Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Synergistic Optimization based Binaural Time-Frequency Masking for Speech Source Localization

Hong Liu, Lulu Wu, Bing Yang
Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, China
hongliu@pku.edu.cn, luluwu@sz.pku.edu.cn, bingyang@sz.pku.edu.cn

***Abstract*— Monaural time-frequency (TF) masking has been demonstrated to advance the performance of binaural speech source localization. However, it fails to consider interaural information, which may result in severe distortion of interaural cues. To mitigate these impacts, this paper presents a novel method for binaural speech source localization based on binaural TF masking. Firstly, the CNN-based binaural TF masking network is designed to suppress the noise and reverberation in TF fragments, which is trained in the independent stage. Then, the resulted binaural TF masking is synergistically refined with the localization network to compensate for the distorted interaural cues. The final source direction is estimated using the trained network. The proposed method is compared with other baseline methods and two-stage models composed by cascade TF masking network and localization network. Experimental results show our method outperforms the other compared methods in the adverse environments with different reverberation time and signal-to-noise ratios.**

***Index Terms*— binaural TF masking, speaker localization, convolutional network, synergistic optimization**

I. Introduction

Speech source localization (SSL) plays an important role for a wide range of applications, such as human-robot interaction and video conference systems [1], [2]. In human-robot interaction, robot can automatically locate speakers by using signals recorded in microphones to realize natural interaction. Binaural SSL based on biological acoustic characteristics has been a prevalent sound localization branch due to its small-sized binaural microphones, a simple binaural SSL model is shown in Fig. 1. Though binaural SSL has been investigated intensively in the past decades, it is still a challenging task in strong noisy and reverberant environments.

According to the auditory scene analysis, human auditory system can separate and locate sounds by synthesizing the time-frequency (TF) relationship, fundamental frequency and harmonic, spatial orientation and other information [3]. The binaural model is studied to mimic this ability, for example by learning the most robust spatial features contained within the simulated head-related transfer function, could therefore greatly improve the localization performance. Over years, many approaches are proposed to achieve favorable performance of binaural SSL, utilizing two primary spatial features, i.e., interaural time difference (ITD) and interaural level difference (ILD) [4]. Generally, speech source is firstly split into a number of frequency bands by a bank of cochlear filters [5]. The ITD and ILD are then estimated in each bands and fed into statistical models like the Gaussian mixture model [6], [7] and artificial network [8], [9] to determine the direction of arrival (DOA) of sources. In order to improve the robustness of interaural cues, reliable frequency bands can be weighted by the maximum likelihood criterion [10], early and late reverberation can be suppressed by the reverberation weighting method [11], frequencies with higher SNR (signal-to-noise ratio) can be emphasized by the TF masking methods [12], [13]. However, these methods typically assume stationary noise, which is an unrealistic assumption in real-world acoustic environments.

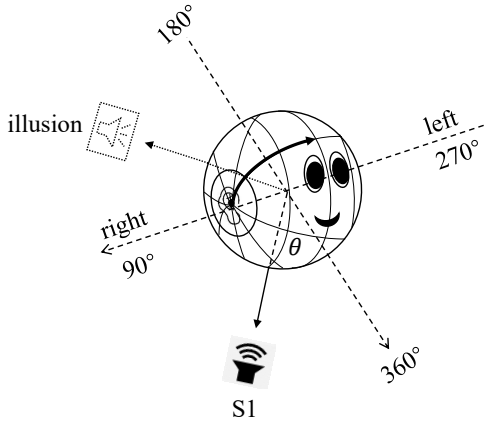

Fig. 1. A simple binaural model with two microphones embedded in the artificial auricles. There are a speech source denoted by S1 and an illusion source symmetric to S1 due to the front-back ambiguity.

Recent deep neural network has shown advanced per-

*This work is supported by National Natural Science Foundation of China (NSFC No.61673030, U1613209), Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (No. ZDSYS201703031405467), National Engineering Laboratory for Video Technology - Shenzhen Division.

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

formances of TF mask estimation, for example, the CNN-based TF masking method reduces the impact of time-varying interferences [14], and the bidirectional long short-term memory (BLSTM) network-based TF masking method improves the performances of generalized cross correlation-, beamforming- and subspace-based localization methods [15], [16]. However, due to the lack of interaural information, these monaural masking methods have the limitation to estimate edge azimuths like 90° and 270°. In other words, the mask of one channel tends to be 0 when estimating edge azimuths, and the interaural cues would be ignored accordingly.

