EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Binaural speaker identification using the equalization-cancelation technique

Masoud Geravanchizadeh and Sina Ghalamiosgouei*

Abstract

In real applications, environmental effects such as additive noise and room reverberation lead to a mismatch between training and testing signals that substantially reduces the performance of far-field speaker identification. As a solution to this mismatch problem, in this paper, a new binaural speaker identification system is proposed which employs the well-known equalization-cancelation technique in its structure. The equalization-cancelation algorithm is employed to enhance the input test speech and alleviate the detrimental effects of noise and reverberation in the speaker identification system. The performance of the proposed speaker identification system is compared with unprocessed identification systems and a traditional binaural speaker identification system from the literature. The proposed system is evaluated in both anechoic and reverberant conditions using different types of noise at various azimuthal positions. Simulation results show the superiority of the proposed method in all experimental conditions.

Keywords: Binaural speaker identification, Equalization cancelation, Computational auditory scene analysis

1 Introduction

Speaker identification (SI) systems aim to extract the embedded speaker information from the speech signal. A typical SI system involves three main stages of feature extraction, speaker modeling, and scoring [1–4].

As the first stage, feature extraction tries to transform the incoming speech signal into a convenient representation for later speaker identification stages. Three widely used features in the SI system are the Mel-frequency cepstral coefficients (MFCCs) [5], Gammatone frequency cepstral coefficients (GFCCs) [6], and perceptual linear predictive (PLP) coefficients [7]. The goal of the speaker modeling stage is to train the models that describe feature distributions of individual speakers. For this purpose, in early studies, Gaussian mixture models (GMMs) were used for speaker modeling in the SI systems. In those systems, the GMM parameters are trained by the expectation-maximization (EM) algorithm [8]. In later works, for efficient training of speaker-related GMM parameters, the combination of GMMs with a universal background model (UBM), known as the GMM-UBM model, was considered [9]. Finally, in the scoring stage of the SI system, identification scores are obtained by calculating the likelihoods of observing feature frames given the speaker model.

While speaker models are constructed using clean speech signals, in real applications, the speaker recognition (SR) is performed by unmatched test signals. Mismatches can be imposed by different causes, including channel and interspeaker variabilities, additive noise, and reverberation. The robust SI systems have the task of improving the recognition performance in unmatched conditions. For this purpose, many robust systems have been introduced to deal with the effect of different mismatches [10, 11]. Ways of tackling with the mismatches generated by the channel distortion have been frequently studied. In this regard, the state-of-the-art SR systems incorporate joint factor analysis (JFA) [12] and i-vector [13–15], as one of its important variants, in their implementations. In recent years, deep neural network (DNN) is also used in the structure of i-vector-based systems [16, 17]. The initial attempts with DNNs for SR have been made in the context of i-vector speaker modeling in terms of computing the phonetic posteriors [18, 19]. To solve the channel mismatch problem, in later studies,

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

a new feature, called the DNN bottleneck feature, was extracted as a substitution for traditional features such as MFCCs [20–23]. To improve the performance of DNN in speaker recognition, data augmentation embedding is used. In this regard, DNN, which is trained to discriminate between speakers, converts variable-length speech to fixed-dimensional embedding called x-vectors [24, 25]. In the model, called generative x-vector, the complementary information of i-vector and x-vector is included [26].

The aim of the aforementioned methods has been the reduction of mismatches created by the telephony systems. However, little attention has been paid to the mismatch effects imposed by employing the SI systems in the far-field conditions. In the far-field conditions, the sensor position is far away from the target speaker. In this condition, the additive noises and room reverberation are the main sources of the mismatches. Compensation methods have been introduced to deal with the effect of this kind of mismatch [10]. Regarding the stage at which they are applied, these methods can be divided into three distinct categories, namely, the methods operating on feature extraction, speaker modeling, or scoring stages, respectively. In the methods of the first category, the noise is removed from the speaker characteristic information directly. Cepstral mean normalization (CMN) [27], relative spectra (RASTA) processing [28], employing multi-taper windows [29, 30], and warping methods [31] are examples of the methods in this group. In the second category of compensation methods, such as parallel model combination [32], the aim is to make the SI system more robust by altering the learned speaker models and employing distortion characteristics. The third category aims at achieving the robustness of the SI system by changing the classifier score at the utterance or frame level. However, all the above-mentioned methods have the drawbacks that they require hard assumptions such as stationarity about the characteristic of the environmental effects and the description of noise explicitly, which leads to poor performance in the far-field speaker identification.

Human listeners perform speaker identification robustly without concerning any assumptions about the distortion characteristics [33]. This has inspired many researchers to introduce robust SI methods based on models of the human auditory system to deal with the mismatch problem. The ability of the human auditory system to separate voices in an environment with multiple sources is referred to as the auditory scene analysis (ASA) [34]. Computational auditory scene analysis (CASA) employs methods inspired by the ASA ability to separate speech in multi-source environments [35]. The techniques of CASA have motivated some researchers in the area of robust speaker identification [36–40]. The auditory system also exploits signals from the right and left ears which gives the ability to perform spatial separation of target speaker signal and interfering sounds [41]. Here, ASA uses the information about the spatial location of sound sources, principally encoded by the interaural time difference (ITD) and the interaural level difference (ILD) cues [34]. As one of the aspects of CASA, the binaural cues can be used to estimate ideal masks to segregate target speech from background noise [42–48]. As one of the binaural speech segregation methods, the mask is estimated by employing a deep neural network (DNN) classification method [47, 48]. The training process of DNNs is based on features extracted from predefined mixture signals and defining the ideal masks as the target. A limitation of such supervised learning methods is that the efficiency of segregation is highly dependent on the quality of training and the amount of training data from various sources. In the speaker identification framework, the work of May et al. [39] suggests the utilization of the binaural scene analysis to deal explicitly with the mismatch problem. In this system, the binaural system is used to simultaneously localize, detect, and identify a predefined number of speakers in the presence of reverberation and interfering noise sources placed at different spatial locations. An important drawback of these separation and identification systems is that they are based on supervised learning strategy and, therefore, depend on prior knowledge of source characteristics, which is a strong limitation to be used for practical applications.

The spatial separation between target speaker and maskers often causes large improvements in speech intelligibility in those environments. The amount of intelligibility gain achieved by the binaural hearing is called binaural masking level difference (BMLD). The equalization-cancelation (EC) model is considered as one of the important and simple computational models of the binaural auditory system. The EC model has been originally developed by Durlach [49] and further improved by Culling and Summerfield [50] to predict BMLD. The original EC model is based on the idea that the auditory system transforms the signals arriving at the two ears so that the masker components are “equalized” (the E process) in both ears, and then the signal in one ear is “canceled” (i.e., subtracted) from that in the other ear (the C process). Culling et al. used the EC model to interpret intelligibility performance in two experiments in a simulated anechoic environment involving multiple speech-shaped noise (SSN) maskers [51]. Beutelmann and Brand applied an extended EC model to predict performance in speech intelligibility tasks in several environments, ranging from anechoic space to a cafeteria hall [52]. The EC idea was further developed by incorporating short-time strategies to predict cases involving nonstationary interferers [53].

An extended version of the EC model was also described and applied to speech intelligibility tasks in the presence of multiple maskers in [54]. Furthermore, Wan et al. developed a short-time version of the extended EC in speech intelligibility experiments in the presence of different maskers, including multiple speech maskers [55]. In another study and inspired by the EC theory, a two-stage binaural speech enhancement with the Wiener filter approach was introduced [56]. Later, an EC-based approach was introduced in the field of speech separation that shows performance superiority to the classical localization-based binaural speech separation systems [57].

The EC model was examined in many binaural processing fields because of its conceptual simplicity and its ability to describe the binaural phenomena. However, so far, it has not been considered as a solution to the mismatch problem in far-field speaker identification systems. In this paper, a new speaker identification system based on the short-time extended EC model is proposed to deal with the mismatch problem imposed by environmental effects in the far-field speaker identification. The backbone of the proposed SI system is the well-known GMM-UBM in which the EC process is applied to the auditory representations of both ears at the testing phase to remove the environmental effects, including noise and reverberation. Then, the output of the EC modeling is given as input to the decision module. The performance of the proposed system is compared with identification systems based on MFCC and GFCC features extracted from unprocessed signals and the traditional binaural speaker identification system of May et al. [39] in different simulated acoustic scenarios.

The structure of the paper is as follows: Section 1 gives a background on the monaural SI system, the auditory feature extraction, and the traditional binaural SI system. Section 2 outlines the main contribution of the paper. Here, the proposed speaker identification is presented along with a detailed explanation of the EC binaural model. In Section 3, speaker identification experiments are conducted to analyze the benefit of using the new binaural model in the SI system. Section 4 summarizes the main findings and concludes the paper.

2 Background

2.1 Monaural speaker identification system

In speaker identification, human speech from an individual is used to identify who that individual is. There are two distinct processing stages. In the first stage, called training (or enrollment), the speech from each known speaker is taken to build (i.e., train) the model for that speaker. In the second stage, called testing, comparison of an unknown source of speech against each of the trained individual speaker models is carried out. In closed-set form identification, the unknown individual belongs to a pre-existing pool or database of speakers (speaker models) and the problem then becomes that of choosing which speaker from the pool the unknown speech is derived from.

As mentioned earlier, in this paper, the SI system based on GMM-UBM is used. Figure 1 shows the core structure of a typical SI system based on GMM-UBM. As illustrated, the block of feature extraction generates features that are used in the training of UBM, in adapting GMMs, and in the testing phase. In the training phase, a universal background model (UBM) is generated by utilizing a large collection of speech utterances and the expectation-maximization (EM) algorithm. The EM algorithm iteratively refines the model parameters by maximizing the likelihood of the resulting UBMs [58]. The speaker-dependent GMM models are obtained by adapting the trained UBM parameters to the speaker-dependent speech material.

