IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 22, 2019, accepted March 11, 2019, date of publication March 18, 2019, date of current version April 10, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2905617

Multitask Learning of Time-Frequency CNN for Sound Source Localization

CHENG PANG[1], HONG LIU[1], (Member, IEEE), AND XIAOFEI LI[2]
[1]Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Beijing 100871, China
[2]INRIA Grenoble Rhône-Alpes, 38330 Montbonnot Saint-Martin, France

Corresponding author: Hong Liu (hongliu@pku.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61673030 and Grant U1613209, and in part by the Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality under Grant ZDSYS201703031405467.

ABSTRACT Sound source localization (SSL) is an important technique for many audio processing systems, such as speech enhancement/recognition and human–robot interaction. Although many methods have been proposed for SSL, it still remains a challenging task to achieve accurate localization under adverse acoustic scenarios. In this paper, a novel binaural SSL method based on time–frequency convolutional neural network (TF-CNN) with multitask learning is proposed to simultaneously localize azimuth and elevation under unknown acoustic conditions. First, the interaural phase difference and interaural level difference are extracted from the received binaural signals, which are taken as the input of the proposed SSL neural network. Then, an SSL neural network is designed to map the interaural cues to sound direction, which consists of TF-CNN module and multitask neural network. The TF-CNN module learns and combines the time–frequency information of extracted interaural cues to generate the shared feature for multitask SSL. With the shared feature, a multitask neural network is designed to simultaneously estimate azimuth and elevation through multitask learning, which generates the posterior probability for candidate directions. Finally, the candidate direction with the highest probability is taken as the final direction estimation. The experiments based on public head-related transfer function (HRTF) database demonstrate that the proposed method achieves preferable localization performance compared with other popular methods.

INDEX TERMS Sound source localization, time-frequency, convolutional neural network, multitask learning.

I. INTRODUCTION

Sound source localization (SSL) is a key component of computational auditory scene analysis, which can be applied to many audio applications, such as hearing-aids, teleconferencing, human-robot interaction, etc [1]–[4]. In the last few decades, various approaches have been proposed for SSL, and they can achieve favorable performance under certain specific acoustic conditions [5]–[7]. Despite decades of research, the task of robustly localizing sound sources in adverse acoustic scenarios still remains a challenging problem for machines.

A great amount of sound localization models have been proposed for SSL under different acoustic conditions. The representative techniques are time difference of arrival (TDOA) via generalized cross correlation (GCC) [8], and high-resolution spectral or beamforming method based on multiple signal classification (MUSIC) [9], steered response power (SRP) [10]. Most of these methods are based on microphone arrays, their performance depends on the array configuration and generally increases with the number of microphones [5]. Unlike microphone array-based approaches, the performance of the human auditory system is very robust against noise and reverberation for SSL through exploring the acoustic signals arriving at both ears. Motivated by the robust sound localization performance of human auditory, the localization based on binaural signals (termed as binaural SSL) has been widely researched in recent years, which has been a prevalent branch of SSL in computational auditory scene analysis (CASA) [6].

In binaural SSL, two primary physical cues are widely used [7], which include interaural time (or phase) difference (ITD and IPD, respectively) and interaural level difference (ILD). The two cues are caused by the sound propagation delay between the two ears and the head shadowing effect. The SSL is achieved according to the mapping

The associate editor coordinating the review of this manuscript and approving it for publication was Jiansong Liu.

relationship between binaural cues and sound direction. The azimuth, elevation and distance of sound source relative to binaural microphones, are used to describe its position in three-dimensional space. In order to mimic the SSL of human auditory, gammatone filter is introduced to process the received binaural signals into a set of narrow-band signals [11]. One classical method to estimate ITD is to search the maximum in the GCC function, nevertheless it is susceptible to reverberation and noise for the assumption of the ideal single-path sound propagation. Different weighting functions are proposed to enhance the estimation of GCC, such as phase transform (PHAT) [8], smoothed coherence transform (SCOT) [12], etc. ILD is obtained through calculating the logarithmic energy ratio between binaural signals, which is proved to be available for SSL alone [13]. A comprehensive review of binaural cues for SSL was shown in [14].

Different from ideal anechoic rooms, realistic indoor environment is generally acoustically disturbed, where the extracted ITD and ILD becomes distorted in general. A combined evaluation of binaural cues has been applied for anechoic SSL [15], where a joint feature space consisting of ITDs and ILDs was constructed based on time-frequency binary mask and trained to localize sound under noisy conditions. A parametric model was proposed to achieve a robust SSL under noisy conditions through combining the estimation of ITD and ILD over frequencies [16]. For human audition, ITD is more robust at low frequencies (lower than 1.5 kHz), whereas ILD is more reliable at high frequencies [17], so the ITD and ILD can operate in complementary ranges of frequencies for SSL. Motivated by this theory, a Bayes-rule based localization framework was proposed to hierarchically combine ITD and ILD for the noisy SSL [18]. In [19], the probability density functions of interaural cues were measured by histograms to perform SSL in nonstationary noise conditions. Based on the interdependency of ITD and ILD, a new binaural feature space was designed for SSL [20]. For reverberation, the multi-path reflections disturb the extraction of interaural cues, which was analyzed in [21]. Many methods have also been proposed to achieve the robust extraction of interaural cues, such as cepstral prefiltering [22], interaural coherence [23], direct-path dominance test [24].

In order to efficiently combine interaural cues for SSL, different localization models have been proposed to achieve robust SSL under different acoustic conditions [25]. A biologically inspired binaural SSL method was proposed through extracting interaural cues from cochleagrams generated by a cochlear model [26]. Model-based methods were proposed to robustly localize sound under noisy and reverberant conditions [27], [28]. Probabilistic model based on normal distribution was proposed to estimate sound direction according to the activity maps of interaural cues [26]. Gaussian mixture model (GMM) was applied to model the binaural feature space for each gammatone subband [20], [29]. The learning-based method with artificial neural network [30] was also introduced to SSL by training the interaural cues in each candidate direction under the acoustic conditions with different signal-to-noise ratios (SNRs) and reverberation times. Although these methods can achieve favorable SSL performance in noisy and reverberant environments, they should be trained for different SNRs and reverberation times, which makes them sensitive to the changes of the room configuration or acoustic conditions used in training.

Recently, along with the development of neural network, machine learning approaches with different types of neural networks have been developed for SSL, such as deep neural network (DNN), convolutional neural network (CNN) [31]. A multilayer perceptron neural network was firstly introduced to model the GCC coefficients weighted by phase transform [32]. GCC features were also input into a probabilistic neural network for robust SSL [33]. A probabilistic neural network was proposed to model interaural cues [34], MUSIC eigenvectors [35] for SSL. CNN was adopted for SSL with short time Fourier transform (STFT) phase as localization feature [36]. In [37], CNN was also applied to a beamformer to improve localization accuracy. Although these neural networks can achieve favorable performance under some acoustic conditions, most of them aim at estimating azimuth in the median plane.

As the sound position information is jointly described by azimuth and elevation, the estimation of the exact sound direction, including elevation, is an essential prerequisite for many other acoustic techniques, such as speech enhancement [38], speech separation [39], etc. For binaural SSL, elevation estimation is essential for many applications. For example, in human-robot interaction, a robot is usually required to localize not only the horizontal direction but also the vertical direction of speakers, since most of commercial robots do not have a similar height with human speakers. However, elevation localization has been rarely considered, because the traditional interaural differences are insufficient for localizing elevation, due to the *"cones of confusion"* exhibiting similar interaural cues [40]. In the past methods, some additional cues are proposed for elevation estimation, such as spectral cues [41], head-related transfer function (HRTF) [42], interaural matching filter [4], etc. As spectral cues are difficult to extract, energies coming from cochlear filter-banks are exploited as well [11]. In recent methods, DNN is applied to estimate azimuth and elevation [43], [44]. In [43], separate neural networks are trained for azimuth and elevation. The localization of azimuth and elevation in [44] is achieved with first-order Ambisonic (FOA) signals obtained by a spherical array. Motivated by these methods, simultaneously localizing azimuth and elevation with binaural signals is investigated in this paper.