In this paper, we approach the end-to-end binaural SSL from the angle of masking interferences in binaural channels. A CNN-based synergistic optimization method is proposed for binaural TF mask estimation and speech source localization. This study makes three contributions. First, different from monaural TF masking, this paper designs a CNN-based method to process interaural information of the signal in binaural channels. Intuitively, the masking method can be regarded as an attention mechanism to highlight robust TF fragments. Second, the binaural TF masking not only suppresses interferences in the direct-path signal but also retains interaural cues, which contributes to high localization accuracy. A key ingredient is the inclusion of directional information in the learning process of TF masking. Third, a two-stage training scheme is proposed to synergistically fine-tune the parameters of the TF masking network and localization network. This two-stage fashion ensures an efficient training process of the TF masking network and the localization network. The rest of this paper is organized as follows. The proposed algorithms are presented in Section II. Experimental setup and evaluation results are reported in Sections III. Section IV concludes this paper.

II. Proposed Method

A. Binaural signal model

Suppose that there is only one target speaker, the physical model for binaural signals in noisy and reverberant environments under the narrowband approximation assumption can be formulated as:

$$Y_m(t,f) = H_{\theta,m}(f) \cdot S(t,f) + V_m(t,f), m \in \{l, r\}, \quad (1)$$

where m represents the microphone index, i.e., left (l) or right (r), t is the time frame index and f is the frequency index, $H_{\theta,m}(f)$ denotes head-related transfer function (HRTF) between the m-th microphone and the speech source $S(t,f)$ at azimuth θ, and $V_m(t,f)$ represents other interferences.

In conventional methods, classical interaural cues, interaural phase difference (IPD) and ILD, are usually extracted from narrowband signals. Actually, the differences of magnitude and phase components between microphone pairs correspond to the ILD and IPD respectively. Instead of using IPD and ILD as inputs, the input matrix composes of the magnitude and phase components of the short-time Fourier transform (STFT) of binaural signals:

$$\mathbf{x} = \begin{bmatrix} x(1,1) & x(1,2) & \cdots & x(1,F) \\ x(2,1) & x(2,2) & \cdots & x(2,F) \\ \vdots & \vdots & \ddots & \vdots \\ x(T,1) & x(T,2) & \cdots & x(T,F) \end{bmatrix}, \quad (2)$$

where T is the number of time frames and F is the number of frequency bins. Each entry of the matrix $\mathbf{x}$ is the permutation of the log-magnitude and the phase components of STFT of each channel:

$$\begin{aligned} x(t,f) = \Big[& 20\log_{10}|Y_l(t,f)|, 20\log_{10}|Y_r(t,f)|, \\ & \angle Y_l(t,f), \angle Y_r(t,f) \Big], \end{aligned} \quad (3)$$

where $|\cdot|$ denotes the magnitude, and $\angle\cdot$ denotes the phase.

B. TF masking model

Although STFT coefficients present the high resolution for spatial cues, there is still a need to assign great attention to reliable TF bins dominated by direct-path target source. There are several typical TF masking methods defined in [17]. The Wiener-like mask has optimal average signal-to-noise ratio (SNR) when the average power statistics are computed for stationary signals, here, it is used to model the TF mask to enhance the contribution of TF points of direct-path signals:

$$\eta_m(t,f) = \frac{10\log_{10}|H^{dp}_{\theta,m}(f) \cdot S(t,f)|^2}{10\log_{10}[|H^{dp}_{\theta,m}(f) \cdot S(t,f)|^2 + |U_m(t,f)|^2]}, \quad (4)$$

where $H^{dp}_{\theta,m}(f)$ represents the direct-path HRTF, $U_m(t,f)|^2$ is the mixture of other interferences, i.e., the noise, early and late reverberation:

$$U_m(t,f) = \overline{H}_{\theta,m}(f) \cdot S(t,f) + V_m(t,f), \quad (5)$$

where $\overline{H}_{\theta,m}(f)$ denotes the early and late reverberation.