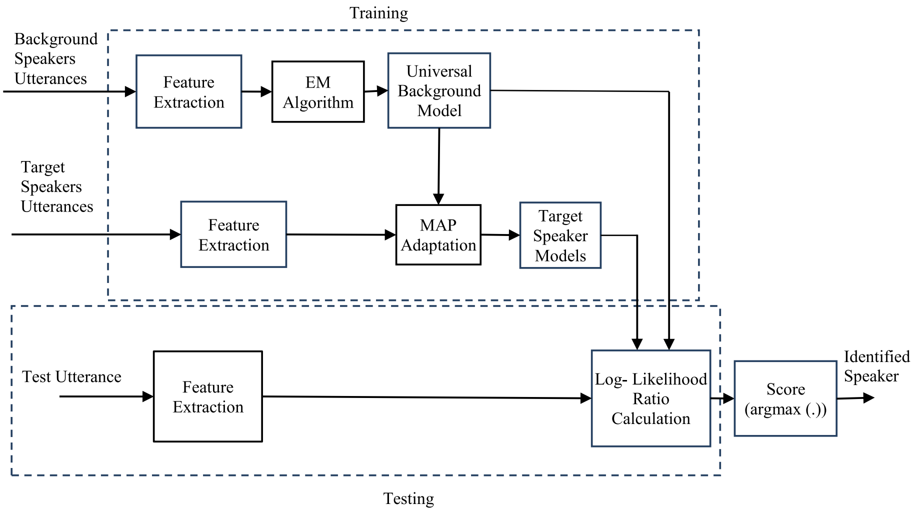

Fig. 1 The block diagram of a typical speaker identification system [10]

After building speaker models in an offline manner, in the testing phase, log-likelihoods of test features given by the GMM models and UBM are calculated. The difference between the likelihoods of GMMs and UBM produces values of scores. Finally, the speaker is identified by searching the argument that has the maximum value of the score.

2.2 Auditory features

The auditory perception of the sound frequency contents for speech signals could be described by nonlinear scales such as Mel and equivalent rectangular bandwidth (ERB) which lead to two feature extraction approaches of MFCC [5] and GFCC [6].

The Mel filterbank is composed of triangular filters where their center frequencies and bandwidths are calculated in the Mel scale. To extract the MFCC features, first, the input signal is decomposed into a time-frequency representation using the Mel filterbank. Then, the representations are compressed by logarithmic function and fed into the discrete cosine transform (DCT) to decorrelate the final MFCC coefficients.

The Gammatone filterbank inspired by psychoacoustical and physiological experiments is one of the standard models of cochlear filtering, which uses ERB-rate scaling to describe center frequencies and bandwidths of the filters. A bank of 32 filters is used with center frequencies ranging from 50 to 4000 Hz or 8000 Hz, depending on the sampling frequency of speech data. In this work, as in [39], a Gammatone filterbank with 32 filters in the range of (50, 8000) Hz was used for the sampling frequency of 16000 Hz.

The impulse response of the Gammatone filterbank is given below [6]:

$$g_c(t) = a\,t^{n-1} \exp(-2\pi b\, ERB_N(f_r)\, t) \times \cos(2\pi f_r t + \phi), \tag{1}$$

where a is the amplitude, n and b are the parameters defining the envelope of the Gamma distribution, f_r is an asymptotic frequency, $ERB_N(f_r)$ is ERB, and ϕ is the initial phase.

Knowing that the filter output retains the original sampling frequency of the input signal, the fully rectified 32-channel filter responses are down-sampled to 100 Hz along the time dimension. This yields a corresponding frame rate of 10 ms, which is used in many short-timespeech feature extraction methods. The resulting responses, called cochleagram, lead to a matrix representing a time-frequency (T-F) decomposition of the input signal [59]. A time frame of the cochleagram representation is called a Gammatone feature (GF). The dimension of a GF vector (here 32) is larger than that of typical feature vectors (e.g., GFCCs) used in a SI system. Additionally, because of the overlap among neighboring filter channels, GFs are largely correlated with each other. To reduce GF dimensionality and decorrelate its components, the DCT operation [60] is applied to produce GFCCs [6] with the dimension of 12.

To take into account the speaking rate of speakers in the SI system, the first and second derivatives of MFCC and GFCC features are computed and used along with the original feature values in the task of speaker identification.

2.3 Traditional binaural speaker identification system

The structure of the traditional binaural SI system introduced by May et al. [39] is illustrated in Fig. 2. This system comprises three important processing stages. In the first stage, the test speech signal is localized using binaural cues, and azimuths related to active sources are determined. In the second stage, called speech detection, the natures of active sources (i.e., speech or non-speech) are identified. The result of the first and second processing stages is a binary mask that is used in the final missing data (MD) SI system.

3 Proposed method

In this paper, a new SI system based on the short-time extended equalization-cancelation method is proposed. The EC processing aims to reduce the mismatches imposed by environmental conditions on the test features. Figure 3 shows the proposed binaural SI system based on the EC process. First, in the training phase, the UBM and GMMs are computed. Then, in the testing stage, the received left and right ear signals are decomposed by the auditory filtering model. For this purpose, the Gammatone filterbank is used. Then, the EC process operates on the Gammatone-filtered signals of the left and right ears. The output of the EC-based model is similar to that of the simulated auditory nerve response and can be converted to an acoustic feature used in the pattern matching unit.

Figure 4 depicts the details of the binaural EC-based model employed in the proposed SI system. The received left and right ear signals from the auditory filtering, $X_{L_i}(t)$ and $X_{R_i}(t)$, are split into time frames of 20 ms with a 10-ms overlap, yielding $X_{L_{i,j}}(t)$ and $X_{R_{i,j}}(t)$, where i and j represent the channel and frame indices, respectively. Assuming $X_{L_i}(t)$ and $X_{R_i}(t)$ as input signals to the EC unit, the output can be computed as [61]:

$$Y_{i,j}(t) = W_j(t)\left\{\frac{1}{\sqrt{\alpha_0(i,j)}}X_{L_i}\left(t+\frac{\tau_0(i,j)}{2}\right) - \sqrt{\alpha_0(i,j)}X_{R_i}\left(t-\frac{\tau_0(i,j)}{2}\right)\right\}, \tag{2}$$

where $W_j(t)$ is the time window obtained at time frame j as:

$$W_j(t) = \begin{cases} 1, & 10j \le t < 10j + K \quad (\text{ms}), \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $K = 20$ represents the length of the window (in ms) for the sampling frequency of $f_s = 16$ kHz, and t is

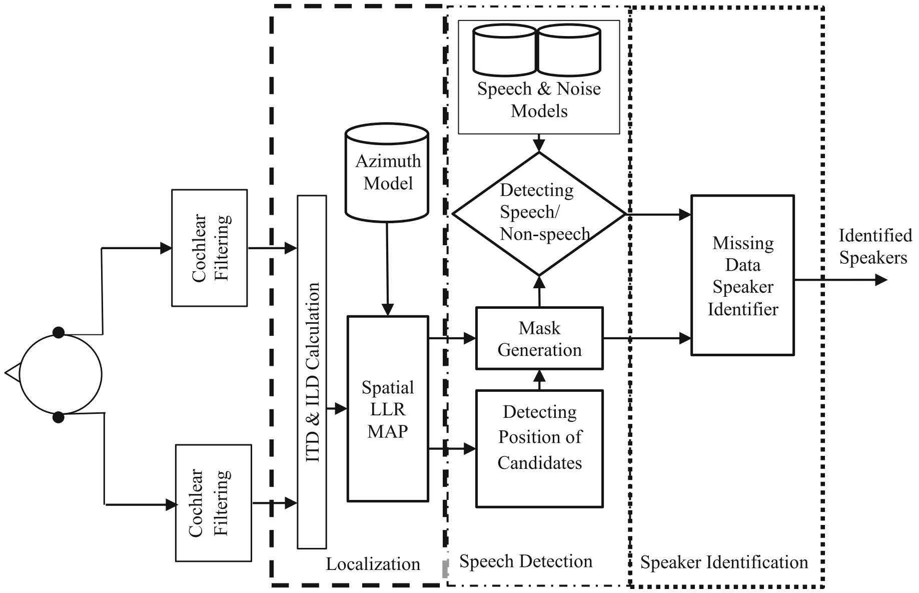

Fig. 2 The block diagram of May *et* al.'s binaural scene analyzer [39]. The system is divided into three main stages: binaural localization, speech detection, and speaker identification

