Joint estimation of binaural distance and azimuth by exploiting deep neural networks

Jiance Ding, Yuxuan Ke, Linjuan Cheng, et al.

Citation: The Journal of the Acoustical Society of America **147**, 2625 (2020); doi: 10.1121/10.0001155

View online: https://doi.org/10.1121/10.0001155

View Table of Contents: https://asa.scitation.org/toc/jas/147/4

Published by the Acoustical Society of America

JASA ARTICLE

Joint estimation of binaural distance and azimuth by exploiting deep neural networks

Jiance Ding,[a)] Yuxuan Ke,[a)] Linjuan Cheng,[a)] Chengshi Zheng,[a),b)] and Xiaodong Li[a)]
Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Science, 100190, Beijing, China

ABSTRACT:
The state-of-the-art supervised binaural distance estimation methods often use binaural features that are related to both the distance and the azimuth, and thus the distance estimation accuracy may degrade a great deal with fluctuant azimuth. To incorporate the azimuth on estimating the distance, this paper proposes a supervised method to jointly estimate the azimuth and the distance of binaural signals based on deep neural networks (DNNs). In this method, the subband binaural features, including many statistical properties of several subband binaural features and the binaural spectral magnitude difference standard deviation, are extracted together as cues to jointly estimate the azimuth and the distance using binaural signals by exploiting a multi-objective DNN framework. Especially, both the azimuth and the distance cues are utilized in the learning stage of the error back-propagation in the multi-objective DNN framework, which can improve the generalization ability of the azimuth and the distance estimation. Experimental results demonstrate that the proposed method can not only achieve high azimuth estimation accuracy but can also effectively improve the distance estimation accuracy when compared with several state-of-the-art supervised binaural distance estimation methods.
https://doi.org/10.1121/10.0001155

(Received 7 November 2019; revised 6 March 2020; accepted 8 April 2020; published online 27 April 2020)

[Editor: Efren Fernandez-Grande] Pages: 2625–2635

I. INTRODUCTION

The human auditory system has a remarkable ability to detect sound direction and sound distance, even in very complex acoustic environments (Blauert, 1996; Hofman *et al.*, 1998; Zhong and Yost, 2015). This localization ability is very important because it allows humans to sense and interact with their surrounding environments. In the last few decades, numerous researchers have developed many binaural localization methods and systems to simulate a human's listening ability to analyze and understand sound events. In this paper, we focus on binaural localization, where both the azimuth and the distance are considered together.

Binaural azimuth and distance localization techniques are aimed at estimating the azimuth and the distance of sound sources using binaural recording microphones, which have wide applications, such as hearing-aids (Farmani *et al.*, 2018), virtual sound reproduction (Liu *et al.*, 2015; Yu *et al.*, 2017), and intelligent robot control (Magassouba *et al.*, 2018). With the progress of techniques in hearable devices, binaural sound source localization has become a popular topic. In the past few decades, many binaural azimuth estimation (BAE) methods and binaural distance estimation (BDE) methods have already been proposed. However, most of these existing BAE and BDE methods only aim at estimating one of the spatial parameters (i.e., either the azimuth or the distance of a point sound source).

Since Jeffress (1948) proposed the coincident theory, many BAE methods have been proposed. Most of the BAE methods estimate the binaural azimuth by using the interaural time/phase differences (ITDs/IPDs) and the interaural level differences (ILDs). Generally, these methods can be categorized into two classes: the traditional unsupervised BAE method and the supervised BAE method. Most traditional unsupervised BAE algorithms usually match the binaural features of binaural signals with templates to obtain the arrival direction of sound sources (Chen *et al.*, 2017; Li and Levinson, 2003; Zhong *et al.*, 2016). Their performance may degrade significantly in noisy and reverberant environments. With the development of machine learning techniques, many supervised BAE algorithms have been proposed in recent years (Deleforge *et al.*, 2015; Ma *et al.*, 2017; May *et al.*, 2016). These supervised methods often learn an affine transformation model from the binaural features to the corresponding azimuths of sound sources using a large training dataset. Previous studies have shown that these supervised algorithms could achieve a promising performance even in noisy and reverberant environments, especially when the training conditions fit well with the test conditions.

Compared with the binaural azimuth localization, the BDE method has not been well studied, and only few BDE methods have been proposed in the past few years. Generally, these BDE methods can also be categorized into two classes: the traditional BDE methods based on computational acoustic

[a)] Also at: University of Chinese Academy of Sciences, 100049, Beijing, China.
[b)] Electronic Email: cszheng@mail.ioa.ac.cn

models (Kolarik *et al.*, 2013; Lu and Cooke, 2010; Zahorik, 2002) and the supervised BDE algorithms based on binaural features mapping (Georganti *et al.*, 2013; Vesa, 2009). Most traditional BDE methods estimate the sound source distance by exploiting the direct-to-reverberant ratio (DRR) of binaural signals. The DRR is not only related to the distance of a sound source, but it is also related to the room characteristics (Jetzt, 1979; Tohyama, 1995). Therefore, in these traditional BDE methods, the knowledge of room characteristics is required, which may not be available in most cases. The supervised BDE algorithms exploit machine learning-based methods to train the relationship between binaural features and sound source distance. These methods could achieve high distance estimation accuracy in matched environments, and they often did not rely on the estimation of the DRR of binaural signals directly.

Theoretically, most binaural features are naturally related to both the azimuth and distance of a point sound source. Thus, it is difficult to completely separate the BDE cues from the BAE cues. According to the theory of human auditory perception, the variation of azimuth has greater impact on binaural features when compared with the variation of distance, and thus the distance estimation performance of the supervised BDE algorithms may degrade significantly when the azimuth changes and vice versa (Georganti *et al.*, 2013; Shi and Xie, 2010; Yu *et al.*, 2012). To improve the distance estimation accuracy, it is reasonable to estimate the distance of the point sound source using binaural signals with the help of its azimuth. Venkatesan and Ganesh (2017) proposed a complete localization method using several binaural signals based on Gaussian Mixture Models (GMMs), which estimated the azimuth and the distance separately, and thus its performance of the distance estimation was still affected by the variation of azimuth. Yiwere and Rhee (2017) proposed to simultaneously estimate the distance and the azimuth using binaural features based on deep neural networks (DNNs), while the performance of the distance estimation was still unsatisfactory in reverberant and noisy environments.

To reduce the adverse impact of the variation of the azimuth on the distance estimation performance, this paper proposes a joint binaural azimuth and distance estimation method by using a multi-objective DNN framework (JBADE-DNN). In the proposed method, both binaural features and statistical properties of binaural signals are extracted and used as the input features to the DNN, and the azimuth and the distance are the output targets of the DNN. The JBADE-DNN can not only estimate the distance and the azimuth simultaneously but can also reduce the negative impact of the variation of the azimuth on the distance estimation using binaural signals. Both simulation and experimental results demonstrate that the JBADE-DNN can not only achieve comparable azimuth estimation accuracy when compared with the state-of-the-art supervised BAE methods but also effectively improve the distance estimation accuracy in reverberant and noisy environments. It is interesting to see that the proposed JBADE-DNN method is much more robust than the competing methods in untrained conditions.

II. SIGNAL MODEL

We assume that the binaural recording microphones are placed at the two ears of a head separately. Let $s(n)$ be a point sound source, then the binaural signals, $x_l(n)$ and $x_r(n)$, can be written as

$$x_{l|r}(n) = h_{l|r}^{(\theta,\varphi,d)}(n) * s(n) + e_{l|r}(n), \tag{1}$$

where $x_{l|r}(n)$ represents either $x_l(n)$ or $x_r(n)$ for compact notations, and * represents the convolution operation. θ, φ, and d represent the azimuth, the elevation, and the distance of the point sound source, respectively. denotes the binaural room impulse responses (BRIRs) from the point sound source to the binaural recording microphones. $e_{l|r}(n)$ denotes the received noise and the internal noise signals. This paper aims at estimating the azimuth and distance of only one point sound source without considering its elevation, and thus $h_{l|r}^{(\theta,\varphi,d)}(n)$ can be simplified to $h_{l|r}^{(\theta,d)}(n)$ in the following for simplicity.