C. DOA encoding

The binaural SSL can be regarded as a multi-label classification task, by assigning each azimuth with probability 0 or 1. However, in reality, adjacent azimuths within certain distances from the true DOA can be considered correct. To release the constraint of one-hot encoding, we follow the Gaussian-like encoding method [18], [19] to encode the output probabilities of azimuths to be Gaussian curves with peaks centered at positions corresponding to the desired azimuths. Such an encoding method is also applicable to multi-source localization. For an I-direction binaural SSL, the output likelihood is encoded into a vector $\{o_i\}$ of I values, which is defined as:

$$o_i = \begin{cases} e^{-d(\theta_i,\theta)^2/\sigma^2} & \text{if source exists} \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

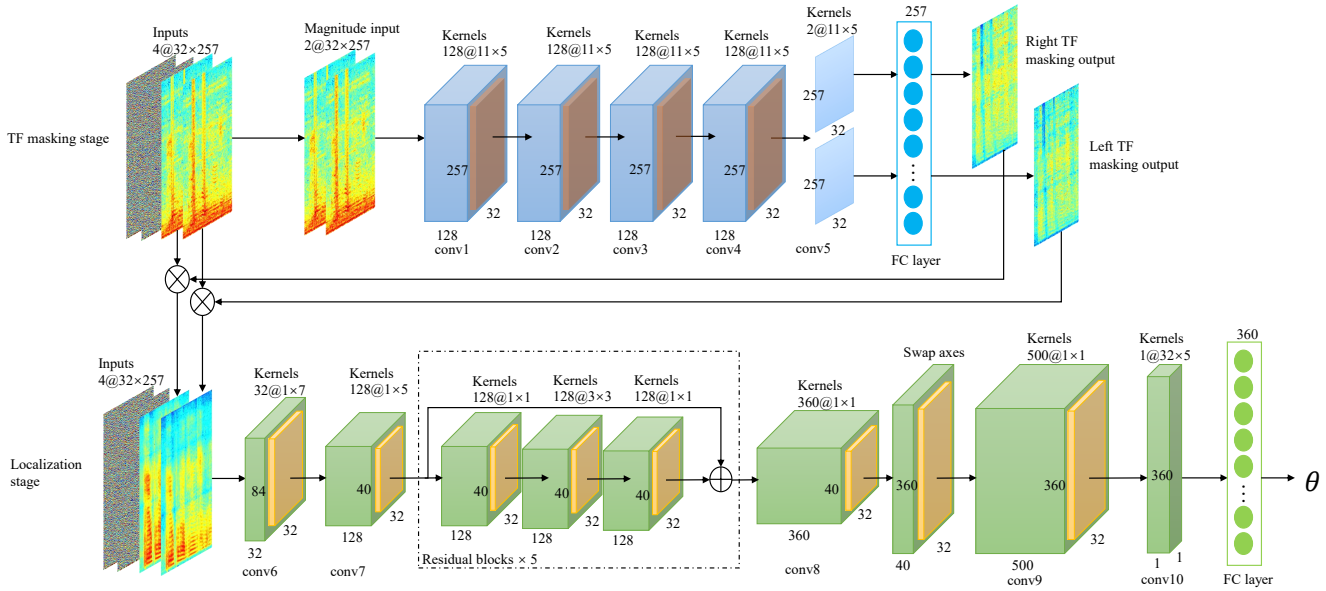

Fig. 2. An overview of the proposed network architecture, including two stages, i.e., TF masking stage and localization stage. Taking a speech signal as an example, 512-point FFT is performed on a signal block containing 32 frames to extract 257-dimensional log-magnitude and phase components.

where θ is the true DOA, $d(.,.)$ denotes the angular distance, and σ is the tolerances between adjacent azimuths. Since there is a closed circular azimuth range in the full horizontal plane, the encoded output should also be in a closed circular range. In the decoding step, the predicted azimuth is retrieved by maximizing the corresponding probability.

D. Network architecture

A schematic diagram of the proposed CNN-based binaural localization system is depicted in Fig. 2. This system includes two stages, namely TF masking stage and localization stage. The TF masking stage yields binaural masks through TF masking network (TFNet) using log-magnitude components of the STFT of two channels. With the help of approximated TF masks, the localization stage feeds the phase and enhanced log-magnitude components to the localization network (DOANet). The details of the whole TF-DOANet are described in the following parts.

TFNet Unlike monaural TF masking, the TFNet takes the log-magnitude components of left and right channels as inputs and outputs binaural TF masks with the same dimensions. The TFNet contains five convolutional layers to simultaneously process both channels. Each layer has a number of kernels with size of 11×5 (time $\times$ frequency) and a rectified linear unit (ReLU) activation function. These convolutional kernels, operating on the time-frequency fragments, learn contextual temporal and spatial information. The number of kernels of each layer is marked before the symbol "@" in Fig. 2. In the last layer, a fully-connected (FC) layer with sigmoidal activation function is designed to regress the masks to the range of [0,1]. The FC layer has the same number of neurons as the dimensions of log-magnitude features, i.e., 257 neurons, and it processes the outputs of each time frame and each channel with the shared weights. Noting that inputs and outputs are of the same dimensions, proper zero-padding is performed in each layer and no max-pooling layer is included.