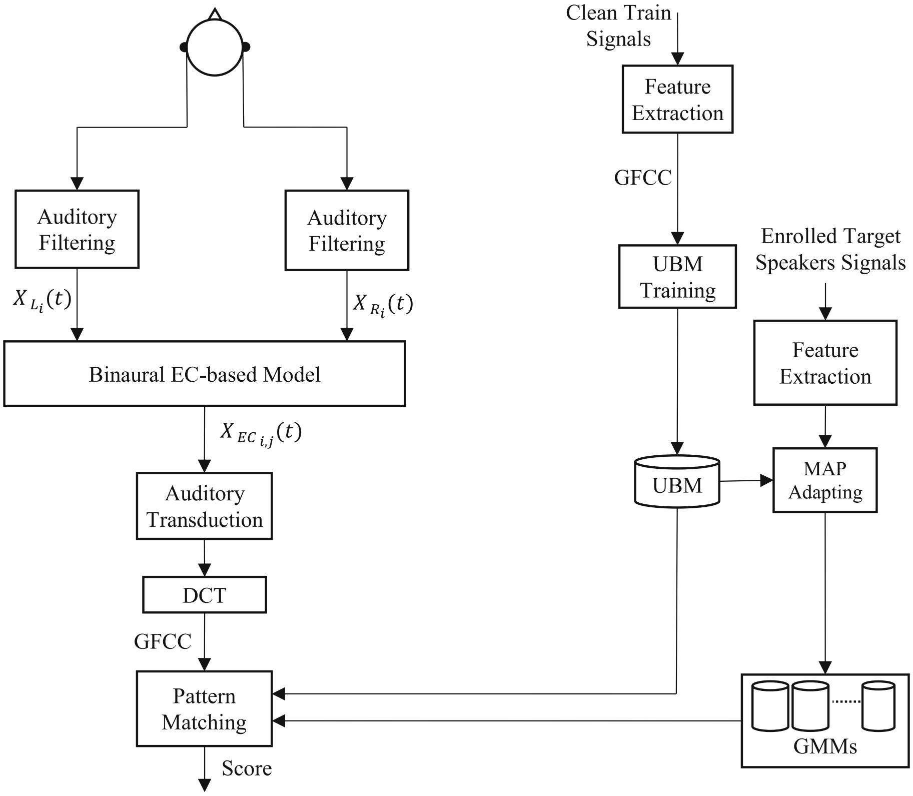

Fig. 3 The structure of the proposed binaural SI system using the EC-based model

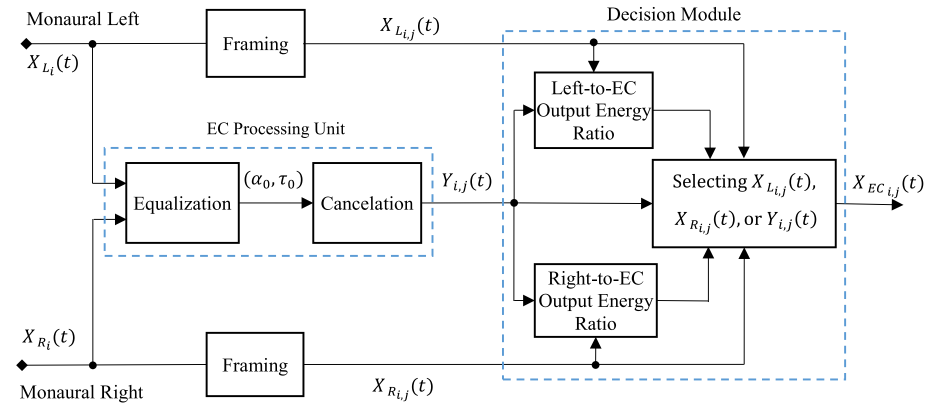

Fig. 4 The block diagram of the binaural EC-based model

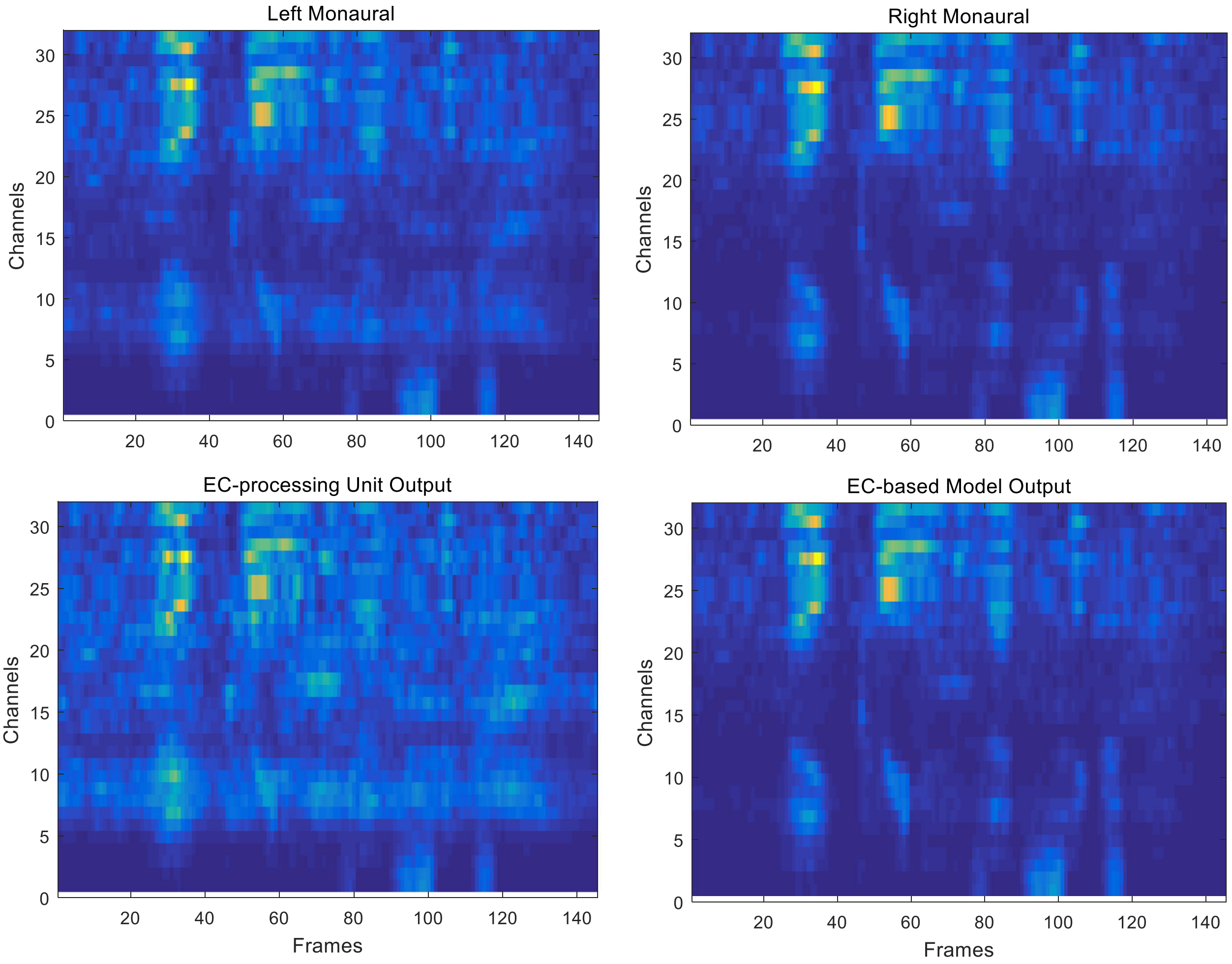

Fig. 5 Cochleagrams of the left monaural, right monaural, and binaural EC-processed signals, and the EC-based model output

the time index (in ms). $\tau_0(i,j)$ is the value that maximizes the cross-correlation function$\rho_{i,\ j}(\tau)$:

$$\tau_0(i,j) = \underset{\tau}{\operatorname{argmax}}\left\{\rho_{i,j}(\tau)\right\}, \quad |\tau| < \frac{\pi}{\omega_i}, \quad (4)$$

with

$$\rho_{i,j}(\tau) = \frac{\int_0^K X_{L_{i,j}}(t) X_{R_{i,j}}(t-\tau)d\tau}{\sqrt{E_{X_{L_{i,j}}} E_{X_{R_{i,j}}}}}, \quad (5)$$

and $\alpha_0(i,j)$ is:

$$\alpha_0(i,j) = \sqrt{\frac{E_{X_{L_{i,j}}}}{E_{X_{R_{i,j}}}}}, \quad (6)$$

where $E_{X_{L_{i,j}}}$ and $E_{X_{R_{i,j}}}$ are the energies of the monaural left and right ear signals.

It is noteworthy that in some applications of the EC algorithm (e.g., [50–55]), the definitions of the parameters $\rho_{i,\ j}(\tau)$ and $\alpha_0(i,j)$ are such that the noise signal is canceled at the output of EC. However, similar to the works in [57, 61, 62], the EC model presented here is based on the modified definitions of $\rho_{i,\ j}(\tau)$ and $\alpha_0(i,j)$ (see Eqs. (5, 6)) to produce the target-canceled signal at the output (see Eq. (2)).

The energies of the framed left and right monaural signals (i.e., $E_{X_{L_{i,j}}}$ and $E_{X_{R_{i,j}}}$) and the output of binaural EC processing unit (i.e., $E_{Y_{i,j}}$) at each T-F unit $(i,\ j)$ are used to select the final output of the EC-based model, $X_{EC_{i,j}}(t)$, in the decision module:

$$X_{EC_{i,j}}(t) = \begin{cases} X_{L_{i,j}}(t) & \left(\frac{E_{X_{L_{i,j}}}}{E_{X_{R_{i,j}}}}\right) \leq 1 \text{ and } \left(\frac{E_{X_{L_{i,j}}}}{E_{Y_{i,j}}}\right) > 0.5, \\ X_{R_{i,j}}(t) & \left(\frac{E_{X_{L_{i,j}}}}{E_{X_{R_{i,j}}}}\right) > 1 \text{ and } \left(\frac{E_{X_{R_{i,j}}}}{E_{Y_{i,j}}}\right) > 0.5, \\ Y_{i,j}(t) & \text{otherwise,} \end{cases} \quad (7)$$

Referring to Eq. (7), some points are worth mentioning. The output of the EC-based model is determined by evaluating two energy ratios; the ratio of the energies of the left and right monaural signals (i.e., $E_{X_{L_{i,j}}}$and $E_{X_{R_{i,j}}}$(and the ratio of the energy of the monaural signal (i.e., $E_{X_{L_{i,j}}}$or $E_{X_{R_{i,j}}}$) and