In this paper, all the azimuth and the distance cues are extracted in the frequency domain and the subband domain. In the frequency domain, Eq. (1) can be rewritten as

$$X_{l|r}(m,k) = H_{l|r}^{(\theta,d)}(k)S(m,k) + E_{l|r}(m,k), \tag{2}$$

where $X_{l|r}(m,k)$, $H_{l|r}^{(\theta,d)}(k)$, $S(m,\ k)$, and $E_{l|r}(m,k)$ are the N_{STFT}-point short-time Fourier transform (STFT) of $x_{l|r}(n)$, $h_{l|r}^{(\theta,d)}(n)$, $s(n)$, and $e_{l|r}(n)$, respectively. m and k are the frame index and the frequency index, respectively. Considering the human auditory characteristics, the binaural signals are often split into several subband signals using the Gammatone filterbanks. In the subband domain, the binaural signals can be given by

$$x_{l|r,c}(m,\acute{n}) = g_c(\acute{n}) * x_{l|r}(m,\acute{n}), \tag{3}$$

where $x_{l|r}(m,\acute{n})$ is the mth frame of $x_{l|r}(n)$, $\acute{n} \in \{0, 1, \ldots, N_{\mathrm{STFT}} - 1\}$, and $g_c(\acute{n})$ is the impulse response of the c-th filter of the Gammatone filterbanks.

III. PROPOSED METHOD

The two most frequently used binaural features in BAE methods are ILD and ITD, while Ma *et al.* (2017) pointed out that the cross-correlation function (CCF) is a more robust binaural cue than ITD for BAE in the presence of noise and reverberation. For the BDE problem, researchers often use binaural features, such as ILD, ITD, CCF, and interaural coherence (IC), as distance estimation cues. While the study in Georganti *et al.* (2013) demonstrated that better performance can be achieved with statistical properties of binaural signals. This may be due to the fact that statistical properties of binaural signals are more robust to noise and reverberation, and their values can be barely affected when the azimuth of the point sound source changes slightly. Almost all binaural features and statistical properties of binaural signals are closely related to both the azimuth and distance of the point sound source. Accordingly, a JBADE-DNN is proposed to estimate the distance and azimuth of a point sound source with binaural signals, simultaneously. For this purpose, many features are extracted as the azimuth and the distance estimation cues, including the ILD, the CCF, and five statistical properties of subband binaural features calculated from the ILD, the ITD, and the IC. In addition, the binaural spectral magnitude difference standard deviation (BSMD-STD) is also computed as an input feature. The procedure of the proposed JBADE-DNN method is summarized in Fig. 1.

A. Binaural subband features extraction

In Fig. 1, the binaural signals are decomposed into 8 auditory channels to extract binaural features in subband domain by using a fourth-order Gammatone filterbank. The center frequencies of the Gammatone filterbanks are equally spaced on the Equivalent Rectangular Bandwidth (ERB) scale between 80 Hz and 8000 Hz, which are about 80 Hz, 260 Hz, 560 Hz, 1030 Hz, 1790 Hz, 2990 Hz, 4920 Hz, and 8000 Hz, respectively. The entire CCF and the ILD of each frequency band are extracted in time domain as subband binaural features (Ma *et al.*, 2017). ILD represents the sound pressure level difference in the two ears, which can be computed by

$$\gamma_{\mathrm{ILD},c}(m) = 20\log_{10}\left(\frac{\sum_{\acute{n}} |x_{l,c}(m,\acute{n})|^2}{\sum_{\acute{n}} |x_{r,c}(m,\acute{n})|^2}\right), \tag{4}$$

where c is the auditory channel index, with $c = 0, 1, \ldots, 7$. CCF is a binaural feature to measure the correlation between the binaural signals in each subband, which is defined as

$$\mathrm{CCF}_c(\tau) = \frac{\sum_{\acute{n}} \left(x_{l,c}(m,\acute{n}) \cdot x_{r,c}(m,\acute{n}+\acute{n}_\tau)\right)}{\sqrt{\left(\sum_{\acute{n}} |x_{l,c}(m,\acute{n})|^2\right) \cdot \left(\sum_{\acute{n}} |x_{r,c}(m,\acute{n})|^2\right)}}, \tag{5}$$

where $\acute{n}_\tau$ is the delayed sampling points, $\tau = \acute{n}_\tau / f_s$ is the time delay corresponding to $\acute{n}_\tau$, and f_s is the sampling rate. τ is constrained to $[-1, 1]$ ms according to the size of the head (Wang and Brown, 2006). For $f_s = 16$ kHz, the entire CCF, with $\tau \in [-1, 1]$ ms, is a 33-dimensional binaural feature vector in each frequency band. Note that the sampling frequency is a fixed value throughout this paper, which is due to the fact that we only consider speech as the point sound source to obtain binaural signals. For speech, the sampling frequency $f_s = 16$ kHz is high enough. When considering the stimuli that include music and noises with various bandwidths, we need to use a higher sampling frequency, and thus the performance of binaural localization may be improved for practical applications.

In summary, the total dimension of binaural subband features is 272 for each frame. These 272-dimensional binaural subband features are averaged using M successive frames to get their mean values, and thus one can also get a 272-dimensional feature vector for every M successive frames.

B. Statistical properties extraction

1. Statistical properties of binaural features

Let $\gamma_{\mathrm{ILD},c}(m)$, $\eta_{\mathrm{ITD},c}(m)$, and $\xi_{\mathrm{IC},c}(m)$ denote the ILD, the ITD, and the IC of the c-th band and the m-th frame, respectively. The binaural subband features of the m-th frame can be given by

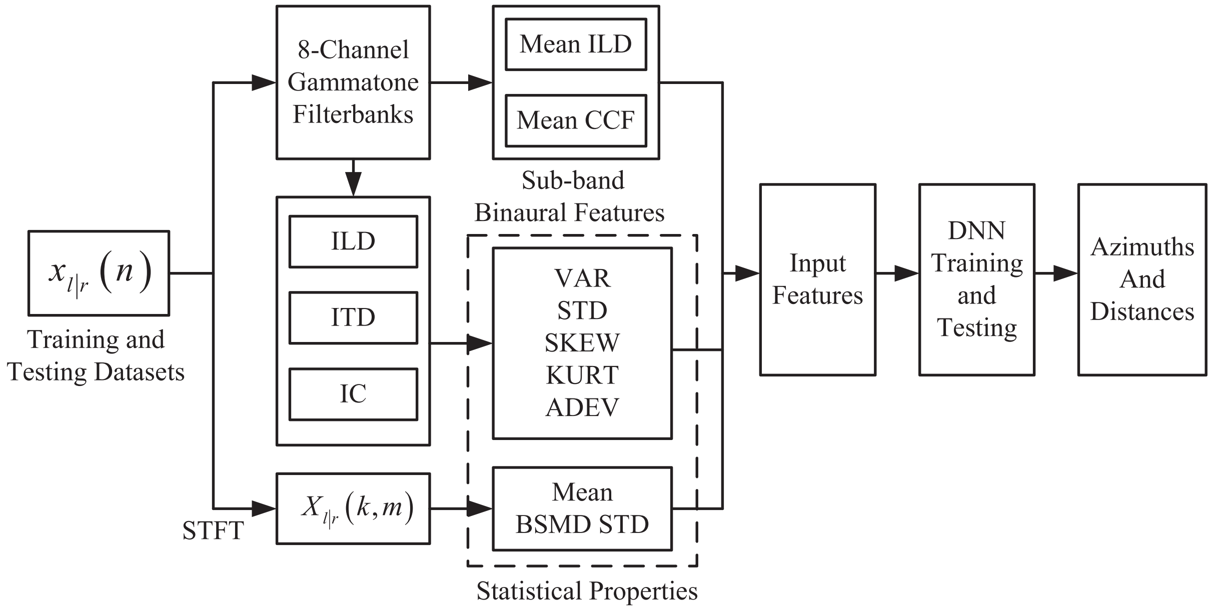

FIG. 1. The procedure of the JBADE-DNN method.