DOANet Before executing the DOANet in the localization stage, the log magnitude is enhanced by the channel-specific TF masks. The raw phase components and the masked magnitude are then rearranged as inputs of the DOANet. A relatively efficient CNN-based sub-network from [20], is taken as the backbone network in our DOANet. The main architecture of DOANet is shown at the bottom of Fig. 2. The kernels in the first two convolutional layers slide along frequency axis with strides of 7 and 5 respectively. In the residual block, the outputs of the previous block and current block are connected to deepen the network. These shortcut connections avoid vanishing information. Temporal information is processed individually in previous convolutional layers except for the last convolutional layer. In the last convolutional layer, the kernel size depends on the number of contextual frames, i.e. $T \times 5$, which enables the feasibility of DOANet to deal with different sizes of signal blocks. In other words, the last layer can be fine-tuned with different size of signal block regardless of parameters of previous layers.

Two-stage training scheme According to the above network configuration, the total size of parameters of the TF-DOANet is 15.7 M, which makes it difficult to train the network in an end-to-end manner. This paper introduces an efficient training scheme that firstly separates the whole

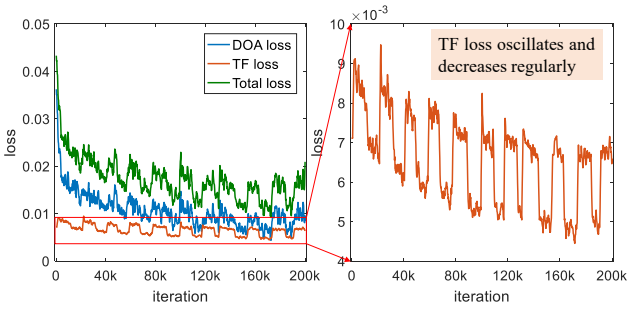

Fig. 3. The iterative TF loss, DOA loss and the cumulative loss. The right figure is the magnification of the TF loss.

network into two stages and then updates the networks one by one from scratch. The TFNet is firstly trained by imposing target binaural TF masks on outputs. The TF masking loss, denoted by $Loss_{tf}$, is computed by the mean square error (MSE) criterion. An optimizer named Adam [21] is used to optimize the network parameters step by step with an initial learning rate of 0.001. The training procedure of TFNet can be early stopped if the TF loss no longer drops and tends to be stable. Second, the TFNet and the DOANet are combined to train in an end-to-end fashion. Here, the DOA loss $Loss_{doa}$ is also computed by the MSE criterion using the true DOA encoding. The TF loss and the DOA loss are aggregated and used to calculate the gradient of the whole network parameters:

$$Loss = Loss_{tf} + Loss_{doa}. \tag{7}$$

The optimizer Adam is also used to fine-tune all parameters in the second training stage with an initial learning rate of 0.015 which is gradually decreased by 0.5 every 3 epochs. The training loss in this stage is depicted in Fig. 3. As the training process proceeding, the DOA loss and the total loss decrease distinctly. The TF loss seems to be convergent, but actually, it continues to oscillate and decline regularly as a result of the collision between noise suppression and cues preservation.

III. Experiments and Analysis

A. Experimental setup

In this section, experiments are carried out with different binaural setups. The head-related impulse responses (HRIRs) are provided by the CIPIC dataset [22] and KEMAR dataset [23] within four source-to-sensor distances (0.5, 1, 2 and 3 m). Source DOAs are considered in the range of $[1°, 360°]$ with a resolution of $1°$. Clean utterances from the TIMIT dataset [24] are used for all experiments, of which 65 sentences are used for training and different 34 sentences are used for testing. A noise-free training set Q_1 is generated by convolving HRIRs with training utterances, the same for the testing set. To simulate the noisy environments, a noisy training set Q_2 is produced by respectively adding three spatially uncorrelated noises (white, m109, f16), which are from the NOISEX dataset [25], to signals of Q_1 at various SNRs. Another noise named babble, which is recorded in a canteen where 100 people spoke simultaneously, is properly truncated and added to the noise-free testing set as diffuse interferences. Based on the image method [26], the Roomsim toolbox [27] is employed to generate binaural room impulse responses (BRIRs) in a simulated enclosure of $(10 \times 6 \times 3)$ m. The subject_021 from the CIPIC dataset is placed at the central position. Speech sources are positioned at a distance of 1.5 m around the subject. In such a reverberant setup, averaged reverberation time ($\overline{RT}_{60}$), (0s, 0.2s and 0.1s, 0.3s, 0.5s, 0.7s) are considered for training and testing respectively. A noisy and reverberant training set Q_3 is produced by adding corresponding noises to reverberant signals generated by convolving BRIRs with clean utterances. These acoustic conditions are summarized in Table I.

