at 20 dB under various reverberant conditions

tures), significantly for all the noise acoustic signals under various SNR values. Also, it produces comparatively better results than GBFB along with Principal Component Analysis (PCA) [28] that has totally 39-dimensional features for the various noise signals especially with the SNR value of -5 dB. The performance results are found to be almost similar for Gabor Hilbert envelope coefficients (GHEC), mean Hilbert envelope coefficients (MHEC) and Gammatone frequency cepstral coefficients (GFCC) for the various noise signals especially with low SNR values.

The computation of Gabor–Hilbert envelope coefficients (GHEC) is based on spectro-temporal cues extraction and log-Mel spectrograms which is believed as a major reason for its robustness as well as better performance. Further, 2-D Gabor filters are high receptive to amplitude and frequency modulation. It should be noted that mean Hilbert envelope coefficients (MHEC) and Gammatone frequency cepstral coefficients (GFCC) are computed by using 36- and 31-dimensional features, respectively. The equal error rate (EER) is also estimated in order to validate the performance of various feature extraction techniques under different reverberant conditions and the results are shown in Fig. 5. The equal error rate is computed by involving *I*-vector-based techniques. It should be noted that both Gabor Hilbert envelope coefficients (GHEC), mean Hilbert envelope coefficients (MHEC) and Gammatone frequency cepstral coefficients (GFCC) show better performance in higher reverberant conditions than other methods. It is observed that the performance of *cepstral* coefficients-based methods decreases as when reverberation increases. However, it is also observed that the proposed GHEC monaural feature shows less performance than MHEC in few cases. The reason is possibly due to its sensitivity towards intrinsic factors, such as speaking rate, speaking effort, style and pitch information.

In this study, text-independent Gaussian mixture model-universal background model (GMM–UBM) as well as *I*-vector-based speaker recognition are validated in the presence of factory noise. The feature extraction includes both mean Hilbert envelope coefficients and Gabor Hilbert envelope coefficients techniques. It should be noted that the GMM–UBM and *I*-vector-based speaker recognition system utilize

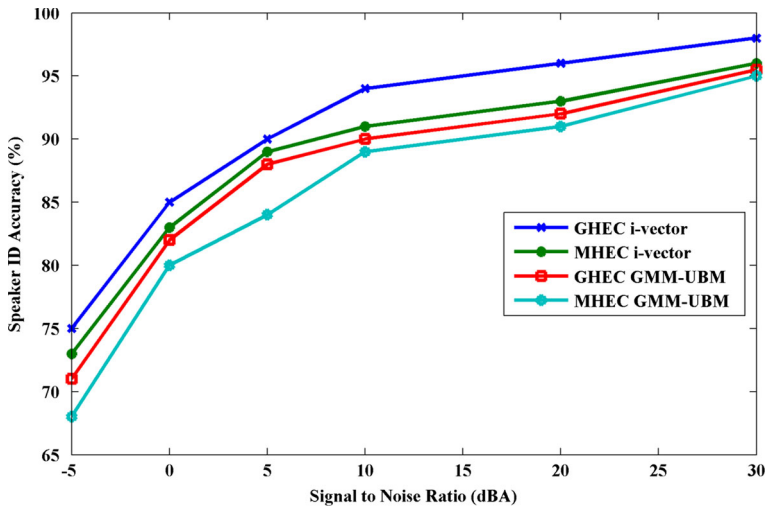


Fig. 6 Performance measure of GHEC and MHEC during joint approach under different acoustic speaker models with single noise source (factory)

a total of 64 Gaussian mixture components for training speech samples. From the results, it is observed that the Gabor Hilbert envelope coefficient (GHEC) shows better performance than mean Hilbert envelope coefficients in joint approach for factory noise at reverberation time of 0.23 s. The above discussed comparison is tested in different SNR values which are shown in Fig. 6.

The joint approach for simultaneous speech segregation and automatic speaker recognition described here may be incorporated in real-time self-autonomous robots. Further, the proposed GHEC monaural feature will be utilized for monaural sound source separation to segregate discriminate information in a multi-talker environment.

