Technical Report OSU-CISRC-2/14-TR06
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftpsite: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2014**
File: **TR06.pdf**
Website: **http://www.cse.ohio-state.edu/research/techReport.shtml**

Binaural Classification for Reverberant Speech Segregation Using Deep Neural Networks

Yi Jiang
Department of Electronic Engineering
Tsinghua University, Beijing 100084, China
jiangyi09@mails.tsinghua.edu.cn

DeLiang Wang
Department of Computer Science and Engineering & Center for Cognitive and Brain Sciences
The Ohio State University, Columbus, OH 43210, USA
dwang@cse.ohio-state.edu

RunSheng Liu
Department of Electronic Engineering
Tsinghua University, Beijing 100084, China
lrs-dee@tsinghua.edu.cn

ZhenMing Feng
Department of Electronic Engineering
Tsinghua University, Beijing 100084, China
fzm@mail.tsinghua.edu.cn

Abstract – Speech signal degradation in real environments mainly results from room reverberation and concurrent noise. While human listening is robust in complex auditory scenes, current speech segregation algorithms do not perform well in noisy and reverberant environments. We treat the binaural segregation problem as binary classification, and employ deep neural networks (DNNs) for the classification task. The binaural features of the interaural time difference and interaural level difference are used as the main auditory features for classification. The monaural feature of gammatone frequency cepstral coefficients is also

used to improve classification performance, especially when interference and target speech are collocated or very close to one another. We systematically examine DNN generalization to untrained spatial configurations. Evaluations and comparisons show that DNN based binaural classification produces superior segregation performance in a variety of multisource and reverberant conditions.

Index Terms – Computational auditory scene analysis (CASA), speech segregation, deep neural networks (DNNs), binary classification, room reverberation.

1 Introduction

The performance gap between human listeners and speech segregation systems remains large in noisy and reverberant environments despite extensive research in speech segregation. A typical auditory environment contains multiple concurrent sources that change their locations constantly and are reflected by the walls and surfaces in a room environment. The auditory system excels in hearing out the target source from a sound mixture under such adverse conditions. Simulating this perceptual ability, or solving the cocktail party problem [8], remains a huge challenge. A solution to the speech segregation problem is essential to an array of applications in hearing prostheses, robust speech recognition, spatial sound reproduction, and mobile communication.

Inspired by human auditory scene analysis [5], computational auditory scene analysis (CASA) [36] approaches the segregation problem on the basis of perceptual principles. A commonly used computational goal in CASA is the ideal binary mask (IBM) [38], which is a two-dimensional matrix of binary labels where 1 indicates that the target signal dominates the corresponding time-frequency (T-F) unit and 0 otherwise. Recent speech perception research shows that IBM segregation produces large improvements of speech intelligibility in noise for normal-hearing listeners [6], [22], [3] and hearing-impaired listeners [2], [37]. Such improvements persist when room reverberation is present [32], [21].

The effectiveness of ideal binary masking implies that the segregation problem may be pursued a binary classification problem, as first formulated by Roman et al. [29, 30] in the binaural domain. The formulation of segregation as supervised classification has recently led to monaural IBM estimation algorithms producing the first demonstrations of speech intelligibility improvements for both normal-hearing [20] and hearing-impaired listeners [12]. It should be noted that these monaural classification algorithms have not considered room reverberation, and tested variations from training noises are limited.

In this study, we address the problem of speech segregation in both noisy and reverberant environments in the binaural setting. A considerable advantage of the classification based approach is that the distinction between monaural and binaural segregation lies only in extracted features, and joint binaural and monaural segregation can be readily addressed by simply concatenating binaural and monaural features. The latter point, we believe, is an important one as such joint analysis is traditionally considered in different stages [26], [34], [41]. Classification based on both monaural and binaural cues would allow an opportunistic use of available cues in a variety of adverse conditions, characteristic of human listening [9]. The proposed classification approach to binaural segregation includes monaural cues in the classification, which are expected to be crucial when target and interfering sources are collocated or close to one another.

As in any classification task, the use of discriminative features is essential for successful classification. Monaural features such as pitch, amplitude modulation spectrogram, mel-frequency

cepstral coefficients, and gammatone frequency cepstral coefficients (GFCCs) have been employed in classification-based segregation [20], [13], [39]. Binaural cues contribute to auditory scene analysis [4, 5]. In particular, the IBM can also be estimated using the binaural cues of interaural time difference (ITD) and interaural level difference (ILD) [29] assuming that target and interfering sources originate from different spatial directions. Binaural mechanisms are also believed to contribute to sequential grouping in reverberant environments [9]. However, when the target and interfering sources are collocated or nearby, binaural cues will not be useful. On the other hand, monaural features are not affected by spatial configuration of sound sources, and can therefore complement binaural segregation. In this report, we primarily employ ITD and ILD cues for classification [29], [24], but also use the monaural cue of GFCC [42] to further enhance binaural segregation. GFCC has been shown to be a good single feature in a recent evaluation [39].

In addition to features, the use of an appropriate classifier is obviously important for T-F unit classification. A variety of classifiers has been explored in classification-based segregation including kernel density estimation [29] and histograms [14] in the binaural domain, and Gaussian mixture models (GMM) [33], [20] , support vector machines [13], multilayer perceptrons (MLP) [19], and deep neural networks (DNNs) [40] in the monaural domain. In this study, we employ DNNs [16] due to their compelling performance in speech and signal processing, including its recent successful use in monaural classification [40].

In the following section, we present an overview of our DNN classification-based binaural speech segregation system. Section III describes how to extract binaural and monaural features and perform DNN classification. The evaluation methodology, including a description of comparison methods, is given in Section IV. We present the evaluation results in Section V, including on trained and untrained source locations. Extensive comparison with several related systems is also presented in this section. We conclude the report in Section VI.

2 System Overview

The proposed DNN classification-based binaural speech segregation system is illustrated in a Fig. 1. The same two auditory filterbanks are used to decompose the left-ear and right-ear input signals into the T-F domain. The output in each frequency channel is then divided into 20 ms T-F units. A T-F unit corresponds to a certain channel in a filterbank at a certain time frame. This peripheral analysis produces a time-frequency representation of the sound mixture.

Binaural features are calculated from each pair of corresponding T-F units in the left-ear and right-ear signal. Monaural features are extracted from the left-ear signal. We extract binaural and monaural features of ITD, ILD and GFCC at the T-F unit level. GFCC features are usually derived at the frame level. By treating the signal in each T-F unit as the input, conventional frame-level feature extraction is then carried out to calculate feature values in

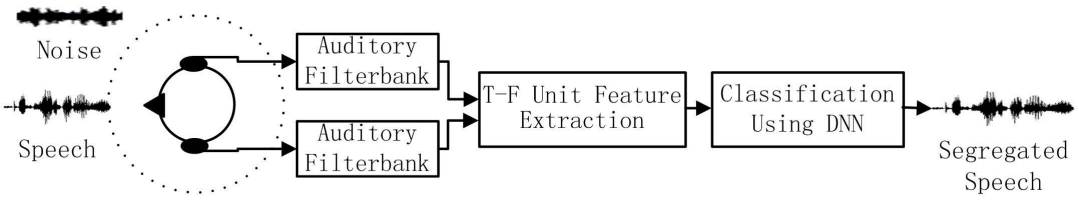


Figure 1: Schematic diagram of the proposed binaural DNN classification system.

each T-F unit [39].