TABLE I

Configuration of training and testing sets.

	Training set	Testing set
Source-to-sensor distance	0.5m, 1m, 2m, 3m, 1.5m	1.5m
Noise types	white, f16, m109	babble
SNR	-10dB:10:20dB	-5dB:10:25dB
$\overline{RT}_{60}$	0s, 0.2s	0.1s, 0.3s, 0.5s, 0.7s
Number of binaural signals	650840	144000

All the binaural signals are resampled to 16 kHz. The leading and trailing zeros are removed from the left and right signals. STFT is performed on every 512 samples with 50% overlapping. The signal block contains 32 consecutive frames, equaling to 528 ms duration. The localization performance is evaluated in terms of two metrics, i.e., the localization accuracy (Acc.) and the mean absolute error (MAE). The predicted azimuth is supposed to be correct if it is within $5°$ away from the true DOA. The MAE is used to measure the angular errors between the true DOA and estimated DOA:

$$MAE(°) = \frac{1}{C}\sum_{i=1}^{C}|\widehat{\theta^i} - \theta^i|, \tag{8}$$

where C is the total number of binaural mixtures, $\widehat{\theta}$ and θ correspond to the true and estimated DOAs. The performance of TFNet is also evaluated in two ways, i.e., signal-to-interference ratio (SIR) and the root-mean-square error (RMSE) of the ILD:

$$SIR = \frac{1}{2}\sum_{i=1}^{2} 20\log_{10}\frac{\sum_{t,f}\eta_i(t,f)\cdot|Y_i(t,f)|}{\sum_{t,f}|Y_i(t,f)| - \eta_i(t,f)\cdot|Y_i(t,f)|}, \tag{9}$$

$$\Delta_{ILD} = \sqrt{\frac{1}{|t|\cdot|f|}\sum_{t,f}(\hat{\lambda}(t,f) - \lambda(t,f))^2}, \tag{10}$$

where $\hat{\lambda}$ and λ denotes the ILD of the enhanced binaural magnitude and the noise-free binaural magnitude, respectively.

TABLE II

PERFORMANCE OF TF MASKING METHODS IN THE REVERBERANT ENVIRONMENTS WITHOUT ADDITIVE NOISE.

$\overline{RT}_{60}$	0.1s		0.3s		0.5s		0.7s	
Measure	SIR	Δ_{ILD}	SIR	Δ_{ILD}	SIR	Δ_{ILD}	SIR	Δ_{ILD}
CNN-TF [14]	15.55	0.3286	16.24	0.3809	15.88	0.3860	15.73	0.3902
BLSTM-TF [15]	17.77	0.8265	15.37	0.1560	15.27	0.2021	15.25	0.2203
TFNet	**21.20**	**0.0635**	**18.57**	**0.1079**	**18.20**	**0.1301**	**18.10**	**0.1404**

B. Contribution of binaural TF masking

The first experiment investigates the contribution of the binaural TF masking. Our TFNet is compared to two network-based monaural TF masking methods, i.e., the CNN-based [14] and the BLSTM-based [15] TF masking. The performances of these masking methods are shown in Table II under different reverberant environments.

The greater value of the SIR means the better interferences suppression, and the less value of the Δ_{ILD} means the better preservation of the ILD. Observation shows that when the SIR of the other two methods is large, the Δ_{ILD} is accordingly large. It is difficult for these monaural masking methods to restore interaural cues while suppressing noise. In turn, keeping only interaural cues is detrimental to binaural SSL, since the similar interaural cues between frontal and back planes would lead to the front-back confusion. Concluding results from this table, our TFNet can effectively suppress interferences and preserve the interaural cue with the help of binaural masking and directional information from DOANet.