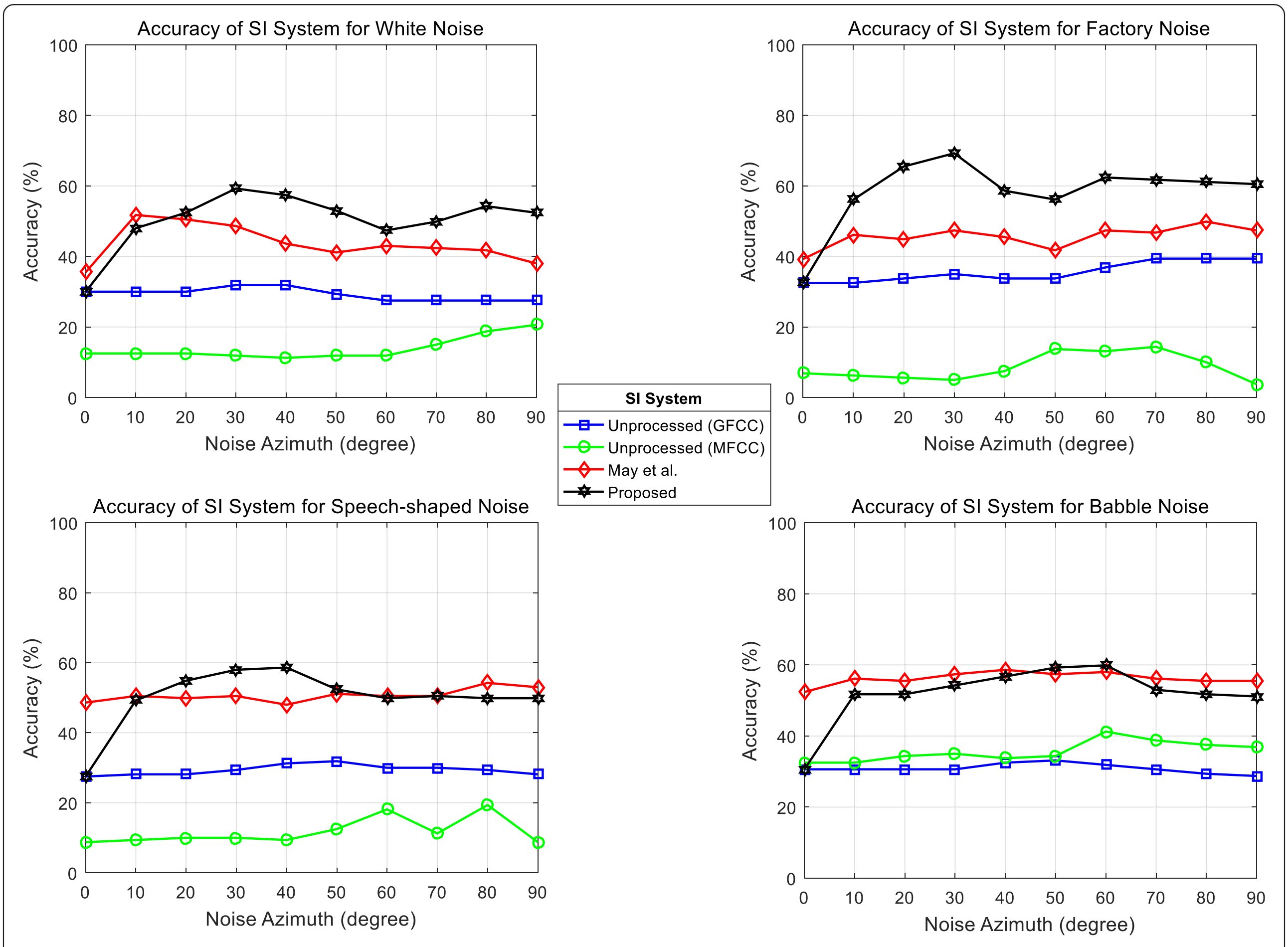

Fig. 6 The accuracy of the SI systems obtained in anechoic and additive noise conditions with four types of noises at SNR = 0 dB

the estimated interference ($E_{Y_{i,j}}$). The following example shows how Eq. (7) works. Suppose the case that the noise source is located on the right side. Knowing that the target signal comes from the frontal azimuthal position, the energy of the right signal is greater than that of the left signal. In this case, $X_{L_{i,j}}(t)$ is considered as a candidate for the output of the EC-based model. However, we cannot assure that the candidate signal is an estimate of the target signal, because $X_{L_{i,j}}(t)$ and $X_{R_{i,j}}(t)$ could both represent the noise signal at the specified T-F unit as well. This is based on the fact that generally, speech has concentrated energy compared to the noise in the T-F representation. Therefore, as a second criterion, the ratio of the energies of the left and residual noise signals (i.e., $E_{X_{L_{i,j}}}/E_{Y_{i,j}}$) is calculated. If this signal-to-noise ratio (SNR) value is larger than 0.5, then the selected signal (i.e., $X_{L_{i,j}}(t)$) is taken as the output of the model. The same argument applies to the justification for selecting $X_{R_{i,j}}(t)$ as the output of the EC-based model in the decision module.

If none of the above conditions are fulfilled, $Y_{i,j}(t)$ is selected as the final output of the model, which is an estimate of the noise signal in that T-F unit. Selecting $Y_{i,j}(t)$ in the model has the effect of flooring the output of the EC-based model to the residual signal, which has been proven to enhance the quality of the source separation system.

Figure 5 represents the cochleagrams of the left monaural, right monaural, and binaural EC-processed signals, and the EC-based model output. Assuming that the clean target and the point Babble noise [63] are located, respectively, at the azimuths of 0° and − 60°, the left and right ear mixture signals are obtained by convolving the clean and noise signals with their corresponding BRIRs and adding them at SNR = 0 dB. As it is obvious from the figure, the output of the model is very similar to the cochleagram of the right ear. This can be justified by the fact that, here, the right ear, called better ear (BE), has the largest SNR as compared to the left ear, and the binaural EC-based model selects the ear signal that is highly correlated to the target.

To obtain the features that serve as input to the SI system, cubic compression (i.e., $\sqrt[3]{(.)}$) and DCT operations are applied to the GFs to obtain the resulting GFCC features.

4 Experiments

The performance of the proposed SI system is assessed in different environmental conditions. For this purpose, speech signals are selected from the Grid database [64].

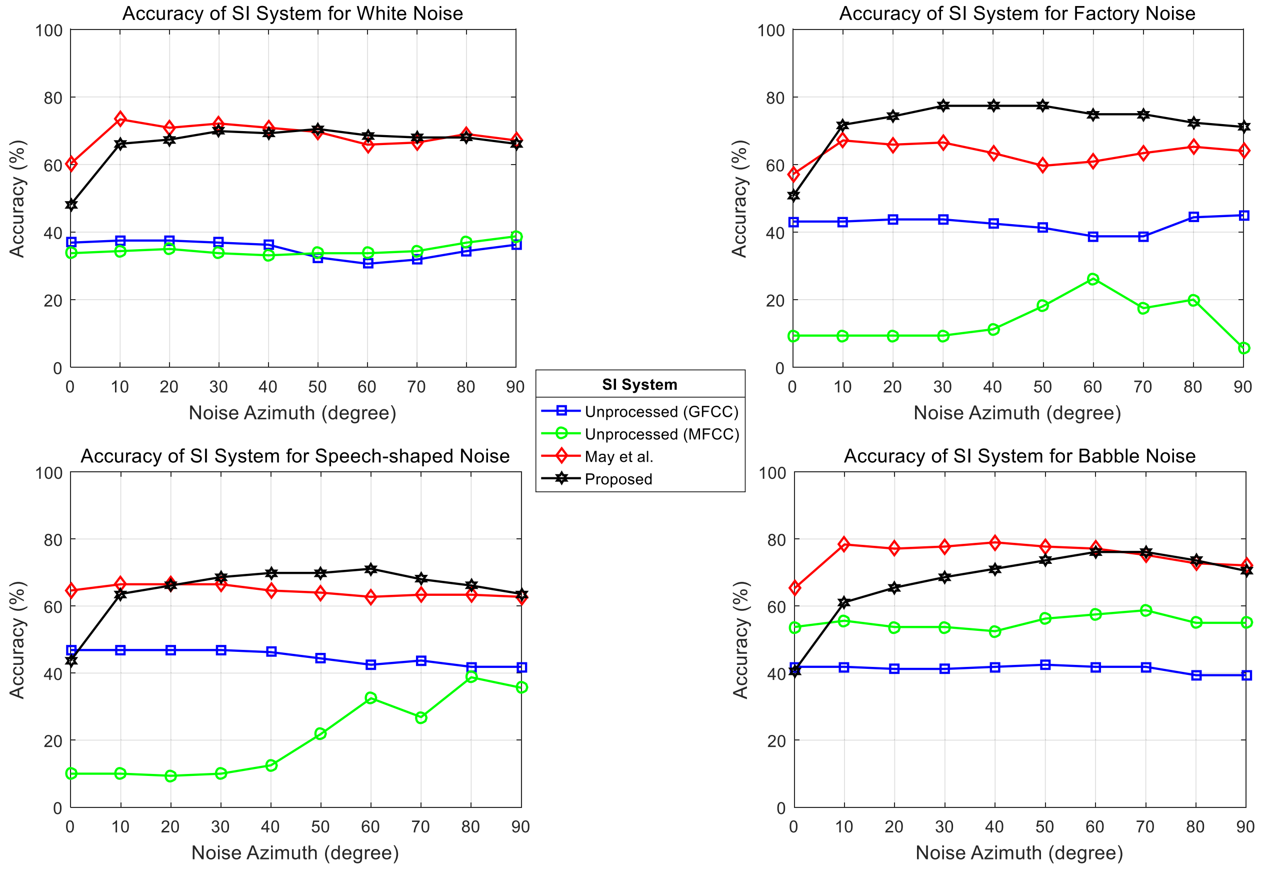

Fig. 7 The accuracy of the SI systems obtained in anechoic and additive noise conditions with four types of noises at SNR = 5 dB

The Grid database consists of 17000 clean utterances spoken by 34 speakers (18 males, 16 females, 500 utterances per speaker). To ensure that there is no overlap between the speech material used for training and testing, the Grid database was randomly split into two sets. The first set consisting of 8500 utterances (250 sentences per speaker) was used to train two gender-dependent UBMs. From the remaining utterances of the second set (250 sentences per speaker), 175 sentences are used to generate GMMs, and the rest is used for the testing stage.

To build the GMM-UBM model, at the first step, the gender-dependent UBMs are constructed using the EM algorithm. Then, the GMM of each speaker is generated by adapting the parameters of the UBM with a relevance factor of 16 [9]. Each of the UBMs is modeled by a GMM with 128 components, and the model of each speaker is obtained by a GMM of 128 components. The GMM-UBMs are implemented by the MSR toolkit [65]. To prevent the underestimation of speech energy due to silent parts, an energy-based voice activity detector (VAD) is employed in the training phase to take into account only signal segments with relevant speech activity [38]. Here, the speech-active segments are defined as those segments which have an energy level within 40 dB of the global maximum.