In this work, a novel binaural SSL method based on time-frequency convolutional neural network (TF-CNN) with multitask learning, is proposed to simultaneously estimate azimuth and elevation under different acoustic conditions. IPD and ILD are extracted from the received binaural signals, then each or both of them are taken as the input of the proposed SSL neural network. TF-CNN is designed to robustly model and combine the interaural cues, which aims

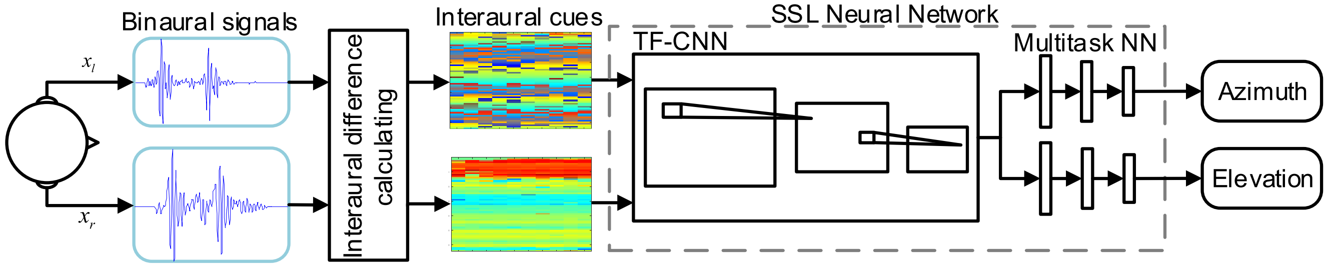


FIGURE 1. Flowchart of the proposed SSL system. Time-frequency interaural cues, i.e. IPD and ILD, are extracted as localization cues. SSL neural network consists of TF-CNN and multitask neural network. Multitask learning is introduced to learn and estimate azimuth and elevation simultaneously.

to learn the time-frequency information of interaural cues, and to generate the shared feature for both azimuth and elevation estimations. With the shared feature, a multitask neural network is adopted to simultaneously estimate azimuth and elevation through multitask learning, which produces the posterior probability for each azimuth and elevation candidates. Finally, the candidate direction with the highest probability is taken as the final direction estimation. Experiments based on the CIPIC HRTF database [45] demonstrate that the proposed method achieves preferable localization performance compared with other popular methods.

The rest of this paper is organized as follows: Section II illustrates the framework of the proposed localization method. In Section III, binaural signal model and interaural cue extraction are illustrated. TF-CNN is explained in Section IV. Section V illustrates the multitask learning strategy for sound source localization. Section VI shows experimental results and analyses, conclusions are drawn in Section VII.

II. PROPOSED LOCALIZATION SYSTEM

In binaural audition, two typical interaural cues (i.e. ITD/IPD and ILD) are commonly used for SSL based on the microphone-array geometry or clustering algorithms. Since original interaural cues are sensitive to strong noise conditions, most of traditional binaural SSL methods focused on extracting robust interaural cues, and how to effectively combine them for SSL. Besides, most of these methods estimated only the azimuth angle, few of them estimated the azimuth and elevation jointly. In this work, we target the problem of localizing azimuth and elevation of a single-sound source using binaural microphones, under various noise conditions. Multitask learning of CNN is introduced to model the original interaural cues for robust joint estimation of azimuth and elevation.

In this paper, SSL is taken as a classification problem by using a SSL neural network consisting of TF-CNN and multitask neural network to model the relationship between time-frequency interaural cues and sound directions (namely azimuth and elevation). The flowchart of the proposed SSL system is shown in Fig. 1, which includes three main components:

- Time-Frequency Feature Extraction: The phase and magnitude of binaural signals are computed by applying STFT to binaural signals. At each time-frequency bin, IPD and ILD are obtained by calculating the difference of phase and magnitude between binaural signals. The extracted IPD and ILD are separately stacked over multiple frames and all frequencies into larger IPD and ILD matrices with a fixed size, which are taken as the input of SSL neural network.
- TF-CNN: TF-CNN is designed to transfer and combine the extracted TF interaural cues over time and frequency domains. TF-CNN learns the time-frequency information by doing the 2D convolutional operation on the input interaural feature, which generates the discriminative shared feature for later multitask SSL.
- Multitask Neural Network: A multitask neural network is designed to simultaneously estimate azimuth and elevation with the shared feature. Multitask learning is introduced to train the proposed SSL neural network through combining the losses of azimuth and elevation estimations. After training, the SSL neural network model is used to estimate the posterior probability for each candidate direction with the input of frame-stacked IPD/ILD features. The candidate direction with maximum posterior probability is taken as the final direction estimate.

Overall, these three components lead to an effective and robust SSL system for both azimuth and elevation localization.

III. BINAURAL MODEL AND CUE EXTRACTION

A. BINAURAL SIGNAL MODEL

The azimuth and elevation of sound source are respectively denoted as θ and φ, which follows the definition of sound direction in the CIPIC HRTF database [45]. The CIPIC HRTF database collected by the U. C. Davis CIPIC Interface Laboratory is used to simulate the binaural signals in this work, which contains HRTFs for 45 different subjects including 27 males, 16 females, and KEMAR with large and small pinnae. For each subject, its HRTFs are measured at source-to-sensors distance of 1 m for 25 different azimuths and 50 different elevations. In the CIPIC HRTF database, the range of azimuth is $[-80^\circ, -65^\circ, -55^\circ, -45^\circ{:}5^\circ{:}45^\circ, 55^\circ, 65^\circ, 80^\circ]$, and elevation ranges from -45° to $+230.625^\circ$ in steps of 5.625°. Note that the angles are defined in interaural-polar coordinates, where the back side of one subject are found at [90°, 230.625°] elevation.

In the binaural hearing scene, let $s(m)$ denote the sound source signal, the binaural signals received by the two "ears" under noisy condition can be modeled as

$$y_i(m) = s(m) \star h_i(m) + v_i(m), \quad i = l, r, \tag{1}$$

where i represents the microphone index, l and r denote the left-ear and right-ear channels, m is the time index, $h_i(m)$ denotes the impulse response from sound source to ears, namely *head-related impulse responses* (HRIRs), $\star$ denotes the time-domain convolution operation, $v_i(m)$ denotes the additional noise. Here, $v_i(m)$ is assumed as a temporally uncorrelated, zero-mean, stationary Gaussian random process. The HRIR $h_i(m)$ involves the effect of dummy head and ears, which varies with sound direction, namely θ and φ.

By applying STFT, (1) is transformed to the time-frequency (TF) domain, which can be formulated as

$$Y_i(\kappa, \omega) = S(\kappa, \omega)H_i(\omega) + V_i(\kappa, \omega), \tag{2}$$

where Y_i, S and V_i are the STFT coefficients of their corresponding time-domain forms, κ is the time frame index, ω denotes the frequency bin index, H_i is the frequency-domain representation of HRIR, namely HRTF.

B. INTERAURAL CUE EXTRACTION

In this part, IPD and ILD are extracted as localization cues, which are taken as the input of SSL neural network. With the TF binaural signals in (2), IPD can be extracted as

$$\phi(\kappa, \omega) = \angle \frac{Y_r(\kappa, \omega)}{Y_l(\kappa, \omega)}, \tag{3}$$

where $\phi(\kappa, \omega)$ denotes the IPD at κ-*th* audio frame and ω-*th* frequency bin. ILD can be calculated as

$$\lambda(\kappa, \omega) = 20\log_{10} \frac{|Y_r(\kappa, \omega)|}{|Y_l(\kappa, \omega)|}, \tag{4}$$

where $\lambda(\kappa, \omega)$ denotes the ILD at κ-*th* audio frame and ω-*th* frequency bin.

The IPD distribution and ILD distribution as a function of azimuth and elevation are shown in Fig. 2. In Fig. 2 (a) and (b), it is obvious that IPD and ILD are changing with the variation of azimuth. When azimuth = 0°, IPD and ILD should theoretically be zero, while there are some small fluctuations along frequency for them. This phenomenon may be mainly caused by that the HRTF used for calculating the IPD and ILD distributions is from subject #21 in the CIPIC HRTF database, which is measured from a real person in a realistic environment. In Fig. 2 (c) and (d), the IPD and ILD distributions at azimuth = 40° are presented, which also have some small fluctuations. It can be observed that the IPD distribution is similar for different elevations, while ILD is more sensitive to elevation variation.