To illustrate the noise-suppress ability of the TFNet, a spectrogram example is shown in Fig. 4. Figures in the first row are the spectrogram of the received binaural signal in the reverberant environment with $\overline{RT}_{60} = 0.1s$ and additive babble noise at SNR = 5 dB. Figures in the other two rows are the corresponding spectrograms of the noise-free signal and the masked signal. Compared with the received signal, the spectrogram of the masked signal is closer to the noise-free signal's. In the third row, the spectrum at low frequencies becomes clearer and less disturbed by noise, while the spectrum at high frequencies is mostly blocked out due to the less energy of high-frequency speech. The influence of the high-frequency spectrum on sound localization is less than that of the low-frequency spectrum, thus the distortion of spectrogram at high frequencies can be negligible.

C. Localization performance

To evaluate our method, three network-based localization methods are selected for the comparison: frequency-dependent DNN [9], BLSTM-based mask-weighted generalized cross correlation with phase transform (GCC-PHAT) [15], and phase-based CNN [28]. In the frequency-dependent DNN, Ma et al. employed multi-layer perceptron for each frequency subband to form a mapping from the joint features of cross-correlation function (CCF) and ILD to source

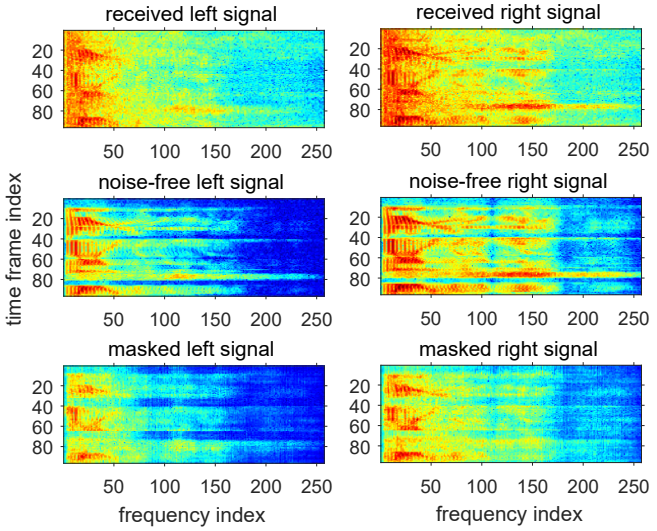

Fig. 4. Spectrogram of the received noisy signal, noise-free signal and masked signal, respectively.

azimuths. Besides, they exploited head movements to identify the front-back confusions, but this auxiliary head movement can also be applied in other methods so that we just compare our method with the frequency-dependent DNN. The GCC-PHAT is the classical method to estimate the time difference of arrival (TDOA). The ideal ratio mask, estimated by the BLSTM using the log-magnitude components, is used to weight the robustness of the TDOA in each TF bin. Since the BLSTM-based GCC-PHAT method only focuses on the half plane, the front and back hemifileds are input as a priori knowledge. The phased-based CNN is originally designed for broadband DOA estimation using the phase spectrum of STFT in the microphone-array setup.

Table III shows the localization accuracy of these methods tested in reverberant rooms with different reverberation time and direct-to-reverberant ratios (DRR). Obviously, our method significantly outperforms others in noisy and reverberant conditions. In the frequency-dependent DNN, the posterior probabilities of azimuths are equally integrated across all frequency bands, which ignores the weights of each frequency band that is dominated by the target source. The performance of frequency-dependent DNN drops dramatically when tested in low-SNR environments. In contrast, our method can achieve more than 90% of localization accuracy under noisy conditions. The BLSTM-based masking method mainly estimates the importance of TF units, which is a prerequisite for accurate localization. The GCC-PHAT improves the localization accuracy by 1% over Ma's method in $\overline{RT}_{60} = 0.5s$ and $0.7s$ reverberant environments, but it fails to locate speech source in different source-to-sensor distance. The phased-based CNN achieves average 70% accuracy in the reverberant conditions without noise, but its best performance is limited to 77% since it ignores the importance of the

TABLE III

Acc.(%) ($\sigma = 5°$) OF THE LOCALIZATION METHODS IN VARYING REVERBERANT ENVIRONMENTS WITH BABBLE NOISE AT DIFFERENT SNRS.