$$\Gamma(m)=\left\{\{\gamma_{\mathrm{ILD},c}(m)\}_{c=0}^{7},\{\eta_{\mathrm{ITD},c}(m)\}_{c=0}^{7},\{\xi_{\mathrm{IC},c}(m)\}_{c=0}^{7}\right\}. \tag{6}$$

For every M frames, we calculate five statistical properties, which include variance (VAR), standard deviarion (STD), skewness (SKEW), kurtosis (KURT), and average deviation (ADEV). Let $B_c(m)$ represent $\{\gamma_{\mathrm{ILD},c}(m)\}$, $\{\eta_{\mathrm{ITD},c}(m)\}$, or $\{\xi_{\mathrm{IC},c}(m)\}$, then these statistical properties of binaural features can be, respectively, computed as follows.

(1) Standard deviation (STD): the standard deviation of binaural subband features of the c-th Gammatone channel is denoted as

$$\sigma_c=\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(B_c(m)-\bar{B}_c(m)\right)^2}, \tag{7}$$

where $\bar{B}_c(m)$ is the mean value of $B_c(m)$ over M frames.

(2) Variance (VAR): the variance of binaural subband features is estimated by

$$\nu_c=\frac{1}{M}\sum_{m=1}^{M}\left(B_c(m)-\bar{B}_c(m)\right)^2. \tag{8}$$

(3) Skewness (SKEW): The skewness of binaural subband features can be computed by

$$\mu_c=\frac{\frac{1}{M}\sum_{m=1}^{M}\left(B_c(m)-\bar{B}_c(m)\right)^3}{\left(\frac{1}{M}\sum_{m=1}^{M}\left(B_c(m)-\bar{B}_c(m)\right)^2\right)^{\frac{3}{2}}}. \tag{9}$$

(4) Kurtosis (KURT): the kurtosis of binaural subband features can be given by

$$\kappa_c=\frac{\frac{1}{M}\sum_{m=1}^{M}\left(B_c(m)-\bar{B}_c(m)\right)^4}{\left(\frac{1}{M}\sum_{m=1}^{M}\left(B_c(m)-\bar{B}_c(m)\right)^2\right)^{2}}. \tag{10}$$

(5) Average deviation (ADEV): the average deviation extracted from binaural subband features is defined as

$$\zeta_c=\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left|B_c(m)-\bar{B}_c(m)\right|}. \tag{11}$$

2. Binaural spectral magnitude difference standard deviation

Based on Georganti *et al.* (2014), the BSMD-STD of binaural signals in room environments can be approximated as

$$\sigma_{\mathrm{BSMD}}\approx 5.57\sqrt{2\frac{1+2\mathrm{DRR}_{\mathrm{dB}}}{\left(1+\mathrm{DRR}_{\mathrm{dB}}\right)^2}}, \tag{12}$$

where $\mathrm{DRR}_{\mathrm{dB}}$ is the DRR of binaural signals in decibels. The DRR of binaural signals is related to both the room characteristics and the sound source distance (Tohyama, 1995), which can be given by

$$\mathrm{DRR}_{\mathrm{dB}}\approx 10\log_{10}\left(\frac{A\bar{\alpha}}{16\pi d^2}\right), \tag{13}$$

where A is the room surface area, $\bar{\alpha}$ is the average sound absorption coefficient of the room, and d is the distance of the point sound source. According to Eqs. (12) and (13), the BSMD-STD is highly correlated with the sound source distance in a room. The details for the calculation of the BSMD-STD feature can be referred to (Georganti *et al.*, 2013, 2014). The BSMD-STD is calculated in each frame, and its mean value for every successive M frames is also calculated as a distance cue.

In summary, for every M successive frames, the total number of statistical properties of binaural features is 121, which forms a 121-dimensional feature vector.

C. Multi-objective DNN framework

A multi-objective DNN framework is introduced to map both the azimuth and the distance cues to their corresponding azimuth and distance of the point sound source. In this paper, the DNN consists of an input layer, three hidden layers, and an output layer. The input layer contains 393 nodes, which is the concatenation of the 272-dimensional binaural subband feature vector and the 121-dimensional statistical feature vector. Each hidden layer contains 256 nodes, and the sigmoid activation function is applied at the hidden layers. The number of nodes in the hidden layer are heuristically chosen through numerous experiments. The output layer contains 2 nodes, which correspond to the azimuth and distance of the point sound source, and the ReLU activation function is applied at the output layer. All of the input features are normalized to be zero mean and unit variance, and white Gaussian noise is injected into the input features to avoid the overfitting problem, where the variance of this additive white Gaussian noise is 10^{-4}. To incorporate the estimation errors of the azimuth and those of the distance on the training errors, we normalize both the azimuths and the distances to the range of $[0,1)$. The loss function for the multi-objective DNN is defined as follows:

$$\mathcal{J}=\sum_{i\in\mathcal{I}}\left(\left(\hat{\theta}(i)-\theta(i)\right)^2+\left(\hat{d}(i)-d(i)\right)^2\right), \tag{14}$$

where $\mathcal{I}$ is the total training data size in the training stage, $\theta(i)$ and $d(i)$ are the normalized true azimuth and the normalized true distance of the i-th signal segment, while $\hat{\theta}(i)$ and $\hat{d}(i)$ denote their corresponding estimation values, respectively. The initial learning rate is set to 0.18, which is

https://doi.org/10.1121/10.0001155

gradually decreased to 0.001 after 100 epochs. The adaptive gradient descent is used to train the multi-objective DNN with the momentum rate set to 0.5 at the first five epochs and 0.9 for the rest. The batch size is 1024. The maximum number of iterations for the DNN is set to 100. To improve the generalization ability, the dropout regularization is also adopted for both the input layer and the hidden layers, where the dropout rate is 10% for the input layer and 20% for the three hidden layers.

D. The training and testing datasets

The training and the testing datasets are built in both simulated and real room environments to train and evaluate the proposed JBADE-DNN method. In the simulated environments, the binaural signals are generated by convolving clean speech signals with BRIRs. The clean speech signals are randomly chosen from the TIMIT dataset (Garofolo, 1979). The head related impulse responses (HRIRs) of the MIT head related transfer function (HRTF) databases (Gardner and Martin, 1994) are used to create the BRIRs in simulated rooms with the image method (Allen and Berkley, 1988). To simulate different reverberant environments, the BRIRs with different azimuths and different distances of the point sound source are generated in four simulated rooms. The dimensions of the four simulated rooms, their reverberation radii, their reverberation times (T_{60}), and the distances from the point sound source to the head for all the rooms are summarized in Table I. In each room, the azimuths of the point sound source are set at $\{-90°, -75°, \ldots, 90°\}$ with the interval of 15°. In the training stage, 1200 clean speech signals are taken from the TIMIT dataset to generate binaural signals to form the training datasets. In the testing stage, 120 clean speech signals are chosen to get the testing datasets for each azimuth and each distance. The multi-objective DNN models are trained and evaluated separately for each room. That is to say, the training dataset contains 46 800 binaural speech signals, and the testing dataset contains 4680 binaural speech signals for each room.