To reduce the dependency of the SI system on the database, the system is simulated 10 times wherein each run of the algorithm the test and train utterances are randomly selected. Then, the simulation results are averaged among all runs of the algorithm.

In the testing phase, the experiments are conducted in the presence of various additive noises, including White, Factory, and Babble noises selected from the Noisex-92 database [63] and Speech-Shaped Noise (SSN) taken from the Oldenburg University webpage [66]. The left and right ear signals are generated by convolving clean and noise test signals with binaural room impulse responses (BRIRs) and mixing them in an additive manner. The BRIRs are generated by using the Roomsim simulation toolkit [67] with the selection of KEMAR as an artificial head [68]. The KEMAR is placed at 1.75 m above the ground in a simulated room of dimensions 6.6 × 8.6 × 3 m^3. The noisy binaural test signals are generated by adding the noises to the left and right target signals at the SNRs of 0, 5, and 10 dB. The SNR of the mixtures is adjusted as the average value at the two ears. For evaluation purposes, the target signal is positioned at 0^o azimuth. The noise source position is gradually changed in steps of 10^o from 0^o to 90^o in radial distance of 1.5 m

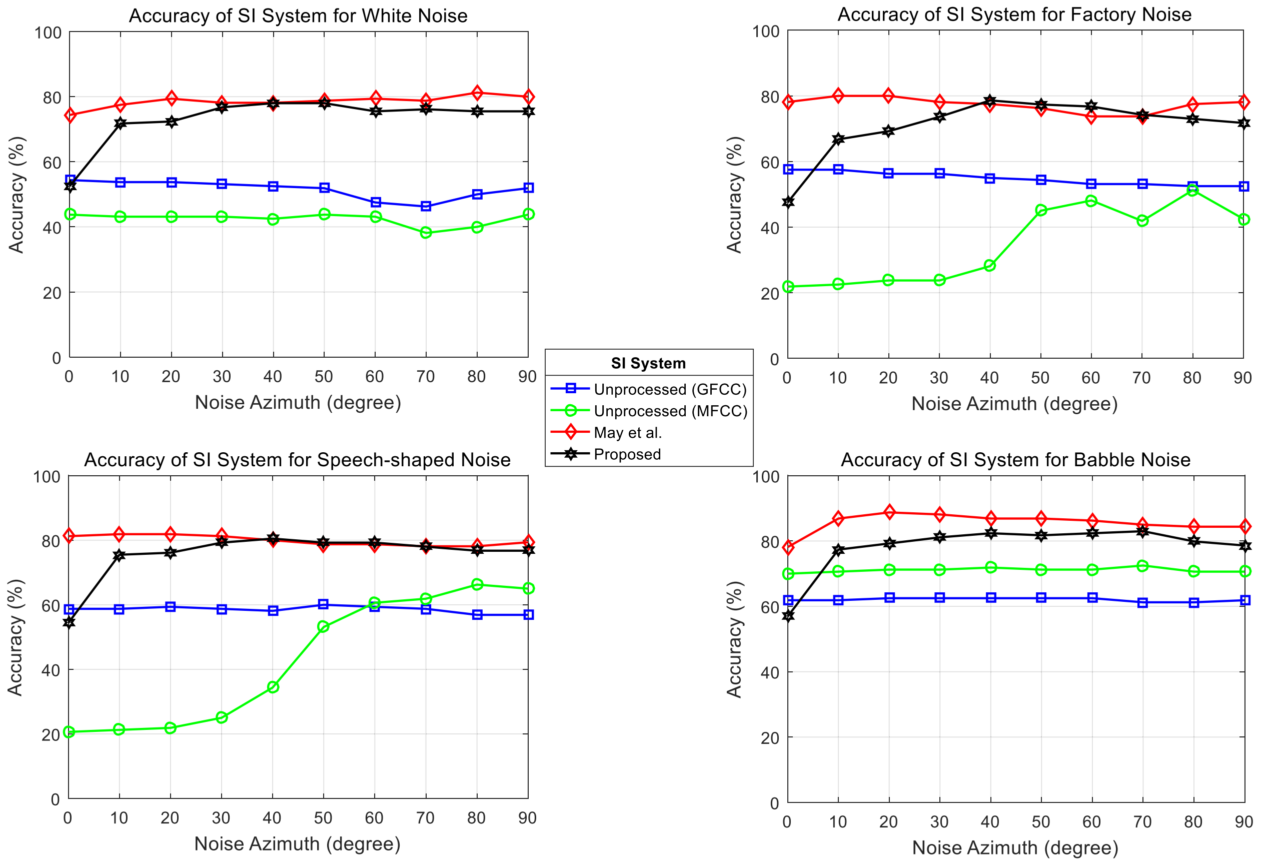

Fig. 8 The accuracy of the SI systems obtained in anechoic and additive noise conditions with four types of noises at SNR = 10 dB

around the listener. The simulated listener is within the critical distance [69] of the target and noise sources. To evaluate systematically the impact of reverberation, the echoic room with T_{60} = 0.29 s is selected for all room boundaries within the room simulation software [67].

4.1 Evaluation criterion

To investigate the performance of the SI system, the recognition accuracy is employed as the performance criterion. The recognition accuracy is defined as the ratio of the number of test speakers detected correctly to the overall number of test utterances.

4.2 Results and discussions

The evaluation results of the proposed SI system are reported in different environmental conditions. The simulations are performed in anechoic and reverberant conditions in the presence of different noise types. For this purpose, the performance of the SI system is investigated by employing 4 types of noises, consisting of White, Factory, SSN, and Babble. The results are obtained for different azimuthal positions of noises.

The proposed binaural SI method ("Proposed") is compared with the traditional SI system of May et al. ("May et al.") and the system with unprocessed inputs ("Unprocessed") using features of MFCC and GFCC. Here, "Unprocessed" means that there is no binaural model that simulates the interaction between the left and right ears. For this purpose, the test feature is obtained by averaging auditory representations of left and right ear signals and applying subsequently the auditory compression and DCT operations. For better modeling of speaker rate in the SI systems, the first and second derivatives of MFCC and GFCC are included in the "Proposed" and the "Unprocessed" systems.

The simulation results of different SI systems in anechoic and reverberant conditions are illustrated in Figs. 6, 7, 8, 9, 10, and 11. Figures 6, 7, and 8 represent the performance evaluation for the anechoic environments at the SNRs of 0, 5, and 10 dB for different noise types. In general, it is seen that

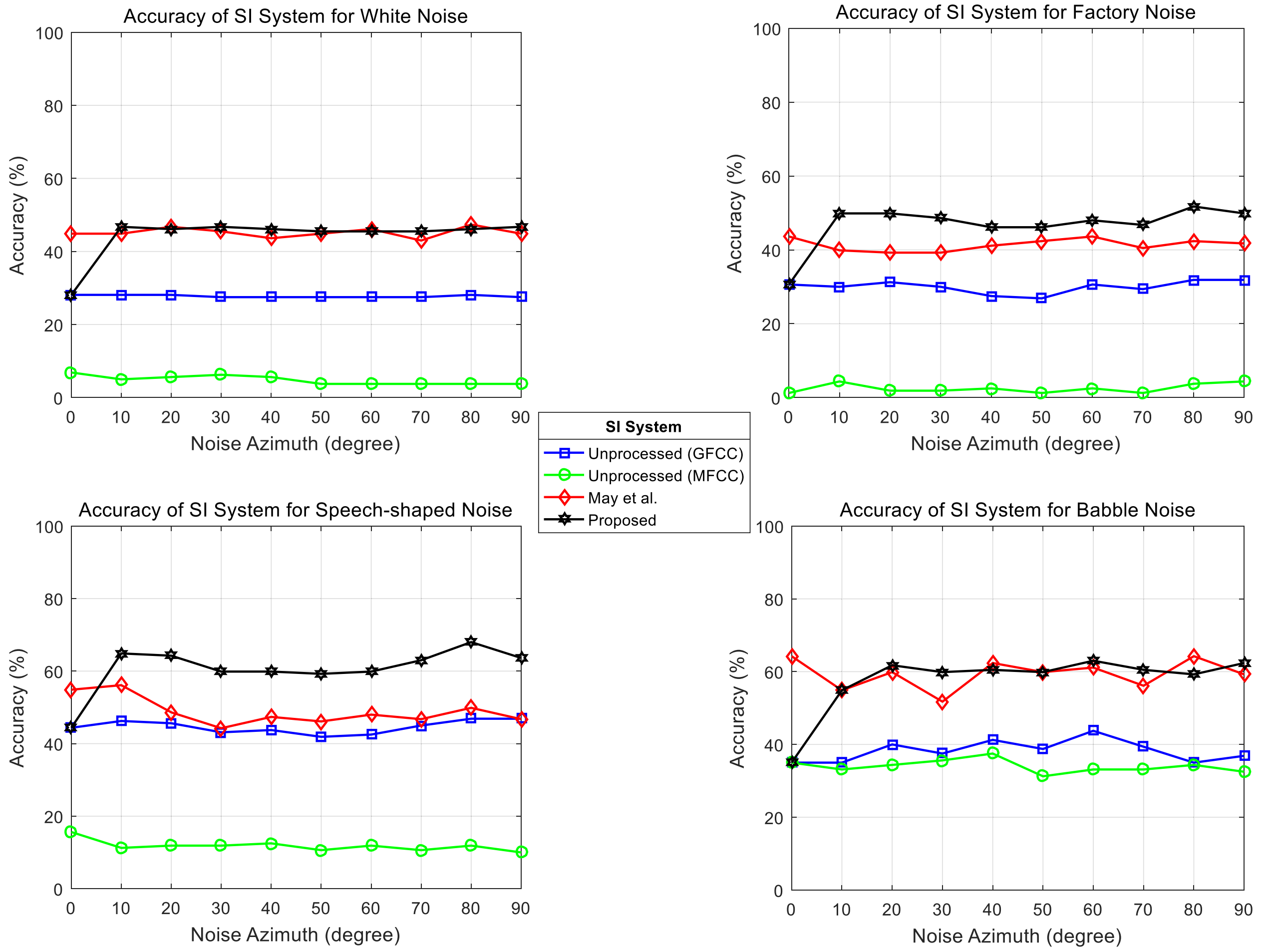

Fig. 9 The accuracy of the SI systems obtained in reverberant (T_{60} = 0.29 s) and additive noise conditions with four types of noises at SNR = 0 dB