We train DNN to utilize the discriminative power of the entire feature set in a noisy and reverberant environment. As binaural and monaural features vary with frequency [29], [15], we train a DNN classifier for each frequency channel. The training labels are provided by the IBM. In testing, the DNN output is interpreted as the posterior probability of a T-F unit dominated by the target and a labeling criterion is used to estimate the IBM. All the T-F units with the target label (unity) comprise the segregated target stream.

3 Feature Extraction and Classification

3.1 Auditory periphery

We use the gammatone filterbank [27] for auditory peripheral processing as shown in Fig. 1. The bandwidths of the gammatone filterbank are set according to equivalent rectangular bandwidths, and a filter's impulse response is described as

$$g(c,t) = \begin{cases} t^{n-1}e^{-2\pi b(f_c)t}\cos(2\pi f_c t), & if\ t \geq 0 \\ 0, & otherwise \end{cases} \tag{1}$$

where c denotes a filter channel, and we use a total of 64 channels for each ear model. The center frequency of the filter, f_c, varies from 50 Hz to 8000 Hz. $b(f_c)$ indicates the bandwidth. The filter order, n, is 4. This peripheral analysis is widely used in CASA.

With the gammatone filterbanks, the input mixture is first decomposed into the time-frequency domain. The response of a filter channel is further transduced by the Meddis model of the auditory nerve [25]. Finally, the signal in each channel is divided into time frames. Here we use 20-ms frame length with 10-ms frame shift. The resulting T-F representation is called a cochleagram [36]. With a 16 kHz sampling rate, the signal in the T-F unit in channel c and frame m, $x(c,m)$, has 320 samples.

3.2 Binaural feature extraction

With the binaural input signals, we extract the two primary binaural features of ITD and ILD. ITD is calculated from the normalized cross-correlation function (CCF) between the two

ear signals, denoted as l, r for left and right ear respectively. CCF, indexed by time lag τ, for a T-F unit pair is described in the following,

$$CCF(c,m,\tau) = \frac{\sum_k (x_{cm,l}(k) - \bar{x}_{cm,l})(x_{cm,r}(k-\tau) - \bar{x}_{cm,r})}{\sqrt{\sum_k (x_{cm,l}(k) - \bar{x}_{cm,l})^2} \sqrt{\sum_k (x_{cm,r}(k-\tau) - \bar{x}_{cm,r})^2}}. \quad (2)$$

In the above equation, τ varies between -1 ms and 1 ms, and k indexes a signal sample in the T-F units. The overbar indicates averaging. For the 16 kHz sampling rate, we obtain 32-dimensional (32D) CCF features for each pair of T-F units.

For comparison, we also calculate a single ITD feature for each T-F unit pair. The ITD is estimated as the lag corresponding to the maximum in the cross-correlation function as [29],

$$ITD(c,m) = \underset{\tau}{\operatorname{argmax}} CCF(c,m,\tau). \quad (3)$$

ILD corresponds to the energy ratio in dB, and is calculated for each unit pair as

$$ILD(c,m) = 10 \cdot \log_{10} \frac{\sum_k x_{cm,l}^2(k)}{\sum_k x_{cm,r}^2(k)}. \quad (4)$$

The above feature gives a single ILD value over the 20-ms frame. We also break the unit feature into two values, each corresponding to a 10-ms duration. We call the resulting two-value feature 2D ILD.

3.3 Monaural feature extraction

To obtain monaural GFCC features, the left-ear unit response, $x_{cm,l}$, is treated as an ordinary signal and first decomposed by the same 64-channel gammatone filterbank. Then, we decimate fully rectified filter responses to 100 Hz along the time dimension, resulting in an effective frame shift of 10 ms. The magnitude of the decimated filter output is then loudness-compressed by a cubic root operation to $G(i,j)$, which is a 2D matrix along frequency and time respectively. Finally, discrete cosine transform (DCT) is applied to the compressed signal to yield GFCC [42],

$$GFCC(c,m,j) = \sqrt{\frac{2}{C}} \cdot \sum_{i=0}^{C-1} G(i,j) \cos[\frac{j\pi}{2C}(2i+1)]. \quad (5)$$

where $C = 64$ refers to the number of frequency channels. The energy of speech signals is distributed towards lower frequencies. As suggested in Zhao et al. [42], we use 36D GFCC features (the first 36 components) for each T-F unit in this report.

The above binaural and monaural features characterize different properties of the speech signal. For classification, the features are concatenated together to form a long feature vector. Depending on features used, we maximally obtain a 70D feature with 32D-CCF, 2D-ILD and 36D-GFCC for each T-F unit pair.

3.4 DNN classification

Each subband DNN classifier consists of an input layer, two hidden layers, and an output layer [40]. The extracted feature vector within each T-F unit pair is used as the DNN input. The real valued input is suitable for modeling acoustic features.

DNN training requires appropriate initialization. It is well known that random initialization is usually unsatisfactory. We follow the approach in [40], where DNN is pre-trained with restricted Boltzmann machines (RBMs). Boltzmann machines are stochastic generative models that can be used to find more abstract representations in input patterns. RBMs are two-layer Boltzmann machines with connections only between the visible and the hidden layer. Visible units corresponding to the input layer are assumed to be Gaussian random variables with unit variance, so the real valued input is first Gaussian normalized and then fed into the DNN. Each hidden layer contains 200 binary neurons, which are Bernoulli random variables. The output layer has only one neuron with a binary label where 1 indicates that the target speech dominates a T-F unit and 0 otherwise.

The joint probability of visible and hidden units is given below,

$$p(v,h) = \frac{exp(-E(v,h))}{Z}. \tag{6}$$

$$E(v,h) = \frac{1}{2}\sum_i (v_i - a_i)^2 - \sum_j b_j h_j - \sum_{i,j} w_{ij} v_i h_j. \tag{7}$$

$$E(v,h) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{i,j} w_{ij} v_i h_j. \tag{8}$$

where v and h denote the visible and the hidden layer, respectively, and Z is called the partition function. E is an energy function, defined in (7) for a Gaussian-Bernoulli RBM for training the first hidden layer, and in (8) for a Bernoulli-Bernoulli RBM for training the other layers. In (7) and (8), v_i and h_j are the ith and jth units of v and h, and a_i and b_j are the biases for v_i and h_j, respectively. In addition, w_{ij} is the symmetric weight between h_j and v_i.

Mini-batch gradient descent with the batch size of 256 is used for training, including a momentum term with the momentum rate set to 0.5. The learning rate for RBM pre-training is set to 0.001 for the first hidden layer, and 0.1 for the other layers. After RMB pre-training, the standard back-propagation algorithm is applied for supervised fine-tuning. Here, the learning rate decreases linearly from 1 to 0.001 in 50 epochs. For more technical discussions and implementation details about DNN training, we refer the interested reader to [16, 40].