$\overline{RT}_{60}$(s)/DRR(dB)	0.1/-1.44					0.3/-2.02					0.5/-2.58					0.7/-3.11				
SNR	-	25 dB	15 dB	5 dB	-5 dB	-	25 dB	15 dB	5 dB	-5 dB	-	25 dB	15 dB	5 dB	-5 dB	-	25 dB	15 dB	5 dB	-5 dB
MA [9]	87.0	84.36	80.42	74.19	70.43	70.05	66.77	60.89	58.43	46.91	59.0	55.72	49.84	47.38	35.86	53.5	50.97	45.09	42.63	31.11
WANG [15]	86.85	83.7	76.0	73.5	65.23	68.75	66.45	57.05	52.05	48.51	60.35	58.05	50.65	49.65	38.38	54.25	51.2	47.8	42.8	34.53
CHAKRABARTY [28]	77.85	77.54	76.94	75.21	61.74	74.26	73.49	70.07	63.4	41.24	73.04	72.15	67.75	56.54	32.35	72.85	71.25	66.42	52.43	29.04
TF-DOANet	**99.93**	**99.75**	**99.81**	**99.36**	**95.90**	**99.24**	**98.57**	**97.92**	**95.10**	**81.72**	**98.40**	**96.71**	**95.36**	**91.10**	**66.90**	**97.64**	**94.82**	**93.64**	**88.03**	**59.12**

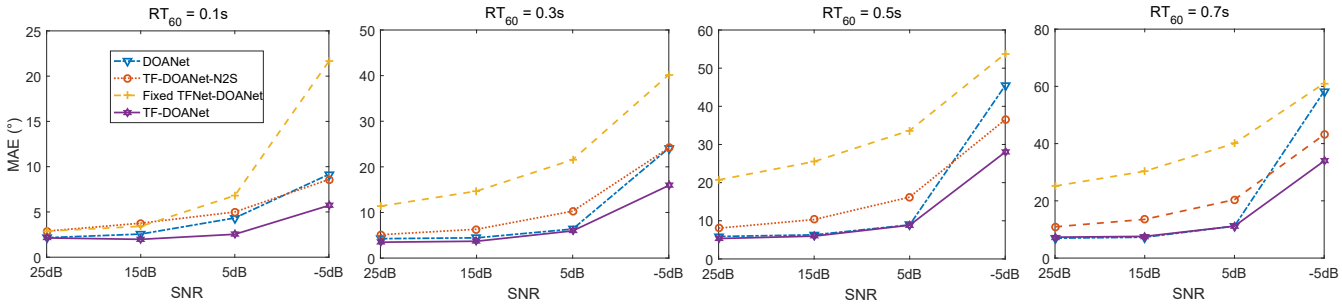

Fig. 5. $MAE.(°)$ ($\sigma = 8°$) of localization methods in the reverberant environments with babble noise at different SNRs.

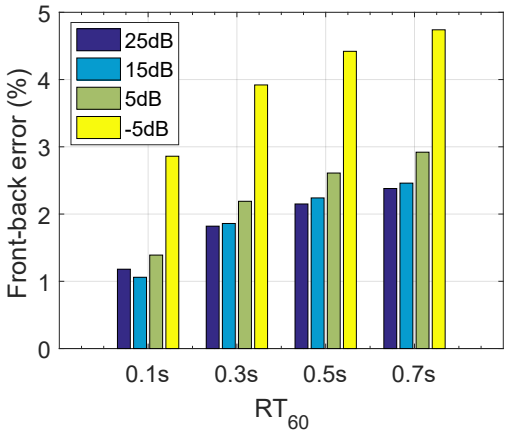

Fig. 6. Front-back error (%) of the TF-DOANet.

magnitude components. In our TF-DOANet, the localization performance is effectively improved by over 30% in strong noisy and reverberant environments, by taking advantages of the binaural masking to suppress noise and reverberation and combining the magnitude with phase components as features.

D. Ablation study

In order to evaluate the role of synergistic optimization for TF masking and binaural SSL, experiments are also conducted to compare with three training schemes:
DOANet: only the DOANet, uses raw STFT as inputs.
TF-DOANet-N2S: the proposed network, is trained in an end-to-end fashion without a two-stage training scheme.
Fixed TFNet-DOANet: the TFNet is firstly trained and its parameters are then fixed in the second training stage.

The MAE results of these methods are illustrated in Fig. 5. In the single-source localization, the MAE is increased along with the decreasing SNR and increasing reverberation time. Compared with other training schemes, the proposed method trained in a two-stage fashion can significantly reduce angular errors, especially in the noisy environments at SNR = -5 dB. Since the parameters of TFNet are fixed in the Fixed TFNet-DOANet, the TFNet trained without directional information would produce wrong TF masks to distort the interaural cues. The TF-DOANet-N2S is trained in an end-to-end fashion without fixing parameters, but there are too many parameters of the whole network to be trained sufficiently. The DOANet has similar performances as ours in high-SNR conditions. However, without the guideline of binaural masking, the MAE of DOANet is increased by more than $5°$ in SNR = -5 dB environments. According to the ablation study, the TF masking is demonstrated to work as an attention mechanism to highlight reliable TF bins. The TF masking network and the localization network are mutually promoted.