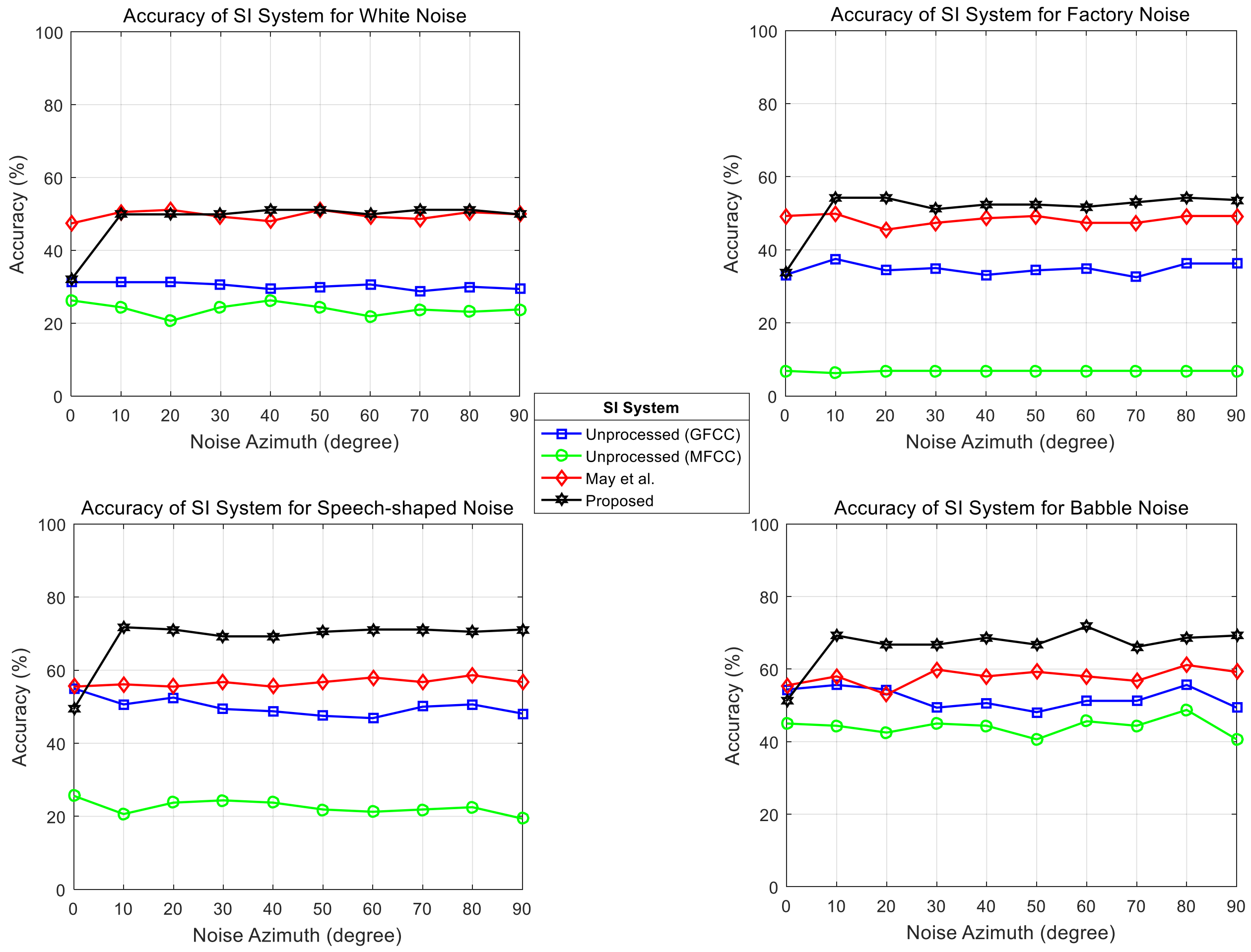

Fig. 10 The accuracy of the SI systems obtained in reverberant (T_{60} = 0.29 s) and additive noise conditions with four types of noises at SNR = 5 dB

the SI systems using binaural processing techniques for the enhancement of the input mixture perform better than those based on unprocessed methods. However, as the level of noise decreases, the performance of the proposed SI method degrades vs. the system of “May et al.” This can be justified by the fact that in contrast to “May et al.,” the proposed EC-based SI model depends highly on the input noise energy to perform the equalization-cancelation procedure satisfactorily. As the value of SNR increases, the contribution of noise at the input of the EC processing unit is lowered which results in decreasing the performance of the proposed model.

The performance comparisons of different SI approaches for the noisy (SNR = 0, 5, 10 dB) and reverberant conditions (T_{60} = 0.29 s) for various types of noises are depicted in Figs. 9, 10, and 11. Once again, it is observed that the SI systems based on unprocessed input signals have the lowest performance as compared with the binaurally processed SI systems (i.e., “Proposed” and “May et al.”). Also, it is seen that the proposed SI model outperforms the SI system of “May et al.” in terms of recognition accuracy. The lower performance of the “May et al.” SI system in the presence of reverberation can be explained by the operation of the speech detection module (refer to Fig. 2). Evidently, the SI method of “May et al.” depends on determining the active source characteristics. Accordingly, in reverberant conditions, the unreliable detected active sources due to the late reflections lead to a challenge in the speech detection module, and consequently, this reduces the identification performance of the system.

Figure 12 shows the averaged accuracies of the SI systems over different noise positions and noise types. The average results also show that the binaural methods achieve superior performance over the unprocessed SI systems.

The results in this diagram confirm those obtained in Figs. 6, 7, 8, 9, 10, and 11. For the anechoic noisy conditions, as the SNR level increases, the performance of “Proposed” gradually decreases in comparison with “May

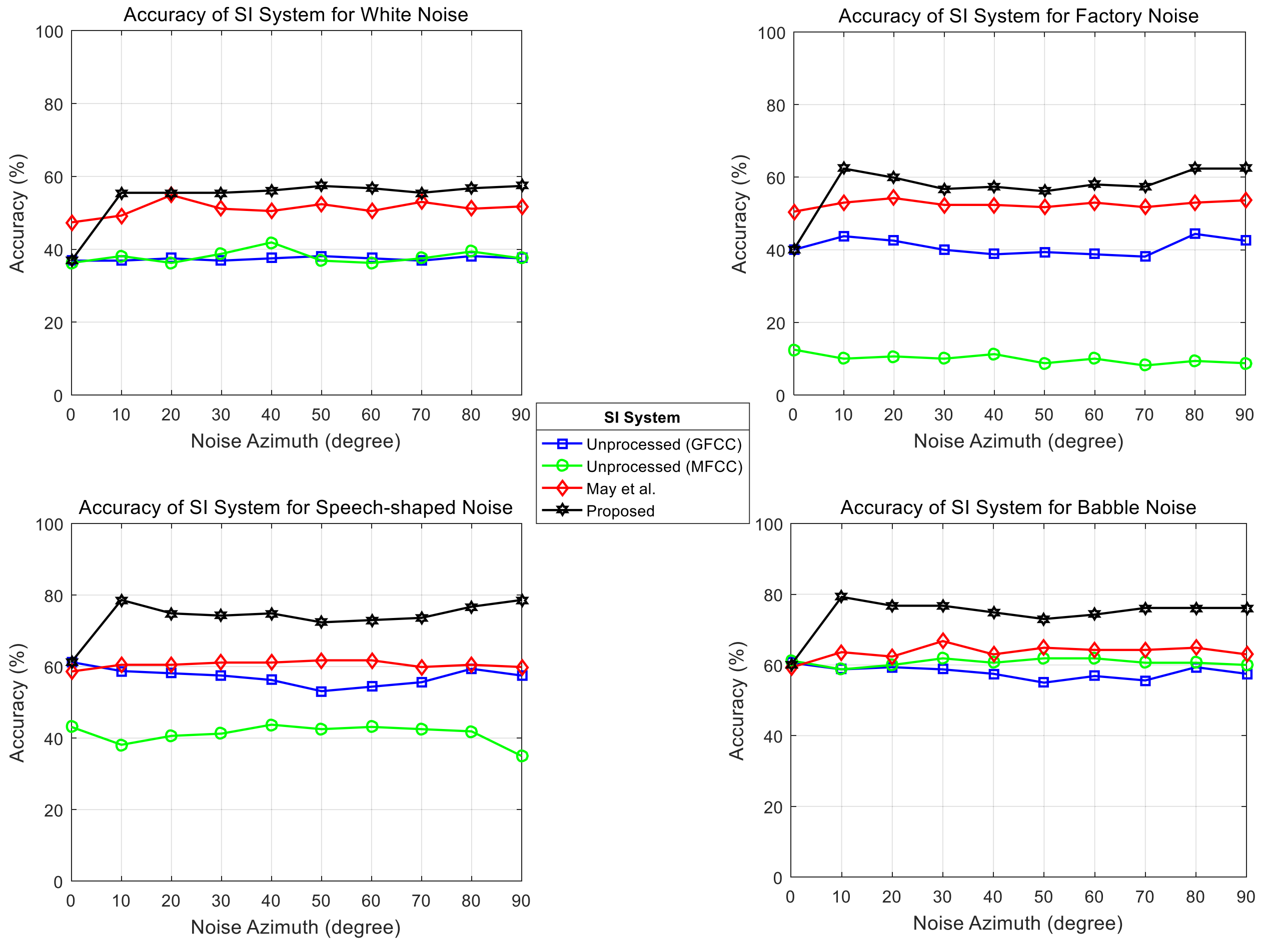

Fig. 11 The accuracy of the SI systems obtained in reverberant (T_{60} = 0.29 s) and additive noise conditions with four types of noises at SNR = 10 dB

et al." For the noisy and reverberant conditions, the proposed model always attains the highest identification accuracy.