4 Evaluation Methodology

4.1 Experimental setup

For both training and evaluation setup, we generate binaural mixtures that simulate pickup of multiple speech sources in a reverberant space. A reverberant signal is generated using binaural impulse responses (BIRs). We use two sets of BIRs to evaluate the proposed system. The ROOMSIM package [7], which uses measured head related transfer functions from the KEMAR dummy head in combination with the image method for simulating room acoustics, is used to generate the first BIR set, referred to as BIR Set A. In addition, we use a recorded BIR set, referred to as BIR Set B, which was collected using the head and torso simulator (HATS) in four reverberant rooms (A, B, C and D) at the University of Surrey [18]. These speech and noise signals are convolved with BIRs to generate individual sources in a room with corresponding reverberation, and summed at each ear to create the binaural mixture input.

In BIR Set A, the dimension of a simulated room is $6m \times 4m \times 3m$ (length, width, height). The position of the listener is fixed at $2.5m \times 2.5m \times 2m$. Reflection and absorption coefficients of the wall surfaces are uniform. The reflection paths of a particular sound source are obtained using the image model for a small rectangular room [1]. The reverberation times (T_{60}) are approximately 0.3s and 0.7s. We also use the anechoic setting as a baseline. All sound sources are presented at the same distance of 1.5 m from the listener (in the available space of each room configuration). We generate BIRs for azimuth angles between 0° and 360°, spaced by 5°. All elevation angles are zero degree. Speech utterances and babble noise are convolved with selected BIRs to generate the mixtures with defined SNRs. These audio signals are originally sampled at 16 kHz. We upsample them to 44.1 kHz to match the sampling rate of the BIRs, and then downsample to 16 kHz for peripheral and subsequent processing.

In BIR Set B, the reverberant rooms of A, B, C and D have different sizes and reflective characteristics, and their reverberation times are 0.32s, 0.47s, 0.68s, and 0.89s, respectively. In this set, BIRs are measured for azimuths between $-90°$ and 90°, spaced by 5°, at a distance of 1.5 m from the HATS. The sampling rate of the BIRs is 16 kHz, and we apply them to speech and noise signals directly.

Training utterances come from the training set of the TIMIT corpus [10], and the test utterances from the test set. Hence there is no overlap between the training and test utterances. The babble noise from the NOISEX corpus [35], about 4 minutes long, is divided into two parts with the first part (106s) used in training and the second part (128s) in testing. Thus there is no overlap in training and test noise segments either. To create a mixture, a noise segment is randomly cut from the training or testing part to match the length of a target utterance.

As described later, our evaluation is conducted in 2-source, 3-source, and 5-source config-

urations. To isolate location-based segregation from localization, we fix the target source at azimuth 0°, i.e. just in front of the dummy head. More details on training configurations will be given in Section V.A. Regardless of configuration, we generate 500 binaural mixtures to train the DNN classifiers, and use 50 sentences to evaluate the performance of the proposed algorithm in each test condition. Irrespective of test SNRs, training mixtures always have 0 dB SNR. The input SNR is measured at the left ear, by treating the reverberant target speech as the target signal in reverberant cases [31].

4.2 Evaluation criterion

To measure the classification-based segregation performance of our system, we use HIT−FA as our main evaluation criterion. The HIT rate is the percent of correctly classified target-dominant T-F units in the IBM, and the FA (false-alarm) rate is the percent of wrongly classified interference-dominant T-F units. The local SNR criterion (LC) in the IBM definition is set to 0 dB. The HIT−FA rate has been shown to be well correlated to human intelligibility [20].

In addition to this measure of classification accuracy, we adopt the IBM-modulated SNR metric to account for the underlying signal energy of each T-F unit. The resynthesized speech from the IBM is used as the ground truth since the IBM is the ground truth of DNN classification [17]:

$$SNR = 10 \cdot \log_{10} \frac{\sum_t s_I(t)^2}{\sum_t (s_I(t) - s_E(t))^2}. \tag{9}$$

Here, s_I and s_E denote the signals resynthesized from the IBM and an estimated IBM, respectively.

4.3 Comparison systems

We compare the performance of the proposed method with four representative binaural separation methods. Roman et al.' method [31] performs binaural segregation in multi-source reverberant environments. They extract the reverberant target signal from a multisource reverberant mixture by utilizing the location information of the target source. Their system combines target cancellation through adaptive filtering and a binary decision to estimate the IBM.

Another comparison system is DUET [28] which is a popular blind source separation method and produces a binary mask. It assumes that the time-frequency representation of speech is sparse, the so-called W-disjoint orthogonality. It can separate an arbitrary number of sources using only two microphones.

The recent system of Woodruff and Wang [41] formulates the IBM estimation problem as

a search through a multisource state space across time, where each multisource state encodes the number of active sources, and the azimuth and the pitch of each active source. A set of MLPs are trained to assign a T-F unit to one of the active sources in each multisource state. They use a hidden Markov model framework to estimate the most probable path through the multisource state space. This system is particularly relevant as it combines binaural and monaural (pitch) cues.

A joint localization and segregation approach [23], dubbed MESSL, uses spatial clustering for source localization. Given the number of sources the system iteratively modifies GMM models of interaural phase difference and ILD to fit the observed data using an expectation-maximization procedure. Across frequency integration is handled by linking the GMM models in individual frequency bands to a principal ITD. In order to compare with the other systems that all produce binary masks as output, we binarize the MESSL output. Note that the binarization does not reduce MESSL's output SNR.

For Roman et al., Woodruff-Wang, and MESSL systems, we use the implementations provided by their respective authors. The DUET implementation comes from its author's book [28]. All comparison system parameters are adjusted to get the optimal results. To run DUET and MESSL, we provide them the correct number of sources.

5 Results

5.1 DNN classification using binaural features only

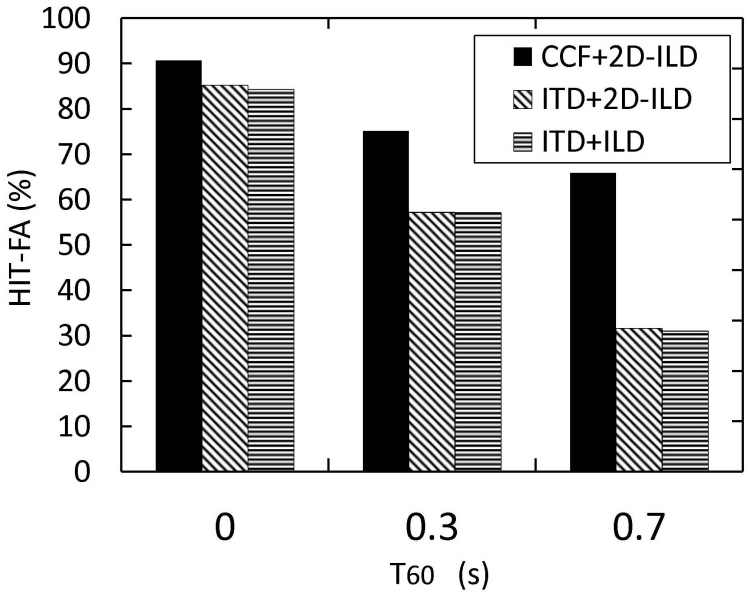


Figure 2: Two-source segregation for trained azimuths at 0-dB SNR.