A well-known challenge for binaural SSL is the front-back ambiguity, as shown in Fig 1. Due to the front-back symmetry of the artificial head, an illusion source may occur on the other side symmetric to the target source. To testify the impact of front-back confusions on binaural source localization, our method is evaluated in different noisy and reverberant environments. The abnormal error rate caused by the front-back confusions is depicted in Fig 6. The sum of localization accuracy and front-back error might be greater than 1, that is because the estimated azimuth and the true azimuth are symmetrical regarding the robotic head and the angular distance is within the tolerances when calculating localization accuracy. Under different reverberation time and different SNRs, the front-back error of our method is always less

than 5%. According to the results, the noise shows greater influence on the front-back confusions than the reverberation.

IV. Conclusions

We have investigated a new synergetic approach for robust speech source localization guided by binaural TF masking. Benefiting from the localization network, the binaural TF masking network can effectively improve the robustness of interaural cues and suppress noise and reverberation in TF points. In turn, with the help of robust interaural cues, the localization performance can also be improved accordingly. An interesting observation is that the TF-DOANet is helpful to identify the front-back ambiguity in binaural sound localization, which further demonstrates the effectiveness of binaural TF masking of suppressing the noise and reverberation. This study also proposed a two-stage training scheme for the TF-DOANet, showing an efficient way to update all parameters. Although the experiments are only conducted on the single-source dataset, our TF-DOANet can also be applied to multi-source localization by modifying the Gaussian-like encoding to that with multiple peaks of azimuth probability. However, there is still limitation of our approach that needs to be solved in future work. Our TF masking network only considers the magnitude components and their corresponding masks, regardless of the phase components. We believe the phase masking would be helpful for improving interaural cues.

References

[1] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1997, pp. 187–190.

[2] B. Kwon, G. Kim, and Y. Park, "Sound source localization methods with considering of microphone placement in robot platform," in *IEEE International Symposium on Robot and Human Interactive Communication*, 2007, pp. 127–130.

[3] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[4] L. Jeffress, "A place theory of sound localization," *IEEE Journal of Comparative and Physiological Psychology*, vol. 61, pp. 468–486, 1947.

[5] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[6] G. R. Karthik and P. K. Ghosh, "Binaural speech source localization using template matching of interaural time difference patterns," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5164–5168.

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

[8] K. Youssef, S. Argentieri, and J.-L. Zarader, "A binaural sound source localization method using auditive cues and vision," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 217–220.

[9] 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 Transactions on Audio, Speech and Language Processing*, vol. 25, no. 12, pp. 2444–2453, 2017.

[10] G. R. Karthik, P. Suresh, and P. K. Ghosh, "Subband weighting for binaural speech source localization." in *INTERSPEECH*, 2018, pp. 861–865.

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

[12] M. I. Mandel, D. P. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Advances in Neural Information Processing Systems*, 2007, pp. 953–960.

[13] X. Wu, D. S. Talagala, W. Zhang, and T. D. Abhayapala, "Spatial feature learning for robust binaural sound source localization using a composite feature vector," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 6320–6324.

[14] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 6125–6129.

[15] Z. Wang, X. Zhang, and D. Wang, "Robust TDOA estimation based on time-frequency masking and deep neural networks," in *INTERSPEECH*, 2018, pp. 322–326.

[16] ——, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 178–188, 2019.

[17] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 708–712.

[18] F. Palmieri, M. Datum, A. Shah, and A. Moiseff, "Learning binaural sound localization through a neural network," in *IEEE Northeast Bioengineering Conference*, 1991, pp. 13–14.

[19] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *IEEE International Conference on Robotics and Automation*, 2018, pp. 74–79.

[20] ——, "Joint localization and classification of multiple sound sources using a multi-task neural network." in *INTERSPEECH*, 2018, pp. 312–316.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference of Learning Representation*, 2014, pp. 1–13.

[22] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 99–102.

[23] H. Wierstorf, M. Geier, and S. Spors, "A free database of head related impulse response measurements in the horizontal plane with multiple distances," in *Audio Engineering Society Convention 130*, 2011.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[25] A. Varga and H. J. 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 Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

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

[28] S. Chakrabarty and E. A. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–10, 2019.