In real environments, the binaural signals are gathered by the head and torso simulator type 4128-C produced by Brüel & Kjær Sound & Vibration (Nærum, Denmark) at different azimuths and different distances from the point sound source to the head in five different rooms. In each room, the experimental setups are shown in Fig. 2. The room sizes, their reverberation radii, their reverberation times (T_{60}), and the distances from the point sound source to the head in the five rooms are summarized in Table II. Similarly, the azimuths of the point sound source are $\{-90°, -75°, \ldots, 90°\}$ with the interval of 15°. At each azimuth and each distance, binaural speech signals are recorded as 300 s long. The first 240 s of binaural signals are cut to use as training datasets, and the last 60 s are used as testing datasets. By doing so, it is obvious that all the testing positions lie on the training position, which is known as the on-grid localization problem. The off-grid localization problem (Ding et al., 2020), where the testing positions are different from the training positions, is not considered in this paper. We focus on studying the impact of noise and reverberation on binaural localization. Additionally, the testing rooms that are different from the training rooms are also considered. Only the binaural signals in Room 1, Room 2, and Room 3 are used as the training and the testing datasets, and all the binaural signals in Room 4 and Room 5 are used as testing signals to evaluate the proposed JBADE-DNN method in untrained environments.

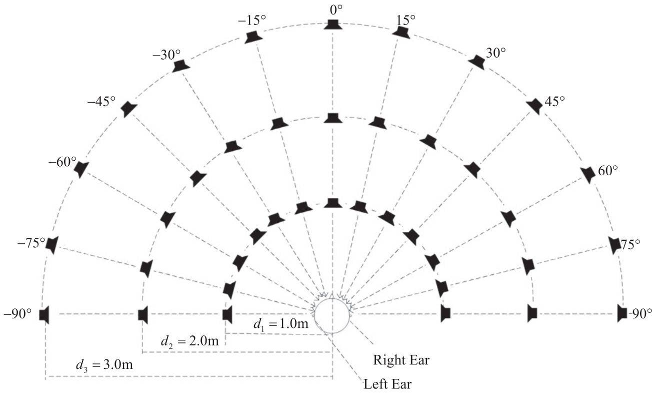

FIG. 2. The experimental setups in real room environments.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

The proposed JBADE-DNN[1] method is evaluated in both simulated and real room environments in this section. The performance of the azimuth estimation of the proposed JBADE-DNN method is compared with the DNN-based BAE method (BAE-DNN[2]) in Ma et al. (2017). For BDE, the proposed JBADE-DNN method is compared with two state-of-the-art supervised BDE algorithms, where one is the GMM-based BDE algorithm based on statistical properties (SP-GMM[3]) (Georganti et al., 2013), and the other is the DNN-based BDE algorithm using binaural features (BF-DNN[4]) (Yiwere and Rhee, 2017).

TABLE I. Room sizes and the distances from the point sound source to the head in simulated room environments. 'r_0' represents the reverberation radius of a room.

	Room size (m^3)	r_0	T_{60} (s)	D (m)
Room A	3.0 × 1.8 × 2.2	0.65	0.1	{0.5, 1.0, 1.5}
Room B	5.0 × 6.4 × 2.9	1.05	0.3	{1.0, 2.0, 3.0}
Room C	10.8 × 10.9 × 3.2	1.55	0.6	{1.0, 2.0, 3.0}
Room D	16.0 × 12.0 × 3.2	1.60	0.9	{1.0, 2.0, 3.0}

TABLE II. Room characteristics and source/receiver distances of five different real rooms.

	Room size (m^3)	r_0	T_{60} (s)	D (m)
Room 1	4.8 × 6.4 × 2.8	1.60	0.4	{1.0, 2.0, 3.0}
Room 2	4.8 × 6.4 × 2.8	0.76	0.6	{1.0, 2.0, 3.0}
Room 3	7.5 × 6.4 × 2.8	0.87	0.8	{1.0, 2.0, 3.0}
Room 4	4.3 × 3.7 × 2.6	0.65	0.5	{1.0, 2.0, 3.0}
Room 5	13.0 × 7.5 × 2.8	1.25	1.1	{1.0, 2.0, 3.0}

The azimuth estimation accuracy is defined as:

$$A_{\text{azi}} = \frac{N_{|\hat{\theta}-\theta \leq 7.5^\circ|}}{N} \times 100\%, \tag{15}$$

where N is the total number of binaural samples, θ and $\hat{\theta}$ are the true azimuth and the estimated azimuth of a point sound source, and $N_{|\hat{\theta}-\theta \leq 7.5^\circ|}$ indicates the total number of binaural samples in which their azimuth estimation errors are smaller than 7.5°.

Similarly, the distance estimation accuracy is defined as:

$$A_{\text{dis}} = \frac{N_{|\hat{d}-d \leq e_d|}}{N} \times 100\%, \tag{16}$$

where d and $\hat{d}$ are the true distance and the estimated distance of a point sound source, and $N_{|\hat{d}-d \leq e_d|}$ is the total number of binaural samples in which their distance estimation errors are smaller than e_d. e_d is set to 0.25 for Room A, while it is set at 0.5 for other rooms.

A. Experimental results in simulated environments

1. Comparison in noise-free and reverberant environments

We first evaluate the performance of the proposed JBADE-DNN method in noise-free and reverberant environments. The binaural signals are generated by convolving clean speech signals with BRIRs in the four simulated rooms. The room characteristics and the distances of sound sources are presented in Table I, and the azimuths of these point sound sources are $\{-90^\circ, -75^\circ, \ldots, 90^\circ\}$. The duration of each binaural signal is about one second for evaluation, and all the binaural features are computed for every 20 ms with 50% overlap. Finally, the azimuth and the distance cues are computed using the entire one-second binaural signals. To improve the robustness of the proposed method, note that we apply voice activity detection (VAD) to remove noise-only frames before computing the input features. In this paper, the frame energy-based VAD is used. For practical applications, many robust VAD methods can be used (see Tan *et al.*, 2020 and references therein). After removing noise-only frames, M indicates the total number of the speech frames of binaural signals. The azimuth estimation accuracy of the JBADE-DNN method and that of the competing BAE method are shown in Fig. 3. One can see that the azimuth estimation accuracy of the proposed JBADE-DNN method is comparable with the BAE-DNN method in all cases, although the BAE-DNN method is slightly better than JBADE-DNN because of more azimuth features and more complicated DNN structures. The distance estimation accuracy of the proposed JBADE-DNN method and that of the compared BDE methods, including the BF-DNN and the SP-GMM, are given in Fig. 4. As shown in Fig. 4, the distance estimation accuracy of the JBADE-DNN method is about 3%–5% higher than the other two competitive BDE methods. This is mainly due to the fact that the JBADE-DNN method estimates the distance and the azimuth of binaural signals simultaneously using the multi-objective DNN framework, which improves the generalization performance of the distance estimation because of the benefits of the multitask learning. For all room conditions, both the azimuth and the distance estimation accuracies are decreased when the reverberation time increases. The main reason is that the variances of binaural features increase with the increase of the reverberation.

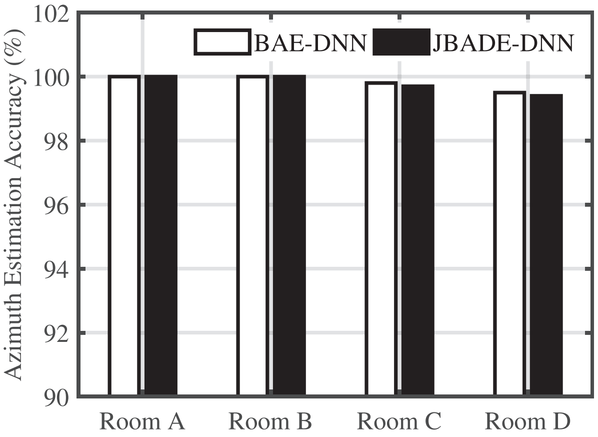

FIG. 3. The azimuth estimation accuracy of the JBADE-DNN method and the BAE-DNN method in noise-free and reverberant environments.