We first examine the case without monaural GFCC features. This also facilitates compar-

ison with other binaural segregation algorithms. We use BIR Set A to train and test CCF and ILD features systematically. First, we train and test our system with one interference at the azimuth of 45° (i.e. to the left side), and the test SNR of 0 dB. Fig. 2 shows the classification results for a few T_{60}s and compares three kinds of binaural features. With reverberation increasing, the results of all feature kinds decrease, and the gap between 34D features (CCF+2D-ILD) and other two lower dimensional features becomes greater. The HIT−FA rate of the 34D features is 6% (absolute) better than the two-value ITD+ILD features in the anechoic condition, and 34% better at $T_{60} = 0.7$s. In heavily reverberant conditions, strong reflections make the target segregation difficult with only two-dimensional binaural features. CCF features are robust to reverberation. In comparison, 2D ILD performs slightly better than 1D ILD.

We present HIT−FA and SNR results for two-source segregation at -5 dB in Table 1. The results are obtained in the anechoic condition with the interference placed at 45°. As in Fig. 2, 34D features yield the best performance. The 32D CCF features provide more detailed information about the binaural input than the 1D ITD feature. 2D ILD also performs slightly better than 1D ILD on all evaluation criteria.

Table 1: Results on two-source segregation at -5 dB for trained azimuths with different kinds of binaural features.

Binaural feature	HIT (%)	FA (%)	HIT−FA (%)	Output SNR (dB)
CCF+2D-ILD	91.13	1.86	89.28	14.21
ITD+2D-ILD	87.17	3.08	84.09	10.13
ITD+1D-ILD	86.97	3.64	83.32	9.60

Fig. 3 illustrates the cochleagram results for a TIMIT test utterance mixed with the babble noise at -5 dB. As shown in the figure, 34D features give the best performance and recover nearly all of the target speech energy even in this very low SNR condition. Because of their superior performance, we will use 34D binaural features in subsequent evaluations.

To examine the performance difference between trained and untrained azimuths, we evaluate the system in 2, 3 and 5 sound sources. In the two-source condition, the single interference is located at 45°. In the three-source condition, the two interfering sources are located at the azimuth angles of 45° and −45°. Finally, in the five-source condition, the four interfering sources are located at the azimuths of 45°, −45°, 135° and −135°.

We train the DNN in two scenarios. In the unmatched training scenario, the interference sources are systematically varied between 0° and 350°, spaced by 10°. More specifically, in 2-source configurations, the single interference is varied systematically. In 3-source configurations, one interference is randomly chosen from the left side and the other interference randomly chosen from the right side. In 5-source configurations, each of the 4 interfering sources is chosen from a unique quadrant (i.e. the 90° range) of the azimuth space, with the 4 quadrants together covering the entire space. In both 3- and 5-source configurations, we

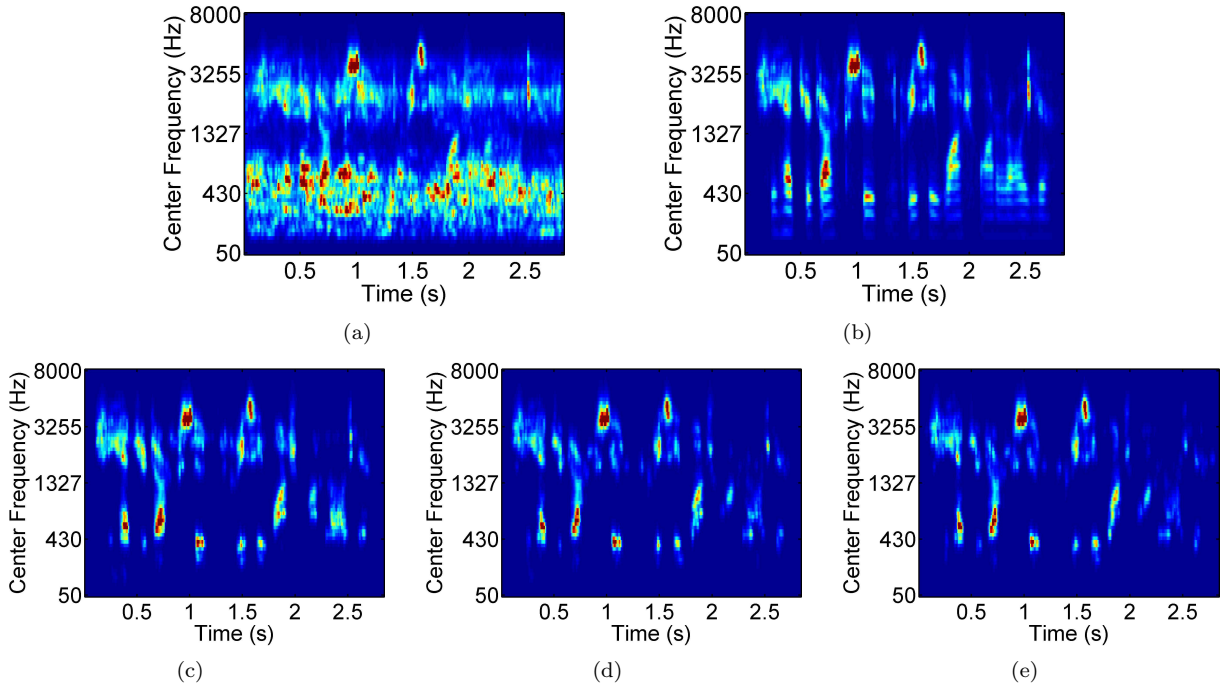


Figure 3: Segregation illustration for a TIMIT utterance mixed with a babble noise at -5 dB. (a) Cochleagram of the mixture. (b) Cochleagram of the target utterance. (c) Cochleagram of separated speech with CCF+2D-ILD features. (d) Cochleagram of separated speech with ITD+2D-ILD features. (e) Cochleagram of separated speech with ITD+1D-ILD features.

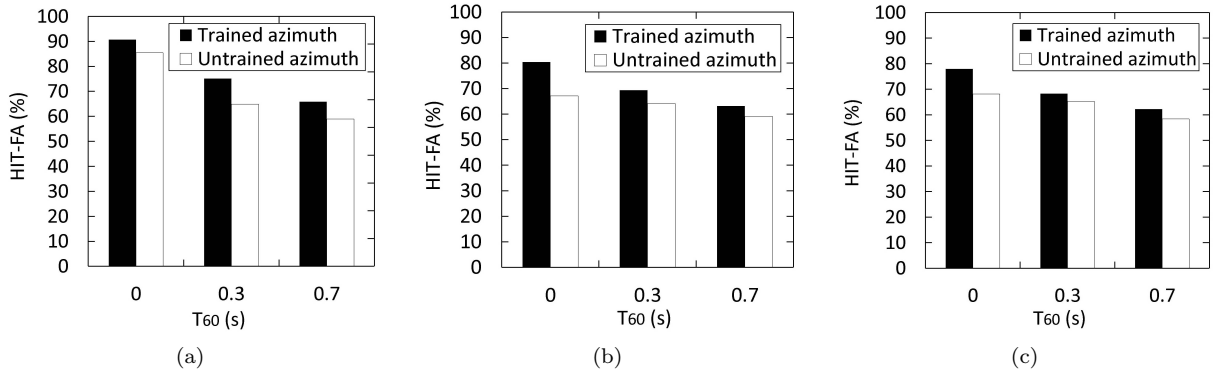
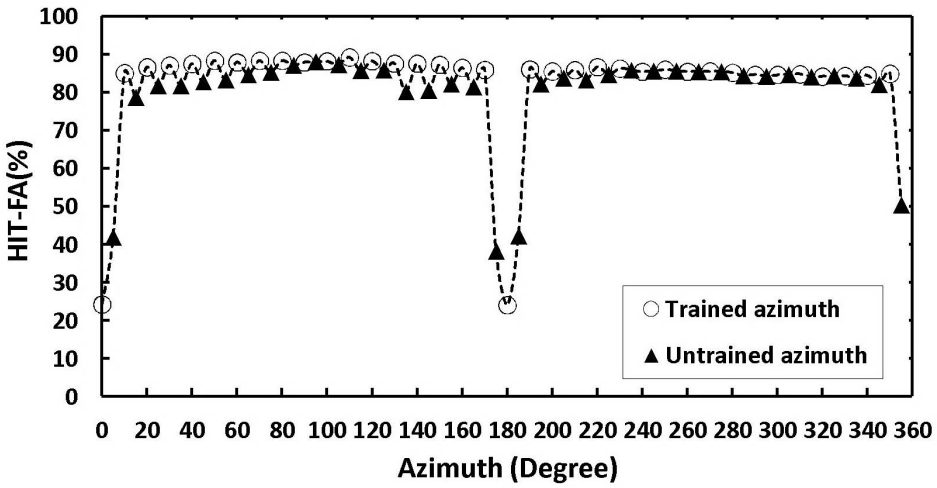