The experiment considers randomly chosen 100 speakers from TIMIT database. The classification results are evaluated when a mixture of two speakers are given as input to the proposed system. The performance of joint binaural speech segregation and automatic speaker recognition is evaluated at different reverberant conditions which are shown in Fig. 7. The proposed system uses LSTM-DRNN-based binaural speech segregation techniques that are observed to be a main reason to produce better evaluation metrics, such as SDR, SIR and SAR by achieving an average of 5 dB in room with reverberation time of 0.89 s than other segregation methods, including DNN. Specifically, LSTM-DRNN uses memory blocks to control the information flow at multiple time scales and it is considered as one of the reasons to yield improved speaker recognition performance even at low signal-to-noise ratio. The proposed binaural speech segregation technique uses optimized deep recurrent structure that outperforms DNN-based speech segregation by employing LSTM-DRNN in the architecture when large numbers of speakers are involved in the training. Spatial cues such as IC, IPD, ILD, and ITD are incorporated into optimized deep recurrent structure for binaural classification to enhance robustness in higher reverberation time. The accuracy, sensitivity and specificity [39] for the proposed system are calculated in terms of confusion matrix and the results (in %) are shown in Table 8.

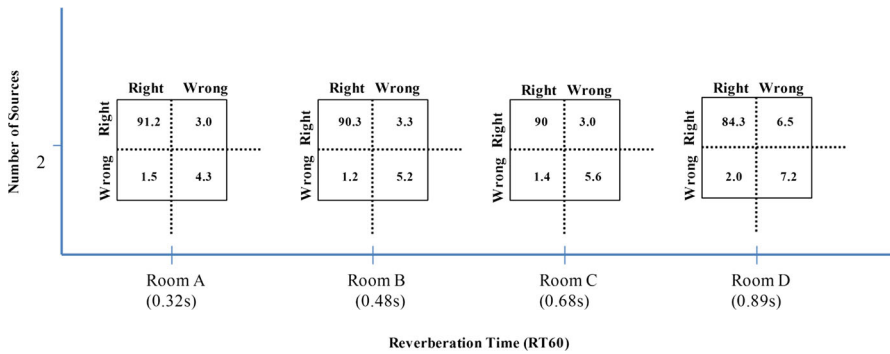


Fig. 7 Performance measure of joint approach for binaural speech segregation and automatic speaker identification in the different reverberant environments

Table 8 Classification results (in %) for the proposed system in the presence of different reverberation conditions

Rooms	Accuracy	Sensitivity	Specificity
A (0.32 s)	95.5	96.81	74.13
B (0.47 s)	95.5	96.47	81.25
C (0.68 s)	95.6	94.14	80
D (0.89 s)	91.5	92.13	78.2

5 Conclusions

The developed framework demonstrates a combination of two modules, namely speech segregation and speaker recognition. The concatenated acoustic cues, such as interaural time and level difference, interaural phase difference, interaural coherence, direct-to-reverberant ratio are successfully incorporated into deep recurrent structure-based joint discriminative model for the separation of input speech mixture. The experiments are carried out by considering different dimensions of spatial cues with deep learning structure. The long short-term memory (LSTM) optimizer is explored to avoid vanishing gradient problem by introducing memory blocks in the recurrent architecture. The proposed model has addressed binaural speech segregation since most real-time applications find speech and interfering sources are located at different positions. The system also proposes a monaural feature, referred as Gabor Hilbert envelope coefficients for speaker recognition system which is found to be robust towards extrinsic variations and implemented by applying temporal envelope extraction. The Hilbert envelope is performed to produce slow varying amplitude modulations in narrow frequency bands which contain spectro-temporal acoustic information. The performance of proposed feature extraction in joint approach is found to be better than other known existing techniques. The spatial cues used for binaural speech segregation are found to be robust even in higher reverberation periods than monaural speech segregation. The proposed deep learning-based binaural speech segregation produced signal-to-distortion ratio (SDR) at an average of 5.85 dB in higher reverberation time (0.89 s) over other accepted conventional models. The proposed monaural feature accuracy is com-

puted by using different speaker models under various noisy reverberant conditions. The system can help to recognize speakers in a multi-talker environment even in the absence of visual information. The developed framework finds several acoustic-related applications, such as intelligent hearing aid devices, hands-free communication, voice interactive systems and audio surveillance. The future work considers optimization of GHEC towards intrinsic variations and also incorporation of combined features of proposed spectral features along with relevant spatial cues in a binaural classification framework. The real-time implementation of the developed algorithm in an automated system has been considered, seriously.