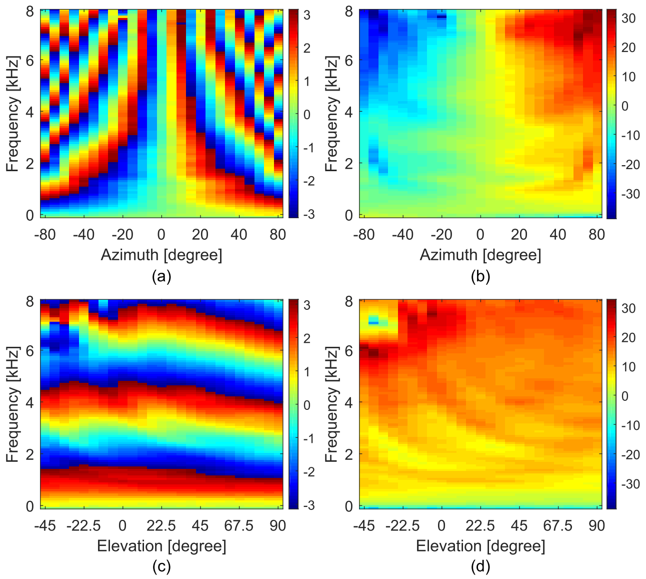


FIGURE 2. IPD distribution (a) and ILD distribution (b) versus azimuth where elevation is 0°, IPD distribution (c) and ILD distribution (d) versus elevation where azimuth is 40°, which are calculated based on the HRTFs of subject #21 in the CIPIC HRTF database.

IV. TIME-FREQUENCY CNN

In this section, the form of input feature for the SSL neural network is firstly explained. Then, the architecture of the time-frequency CNN module is illustrated to generate the shared feature for both azimuth and elevation estimation.

A. INPUT FEATURE

With the extracted time-frequency IPD and ILD, they are stacked over constant K time frames and all of the frequency bins. Without loss of generality, let $\kappa \in 1, 2, \cdots, K$ denote the constant K frames, and $\omega \in 1, 2, \cdots, F$ all the frequency bins. So the input IPD feature can be formulated as

$$\boldsymbol{\phi} = \begin{bmatrix} \phi(1,1) & \phi(1,2) & \cdots & \phi(1,F) \\ \phi(2,1) & \phi(2,2) & \cdots & \phi(2,F) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(K,1) & \phi(K,2) & \cdots & \phi(K,F) \end{bmatrix}$$

In the same way, the input ILD feature is formulated as

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda(1,1) & \lambda(1,2) & \cdots & \lambda(1,F) \\ \lambda(2,1) & \lambda(2,2) & \cdots & \lambda(2,F) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda(K,1) & \lambda(K,2) & \cdots & \lambda(K,F) \end{bmatrix}.$$

According to the definition of IPD matrix $\boldsymbol{\phi}$ and ILD matrix $\boldsymbol{\lambda}$, the two interaural features involve the time and frequency information with the size of $K \times F$. In this work, interaural features are extracted from the binaural signals with sampling rate of 16 kHz using STFT with a hamming window. The window length is 40 ms (640 samples) with a hop length of 20 ms. The interaural features are extracted for the frequency bins from 0 to 320, which represent the

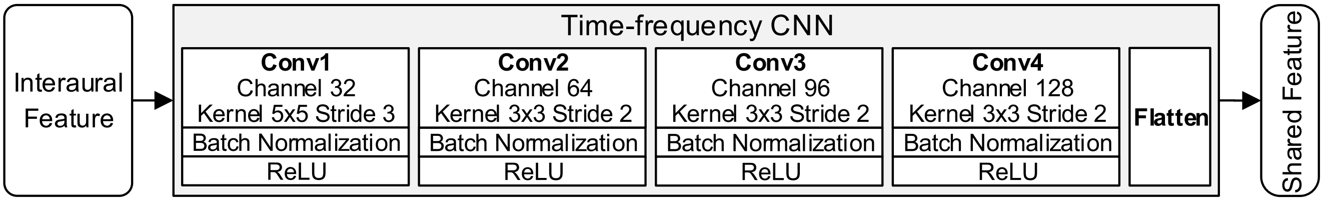


FIGURE 3. Architecture of time-frequency CNN module includes four convolutional layers, four batch normalization layers followed by ReLU activation, and one flatten layer, which generates the shared feature for azimuth and elevation localization.

frequencies from 0 to 8 kHz. Here, the time-domain input size is 200 ms corresponding to 10 frames. In this work, the IPD matrix $\boldsymbol{\phi}$, ILD matrix $\boldsymbol{\lambda}$ and their concatenation along time (labeled as IPD+ILD) are taken as the input feature, respectively. The size of the input interaural feature $K \times F$ is 10 and 321 for IPD matrix or ILD matrix, 20 and 321 for IPD+ILD. Let $\boldsymbol{A}$ denote the input of the SSL neural network, $\boldsymbol{A}$ can be $\boldsymbol{\phi}$, $\boldsymbol{\lambda}$ and $[\boldsymbol{\phi}, \boldsymbol{\lambda}]$.

B. TF-CNN MODULE

With the interaural features constructed in Section IV-A as input, a 2D CNN is proposed to model the time-frequency information of the input feature, which is called time-frequency CNN (TF-CNN). The architecture of TF-CNN is shown in Fig. 3, which includes four convolution layers with different numbers of filters (namely 'Channel' in Fig. 3), four batch normalization (BN) layers and one flatten layer. Rectified Linear Unit (ReLU) activation [46] is used after each batch normalization layer. The kernel size of the convolution layer is presented as $R \times S$, where R and S represent the dimensions of time and frequency, respectively.

As shown in Fig. 3, the interaural feature is first put into a 2D convolution layer with squared kernel size of 5×5 and stride of 3. Then, a batch normalization [47] layer is used to improve the stability of the SSL network. After the batch normalization operation, a 2D convolution layer with squared kernel size of 3×3 and stride of 2, and a batch normalization layer are used to weight the input interaural features. Next, the same 2D convolution and batch normalization operations are repeated twice with different numbers of filters. ReLU activation is utilized after each batch normalization layer. Finally, a flatten layer is used to flatten the output of the previous layer to a feature vector. The feature vector output from the TF-CNN module is taken as the shared feature for the following multitask SSL (will be presented in Section V).Let $Conv_j$ denote the *j-th* convolution block in the TF-CNN module, namely $Conv_j(\cdot) = \text{ReLU}(\text{BN}(\text{Conv}j(\cdot)))$, where Conv$j(\cdot)$ denotes the convolution layer shown in Fig. 3, $j = 1, 2, 3, 4$. The shared feature generated by the TF-CNN module can be mathematically presented as

$$\boldsymbol{F}_s = \text{Flatten}(Conv_4(Conv_3(Conv_2(Conv_1(\boldsymbol{A}))))). \quad (5)$$

In the 2D convolution layers, small-size local filters are applied to learn the correlation of interaural features across neighboring time frames and frequency bins. The squared 2D local filters in TF-CNN are able to potentially improve the robustness of SSL system. Regarding the number of convolution layers, we have done various pilot experiments with multitask learning for both azimuth and elevation estimation. The number of convolution layers is set to 3, 4 and 5, respectively. The training (and also prediction) cost in terms of computational and memory resources is linearly proportional to the number of layers. The localization accuracy (averaged over azimuth and elevation) of the 4-layer network exceeds the one of the 3-layer network by about 4%. The accuracy of the 5-layer network exceeds the one of the 4-layer network less than 1%. Thence, we use 4 convolution layers in this work.

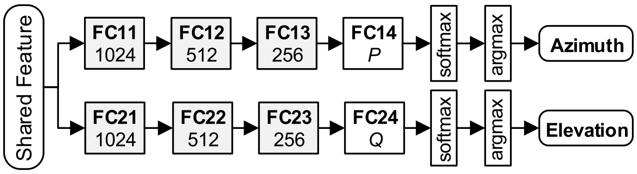


FIGURE 4. Architecture of multitask neural network for azimuth and elevation estimation.

V. MULTITASK SOUND SOURCE LOCALIZATION

A. MULTITASK NEURAL NETWORK FOR SSL

Multitask learning is able to train a 'universal' model for several different but related tasks using a shared representation, which has been applied to many speech/audio processing systems, such as ASR [48], speech enhancement [49]. In multitask learning, internal representations learned for one task can be helpful for the other related tasks, and vice versa. In this work, the multitask for SSL includes azimuth localization and elevation localization.