5 Conclusions

It is known that the performance of the far-field speaker identification is reduced in real environmental conditions due to the mismatch between training and testing features. In this paper, a new binaural speaker identification system is proposed which employs a short-time extended EC model to tackle the mismatch problem by removing the detrimental effects of noise and reverberation from the input mixture signal. The proposed speaker identification system uses the GMM-UBM structure as the speaker modeling and a binaural EC-based model as a speech separation system that processes auditory representations of both ears to remove noise and reverberation from the input signal.

The binaural EC-based model incorporates an EC processing unit and a decision module. First, in the EC processing unit, an estimate of the residual signal (i.e., interference) is computed by canceling the target signal from the mixture. Then, in the decision module, an estimate of the target signal from the input mixture signal is determined using the energies of the monaural left ear, monaural right ear, and the estimated residual signals. The advantage of the EC-based method is its simplicity which makes it easy to employ the spatial information for identifying the target speaker in complex auditory scenes.

To assess the efficiency of the proposed binaural SI system, the performance of the model is compared with those of the unprocessed and a baseline SI system from the literature. The experiments are conducted in anechoic and reverberant conditions using different types of noises. The simulation results show that the proposed binaural EC-based SI system outperforms its unprocessed counterpart in both experimental conditions. Moreover, in reverberant and low SNR scenarios, the proposed system has superior performance in comparison with the mask-based binaural SI system of "May et al." used as the baseline.

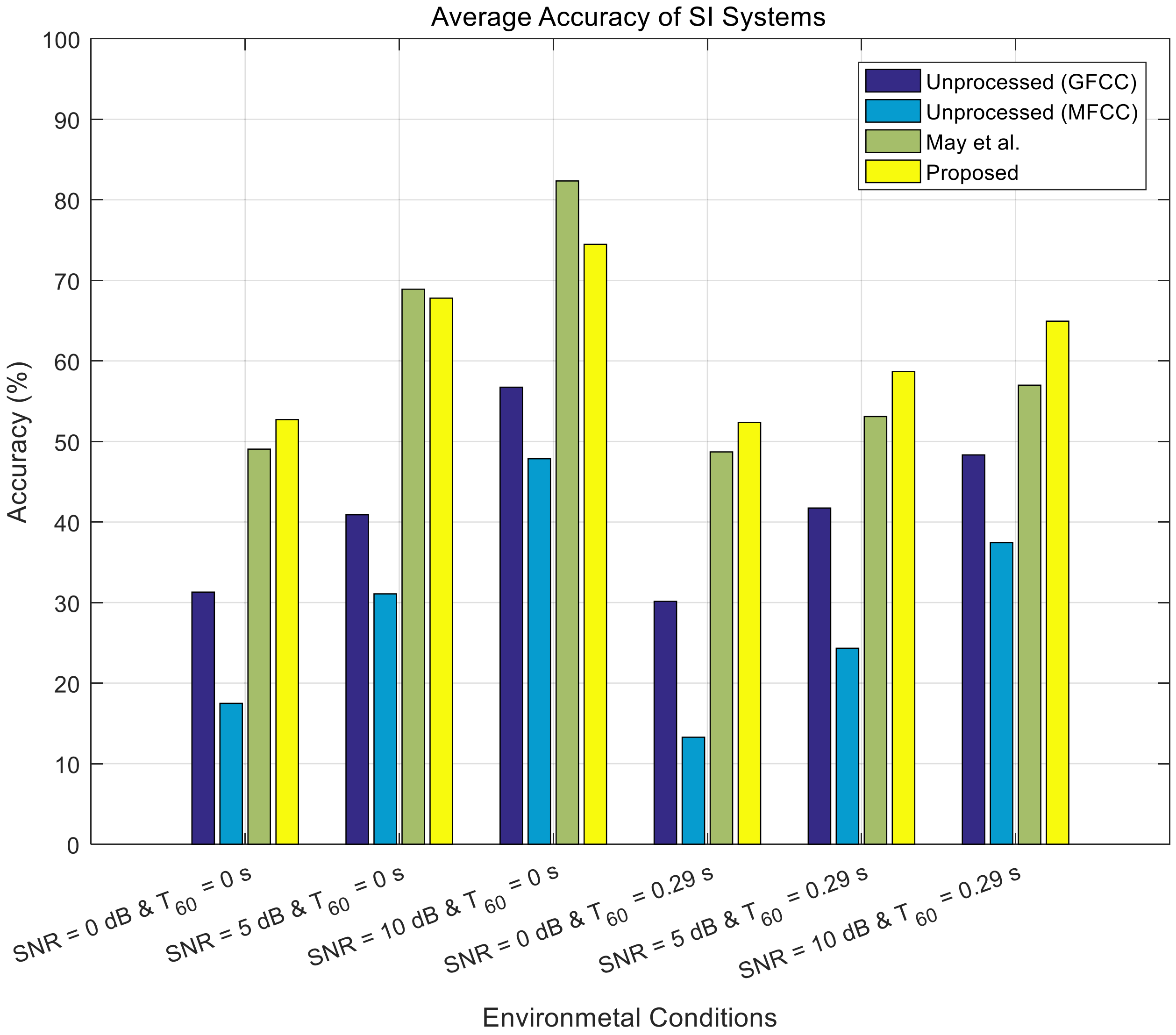

Fig. 12 The averaged accuracies of the SI systems over different noise positions and noise types obtained in reverberant (T_{60} = 0.29 s) and additive noise conditions (SNRs = 0, 5, 10 dB)

It is known that human listeners identify the target speaker robustly in different environmental conditions. In this paper, an auditory model was proposed to remove the undesired environmental effects from the input mixture signal. In dealing with the mismatched problem, it remains to explore the benefits of other binaural auditory models in the proposed SI system for more realistic situations such as cocktail party environments. Moreover, simulating the speaker identification performance of the human is a way to introduce new auditory-based speaker modeling that improves the overall performance of the traditional SI systems in real environmental conditions. Therefore, as future work, the authors plan to design modern auditory-based speaker identification systems and evaluate their performance by conducting listening tests. As a common evaluation procedure of CASA systems, such listening tests are also important in exploring the limitations of the new SI models, and thereby, trying to achieve the human auditory SI performance.

Abbreviations

ASA: Auditory scene analysis; BE: Better ear; BMLD: Binaural masking level difference; CASA: Computational auditory scene analysis; CMN: Cepstral mean normalization; DCT: Discrete cosine transform; DNN: Deep neural network; EC: Equalization-cancelation; EM: Expectation maximization; ERB: Equivalent rectangular bandwidth; GF: Gammatone feature; GFCC: Gammatone frequency cepstral coefficient; GMM: Gaussian mixture model; ILD: Interaural level difference; ITD: Interaural time difference; JFA: Joint factor analysis; MFCC: Mel-frequency cepstral coefficient; PLP: Perceptual linear predictive; RASTA: Relative spectra; SI: Speaker identification; SNR: Signal-to-noise ratio; SSN: Speech-shaped noise; T-F: Time-frequency; UBM: Universal background model

Acknowledgements
Not applicable.

Authors' contributions
MG and SG worked collaboratively on this study. MG and SG have jointly participated in proposing the ideas, discussing the results, and writing and proofreading the manuscript. SG carried out the implementation of the algorithms and the experiments. The authors read and approved the final manuscript.

Authors' information
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, Iran, where he is currently an associate professor. His research interests include binaural signal processing and modeling, 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.
Sina Ghalamiosgouei received the B.Sc. and M.Sc. degrees both in Electrical Engineering from the University of Tabriz in 2009 and 2011, respectively. He is currently pursuing a Ph.D. degree in the Faculty of Electrical and Computer Engineering, at the University of Tabriz, Tabriz, Iran. His current research interests are focused on speech enhancement, speaker identification, computational auditory scene analysis, and binaural signal processing.

Funding
Not applicable.

Availability of data and materials
The datasets used in this paper are publicly available and include the Grid database [64], Noisex-92 (White, Factory, Babble) [63], and speech-shaped noise [66].

Competing interests
The authors declare that they have no competing interests.

Received: 20 March 2020 Accepted: 16 November 2020
Published online: 03 December 2020