Acknowledgements The authors wish to thank Department of Science and Technology for awarding a project under Cognitive Science Initiative Programme (DST File No. SR/CSI/09/2011) through which the work has been implemented. Also, authors are very much grateful to the anonymous reviewers for their valuable and constructive suggestions that improved the quality of the manuscript.

References

1. R. Abdipour, A. Akbari, M. Rahmani, B. Nasersharif, Binaural source separation based on spatial cues and maximum likelihood model adaptation. *Digit. Signal Proc.* **36**, 174–183 (2015)
2. A.K.H. Al-Ali, D. Dean, B. Senadji, V. Chandran, G.R. Naik, Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions. *IEEE Access* **99**, 1–1 (2017). <https://doi.org/10.1109/ACCESS.2017.2728801>
3. A. Alinaghi, W. Wang, P. J. B. Jackson, Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation, in *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Process. (ICASSP)* (2013), pp. 684–688
4. A. Alinaghi, P.J. Jackson, Q. Liu, W. Wang, Joint mixing vector and binaural model based stereo source separation. *IEEE ACM Trans. Audio Speech Lang. Process. (TASLP)* **22**(9), 1434–1448 (2014)
5. X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, Speaker diarization: a review of recent research. *IEEE Trans. Audio Speech Lang. Process.* **20**(2), 356–370 (2012)
6. F. Asano, H. Asoh, K. Nakadai, Sound source localization using joint Bayesian estimation with a hierarchical noise model. *IEEE Trans. Audio Speech Lang. Process.* **21**(9), 1953–1965 (2013)
7. A. Bednar, F.M. Boland, E.C. Lalor, Different spatio-temporal electroencephalography features drive the successful decoding of binaural and monaural cues for sound localization. *Eur. J. Neurosci.* **45**(5), 679–689 (2017)
8. J. Chen, Y. Wang, D.L. Wang, A feature study for classification-based speech separation at low signal-to-noise ratios. *IEEE ACM Trans. Audio Speech Lang. Process.* **22**(12), 1993–2002 (2014)
9. N. Dehak, R. Dehak, J. Glass, D. Reynolds, P. Kenny, Cosine similarity scoring without score normalization techniques. in *Proceedings of Odyssey 2010 - The Speaker and Language Recognition Workshop (Odyssey, 2010)*, pp. 71–75
10. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* **19**(4), 788–798 (2011)
11. J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus* (Linguistic Data Consortium, Philadelphia, 1993)
12. K. Han, D.L. Wang, A classification based approach to speech segregation. *J. Acoust. Soc. Am.* **132**(5), 3475–3483 (2012)
13. Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, Y. Haneda, Estimating direct-to-reverberant energy ratio using D/R spatial correlation matrix model. *IEEE Trans. Audio Speech Lang. Process.* **19**(8), 2374–2384 (2011)
14. Y. Hu, P. Loizou, Subjective evaluation and comparison of speech enhancement algorithms. *Speech Commun.* **49**, 588–601 (2007)
15. P.S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smaragdis, Deep learning for monaural speech separation, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2014), pp. 1562–1566