The architecture of the multitask learning network is shown in Fig. 4, which includes two branches for azimuth and elevation estimation, respectively. In the upper branch of Fig. 4, four fully connected (FC) layers, i.e. [FC11, FC12, FC13] with ReLU activation and FC14, are used to combine the local structures in the shared feature learned by TF-CNN for azimuth estimation, whose output size is the number of azimuth candidates. Since the number of azimuth candidates is same with elevation in this work, the same neural network architecture is designed for elevation. As shown in the lower part of Fig. 4, four FC layers, i.e. [FC21, FC22, FC23] with ReLU activation and FC24, are utilized to combine the learned local structures in the same shared feature for elevation estimation, whose output size is the number of elevation candidates. The outputs of FC14 and FC24 can be respectively formulated as

$$\boldsymbol{F}_a = \text{FC14}(\text{FC13}(\text{FC12}(\text{FC11}(\boldsymbol{F}_s)))), \tag{6}$$
$$\boldsymbol{F}_e = \text{FC24}(\text{FC23}(\text{FC22}(\text{FC21}(\boldsymbol{F}_s)))). \tag{7}$$

Then, $\boldsymbol{F}_a$ and $\boldsymbol{F}_e$ are passed through two softmax layers, respectively. The outputs of the two softmax layers are taken as the posterior probabilities for the candidates of azimuth and elevation, respectively, which can be obtained as

$$\mathcal{P}(\theta_p|\boldsymbol{A}, \boldsymbol{W}_a) = \frac{e^{[\boldsymbol{F}_a]_p}}{\sum_{n=1}^{P} e^{[\boldsymbol{F}_a]_n}}, p = 1, 2, \cdots, P, \tag{8}$$
$$\mathcal{P}(\varphi_q|\boldsymbol{A}, \boldsymbol{W}_e) = \frac{e^{[\boldsymbol{F}_e]_p}}{\sum_{n=1}^{Q} e^{[\boldsymbol{F}_e]_n}}, q = 1, 2, \cdots, Q, \tag{9}$$

where $[\cdot]_n$, $[\cdot]_p$ and $[\cdot]_q$ separately represent the *n-th*, *p-th* and *q-th* element in the vector, $\boldsymbol{W}_a$ and $\boldsymbol{W}_e$ denote the the learnable weight matrix of the SSL neural network for azimuth and elevation estimation, θ_p and φ_q denote the *p-th* candidate azimuth and *q-th* candidate elevation, P and Q represent the number of azimuth and elevation candidates, respectively. The final azimuth/elevation estimation is achieved by finding the candidate direction with the maximum posterior probability, which can be achieved by

$$\hat{\theta} = \underset{p}{\text{argmax}}\ \mathcal{P}(\theta_p|\boldsymbol{A}, \boldsymbol{W}_a), \tag{10}$$
$$\hat{\varphi} = \underset{q}{\text{argmax}}\ \mathcal{P}(\varphi_q|\boldsymbol{A}, \boldsymbol{W}_e), \tag{11}$$

where $\hat{\theta}$ and $\hat{\varphi}$ denote the estimated azimuth and elevation, respectively.

In the following experimental section, for comparison purpose, we also test the single-task training scheme, in which the training is applied to the single branch of the multitask network for individual aimuth/elevation estimation. The azimuth/elevation can be individually achieved by (10)/(11). In the context, the TF-CNN with multitask neural network is called multitask TF-CNN, and the TF-CNN with single branch of multitask neural network is called single-task TF-CNN.

B. LOSS FUNCTION

For the single task of azimuth (or elevation) estimation, the SSL neural network is trained by minimizing the cross-entropy loss between the predicted azimuth (or elevation) and the ground truth. For single-task TF-CNN, the cross-entropy function [50] used for training is formulated as

$$\mathcal{L}_a = -\sum_{p=1}^{P} \left[t_a \log\left(\mathcal{P}(\theta_p|\boldsymbol{A}, \boldsymbol{W}_a)\right)\right], \tag{12}$$
$$\mathcal{L}_e = -\sum_{q=1}^{Q} \left[t_e \log\left(\mathcal{P}(\varphi_q|\boldsymbol{A}, \boldsymbol{W}_e)\right)\right], \tag{13}$$

where $\mathcal{L}_a$ and $\mathcal{L}_e$ denote the azimuth and elevation estimation loss, t_a and t_e denote the ground-truth azimuth and elevation labels, respectively.

Sound direction can be represented by azimuth and elevation, which can be simultaneously estimated by using multitask learning [51], [52]. In order to achieve the multitask learning, the losses of azimuth and elevation estimations are jointly minimized. For multitask TF-CNN, the parameters of the whole neural network for SSL, denoted as Θ, are randomly initialized between -1 and 1, which are trained by optimizing the following combined loss function using back-propagation:

$$\min_{\Theta} \left(\alpha \mathcal{L}_a + (1-\alpha)\mathcal{L}_e\right), \tag{14}$$

where α is the mixing weight with the value ranging from 0 to 1. In the same way, the single-task TF-CNN for azimuth (or elevation) estimation is trained with the corresponding single loss function $\mathcal{L}_a$ (or $\mathcal{L}_e$).

VI. EXPERIMENTS AND ANALYSES

A. EXPERIMENTAL SETUP

1) THE DATASET

To evaluate the effectiveness of the proposed method, the subject #21 (i.e., Kemar head) in the CIPIC HRTF database [45] is used to simulate the SSL environment in the following experiments. The SSL in the front area is considered in this work, the range of azimuth localization is from $-80°$ to $80°$, and elevation localization from $-45°$ to $+90°$. The number of candidate azimuth and elevation for localization is 25 and 25, respectively, which leads to a total of $25 \times 25 = 625$ directions.

Audio signals from the TIMIT dataset [53] are taken as the sound source signals, which are convolved with the HRIRs to generate the binaural signals. Four types of spatially uncorrelated noises (white Gaussian noise, speech babble noise, pink noise and f16 noise) from Noisex92 database [54] are used as interference signals to generate the noisy binaural signals, which are directly added to binaural signals with signal-to-noise ratios (SNRs) ranging from -5 dB to 35 dB in steps of 5 dB. The sampling rate of binaural signals is 16 kHz. For each SNR, 15 sets of 625 binaural signals were generated as training data by randomly selecting 15 different speech signals from the train set from TIMIT dataset and convolving each of these 15 signals with each of the 625 BRIRs. Similarly, with the audio signals from the test set in TIMIT dataset, 5 sets and 10 sets are generated as validation and test data. In this way, the speakers and utterances used for training and test in our SSL system are different, namely our SSL experiments are speaker/content-independent.

In order to evaluate the robustness of the proposed localization system for different types of noises, cross-validation is conducted: i) the data of [0:10:30] dB are used for training and evaluation, and the data of [-5:10:35] dB for test; ii) the data from the three of four types of noises, i.e. Gaussian,

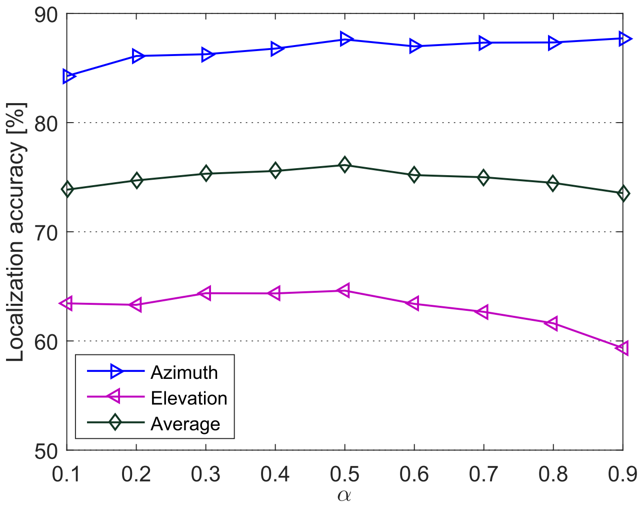


FIGURE 5. The localization accuracy of azimuth, elevation and the average of them for the multitask TF-CNN method with IPD+ILD input, as a function of α. The accuracy is averaged over four types of spatially uncorrelated noise and over all test SNRs, under the noise-unmatched condition.

babble, pink and f16, are used to train and validate the proposed SSL neural network, the data of remaining one is used to test the adaptability of the trained SSL neural network. In addition to the cross-validation experiments (under the noise-unmatched condition), we also test the noise-matched condition, namely the same type of noise is used for training and test. Note that the noise-matched condition also uses the different SNRs for training and test, as for the cross-validation experiments. Similar with the setting for the spatially uncorrelated noise, the spatially diffuse noise version of the above four types of noise are also tested. The diffuse noise is generated by convoluting different noise slices with the HRIRs of all 625 directions, and then summing up them.

2) EVALUATION SETTING

The proposed SSL neural network is trained with IPD, ILD and IPD+ILD under different noise conditions by single-task (azimuth/elevation) training and multi-task (azimuth and elevation) training. The localization performance is measured by localization accuracy with the 0° localization error, namely one source is said to be correctly localized only when the estimated direction index is identical to the ground truth index.