2. Comparison in noisy environments

In this part, we evaluate the performance of the proposed JBADE-DNN method in noisy environments. The binaural recordings are generated similarly to Sec. IV A 1. Diffused noises are injected into binaural signals to simulate different noisy conditions. To simulate a diffuse noise field, we generate 72 independent white Gaussian noise processes having the same powers and then convolve them with the HRIRs taken from the MIT HRTF database at 72 azimuths with the interval of about 5°, and then these 72 binaural signals are summed up together. Note that the sampling rate of

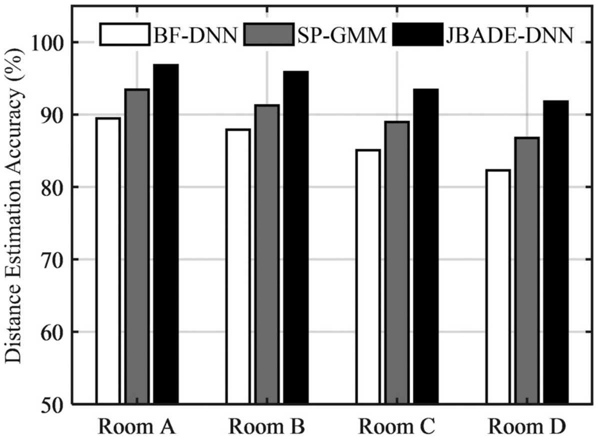

FIG. 4. The distance estimation accuracy of the proposed JBADE-DNN method and that of the BF-DNN method as well as SP-GMM method in noise-free and reverberant environments.

the white Gaussian noise processes are 16 kHz, and the HRIRs need to be resampled to 16 kHz before convolution. The signal-to-noise ratio (SNR) of the mixed signals varies from −10 dB to 20 dB with the interval of 10 dB.

The azimuth estimation accuracy of the JBADE-DNN method and that of the BAE-DNN method in different noisy environments are presented in Fig. 5.

One can see from Fig. 5 that both the JBADE-DNN method and the BAE-DNN method are robust in azimuth estimation in noisy environments. The distance estimation accuracy of the JBADE-DNN method and that of the competing BDE methods in noisy environments are given in Fig. 6, which shows that the JBADE-DNN method can achieve higher distance estimation accuracy and is more robust in noisy environments. Remarkably, the distance estimation performance gaps between the JBADE-DNN method and the competitive BDE methods in high SNRs are larger than the performance gaps in low SNRs, which indicates that better performance can be achieved when the statistical properties are also used. In the following, we study the importance of different features on localization by simulation.

3. Analysis of features on localization in the JBADE-DNN method

To study the importance of each binaural feature and its statistical properties on binaural azimuth and distance estimation separately, several experiments are conducted in simulated noisy and reverberant environments. In each experiment, only some subband features or statistical properties are chosen to estimate the azimuth and distance of the point sound source using the DNN framework. The simulated BRIRs of Room C are chosen to build the training and the testing datasets for each experiment, and then diffused noises are injected into binaural signals to simulate noisy environments. The SNR of the mixed signals in all experiments is fixed to 0 dB. All the other setups in these experiments are similar to Sec. IV A 1. Table III presents the azimuth and the distance estimation accuracy of all the experiments when using different input features, where ‘CCF(264)’ denotes the 264-dimensional subband CCF features, ‘ILD(8)’ means the 8-dimensional subband ILD features, ‘SP_ILD(40)’, ‘SP_ITD(40)’, and ‘SP_IC(40)’ represent the 40-dimensional statistical properties of subband ILD, ITD, and IC, respectively, and ‘BSMD-STD(1)’ is the 1-dimensional BSMD-STD feature.

Table III shows that different features lead to notably different performance on the binaural azimuth and distance estimation in the JBADE-DNN method. For the binaural azimuth estimation, the most effective features are the 264-dimensional subband CCF features, followed by the 8-dimensional subband ILD features, and all the statistical properties are not so important to the binaural azimuth estimation. This result is consistent with the previous study in

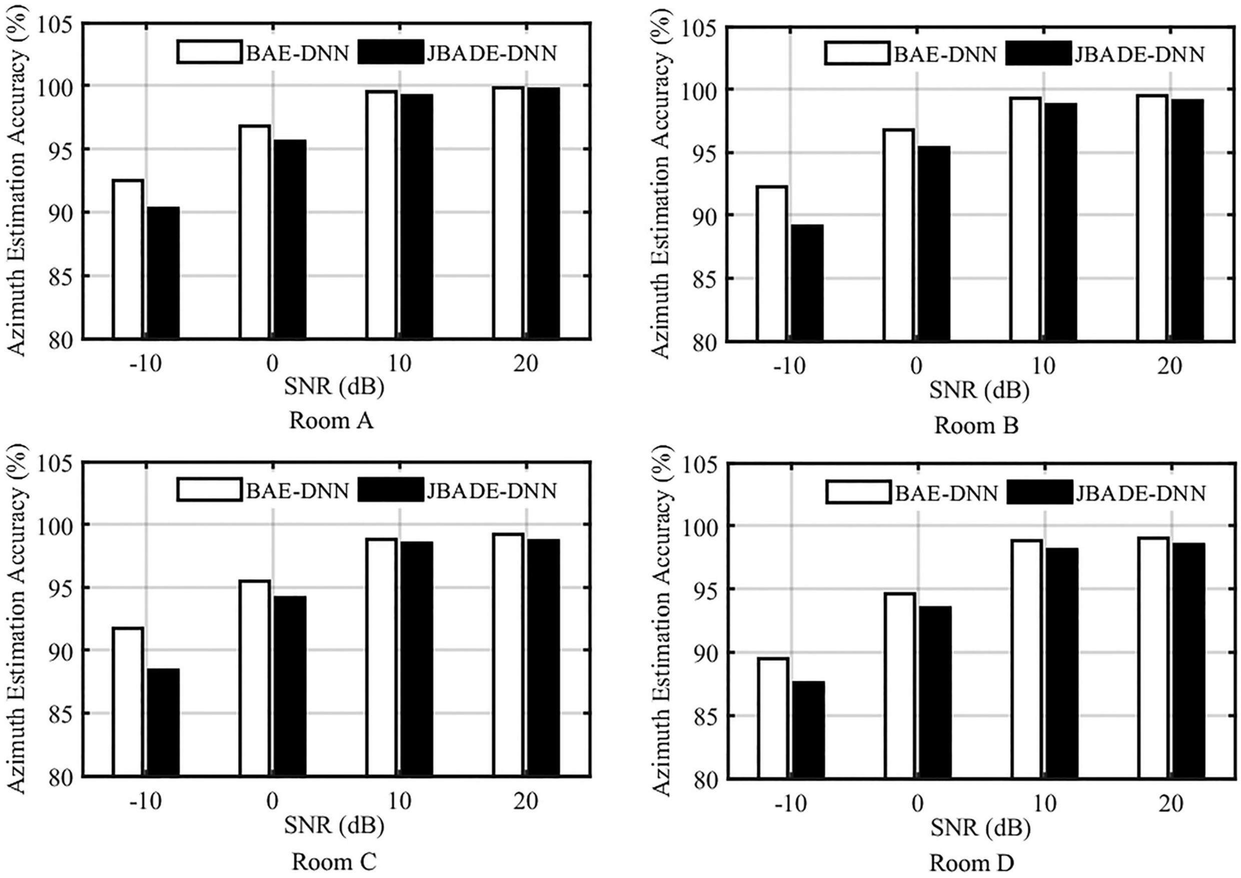

FIG. 5. The azimuth estimation accuracy of the JBADE-DNN method and that of the BAE-DNN method in different noisy environments.