16. P.S. Huang, M. Kim, M. Hasegawa-johnson, P. Smaragdis, Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE ACM Trans. Audio Speech Lang. Process.* **23**(12), 1–12 (2015)
17. C. Hummersone, R. Mason, T. Brookes, Dynamic precedence effect modeling for source separation in reverberant environments. *IEEE Trans. Audio Speech Lang. Process.* **18**(7), 1867–1871 (2010)
18. M. Jeub, M. Schäfer, P. Vary, A binaural room impulse response database for the evaluation of dereverberation algorithms, in *Proceedings of International Conference on Digital Signal Processing (DSP)* (2009), pp. 1–4
19. Y. Jiang, D.L. Wang, R.S. Liu, Z.M. Feng, Binaural classification for reverberant speech segregation using deep neural networks. *IEEE Trans. Audio Speech Lang. Process.* **22**(12), 2112–2121 (2014)
20. Z. Jin, D.L. Wang, A supervised Learning Approach to monaural segregation of reverberant speech. *IEEE Trans. Audio Speech Lang. Process.* **17**(4), 625–638 (2009)
21. A. Kanagasundaram, R. Vogt, D.B. Dean, S. Sridharan, M.W. Mason, I-vector based speaker recognition on short utterances, in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (ISCA)* (2011), pp. 2341–2344
22. A. Kanagasundaram, D. Dean, S. Sridharan, R. Vogt, I-vector based speaker recognition using advanced channel compensation technique. *Comput. Speech Lang.* **28**(1), 121–140 (2014)
23. T. Kinnunen, H. Li, An overview of text-independent speaker recognition: from features to supervectors. *Speech Commun.* **52**(1), 12–40 (2010)
24. A. Kohlrausch, J. Braasch, D. Kolossa, J. Blauert, *The Technology of Binaural Listening* (Springer, Berlin, 2013)
25. G. Kovács, L. Tóth, Dirk Van Compernelle, selection and enhancement of Gabor filters for automatic speech recognition. *Int. J. Speech Technol.* **18**(1), 1–16 (2014)
26. M. Kuster, Estimating the direct-to-reverberant energy ratio from the coherence between coincident pressure and particle velocity. *J. Acoust. Soc. Am.* **130**(6), 3781–3787 (2011)
27. S.M. Lajevardi, Z.M. Hussain, Automatic facial expression recognition: feature extraction and selection. *SIVIP* **6**(1), 159–169 (2012)
28. H. Lei, B.T. Meyer, N. Mirghafori, Spectro-temporal Gabor features for speaker recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2012), pp. 4241–4244
29. J. Li, L. Deng, Y. Gong, R. Haeb-Umbach, An overview of Noise–Robust automatic speech recognition. *IEEE ACM Trans. Audio Speech Lang. Process.* **22**(4), 745–777 (2014)
30. Y.C. Lu, M. Cooke, Binaural distance perception based on direct-to-reverberant energy Ratio, in *Proceedings of International Workshop on Acoust. Echo and Noise Control*, 2008, pp. 1793–1805
31. Y.C. Lu, M. Cooke, Binaural estimation of sound source distance via the direct reverberant energy ratio for static and moving sources. *IEEE Trans. Audio Speech Lang. Process.* **18**(7), 793–1805 (2010)
32. A. Maas, Q.V. Le, T.M. O’Neil, O. Vinyals, P. Nguyen, A.Y. Ng, Recurrent neural networks for noise reduction in robust ASR, in *Proceedings of 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (2012), pp. 22–25
33. M.I. Mandel, S. Bressler, B. Shinn-Cunningham, D.P.W. Ellis, Evaluating source separation algorithms with reverberant speech. *IEEE Trans. Audio Speech Lang. Process.* **18**(7), 1872–1883 (2010)
34. T. May, S. Van de Par, A. Kohlrausch, 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 (2012)
35. P. Mowlaee, R. Saeidi, M.G. Christensen, Z.H. Tan, T. Kinnunen, P. Franti, S.H. Jensen, A joint approach for single-channel speaker identification and speech separation. *IEEE Trans. Audio Speech Lang. Process.* **20**(9), 2586–2601 (2012)
36. G.R. Naik, Measure of quality of source separation for sub-and super-Gaussian audio mixtures. *Informatica* **23**(4), 581–599 (2012)
37. G.R. Naik, W. Wang, Audio analysis of statistically instantaneous signals with mixed Gaussian probability distributions. *Int. J. Electron.* **99**(10), 1333–1350 (2012)
38. G.R. Naik, W. Wang, *Blind Source Separation: Advances in Theory, Algorithms and Applications* (Springer, Heidelberg, 2014)
39. S. Nandini, Md Sahidullah, G. Saha, Lung sound classification using cepstral-based statistical features. *Comput. Biol. Med.* **75**, 118–129 (2016)
40. M. Raspaud, H. Viste, G. Evangelista, Binaural source localization by joint estimation of ILD and ITD. *IEEE Trans. Audio Speech Lang. Process.* **18**(1), 68–77 (2010)