In this work, the training and evaluations for all networks are conducted by Tensorflow [55] using one NVIDIA GeForce Titan XP GPU. The batch size is set to 128 for all the experiments. The stochastic gradient descent [56] with a momentum of 0.9 is adopted for training the SSL neural network. The learning rate is set to 0.0001.

In order to determine the optimal mixing weight α, different values of α were evaluated with multitask TF-CNN based on IPD+ILD under different spatially uncorrelated noise conditions. The localization results are shown in Fig. 5. It can be seen that the best localization performance is achieved when $\alpha = 0.5$, which may be due to the same number of azimuth and elevation candidates, as well as the same neural network structure for azimuth and elevation estimation. Therefore, the mixing weight α is set to 0.5 in the following experiments.

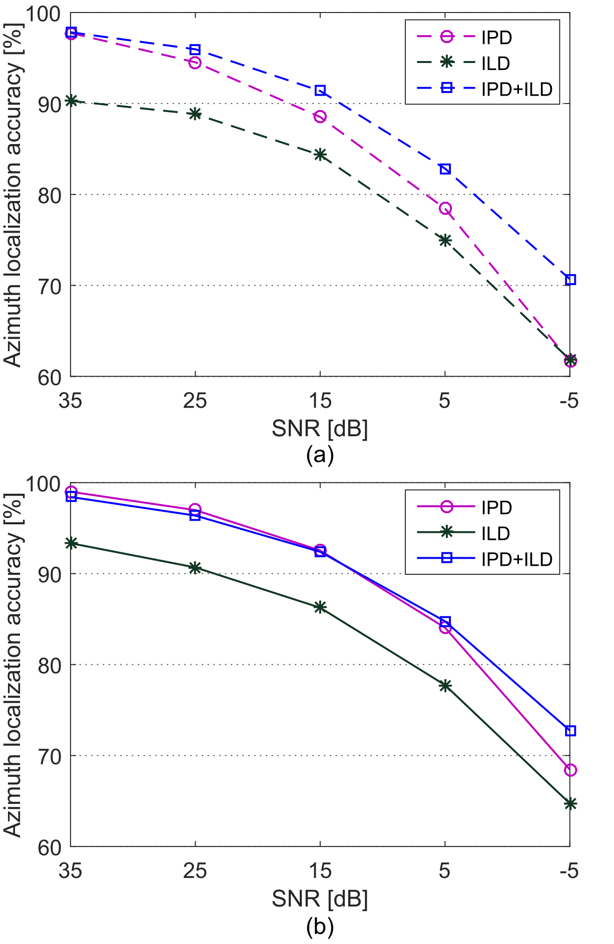


FIGURE 6. Average azimuth localization accuracy over four types of spatially uncorrelated noise, under the noise-unmatched condition. (a) single-task training (b) multitask training, with IPD, ILD and IPD+ILD, respectively.

B. AZIMUTH LOCALIZATION RESULTS

1) RESULTS FOR SINGLE-TASK AND MULTITASK TRAINING

The average azimuth localization accuracy over the four types of spatially uncorrelated noise with different SNRs is shown in Fig. 6, under the noise-unmatched condition. Fig. 6 (a) presents the results of single-task training with IPD, ILD and IPD+ILD as localization features. It can be seen that the average azimuth localization performance for each type of localization feature degrades with the decrease of SNR, since the strong noise seriously harms the extraction of interaural features. The IPD feature achieves better azimuth localization performance than ILD, which is due to that IPD is more discriminative than ILD for azimuth localization, as shown in Fig. 2. Through taking both IPD and ILD as the input feature, better azimuth localization accuracy is achieved compared with each of IPD or ILD, especially under strong noise conditions, which demonstrates that TF-CNN can effectively combine IPD and ILD

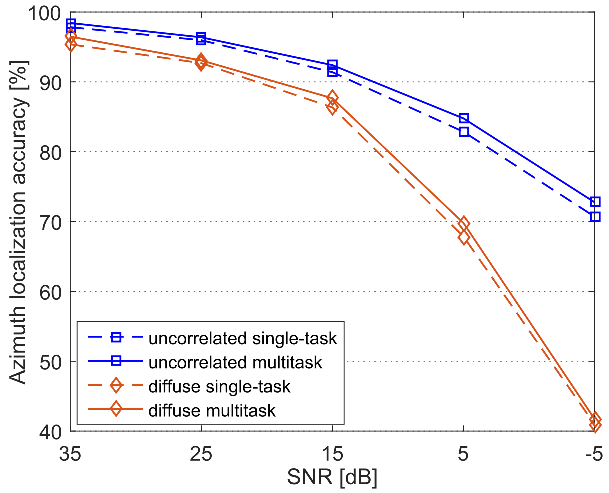


FIGURE 7. Average azimuth localization accuracy over four types of spatially uncorrelated noise and diffuse noise, for single-task and multitask TF-CNN based on IPD+ILD under the noise-unmatched condition.

for azimuth localization. Fig. 6 (b) shows the results of multitask training. It can be seen that the azimuth localization performance is obviously improved by multitask TF-CNN compared with the single-task TF-CNN. For example, the azimuth localization accuracy of single-task TF-CNN with IPD is 88.52% at SNR=15 dB, while that of multitask TF-CNN reaches 92.55%. IPD achieves similar performance with IPD+ILD through multitask TF-CNN when SNR≥5 dB. It can also be seen that IPD obtains slightly better azimuth localization performance than ILD when SNR>15 dB. This phenomenon may be caused by that, for multitask training with IPD+ILD, the trained network gives more emphasis to ILD (relative to single-task training), which will improve the performance for elevation estimation.

Fig. 7 shows the average azimuth localization accuracy over four types of spatially uncorrelated noise and diffuse noise with different SNRs, for single-task and multitask TF-CNN based on IPD+ILD under the noise-unmatched condition. It can be seen that, based on IPD+ILD, multitask TF-CNN achieves better performance than single-task TF-CNN under both spatially uncorrelated and diffuse noise conditions, which mainly owes to the multitask learning for joint azimuth and elevation estimation. The performance gap between multitask TF-CNN and single-task TF-CNN is small, which is attributed to the fusion estimation based on IPD and ILD.

2) RESULTS FOR EACH TYPE OF NOISE

The azimuth localization results of multitask TF-CNN for each type of noise with different SNRs are shown in Fig. 8. In detail, Fig. 8 (a) shows the results under spatially uncorrelated noise conditions. It can be seen that the azimuth localization accuracy of TF-CNN exceeds 90% for the four types of noise when SNR≥25 dB, while degrades with the decrease of SNR. The azimuth localization performance for the four types of noise are similar at SNR=35 dB. For other SNRs, the performance ranking for the four types of noise is: babble, f16, pink and Gaussian. This phenomenon may be caused by the wide-band property of white Gaussian noise, which makes the local filter in convolutional layers difficult to compensate the information in the high frequency region. Fig. 8 (b) shows the results under diffuse noise condition. Compared with the spatially uncorrelated noise case, the performance degrades more rapidly under the diffuse noise condition with the decrease of SNR. The reason is that the spatial correlation of diffuse noise reduces the accuracy of interaural feature extraction, especially for the IPD.

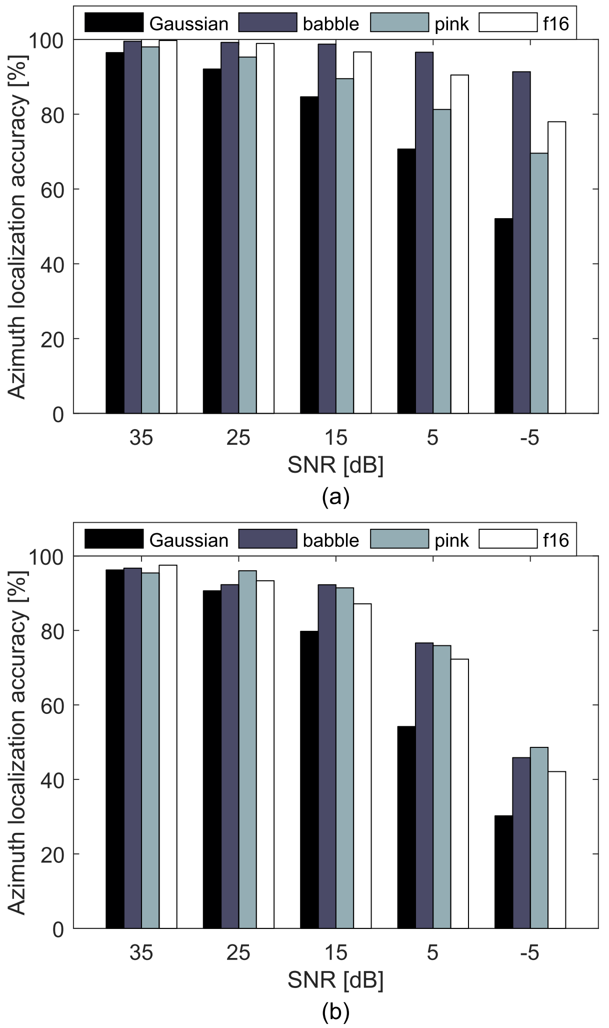


FIGURE 8. Azimuth localization accuracy of multitask TF-CNN based on IPD+ILD under (a) spatially uncorrelated and (b) diffuse noise conditions, and under the noise-unmatched condition.