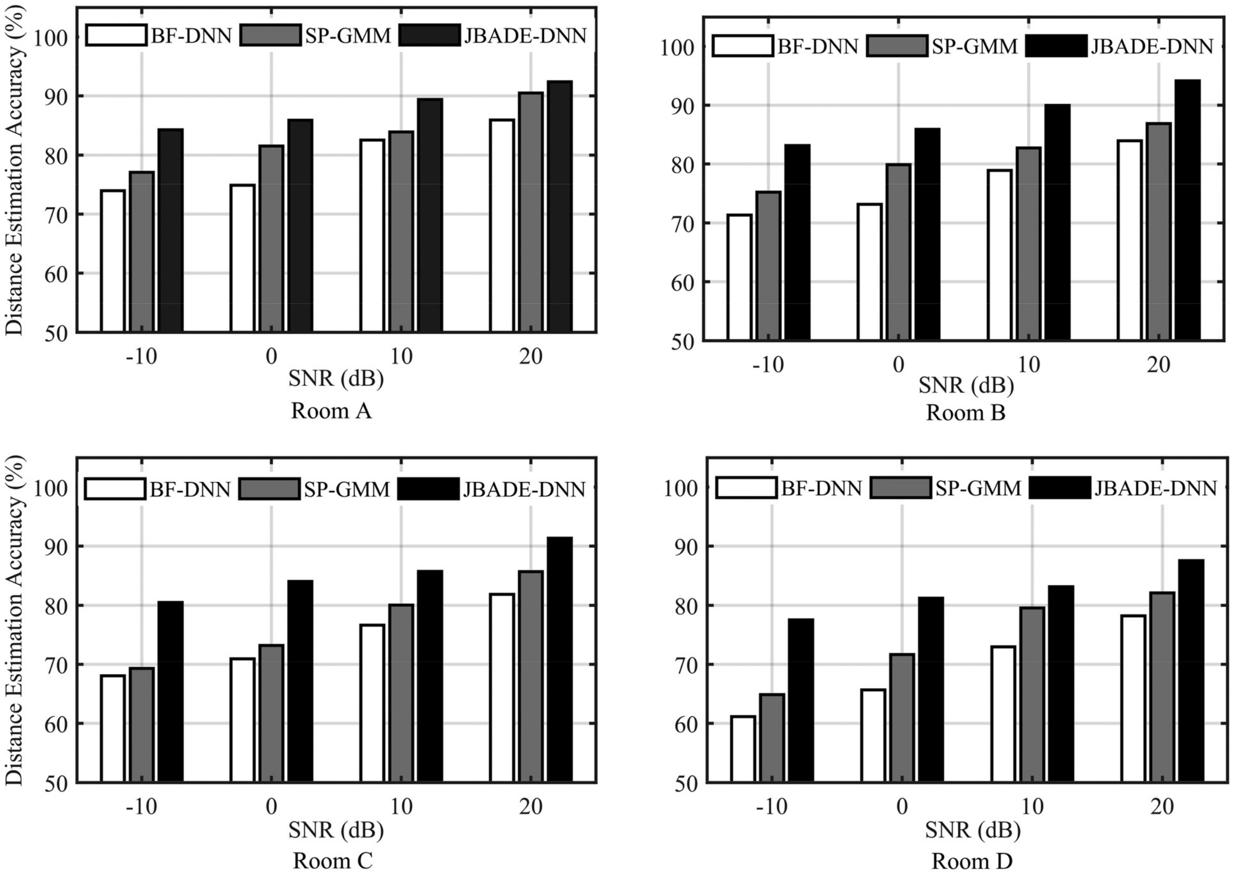

FIG. 6. The distance estimation accuracy of the JBADE-DNN method and the competing BDE methods in different noisy environments.

Ma *et al.* (2017). It is interesting to see that the order of these features, ranked by their importance for the binaural distance estimation, are quite different from their orders for the binaural azimuth estimation. For the binaural distance estimation, the orders of these features are the 8-dimensional subband ILD features, the 264-dimensional subband CCF features, the 1-dimensional BSMD-STD feature, the 40-dimensional statistical properties of subband ILD, the 40-dimensional statistical properties of subband IC, and the 40-dimensional statistical properties of subband ITD. Among these features, both of the subband CCF and subband ILD features are important to the binaural azimuth and distance estimation. Obviously, the statistical properties of subband binaural features are more important to the binaural distance estimation than the binaural azimuth estimation, which is consistent with the study in Georganti *et al.* (2013).

TABLE III. The azimuth and the distance estimation accuracy with different input features in simulated noisy and reverberant environments. '&' means "and" in this paper.

Input features	BAE	BDE
CCF (264)	92.6%	68.7%
ILD (8)	72.4%	70.9%
SP_ILD (40)	52.4%	51.7%
SP_ITD (40)	40.3%	35.3%
SP_IC (40)	39.2%	44.2%
BSMD-STD (1)	34.1%	59.4%
CCF&ILD (272)	96.2%	80.7%
CCF&BSMD-STD (265)	90.7%	81.9%
ILD&BSMD-STD (9)	68.9%	79.3%
All (393)	95.5%	84.2%

The last four rows of Table III list the azimuth and the distance estimation performance when exploiting the combinations of subband features or statistical properties. For the binaural azimuth estimation, the combinations of the subband CCF and the subband ILD can achieve the highest azimuth estimation accuracy. For the binaural distance estimation, all the four feature combinations can significantly improve the distance estimation performance.

4. Comparison in untrained environments

The previous experiments evaluate the JBADE-DNN method in matched conditions. While it is important to examine whether the proposed method can be generalized to untrained room conditions, in this experiment, the training model is obtained by only one of the rooms or parts of the rooms in Table I, and the testing datasets are established in another one. Figures 7 and 8 present the azimuth and the distance estimation accuracy in unmatched conditions, respectively. The combinations of different rooms for training and testing are shown in Figs. 7 and 8, where 'Room B(C)'

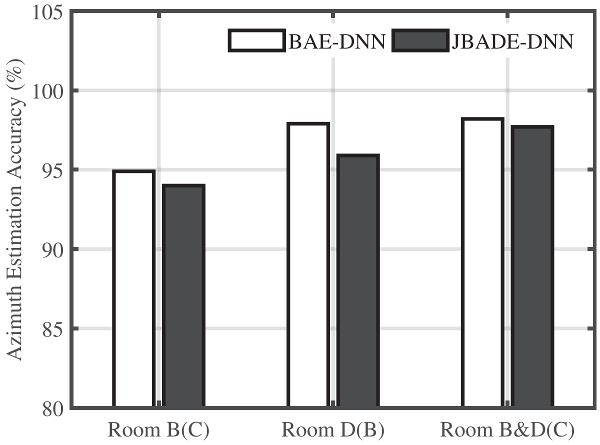

FIG. 7. The azimuth estimation accuracy in different unmatched room environments.

means the model is trained using the datasets recorded in Room B, while the testing datasets are recorded in Room C. Analogously, 'Room D(B)' and 'Room B&D(C)' mean the experiments in other unmatched conditions.