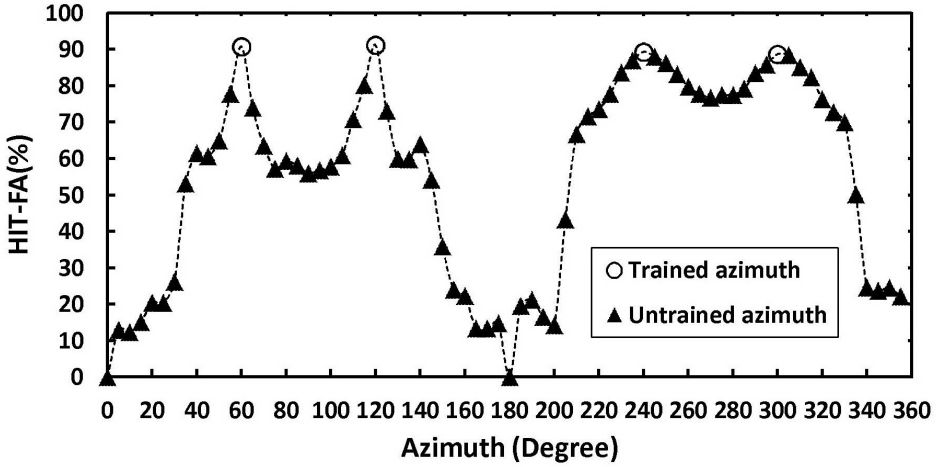
Figure 4: HIT−FA performance at trained and untrained azimuths in anechoic and two reverberant conditions. We train and test with 0-dB mixtures. (a) 2-source segregation. (b) 3-source segregation. (c) 5-source segregation.

see that all multiples of 10° of the azimuth space have been used during training. In this unmatched training scenario, test (evaluation) results are obtained from untrained interference locations. In the matched training scenario, test interference locations are the same as used in training the DNN. Figure 4 shows the classification results in both scenarios. As shown in the figure, the performance gap between trained and untrained azimuths is not large. In the two-source condition, the untrained-azimuth results are lower than the trained-azimuth results by 7.49% in HIT−FA. This average HIT−FA gap is 7.55% in the three-source condition, and

5.54% in five-source condition.



(a)



(b)

Figure 5: HIT−FA performance for two-source segregation at various interference training azimuths and 0-dB SNR. (a) 36 interference azimuths are used in training. (b) 4 interference azimuths are used in training.

To more closely compare between trained and untrained azimuths, Fig. 5 shows 2-source segregation results in the anechoic condition by systematically varying training and test azimuths. In Fig. 5(a), the interference azimuth used in training varies between 0° and 350°, spaced by 10°. In testing, we place the interference at the azimuths between 0° and 355° in 5° steps. In this way, half of the interference azimuths are used in training whereas the other half are not. As shown in Fig. 5(a), the HIT−FA rates are above 80% for most interference azimuths and close to 90% for some azimuths. When the interference locations are close or opposite to the target sound, at azimuths of 0°, 5°, 175°, 180°, 185° and 355°, the HIT−FA rates are down to as low as 25%. This is to be expected as the proposed system operates on the basis of binaural cues only, which have trouble distinguishing an azimuth in the front from

its mirror azimuth in the back. Overall, the trained locations yield a little higher HIT−FA rates than the nearby untrained locations. At the better ear side (i.e. right side), for the interference located between 185° and 355°, the performance differences between trained and untrained locations are small. In Fig. 5(b), we train our system at 4 interference azimuths of 60°, 120°, 240° and 300°, but evaluate interference azimuths at every 5°. As expected, these trained locations produce the four peaks of HIT−FA rates, which gradually decrease as the test interference moves away from the trained locations. Comparing the results in Fig. 5(a) and Fig. 5(b), it is clear that the more the trained angles cover the azimuth space, the better the trained system performs at untrained angles.

Table 2: Two-source binaural segregation results with respect to input SNR.

Input SNR (dB)	HIT (%)	FA (%)	HIT−FA (%)	Accuracy (%)	Output SNR (dB)
-15	80.96	2.06	78.91	97.00	1.79
-10	86.46	4.16	82.30	94.86	6.46
-5	89.10	4.40	84.70	94.43	10.72
0	92.68	7.23	85.45	92.74	14.34
5	94.30	9.41	84.89	92.03	17.29
10	94.64	10.89	83.75	91.91	18.45

The next evaluation tests the system performance by varying the input SNR. In this evaluation we use the babble noise at azimuth between 0° and 350° spaced by 10° to train the DNNs. Then an untrained interference angle at 45° is used to test the system. No reverberation is considered. Note that only the input SNR of 0 dB is used in training. The classification and SNR results are shown in Table 2. The proposed system produces excellent performance in terms of HIT−FA and SNR. As the input SNR decreases, the HIT−FA rate decreases gradually. With the input SNR of -15 dB, the HIT−FA rate of 78.91% is still high; as a reference, this result is higher than the monaural segregation method at -5 dB SNR [20]. Our informal listening indicates that we can recognize segregated speech in this very low SNR condition.

We now compare our classification system and three related systems in Table 3. The test results from our system in the anechoic condition are generated from the untrained interference azimuths. Note that the input SNR of -5 dB is not used in training. The proposed system produces the best results in all test conditions. The MESSL results are better than those of the other two comparison systems, both of which also produce improved SNRs in all test conditions.