3) NOISE-MATCHED/UNMATCHED RESULTS

To evaluate the generalization capability of the proposed localization system with respect to the noise type, we compare the localization results under noise-matched and noise-unmatched conditions, which are shown in Fig. 9. We remind that noise-matched refers to that training and test use the

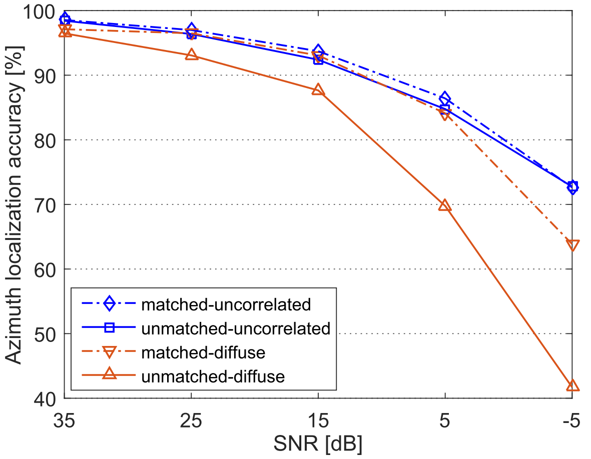


FIGURE 9. Azimuth localization accuracy (averaged over four types of noise) of multitask TF-CNN under noise-matched and noise-unmatched conditions, respectively.

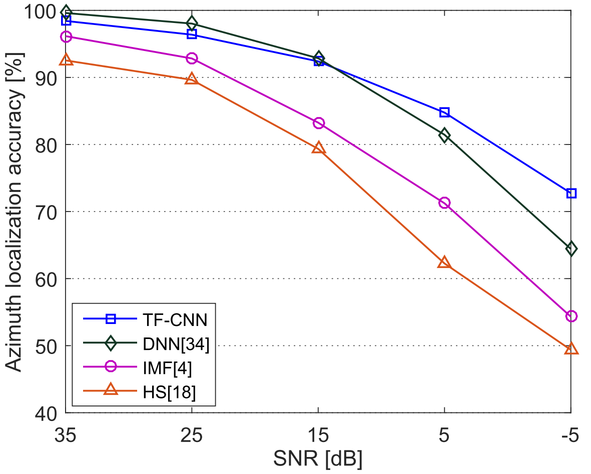


FIGURE 10. Comparing the azimuth estimation accuracy with three baseline methods under the spatially uncorrelated noise conditions. The proposed method, i.e. multitask TF-CNN, uses the IPD+PLD feature, and is tested under the noise-unmatched condition.

same type of noise, while noise-unmatched refers to that training and test use different types of noises, i.e. Gaussian, babble, pink and f16. For the spatially uncorrelated noise, the performance of the noise-matched and noise-unmatched cases are comparable, which demonstrates the good noise type generalization capability of the proposed system when the noise is spatially uncorrelated. However, for the diffuse noise, the noise-matched case noticeably outperforms the noise-unmatched case, which indicates the bad noise type generalization capability. To overcome this, the neural network training should use many different types of noises to cover the unseen test noise type as much as possible.

4) COMPARISON WITH BASELINE METHODS

To evaluate the effectiveness of the proposed method, three baseline methods are compared, i.e. DNN [34], Interaural Matching Filter (IMF) [4] and Hierarchical System (HS) [18].

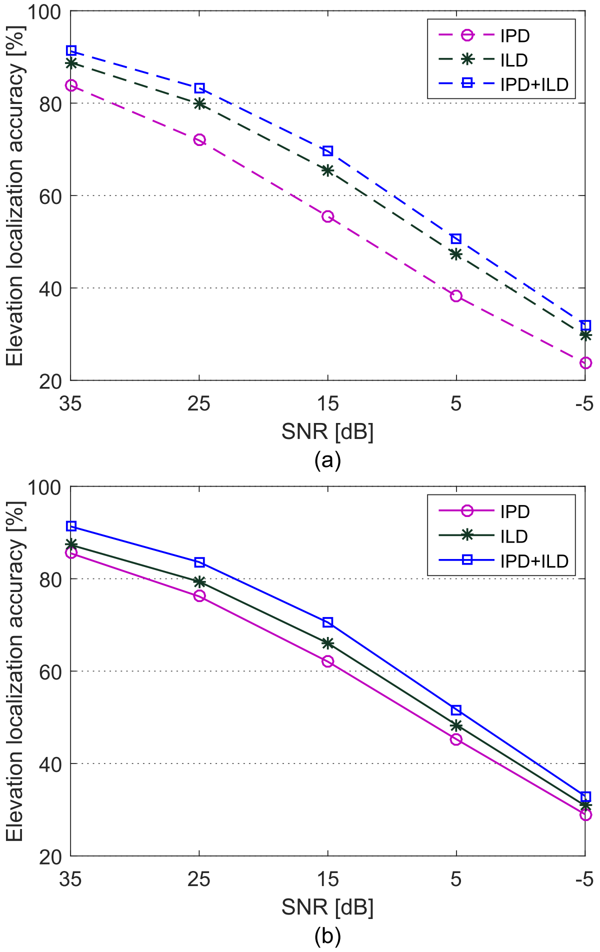


FIGURE 11. Average elevation localization accuracy over four types of spatially uncorrelated noise, under the noise-unmatched condition. (a) single-task training (b) multitask training, with IPD, ILD and IPD+ILD, respectively.

DNN uses the GCC-PHAT coefficients and ILD computed from the outputs of gammatone filters as localization cues. Note that, different from the proposed method that SSL is conducted for each 200 ms signal segment, DNN uses the whole signal sequence. IMF and HS are the hierarchical methods. The former uses ITD, ILD and IMF, and the latter uses ITD and ILD, as localization cues.

Fig. 10 presents the comparison of azimuth localization accuracy. The HS-based method gets the lowest azimuth estimation performance among the three methods. The IMF-based method obtains better localization performance than HS, since the extra localization feature over HS, i.e. IMF, is effective to discriminate directions. The proposed method prominently outperforms IMF for all the SNR conditions, although IMF uses ITD, ILD and IMF, the proposed method only uses IPD (equivalent to ITD) and ILD. The DNN-based method performs slightly better than the proposed method when SNR>15 dB, but noticeably worse than the proposed method under low SNR conditions, even if DNN-based method uses the whole signal sequence. Overall, the proposed method shows great performance superiority than other

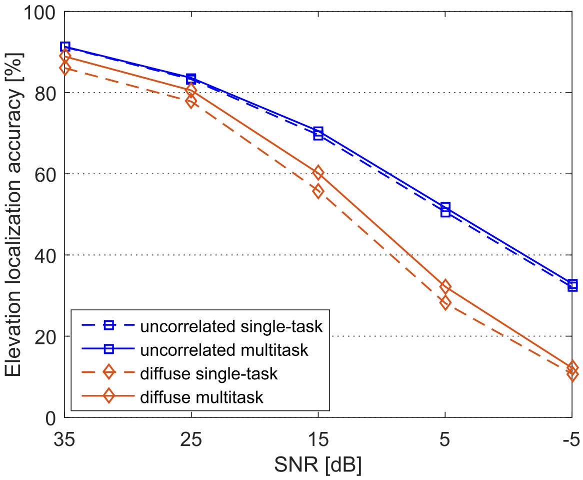


FIGURE 12. Average elevation localization accuracy over four types of spatially uncorrelated noise and diffuse noise, for single-task and multitask TF-CNN based on IPD+ILD under the noise-unmatched condition.

three methods, especially under strong noise conditions. This preferable azimuth localization performance of TF-CNN is attributed to two aspects: one is the local filters in CNN, which can compensate the information influenced by low SNR, since the weight sharing in CNN provides robustness to local distortions in the input; the other is multitask training and localization, which combines and benefits both azimuth and elevation localization.