References

1. J.P. Campbell Jr., Speaker recognition: a tutorial. Proc. IEEE **85**, 1437–1462 (1997)
2. S. Furui, in *Advances in Biometrics*. 40 Years of Progress in Automatic Speaker Recognition (Springer, Berlin, Heidelberg, Germany, 2009), pp. 1050–1059
3. H. Beigi, *Fundamentals of speaker recognition* (Springer Science & Business Media, Boston, MA, USA, 2011)
4. J.H. Hansen, T. Hasan, Speaker recognition by machines and humans: a tutorial review. IEEE Signal Proc. Magazine **32**, 74–99 (2015)
5. S.B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustics Speech Signal Process. **28**, 357–366 (1980)
6. Y. Shao, D. Wang, in *IEEE International Conference on Acoustics Speech, and Signal Processing (ICASSP)*. Robust speaker identification using auditory features and computational auditory scene analysis (IEEE, Las Vegas, NV, USA, 2008), pp. 1589–1592
7. H. Hermansky, Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. **87**, 1738–1752 (1990)
8. D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. **3**, 72–83 (1995)
9. D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models. Digit. Signal Process. **10**, 19–41 (2000)
10. R. Togneri, D. Pullella, An overview of speaker identification: accuracy and robustness issues. IEEE Circuits Syst. Magazine **11**, 23–61 (2011)
11. Y. Gong, in *IEEE International Conference on Acoustics Speech, and Signal Processing (ICASSP)*. Noise-robust open-set speaker recognition using noise-dependent Gaussian mixture classifier (IEEE, Orlando, FL, USA, 2002), vol. 1, pp. I-133–I-136
12. P. Kenny, P. Dumouchel, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Disentangling speaker and channel effects in speaker verification (IEEE, Montreal, Que., Canada, 2004), vol. 1, pp. 37–40
13. 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)
14. N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika, F. Castaldo, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Support vector machines and joint factor analysis for speaker verification (IEEE, Taipei, Taiwan, 2009), pp. 4237–4240
15. R. Saeidi, J. Pohjalainen, T. Kinnunen, P. Alku, Temporally weighted linear prediction features for tackling additive noise in speaker verification. IEEE Signal Process. Lett. **17**(6), 599–602 (2010)
16. K. Lee, V. Hautamäki, T. Kinnunen, A. Larcher, C. Zhang, A. Nautsch, T. Stafylakis, G. Liu, M. Rouvier, W. Rao, *The I4U mega fusion and collaboration for NIST speaker recognition evaluation* (2016)
17. P.A. Torres-Carrasquillo, F. Richardson, S. Nercessian, D.E. Sturim, W.M. Campbell, Y. Gwon, S. Vattam, N. Dehak, S.H.R. Mallidi, P.S. Nidadavolu, in *INTERSPEECH*. The MIT-LL, JHU and LRDE NIST 2016 speaker recognition evaluation system (2017), pp. 1333–1337
18. Y. Lei, N. Scheffer, L. Ferrer, M. McLaren, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A novel scheme for speaker recognition using a phonetically-aware deep neural network (IEEE, Florence, Italy, 2014), pp. 1695–1699
19. P. Kenny, T. Stafylakis, P. Ouellet, V. Gupta, M.J. Alam, in *Odyssey*. Deep neural networks for extracting Baum-Welch statistics for speaker recognition (2014), pp. 293–298
20. F. Richardson, D. Reynolds, N. Dehak, Deep neural network approaches to speaker and language recognition. IEEE Signal Process. Lett. **22**(10), 1671–1675 (2015)
21. M. McLaren, Y. Lei, L. Ferrer, in *IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Advances in deep neural network approaches to speaker recognition (IEEE, Brisbane, QLD, Australia, 2015), pp. 4814–4818
22. J. Yang, S. Zhang, W. Jin, in *Proceedings of the 20th ACM international conference on Information and knowledge management*. Delta: indexing and querying multi-labeled graphs (2011), pp. 1765–1774
23. P. Matějka, O. Glembek, O. Novotný, O. Plchot, F. Grézl, L. Burget, J.H. Cernocký, in *IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Analysis of DNN approaches to speaker identification (IEEE, Shanghai, China, 2016), pp. 5100–5104
24. C. Kim, R.M. Stern, in *Eleventh Annual Conference of the International Speech Communication Association*. Nonlinear enhancement of onset for robust speech recognition (2010)
25. Z. Zhang, L. Wang, A. Kai, T. Yamada, W. Li, M. Iwahashi, Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification. EURASIP J. Audio Speech Music Process. **1**, 12 (2015)
26. L. Xu, R.K. Das, E. Yılmaz, J. Yang, H. Li, in *2018 IEEE Spoken Language Technology Workshop (SLT)*. Generative x-vectors for text-independent speaker verification (2018), pp. 1014–1020
27. S. Furui, Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoustics Speech Signal Process. **29**, 254–272 (1981)
28. H. Hermansky, N. Morgan, A. Bayya, P. Kohn, in *IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP)*. RASTA-PLP speech analysis technique (IEEE, San Francisco, CA, USA, 1992), pp. 121–124
29. M.J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, D. O'Shaughnessy, Multitaper MFCC and PLP features for speaker verification using i-vectors. Speech Commun. **55**, 237–251 (2013)
30. T. Kinnunen, R. Saeidi, F. Sedlák, K.A. Lee, J. Sandberg, M. Hansson-Sandsten, et al., Low-variance multitaper MFCC features: a case study in robust speaker verification. IEEE Trans. Audio Speech Lang Process. **20**, 1990–2001 (2012)
31. J. Pelecanos, S. Sridharan, in *A speaker Odyssey, the speaker recognition workshop*. Feature warping for robust speaker verification (2001)
32. M. Gales, S. Young, in *IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP)*. An improved approach to the hidden Markov model decomposition of speech and noise (IEEE, San Francisco, CA, USA, 1992), vol. 1, pp. 233–236

33. A. Schmidt-Nielsen, T.H. Crystal, Speaker verification by human listeners: experiments comparing human and machine performance using the NIST 1998 speaker evaluation data. Digit. Signal Process. **10**, 249–266 (2000)
34. D. Wang, G.J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications* (Wiley-IEEE Press, Hoboken, N.J., USA, 2006)
35. A.S. Bregman, *Auditory scene analysis: The perceptual organization of sound* (MIT press, Cambridge, MA, USA, 1994)
36. X. Zhao, Y. Shao, D. Wang, CASA-based robust speaker identification. IEEE Trans. Audio Speech Lang. Process. **20**, 1608–1616 (2012)
37. X. Zhao, Y. Wang, D. Wang, Robust speaker identification in noisy and reverberant conditions. IEEE/ACM Trans. Audio Speech Lang. Process. **22**, 836–845 (2014)
38. T. May, S. van de Par, A. Kohlrausch, Noise-robust speaker recognition combining missing data techniques and universal background modeling. IEEE Trans. Audio Speech Lang. Process. **20**, 108–121 (2012)
39. 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**, 2016–2030 (2012)
40. M.A. Islam, W.A. Jassim, N.S. Cheok, M.S.A. Zilany, A robust speaker identification system using the responses from a model of the auditory periphery. PloS One **11**, e0158520 (2016)
41. M.L. Hawley, R.Y. Litovsky, J.F. Culling, The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. J. Acoust. Soc. Am. **115**, 833–843 (2004)
42. N. Roman, D. Wang, G.J. Brown, Speech segregation based on sound localization. J. Acoust. Soc. Am. **114**, 2236–2252 (2003)
43. M.I. Mandel, R.J. Weiss, D.P. Ellis, Model-based expectation-maximization source separation and localization. IEEE Trans. Audio Speech Lang. Process. **18**, 382–394 (2010)
44. Y. Jiang, D. Wang, R. Liu, in *Fifteenth Annual Conference of the International Speech Communication Association*. Binaural deep neural network classification for reverberant speech segregation (2014)
45. Y. Jiang, D. Wang, R. Liu, Z. Feng, Binaural classification for reverberant speech segregation using deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) **22**, 2112–2121 (2014)
46. R. Venkatesan, A.B. Ganesh, Binaural classification-based speech segregation and robust speaker recognition system. Circuits, Syst. Signal Process. **37**, 3383–3411 (2018)
47. X. Zhang, D. Wang, Deep learning based binaural speech separation in reverberant environments. IEEE/ACM Trans. Audio Speech Lang. Process. **25**, 1075–1084 (2017)
48. P. Dadvar, M. Geravanchizadeh, Robust binaural speech separation in adverse conditions based on deep neural network with modified spatial features and training target. Speech Commun. **108**, 41–52 (2019)
49. N.I. Durlach, Equalization and cancellation theory of binaural masking-level differences. J. Acoust. Soc. Am. **35**, 1206–1218 (1963)
50. J.F. Culling, Q. Summerfield, Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay. J. Acoust. Soc. Am. **98**, 785–797 (1995)
51. J.F. Culling, M.L. Hawley, R.Y. Litovsky, The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. J. Acoust. Soc. Am. **116**, 1057–1065 (2004)
52. R. Beutelmann, T. Brand, Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. **120**, 331–342 (2006)
53. R. Beutelmann, T. Brand, B. Kollmeier, Revision, extension, and evaluation of a binaural speech intelligibility model. J. Acoust. Soc. Am. **127**, 2479–2497 (2010)
54. R. Wan, N.I. Durlach, H.S. Colburn, Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers. J. Acoust. Soc. Am. **128**, 3678–3690 (2010)
55. R. Wan, N.I. Durlach, H.S. Colburn, Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers. J. Acoust. Soc. Am. **136**, 768–776 (2014)
56. J. Li, S. Sakamoto, S. Hongo, M. Akagi, Y. Suzuki, Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication. Speech Commun. **53**, 677–689 (2011)
57. J. Mi, M. Groll, H.S. Colburn, Comparison of a target-equalization-cancellation approach and a localization approach to source separation. J. Acoust. Soc. Am. **142**, 2933–2941 (2017)
58. J.A. Bilmes, A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Int. Comput. Sci. Inst **4**, 126 (1998)
59. D.S. Brungart, P.S. Chang, B.D. Simpson, D. Wang, Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. J. Acoust. Soc. Am. **120**, 4007–4018 (2006)
60. A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing* (Pearson Higher Education, Inc., Upper Saddle River, N.J., USA, 2010)
61. J. Mi, H.S. Colburn, A binaural grouping model for predicting speech intelligibility in multitalker environments. Trends Hear. **20**, 1–12 (2016)
62. Mi, J.: Acoustic source separation based on target equalization-cancellation. PhD Dissertation, Boston Univesity (2018).
63. H. Varga, M. Tomlinson, D. Jones, *The NOISEX–92 study on the effect of additive noise on automatic speech recognition* (Tech. Rep., DRA Speech Res. Unit, Malvern England, 1992)
64. M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. **120**, 2421–2424 (2006)
65. S.O. Sadjadi, M. Slaney, L. Heck, MSR identity toolbox v1. 0: A MATLAB toolbox for speaker-recognition research. Speech Lang. Process. Tech. Comm. Newsl. **1.4**, 1–32 (2013)
66. B. Kollmeier, and V. Hohmann. (2016). http://medi.uni-oldenburg.de/download/ICRA/index.html.
67. S.M. Schimmel, M.F. Muller, N. Dillier, in *IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP).* A fast and accurate "shoebox" room acoustics simulator (IEEE, Taipei, Taiwan, 2009), pp. 241–244
68. W.G. Gardner, K.D. Martin, HRTF measurements of a KEMAR. J. Acoust. Soc. Am. **97**, 3907–3908 (1995)
69. H. Dillon, *Hearing aids* (Thieme Medical Publishers, Inc., NY, USA, 2012)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.