5.2 Incorporation of monaural features

We first evaluate whether GFCC features enhance classification performance. The first feature set is 34D binaural-only features, and the second feature set includes 36D monaural GFCC features to form 70D joint binaural and monaural features in each T-F unit pair. Fig. 6

Table 3: SNR (dB) performance comparisons in multisource segregation with no reverberation and the input SNR of -5 dB

No. of sound sources	Proposed	Roman et al.	MESSL	DUET
2	10.72	3.37	5.86	2.17
3	4.65	1.16	3.06	1.78
5	3.96	0.5	3.37	1.59

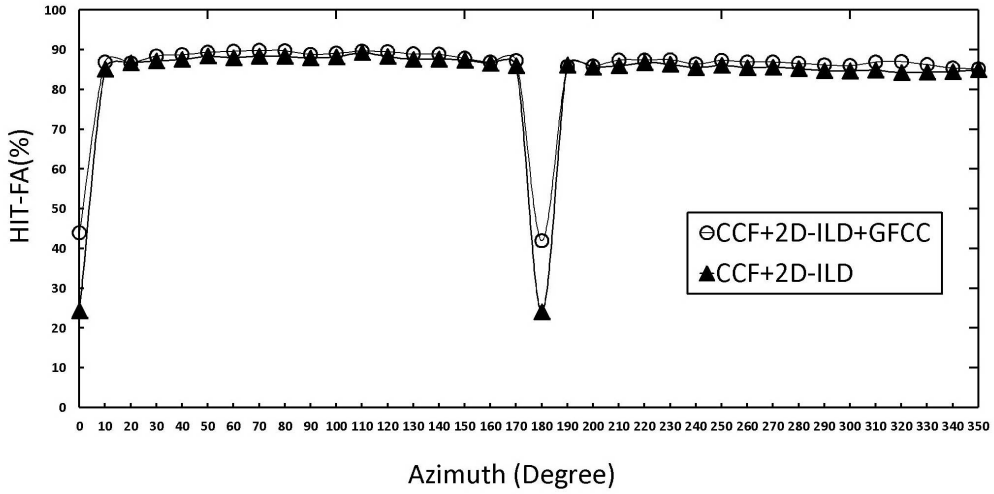


Figure 6: HIT−FA performance for two-source segregation on the 0-dB test set.

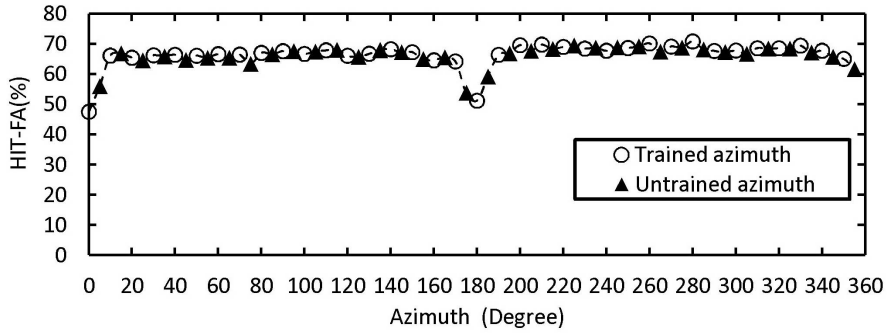
compares two-source segregation in the anechoic condition where interference azimuth varies in the training and test between 0° and 350°, spaced by 10°. As shown in the figure, the joint feature set gives the better performance at all interference azimuths. When the interference is close to the target speech, i.e. at 180°, the HIT−FA rate of the binaural feature set drops to 24.11%, and the joint feature set improves the results to 42.40%, or by nearly 20%. Similar improvement occurs at the interference azimuth of 0°. When the interference is 10 degrees or more away from the target speech, the joint feature set performs slightly better (about one percent).

With reverberation time $T_{60} = 0.3$s, we evaluate the proposed and comparison systems in 2, 3 and 5 source conditions. This comparison also includes the Woodruff-Wang algorithm [41], which is designed for reverberant source segregation and incorporates a monaural pitch cue. The SNR results are given in Table 4. Our system produces the best results in all test conditions, almost 5 dB better than the other systems. The performance of the proposed system is not affected by the number of the interfering sources. All of the comparison systems also produce SNR improvements in all test conditions. Compared to TABLE 3, reverberation drops the SNR performance of the comparison systems by about 2 dB.

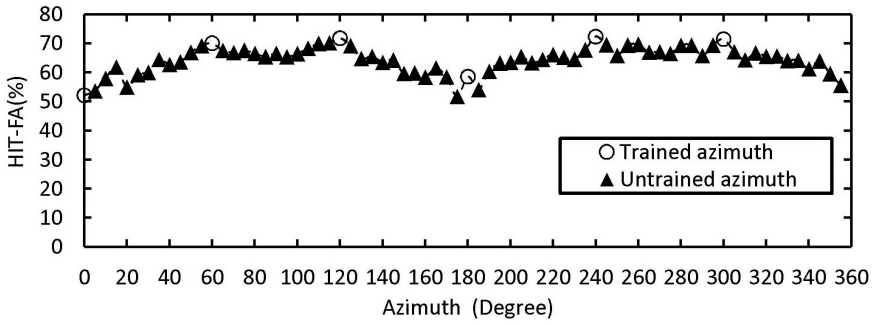
Next, we use BIR Set A with T_{60} of 0.3s to test the generalization of the 70D joint feature set in the reverberant condition. As in Fig. 5(a), we use the interference azimuths between 0° and 350° spaced by 10° to train the DNNs. We then place the interference at the azimuths

Table 4: SNR (dB) performance comparisons in multisource segregation with $T_{60} = 0.3$s and the input SNR of -5 dB

No. of sound sources	Proposed	Woodruff-Wang	Roman et al.	MESSL	DUET
2	5.53	1.58	-2.06	2.73	0.14
3	5.42	0.17	-1.61	-0.23	0.49
5	5.53	0.92	-2.14	0.55	0.54



(a)



(b)

Figure 7: HIT−FA performance for two-source segregation at various interference training azimuths with joint features in the reverberant condition at 0 dB. (a) 36 interference azimuths are used in training. (b) 6 interference azimuths are used in training.

between 0° and 355° in 5° steps to evaluate the trained system. As shown in Fig. 7(a), the HIT−FA rates are above 47% at all interference azimuths and close to 70% for most of the test azimuths. When the interference azimuths are close to the target sound or its mirror angle, at azimuths of 0°, 5°, 175°, 180°, 185° and 355°, the HIT−FA rates are down to 50%. Note that, in this reverberant condition the untrained locations yield similar HIT−FA rates to the nearby trained locations. The disappearance of the small gap seen in Fig. 5(a) is due to the use of GFCC features, which are insensitive to azimuth. In Fig. 7(b), we train the system at 6 azimuths of 0°, 60°, 120°, 240°, 300° and 360°. This way of training produces the four high peaks of HIT−FA at the trained azimuths of 60°, 120°, 240°, and 300°. The HIT−FA rates decrease as the test interference locations move away from the trained azimuths. Comparing the results in Fig. 7(a) and Fig. 7(b), it is clear that with more trained angles, the trained system performs better at untrained angles, similar to Fig. 5.

We now compare the proposed system with the four comparison systems in the 5-source environment with different levels of reverberation. We use BIR Set A with $T_{60} = 0.3$s and 0.7s in addition to the anechoic condition. The SNR results from our algorithm and the comparison methods are plotted in Fig. 8. As shown in the figure, the joint feature DNN classification system yields the best results at all reverberation levels. When reverberation increases, the performance of the proposed system decreases rather gradually from 10.37 dB to 7.49 dB. The joint features perform 2 dB better than binaural-only features. The performance gap between our system and comparison systems becomes larger in reverberant conditions. In the anechoic condition, the MESSL and Woodruff-Wang methods produce 7.54-dB and 7.45-dB SNR improvements, respectively, which are better than Roman et al. (4.10 dB) and DUET (5.41 dB). But they drop more quickly as T_{60} increases (2.70 dB and 2.20 dB improvements at $T_{60} = 0.7$s). In heavily reverberant conditions, the four comparison systems show similar results.