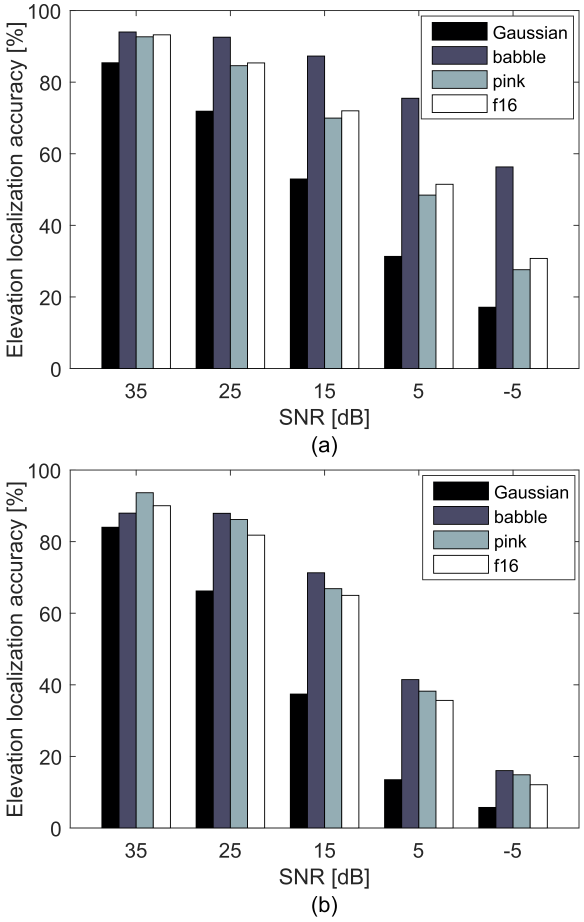
C. ELEVATION LOCALIZATION RESULTS

In this section, the multitask TF-CNN model used for azimuth localization is directly used to estimate elevation, the single-task TF-CNN model for elevation localization is trained with the corresponding single cross-entropy loss.

1) RESULTS FOR SINGLE-TASK AND MULTITASK TRAINING

The average elevation results for different SNRs are shown in Fig. 11. Fig. 11 (a) presents the results for single-task TF-CNN. Similar with the azimuth localization, the elevation localization accuracy of TF-CNN decreases with the increasing noise intensity. However, the TF-CNN with ILD as localization cue performs better elevation localization than IPD, which is due to that ILD is more discriminative than IPD for elevation localization. The single-task TF-CNN gets the best elevation localization performance by combining IPD and ILD as localization feature, which also demonstrates that TF-CNN can effectively combine IPD and ILD for elevation localization.

In addition to the single-task TF-CNN, multitask TF-CNN is also applied, whose results are shown in Fig. 11 (b). It can be observed that the elevation localization performance for different interaural cues is improved with the multitask TF-CNN. The elevation localization accuracy of the single-task TF-CNN with IPD is lower than 40% at SNR = 5 dB, while that of multitask TF-CNN reaches over 40%. This phenomenon verifies that the multitask learning can take the full advantage of interaural features for elevation localization. Fig. 12 shows the average elevation localization accuracy over four types of spatially uncorrelated noise and diffuse noise with different SNRs, for single-task and multitask TF-CNN based on IPD+ILD under the noise-unmatched condition. It can be seen that multitask TF-CNN outperforms single-task TF-CNN under both spatially uncorrelated and diffuse noise conditions, and this advantage is similar to the azimuth localization.



(a)

(b)

FIGURE 13. Elevation localization accuracy of multitask TF-CNN based on IPD+ILD under (a) spatially uncorrelated and (b) diffuse noise conditions, and under the noise-unmatched condition.

2) RESULTS FOR EACH TYPE OF NOISE

The elevation localization results for each type of noise with different SNRs are shown in Fig. 13. The results under the spatially uncorrelated noise condition are shown in Fig. 13 (a). It can be observed that the elevation localization accuracy reaches above 80% at SNR = 35 dB for white Gaussian noise, and it reaches above 90% for other three types of noise. Similar with the azimuth localization, the performance ranking for the four types of noise is: babble, f16, pink and Gaussian.

For the diffuse noise condition, the elevation localization performance for each type of noise with different SNRs are shown in Fig. 13 (b). It can be observed that, for white

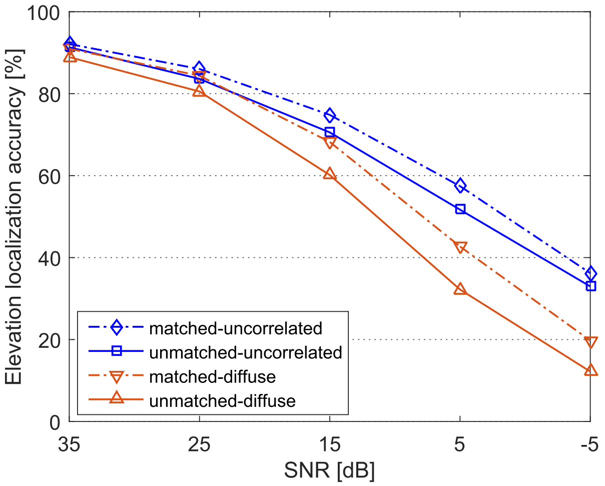


FIGURE 14. Elevation localization accuracy (averaged over four types of noise) of multitask TF-CNN under noise-matched and noise-unmatched conditions, respectively.

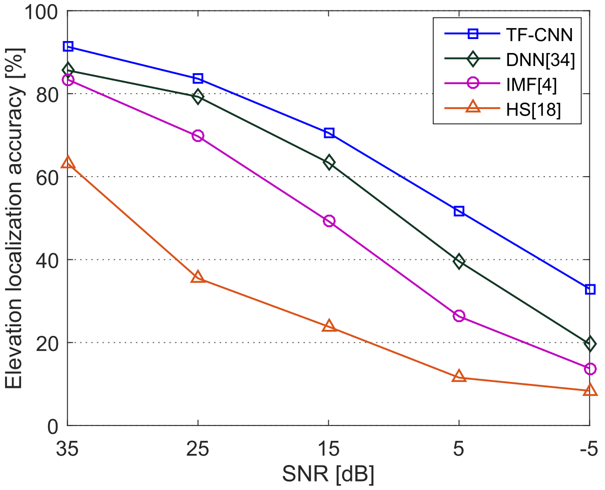


FIGURE 15. Comparing the elevation estimation accuracy with three baseline methods under the spatially uncorrelated noise conditions. The proposed method, i.e. multitask TF-CNN, uses the IPD+PLD feature, and is tested under the noise-unmatched condition.

Gaussian noise, the elevation localization performance also reaches above 80% at SNR = 35 dB, and degrades sharply with the decrease of SNR, which is similar with the uncorrelated noise case. TF-CNN achieves similar performance for each SNR for other three types of noise. Compared with the spatially uncorrelated noise case, the elevation localization performance has a larger degradation with the decrease of SNR under the diffuse noise condition, due to the spatial correlation of diffuse noise.

3) NOISE-MATCHED/UNMATCHED RESULTS

With the same noise-matched and noise-unmatched settings for azimuth localization, the average elevation localization results are shown in Fig. 14. It is not surprising that the noise-matched setting outperforms the noise-unmatched setting for all the conditions. However, under the diffuse noise condition, different from the azimuth localization results that the performance gap between the two settings is very large, the elevation performance gap is much smaller between the two settings. This means that, under the diffuse noise condition, the noise type generalization capability of the proposed SSL system is good for the elevation localization. Since ILD is the dominant cue for elevation, this phenomenon indicates that the ILD extraction does not largely rely on the noise type.

4) COMPARISON WITH BASELINE METHODS

The comparison of elevation localization performance between the baseline methods and the proposed method is shown in Fig. 15. Although DNN is only used for azimuth estimation in [34], we changed its training target for elevation estimation. The proposed method systematically outperforms the DNN-based method for elevation estimation. The main reason is the use of the multitask training in the proposed method. Besides, the DNN is possibly not very suitable for elevation estimation. It can be seen that the proposed method prominently outperforms the other three methods, and the superiority is similar to the azimuth localization results.