41. D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models. *Digital Signal Process.* **10**, 19–41 (2000)
42. S.O. Sadjadi, J.H.L. Hansen, Blind spectral weighting for robust speaker identification under reverberation mismatch. *IEEE Trans. Audio Speech Lang. Process.* **22**(5), 937–945 (2014)
43. S.O. Sadjadi, J.H.L. Hansen, Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification. *Speech Commun.* **17**, 138–148 (2015)
44. S.O. Sadjadi, T. Hasan, J.H. Hansen, Mean Hilbert envelope coefficients (MHEC) for robust speaker recognition, in *INTERSPEECH*. (2012), pp. 1696–1699
45. M.R. Schädler, B. Kollmeier, Separable spectro-temporal Gabor filter bank features: reducing the complexity of robust features for automatic speech recognition. *J. Acoust. Soc. Am.* **134**(4), 2047–2059 (2015)
46. M.R. Schädler, B.T. Meyer, B. Kollmeier, Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. *J. Acoust. Soc. Am.* **131**(5), 4134–4151 (2012)
47. Y. Shao, S. Srinivasan, Z. Jin, D. Wang, A computational auditory scene analysis system for speech segregation and robust speech recognition. *Comput. Speech Lang.* **24**(1), 77–93 (2010)
48. C. Spille, M. Dietz, V. Hohmann, Using binaural processing for automatic speech recognition in multi-talker scenes, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2013), pp. 7805–7809
49. I. Trowitzsch, J. Mohr, Y. Kashef, K. Obermayer, Robust detection of environmental sounds in binaural auditory scenes. *IEEE ACM Trans. Audio Speech Lang. Process.* **25**(6), 1344–1356 (2017)
50. R. Venkatesan, A. Balaji Ganesh, Full sound source localization of binaural signals, in *International conference on Wireless Communication, Signal Processing and Networking*, 2017 (**Accepted**)
51. Y. Wang, K. Han, D.L. Wang, Exploring monaural features for classification-based speech segregation. *IEEE Trans. Audio Speech Lang. Process.* **21**(2), 270–279 (2013)
52. R.J. Weiss, Michael I. Mandel, Daniel P.W. Ellis, Combining localization cues and source model constraints for binaural source separation. *Speech Commun.* **53**(5), 606–621 (2011)
53. J. Woodruff, D.L. Wang, Binaural localization of multiple sources in reverberant and noisy environments. *IEEE Trans. Audio Speech Lang. Process.* **20**(5), 1503–1512 (2012)
54. J. Woodruff, D.L. Wang, Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues. *IEEE Trans. Audio Speech Lang. Process.* **21**(4), 806–815 (2013)
55. S.N. Wrigley, G.J. Brown, Binaural speech separation using recurrent timing neural networks for joint F0-localisation, in: *Machine Learning for Multimodal Interaction*, (2008), pp. 271–282
56. F. Xiong, B.T. Meyer, N. Moritz, R. Rehr, J. Anemüller, T. Gerkmann, S. Doclo, S. Goetze, Front-end technologies for robust ASR in reverberant environments-spectral enhancement-based dereverberation and auditory modulation filterbank features. *EURASIP J. Adv. Signal Process.* **70**(1), 1–18 (2015)
57. Y. Yu, W. Wang, P. Han, Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural network. *J Audio Speech Music Proc.* (2016). <https://doi.org/10.1186/s13636-016-0085-x>
58. X. Zhang, D. Wang, Deep learning based binaural speech separation in reverberant environments. *IEEE ACM Trans. Audio Speech Lang. Process.* **25**(5), 1075–1084 (2017)
59. X. Zhao, Y. Shao, D.L. Wang, CASA based robust speaker identification. *IEEE Trans. Audio Speech Lang. Process.* **20**(51), 608–1616 (2012)
60. X. Zhao, Y. Wang, D.L. Wang, Robust speaker identification in noisy and reverberant conditions. *IEEE Trans. Audio Speech Lang. Process.* **22**(4), 836–845 (2014)