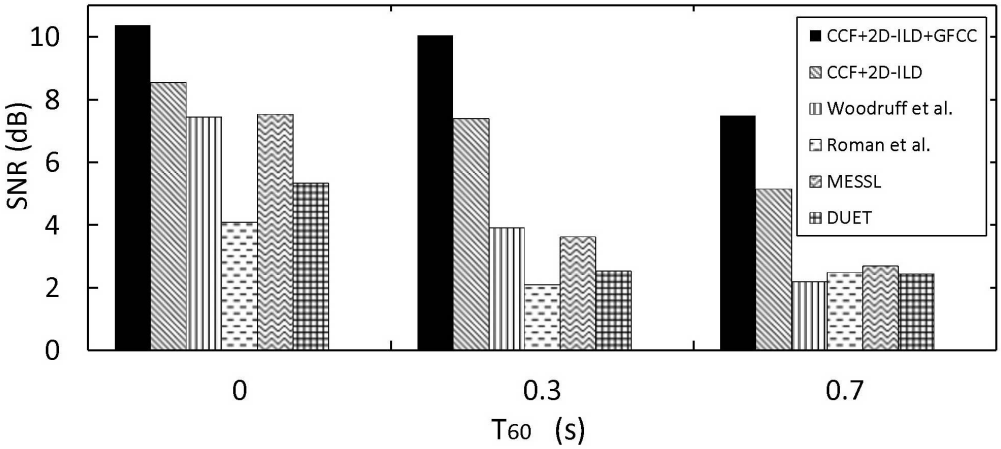


Figure 8: SNR comparisons in the 5-source environment where speech utterances are mixed with the babble noise at 0 dB.

5.3 Evaluation with recorded BIRs

In the following experiments, we use the measured BIR Set B to evaluate our system for 2-source segregation. The babble noise located between $-90°$ and $90°$ spaced by $10°$ is used to train the DNNs. We first compare the binaural-only feature set and the joint feature set in the four reverberant rooms. The noise is located at the untrained azimuth of $15°$, producing 0 dB mixtures. As shown in Fig. 9, the HIT$-$FA rate difference between these two feature sets is, on average, 3.2%. The maximum gap is 5.86% in Room B with $T_{60} = 0.47$s.

Fig. 10 illustrates the segregation results for a TIMIT test utterance mixed with babble noise at 0 dB in Room C with $T_{60} = 0.68$s. The joint features recover most of the target speech in this condition, producing a similar cochleagram to that of the target speech.

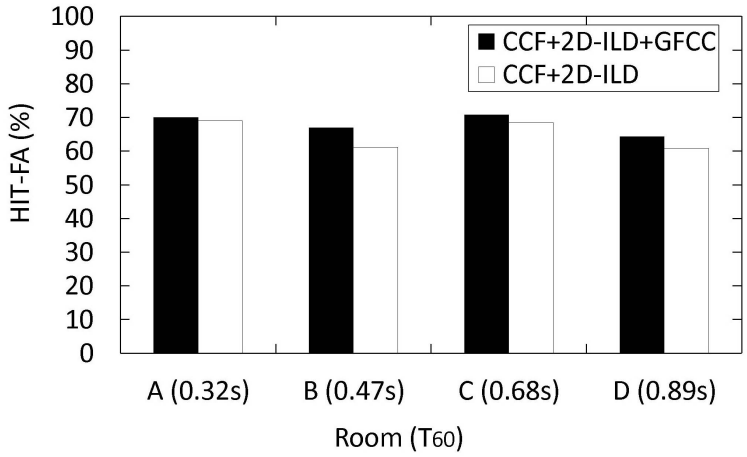
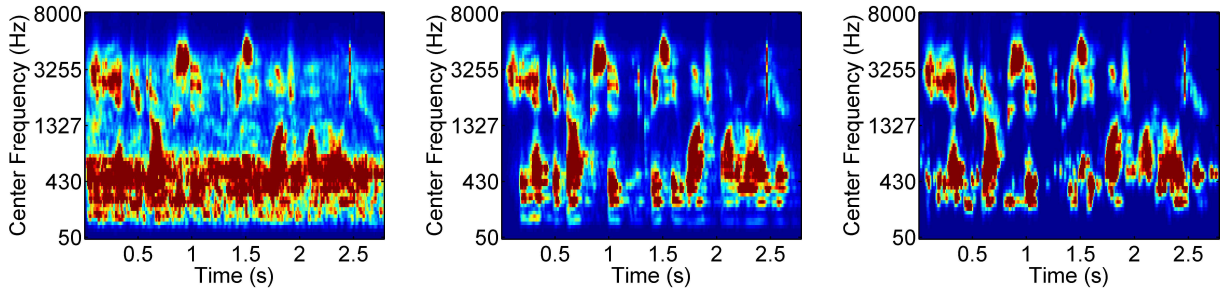


Figure 9: Two-source segregation with binaural-only and binaural-monaural features in four reverberant rooms at the input SNR of 0 dB.



(a) (b) (c)

Figure 10: Segregation illustration for a TIMIT utterance mixed with babble noise in Room C at 0-dB SNR. (a) Cochleagram of the reverberant mixture. (b) Cochleagram of the reverberant target utterance. (c) Cochleagram of separated speech.

We next present more detailed results of the DNN classification system with joint features at the untrained interference angle of 45° in Table 5. As shown in the table, the proposed system produces strong performance in terms of both HIT−FA and SNR. As reverberation increases, the HIT−FA rate decreases only gradually. Even in Room D with T_{60} of 0.89s, the HIT−FA is still high. Comparing with the results in Fig. 9, we note that the larger azimuth separation in Table 5 increases the HIT−FA rate.

Table 6 shows SNR comparisons for the test mixtures at -5 dB. The test azimuth of the babble noise is the untrained 15°. Consistent with the results using simulated BIRs, the proposed system gives the best results in all conditions. Woodruff-Wang and MESSL outperform the other two systems in most of the conditions.

We have also compared the proposed system with the others for 0-dB mixtures with the interference located at 15° or 45°. Similar SNR improvements are obtained as for -5 dB mixtures in Table 6. With interference farther away from the target speech, the performance increases as concluded in Section V.B, with the only exception of the Roman et al. method

Table 5: Two-source segregation results in four reverberant rooms at the input SNR of 0 dB.

Room	HIT (%)	FA (%)	HIT−FA (%)	Accuracy (%)	Output SNR (dB)
A	83.02	10.51	72.52	87.67	11.56
B	80.07	10.10	69.97	87.04	9.47
C	82.94	8.88	74.06	88.66	10.56
D	78.79	13.72	65.06	83.92	7.98

Table 6: SNR comparisons in two-source segregation using measured impulse responses from four reverberant rooms at the input SNR of -5 dB. T_{60} (in s) in each room is listed in parentheses.