VII. CONCLUSION

In this paper, a novel binaural sound source localization method based on time-frequency convolutional neural network (TF-CNN) with multitask learning was proposed to simultaneously estimate azimuth and elevation. IPD and ILD extracted from the received binaural signals are taken as localization feature, which are combined for both azimuth and elevation localization. For single interaural feature, IPD is more effective for azimuth localization, and ILD is more effective for elevation localization. TF-CNN robustly modeled the noise-influenced interaural features by learning their time-frequency information. Besides, TF-CNN also behaves effectively for the unseen noise, except for the azimuth localization under the diffuse noise condition. Multitask learning simultaneously estimated azimuth and elevation with the same neural network, which is demonstrated to effectively improve the localization performance over the single-task training scheme. Experiments based on the CIPIC HRTF database under spatially uncorrelated and diffuse noise conditions demonstrated that the proposed method achieves preferable localization performance compared with other popular methods. Since the proposed method only considers the azimuth and elevation localization of a single sound source under noise conditions, the future works may focus on the localization under reverberant conditions and multi-sound source localization.

REFERENCES

[1] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.

[2] C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 538–548, Apr. 2008.

[3] S. M. Kim and H. K. Kim, "Direction-of-arrival based SNR estimation for dual-microphone speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 2207–2217, Dec. 2014.

[4] J. Zhang and H. Liu, "Robust acoustic localization via time-delay compensation and interaural matching filter," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4771–4783, Sep. 2015.

[5] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. Berlin, Germany: Springer, 2001, pp. 157–180.

[6] R. M. Stern, G. J. Brown, D. Wang, D. Wang, and G. Brown, "Binaural sound localization," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. New York, NY, USA: Wiley, 2006, pp. 147–185.

[7] T. May, S. van de Par, and A. Kohlrausch, "Binaural localization and detection of speakers in complex acoustic scenes," in *The Technology of Binaural Listening*. Berlin, Germany: Springer, 2013, pp. 397–425.

[8] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[9] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[10] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *Proc. IEEE Int. Conf. Robot. Autom.*, vol. 1, Apr./May 2004, pp. 1033–1038.

[11] K. Youssef, S. Argentieri, and J.-L. Zarader, "A binaural sound source localization method using auditive cues and vision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2012, pp. 217–220.

[12] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform," *Proc. IEEE*, vol. 61, no. 10, pp. 1497–1498, Oct. 1973.

[13] S. T. Birchfield and R. Gangishetty, "Acoustic localization by interaural level difference," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Mar. 2005, pp. 1109–1112.

[14] K. Youssef, S. Argentieri, and J.-L. Zarader, "Towards a systematic study of binaural cues," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 1004–1009.

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

[16] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 68–77, Jan. 2010.

[17] L. A. Jeffress, "A place theory of sound localization," *J. Comparative Physiol. Psychol.*, vol. 41, no. 1, pp. 35–39, 1948.

[18] D. Li and S. E. Levinson, "A Bayes-rule based hierarchical system for binaural sound source localization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, Apr. 2003, pp. 521–524.

[19] J. Nix and V. Hohmann, "Sound source localization in real sound fields based on empirical statistics of interaural parameters," *J. Acoust. Soc. Amer.*, vol. 119, no. 1, pp. 463–479, 2006.

[20] T. May, S. Van De Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[21] C. M. Zannini, R. Parisi, and A. Uncini, "Binaural sound source localization in the presence of reverberation," in *Proc. IEEE Int. Conf. Digit. Signal Process.*, Jul. 2011, pp. 1–6.

[22] R. Parisi, F. Camoes, M. Scarpiniti, and A. Uncini, "Cepstrum prefiltering for binaural source localization in reverberant environments," *IEEE Signal Process. Lett.*, vol. 19, no. 2, pp. 99–102, Feb. 2012.

[23] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3075–3089, 2004.

[24] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1494–1505, Oct. 2014.

[25] C. Pang, H. Liu, J. Zhang, and X. Li, "Binaural sound localization based on reverberation weighting and generalized parametric mapping," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 8, pp. 1618–1632, Aug. 2017.

[26] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner, "A probabilistic model for binaural sound localization," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 36, no. 5, pp. 982–994, Oct. 2006.

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

[28] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Commun.*, vol. 53, no. 5, pp. 592–605, 2011.

[29] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1503–1512, Jul. 2012.

[30] K. Youssef, S. Argentieri, and J.-L. Zarader, "A learning-based approach to robust binaural sound localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 2927–2932.

[31] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Localizing speakers in multiple rooms by using deep neural networks," *Comput. Speech Lang.*, vol. 49, pp. 83–106, May 2018.

[32] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2015, pp. 2814–2818.

[33] Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilistic neural network," *IEEE Trans. Ind. Electron.*, vol. 65, no. 8, pp. 6403–6413, Aug. 2018.

[34] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Trans. Audio, Speech Language Process.*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.

[35] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 405–409.

[36] S. Chakrabarty and E. A. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2017, pp. 136–140.

[37] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 103–116, Apr. 2018.

[38] P. Pertilä and J. Nikunen, "Microphone array post-filtering using supervised machine learning for speech enhancement," in *Proc. Interspeech*, 2014, pp. 2675–2679.

[39] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 1, Apr. 2007, pp. 41–44.

[40] J. Braasch, S. Clapp, A. Parks, T. Pastore, and N. Xiang, "A binaural model that analyses acoustic spaces and stereophonic reproduction systems by utilizing head rotations," in *The Technology of Binaural Listening*. Berlin, Germany: Springer, 2013, pp. 201–223.

[41] T. Rodemann, G. Ince, F. Joublin, and C. Goerick, "Using binaural and spectral cues for azimuth and elevation localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2008, pp. 2185–2190.

[42] X. Wu, D. S. Talagala, W. Zhang, and T. D. Abhayapala, "Binaural localization of speech sources in 3-D using a composite feature vector of the HRTF," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2015, pp. 2654–2658.

[43] R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, and S. Goetze, "On sound source localization of speech signals using deep neural networks," in *Proc. Deutsche Jahrestagung Akustik (DAGA)*, 2015, pp. 1510–1513.

[44] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. IEEE Eur. Signal Process. Conf.*, Sep. 2018, pp. 1462–1466.

[45] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2001, pp. 99–102.

[46] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[48] R. Giri, M. L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2015, pp. 5014–5018.

[49] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Proc. Interspeech*, 2015, pp. 5014–5018.

[50] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inf. Theory*, vol. 26, no. 1, pp. 26–37, Jan. 1980.

[51] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

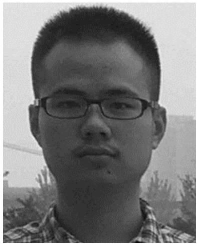
[52] S. Ruder. (2017). "An overview of multi-task learning in deep neural networks." [Online]. Available: https://arxiv.org/abs/1706.05098

[53] J. S. Garfolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus," *Nat. Inst. Standards Technol.*, 1993. [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[54] A. Varga and H. J. M. 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, Jul. 1993.

[55] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, vol. 16, 2016, pp. 265–283.

[56] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing*, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362.



CHENG PANG received the B.E. degree in mechatronic engineering, in 2013. He is currently pursuing the Ph.D. degree with the School of Electronics Engineering and Computer Science (EECS), Peking University (PKU), China. His current research interests include speech and audio signal processing, with a focus on sound source localization, speech enhancement, and speech separation.

HONG LIU received the Ph.D. degree in mechanical electronics and automation, in 1996. He is currently a Full Professor with the School of EECS, Peking University (PKU), China. He is also the Director of the Open Lab on Human Robot Interaction, PKU. He has published more than 150 papers. His research interests include computer vision and robotics, image processing, and pattern recognition. He was a recipient of the Chinese National Aero-Space Award, the Wu Wenjun Award on Artificial Intelligence, and the Excellence Teaching Award. He received the Candidates of Top Ten Outstanding Professors in PKU. He has been selected as a Chinese Innovation Leading Talent supported by the National High-level Talents Special Support Plan, since 2013. He is a Vice President of the Chinese Association for Artificial Intelligent (CAAI) and a Vice Chair of the Intelligent Robotics Society, CAAI. He has served as a Keynote Speaker, a Co-Chair, the Session Chair, and a PC Member for many important international conferences, such as IEEE/RSJ IROS, IEEE ROBIO, IEEE SMC, and IIHMSP. Recently, he serves as a Reviewer for many international journals such as *Pattern Recognition*, the IEEE Transactions on Signal Processing, and the IEEE Transactions on Pattern Analysis and Machine Intelligence.

XIAOFEI LI received the Ph.D. degree in electronics from Peking University, Beijing, China, in 2013. He is currently a Postdoctoral Researcher with INRIA (French Computer Science Research Institute), Montbonnot Saint-Martin, France. His research interests include multimicrophone speech processing for sound source localization, separation and dereverberation, single-microphone signal processing for noise estimation, voice activity detection, and speech enhancement.