Figure 7 shows that both the JBADE-DNN method and the BAE-DNN method have high generalization ability in azimuth estimation. This is because the change of azimuth has higher impact on the mean binaural features than the change of room characteristics. One can see in Fig. 8 that the distance estimation accuracy of the JBADE-DNN method is about 13% − 18% higher than that of the competing BDE methods in unmatched conditions. The multi-objective DNN framework in the JBADE-DNN method has higher generalization performance than the other two competing methods. Moreover, both the binaural features and the statistical cues contain distance information, which improves the robustness of the JBADE-DNN method in distance estimation. It also can be seen that the distance estimation accuracy is higher if the room reverberation radii of the training room and the testing room are closer to each other. This is due to the fact that the extracted features, such as ILD, ITD, and statistical cues, at the same distances and azimuths in different rooms are more similar when the room reverberation radii are closer.

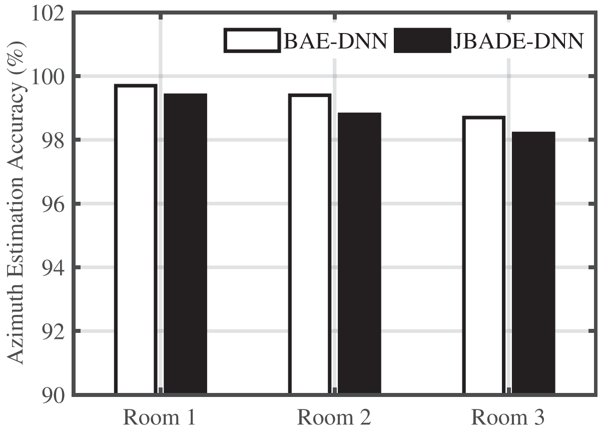

FIG. 9. The azimuth estimation accuracy of the JBADE-DNN method and the BAE-DNN method in real environments.

B. Experimental results in real environments

1. Comparison in real room environments

In this part, we conduct an experiment to further evaluate the JBADE-DNN method in real room environments. The details of the training datasets and the testing datasets are described in Sec. III D. Both the training signals and the testing signals are spilt into 1-s duration segments with 50% overlap to extract the azimuth and distance cues. In each room, the training datasets contain 18 720 segments, and the testing datasets contain 4680 segments. Figure 9 presents the azimuth estimation accuracy of the JBADE-DNN method and that of the BAE-DNN method in Room 1, Room 2, and Room 3, respectively. The distance estimation accuracy of the JBADE-DNN method and that of the competing BDE methods in Room 1, Room 2, and Room 3 are shown in Fig. 10.

It can be seen in Figs. 9 and 10 that the azimuth estimation accuracy and the distance estimation accuracy in real rooms are very close to the results in Sec. IV A 1, which

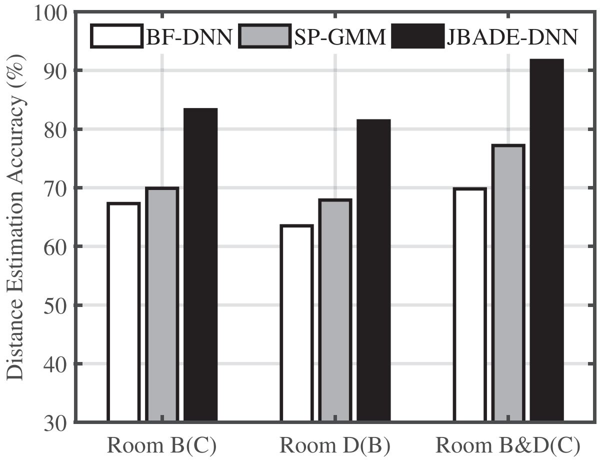

FIG. 8. The distance estimation accuracy in different unmatched room environments.

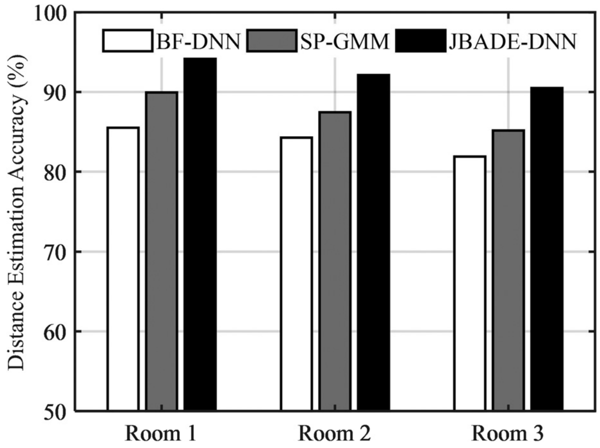

FIG. 10. The distance estimation accuracy of the JBADE-DNN method and the competing BDE methods in real environments.

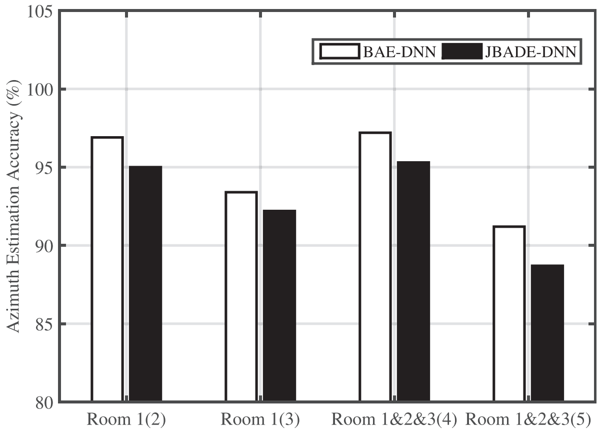

FIG. 11. The azimuth estimation accuracy in various unmatched real environments.

proves the validation of the proposed JBADE-DNN method. The azimuth and the distance estimation performance of all supervised BAE and BDE methods slightly decrease in the real environments because of several factors, such as the influence of the scattering field in real rooms, the measurement errors and so on. The effectiveness of the proposed method demonstrates that the distance estimation performance can be improved by joint estimation of the azimuth.

2. Comparison in unmatched real environments

Similar to Sec. IV A 3, this part conducts an experiment to evaluate the generalization ability of the JBADE-DNN in untrained real environments. The room characteristics and true sound source distances are shown in Table II, the true azimuths of the sound source are $\{-90^\circ, -75^\circ, \ldots, 90^\circ\}$ with the interval of 15°. The azimuth estimation accuracy and the distance estimation accuracy in various unmatched conditions are presented in the last two columns of Figs. 11 and 12. The combinations of different rooms for training and testing are shown in Figs. 11 and 12, where ‘Room 1(2)’ represents the training datasets that are collected in Room 1 and the testing datasets that are collected in Room 2, which is analogous to ‘Room 1(3)’, ‘Room 1&2&3(4)’ and ‘Room 1&2&3(5)’.

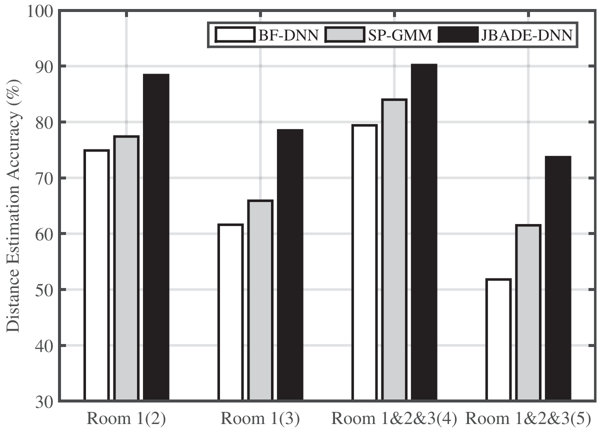

FIG. 12. The distance estimation accuracy in various untrained real environments.

As shown in Figs. 11 and 12, the proposed JBADE-DNN method has higher generalization ability in azimuth estimation than the distance estimation. The proposed JBADE-DNN method has slightly lower azimuth estimation performance than the BAE-DNN method, while it significantly improves the distance estimation performance in unmatched real environments. Moreover, both the azimuth estimation accuracy and the distance estimation accuracy of all supervised BAE and BDE methods are higher when the room reverberation radii of the training room and the testing room are closer.