SYSTEM	A (0.32)	B (0.47)	C (0.68)	D (0.89)
Proposed	4.56	2.80	1.61	1.15
Woodruff-Wang	1.75	1.50	0.95	0.65
Roman et al.	-1.13	0.97	-0.45	0.15
MESSL	1.86	0.70	0.63	0.58
DUET	-2.45	-3.27	-2.96	-3.72

that shows little change as this method uses adaptive filtering to segregate speech.

6 Concluding Remarks

In this study, we have proposed a DNN-based classification algorithm with joint binaural and monaural features for binaural speech segregation in reverberant environments. To our knowledge, this is the first study that introduces deep neural networks to binaural segregation. The evaluation results show that the proposed system achieves substantially better results than four representative binaural separation algorithms. Even at very low input SNRs and with strong reverberation, the proposed system yields excellent segregation performance, which decreases only gradually with increased room reverberation.

The results from our evaluation indicate encouraging generalization to untrained spatial configurations. This is important for supervised learning algorithms. Dependency on trained configurations is a main limitation of the first supervised classification method of Roman et al. [29], [30] for binaural segregation. The key to overcome this limitation is to train with a variety of configurations and the apparent generalization ability of deep neural networks. Training with a variety of configurations also allows the system to perform binaural segregation without sound localization, in contrast to localization-based segregation [36].

We believe that the classification framework is a very promising direction for future development [13]. In this framework, for example, it is straightforward to include monaural features to complement binaural features for improved segregation, especially when the target and interfering sources are either collocated or close to one another. We can expect further improvements by including more binaural and monaural features (see e.g. [39]). The seamless integration of binaural and monaural cues in the classification framework provides

a natural way for the system to leverage whatever discriminant features that exist in a particular environment to segregate the target signal, a characteristic of human auditory scene analysis [5], [9].

Acknowledgements

This work was performed while the first author was a visiting scholar at the Ohio State University. The authors wish to thank the Ohio Supercomputing Center for providing computing resources. The authors also thank Michael Mandel, Nicoleta Roman, Yuxuan Wang, and John Woodruff for making implementations of their algorithms available to us. Wang's research was supported in part by an AFOSR grant (FA9550-12-1-0130).

References

[1] J. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943-950, 1979.

[2] M.C. Anzalone, L. Calandruccio, K. A. Doherty, and L.H. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear Hear*, vol. 27, no. 5, pp. 480-492, 2006.

[3] D.G. Sinex, "Recognition of speech in noise after application of time-frequency masks: Dependency on frequency and threshold parameters," *J. Acoust. Soc. Amer.*, vol. 133, no. 4, pp. 2390-2396, 2013.

[4] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization.* Cambridge, MA: MIT Press, 1997.

[5] A.S. Bregman, *Auditory Scene Analysis.* Cambridge, MA: MIT Press, 1990.

[6] D.S. Brungart, P.S. Chang, B.D. Simpson, and D.L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4007-4018, 2006.

[7] D.R. Campbell, K.J. Palo, and G.J. Brown, "A MATLAB simulation of "shoebox" room acoustics for use in research and teaching," *Computing and Information Systems Journal*, vol. 9, no. 3, pp. 48-51, 2005.

[8] E.C. Cherry, *On human communication.* Cambridge, MA: MIT Press, 1957.

[9] C.J. Darwin, "Listening to speech in the presence of other sounds," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1493, pp. 1011-1021, 2008.

[10] J. Garofolo, et al. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, Philadelphia, PA, 1993

[11] S. Harding, J. Barker, and G.J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 58-67, 2006.

[12] E.W. Healy, S.E. Yoho, Y.X. Wang, and D.L. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust Soc. Amer.*, vol. 134, no. 4, pp. 3029-3038, 2013.

[13] K. Han and D.L. Wang, "A classification based approach to speech segregation," *J. Acoust. Soc. Amer.*, vol. 132, no. 5, pp. 3475-3483, 2012.

[14] S. Harding, J. Barker, and G.J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 58-67, 2006.

[15] L.M. Heller and V.M. Richards, "Binaural interference in lateralization thresholds for interaural time and level differences," *J. Acoust. Soc. Amer.*, vol. 128, no. 1, pp. 310-319, 2010.

[16] G.E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Compt.*, vol. 18, no. 7, pp. 1527-1554, 2006.

[17] G.N. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networ.*, vol. 15, no. 5, pp. 1135-1150, 2004.

[18] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867-1871, 2010.

[19] Z.Z. Jin and D.L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 625-638, 2009.

[20] G. Kim, Y. Lu, Y. Hu, and P.C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1486-1494, 2009.

[21] K. Kokkinakis, O. Hazrati, and P. C.Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants," *J. Acoust. Soc. Amer.*, vol. 129, no. 5, pp. 3221-3232, 2011.

[22] N. Li and P.C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, no. 3, pp. 1673-1682, 2008.

[23] M.I. Mandel, R.J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382-394, 2010.

[24] T. May, S. Van, and 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, and Lang. Proc.*, vol. 20, no. 7, pp. 2016-2030, 2012.

[25] R. Meddis, "Simulation of auditory neural transduction: Further studies," *J. Acoust. Soc. Amer.*, vol. 83, no. 3, pp. 1056-1063, 1988.

[26] T. Nakatani, M. Goto and H.G. Okuno. "Localization by harmonic structure and its application to harmonic sound stream segregation," In *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Process.*, 1996, pp. 653-656.

[27] R.D. Patterson, I. Nimmo , J. Holdsworth, and P. Rice, "SVOS final report, part B: Implementing a gammatone filterbank," Rep. 2341, *MRC Applied Psychology Unit*, 1988.

[28] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation*, S. Makino, T. Lee and H. E. Sawada, Ed. New York: Springer, 2007.

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

[30] N. Roman, D.L. Wang, and G.J. Brown, "A classification-based cocktail-party processor," in *Proc. Advances in Neural Information Processing Systems*, 2003, pp. 1425-1432.

[31] N. Roman, S. Srinivasan, and D.L. Wang, "Binaural segregation in multisource reverberant environments.," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4040-4051, 2006.

[32] N. Roman and J. Woodruff, "Intelligibility of reverberant noisy speech with ideal binary masking," *J. Acoust. Soc. Amer.*, vol. 130, no. 4, pp. 2153-2161, 2011.

[33] M.L. Seltzer, B. Raj, and B. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Commun.*, vol. 3, no. 4, pp. 379-393, 2004

[34] A. Shamsoddini and P.N. Denbigh, "A sound segregation algorithm for reverberant conditions," *Speech Commun.*, vol. 33, no. 3, pp. 179-196, 2001.

[35] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247-251, 1993.

[36] D.L. Wang and G.J. Brown, Ed. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications.* Hobboken, NJ Jersey: WILEY-IEEE Press, 2006.

[37] D.L. Wang, U. Kjems, M. Pedersen, J. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Amer.*, vol. 125, no. 4, pp. 2336-2347, 2009.

[38] D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Boston, MA: Kluwer, 2005.

[39] Y.X. Wang, K. Han and D.L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 270-279, 2013.

[40] Y.X. Wang and D.L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 7, pp. 1381-1390, 2013.

[41] J. Woodruff and D.L. Wang, "Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 4, pp. 806-815, 2013.

[42] X.J. Zhao, Y. Shao, and D.L. Wang, "CASA-based robust speaker identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1608-1616, 2012.