V. CONCLUSIONS

This work proposes a JBADE-DNN that employs both binaural features and statistical properties as input features, and that learns the relationship between input features and their corresponding azimuth and distance of a point sound source through large training datasets by exploiting a multi-objective DNN framework. The proposed method can achieve higher generalization performance because of the benefits of the multitask learning. Furthermore, in the proposed method, both the azimuth and the distance cues are utilized in the learning stage of the error back-propagation in the multi-objective DNN framework, which can reduce the negative impact of the variation of azimuth on the distance estimation. Both simulated and real experimental results demonstrate that the azimuth estimation performance of the JBADE-DNN is very close to the BAE-DNN algorithm, while the distance estimation performance of the proposed JBADE-DNN is much better than the competing state-of-the-art supervised BDE methods, especially in untrained conditions.

ACKNOWLEDGMENTS

This work was supported by the National Science Fund of China under Grant Nos. 11974086 and 61571435.

[1] Code repository at *https://github.com/ahency/JBADE-DNN_SourceCode_Matlab_new*.

[2] The ‘BAE-DNN’ method is replicated with the open-source code of MATLAB toolbox for DNN-based speech separation at *https://github.com/Perception-and-Neurodynamics-Laboratory/Matlab-toolbox-for-DNN-based-speech-separation*.

[3] The ‘SP-GMM’ method is replicated with the open-source code of MATLAB toolbox for Supervised Binaural Mapping (SBM) at *https://team.inria.fr/perception/supervised-binaural-mapping/*.

[4] The ‘BF-DNN’ method is also replicated with the open-source code at *https://github.com/Perception-and-Neurodynamics-Laboratory/Matlab-toolbox-for-DNN-based-speech-separation*.

Allen, J. B., and Berkley, D. A. (**1979**). “Image method for efficiently simulating small-room acoustics,” J. Acoust. Soc. Am. **65**(4), 943–950.

Blauert, J. (**1997**). *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, Massachusetts).

Chen, P., Liu, H., Zhang, J., and Li, X. (**2017**). “Binaural sound localization based on reverberation weighting and generalized parametric mapping,” IEEE Trans. Audio Speech Lang. Process. **25**(8), 1618–1632.

Deleforge, A., Horaud, R., Schechner, Y. Y., and Girin, L. (**2015**). "Co-localization of audio sources in images ysing binaural features and locally linear regression," IEEE Trans. Audio Speech Lang. Process. **23**(4), 718–731.

Ding, J., Li, J., Zheng, C. S., and Li, X. D. (**2020**). "Wideband sparse Bayesian learning for off-grid binaural sound source localization," Signal Process. **166**(1), 107250.

Farmani, M., Pedersen, M. S., Tan, Z. H., and Jensen, J. (**2017**). "Informed sound source localization using relative transfer functions for hearing aid applications," IEEE Trans. Audio, Speech, Lang. Process. **25**(3), 611–623.

Gardner, B., and Martin, K. (**1994**). "HRTF measurements of a KEMAR dummy-head microphone," MIT Media Lab. Perceptual Computing-Technical Report 1–7.

Garofolo, J. S. (**1979**). DARPA TIMIT acoustic-phonetic speech database, National Institute of Standards and Technology (NIST), Gaithersburg, MD.

Georganti, E., May, T., van de Par, S., and Mourjopoulos, J. (**2013**). "Sound source distance estimation in rooms based on statistical properties of binaural signals," IEEE Trans. Audio Speech Lang. Process. **21**(8), 1727–1741.

Georganti, E., Mourjopoulos, J., and van de Par, S. (**2014**). "Room statistics and direct-to-reverberant ratio estimation from dual-channel signals," in *Proceedings of IEEE International Conference on Acoustics, Speech Signal Processing*, pp. 4713–4717.

Hofman, P. M., Riswick, J. G., and Opstal, A. (**1998**). "Relearning sound localization with new ears," Nau. Neurosci. **1**(5), 417–421.

Jeffress, L. A. (**1948**). "A place theory of sound localization," J. Comp. Physiol. Psychol. **41**(1), 35–39.

Jetzt, J. J. (**1979**). "Critical distance measurement of rooms from the sound energy spectral response," J. Acoust. Soc. Am. **65**(5), 1204–1211.

Kolarik, A. J., Cirstea, S., and Pardhan, S. (**2011**). "Perceiving auditory distance using level and direct-to-reverberant ratio cues," J. Acoust. Soc. Am. **130**, 2545.

Li, D., and Levinson, S. E. (**2003**). "A Bayes-rule based hierarchical system for binaural sound source localization," in *Proceedings of IEEE International Conference on Acoustics, Speech Signal Processing*, Vol. 5, pp. 521–524.

Liu, Y., Xie, B., Mai, H., and Chen, J. (**2015**). "Efficient algorithm and localization experiment on spherical microphone array recording and binaural rendering," J. Acoust. Soc. Am. **138**, 1833–1833.

Lu, Y. C., and Cooke, M. (**2010**). "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," IEEE Trans. Audio Speech Lang. Process. **18**(7), 1793–1805.

Ma, N., May, T., and Brown, G. J. (**2017**). "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," IEEE Trans. Audio Speech Lang. Process. **25**(12), 2444–2453.

Magassouba, A., Bertin, N., and Chaumette, F. (**2018**). "Exploiting the distance information of the interaural level difference for binaural robot motion control," IEEE Robot. Auto. Lett. **3**(3), 2048–2055.

May, T., Ma, N., and Brown, G. J. (**2015**). "Robust localization of multiple speakers exploiting head movements and multi-conditional training of binaural cues," in *Proceedings of IEEE International Conference on Acoustics, Speech Signal Processing*, pp. 2679–2683.

Shi, B., and Xie, B. (**2010**). "The cross-correlation of feeding signals and spatial impression in surround sound reproduction," Chinese J. Acoust. **29**(3), 308–320.

Tan, Z., Sarkar, A. K., and Dehak, N. (**2020**). "rVAD: An unsupervised segment-based robust voice activity detection method," Comput. Speech Lang. **59**, 1–21.

Tohyama, M. (**1995**). *The Nature and Technology of Acoustic Space* (Academic Press, New York).

Venkatesan, R., and Ganesh, A. B. (**2017**). "Full sound source localization of binaural signals," in *Proceedings of IEEE International Conference on Wireless Communications Signal Processing and Networking*, pp. 213–217.

Vesa, S. (**2009**). "Binaural sound source distance learning in rooms," IEEE Trans. Audio Speech Lang. Process. **17**(8), 1498–1507.

Wang, D., and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (IEEE Press, Wiley Interscience).

Yiwere, M., and Rhee, E. J. (**2017**). "Distance estimation and localization of sound sources in reverberant conditions using deep neural networks," Int. J. Appl. Eng. Res. **12**(22), 12384–12389.

Yu, G., Wu, Y., and Xie, B. (**2017**). "Perceptual evaluation on the influence of individualized near-field head-related transfer functions on auditory distance localization," J. Acoust. Soc. Am. **141**, 3536–3536.

Yu, G., Xie, B., and Rao, D. (**2012**). "Near-field head-related transfer functions of an artificial head and its characteristics," Acta Acust. **37**(4), 378–385 (in Chinese).

Zahorik, P. (**2002**). "Assessing auditory distance perception using virtual acoustics," J. Acoust. Soc. Am. **111**, 1832–1846.

Zhong, X., Sun, L., and Yost, W. (**2016**). "Active binaural localization of multiple sound sources," Robot. Auton. Syst. **85**, 83–92.

Zhong, X., and Yost, W. A. (**2015**). "Multiple sound source localization when sounds are stationary or rotating," J. Acoust. Soc. Am. **137**(4), 2228–2228.