Received: 30 September 2020 | Revised: 15 December 2020 | Accepted: 25 December 2020 | CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12010

ORIGINAL RESEARCH

Head-related transfer function–reserved time-frequency masking for robust binaural sound source localization

Hong Liu[1] | Peipei Yuan[1] | Bing Yang[1] | Ge Yang[2] | Yang Chen[3]

[1]Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Shenzhen, China

[2]School of Artificial Intelligence, Chongqing University of Technology, Chongqing, China

[3]Yanka Kupala State University of Grodno, Grodno, Belarus

Correspondence

Peipei Yuan, Room A328, Building A, Lishui Road No.2199, Nanshan District, Shenzhen City, 518000, Guangdong Province
Email: peipeiyuan@pku.edu.cn

Funding information

National Natural Science Foundation of China, Grant/Award Numbers: 61673030, U1613209; National Natural Science Foundation of Shenzhen, Grant/Award Number: JCYJ20190808182209321

Abstract

Various time-frequency (T-F) masks are being applied to sound source localization tasks. Moreover, deep learning has dramatically advanced T-F mask estimation. However, existing masks are usually designed for speech separation tasks and are suitable only for single-channel signals. A novel complex-valued T-F mask is proposed that reserves the head-related transfer function (HRTF), customized for binaural sound source localization. In addition, because the convolutional neural network that is exploited to estimate the proposed mask takes binaural spectral information as the input and output, accurate binaural cues can be preserved. Compared with conventional T-F masks that emphasize single speech source–dominated T-F units, HRTF-reserved masks eliminate the speech component while keeping the direct propagation path. Thus, the estimated HRTF is capable of extracting more reliable localization features for the final direction of arrival estimation. Hence, binaural sound source localization guided by the proposed T-F mask is robust under noisy and reverberant acoustic environments. The experimental results demonstrate that the new T-F mask is superior to conventional T-F masks and lead to the better performance of sound source localization in adverse environments.

1 | INTRODUCTION

Binaural sound source localization (SSL) aims to determine the azimuth, elevation or distance between the sound source and the center of the microphone array, which uses binaural microphones mounted on the left and right sides of the robot head. SSL is valuable for research and a variety of applications, such as speech enhancement, speech separation, speech recognition, human–robot interaction, teleconferencing, and hearing aids [1–5].

Plenty of approaches have been proposed to estimate the direction of arrival (DOA), most of which are composed of two steps. In the first step, the localization features are extracted from the received waveform signals or spectral information. For binaural SSL, binaural cues in a biomimetic way are commonly used, including interaural time difference (ITD), interaural phase difference (IPD) and interaural level difference (ILD). In detail, ITD describes the time difference of a sound source arriving at binaural microphones, whereas IPD refers to the phase difference of a sound wave reaching each ear. Moreover, ILD represents the level difference of a sound source between binaural signals. In the second stage, the DOA is estimated according to the mapping relationship between input features and DOA. During that stage, many methods can be used to establish this mapping relationship, such as peak searching [6] and template matching [7, 8]. In addition, probabilistic statistic models can be established using some methods such as gaussian mixture models [9] and deep learning–based models [10].

Based on an auditory front end, the complex interaction of ITDs and ILDs is built by a probabilistic model trained under various acoustic conditions [11]. In the meantime, DNNs are used to structure the relationship between the source azimuth and binaural cues, including the complete cross-correlation function (CCF) and ILDs [10]. Most previous methods tried to exploit information in the binaural

cue pairs and capture the complicated mapping relationship. However, the performance of traditional localization feature extraction approaches may seriously degrade in the presence of reverberation and noise. Therefore, it is necessary to add an extra step (weighting or enhancing the localization features) before the estimation of the DOA. With the cooperation of binaural coherence [12], the coherence test [13], a time-frequency(T-F) mask [14, 15], and so on, the features are finally obtained from more reliable T-F units. In addition, accurate binaural cues can be directly extracted from the received waveform signals or spectrum [16, 17].

Because of the eminent learning ability of deep neural networks, T-F masks are capable of guiding the SSL to focus on single source–dominated T-F units [14, 18]. Various T-F masks designed manually are listed in Wand and Chen [19]. A ideal binary mask(IBM) and ideal ratio mask(IRM) are two popular masks for speech separation. The value of IBM is either 1 or 0, depending on whether the signal-to-noise ratio (SNR) is low enough in the specific two-dimensional T-F representation. The IRM is a soft mask that is similar to the IBM. It evaluates the ratio between the speech energy and the summation of the clean speech energy and noise energy. According to the definition of the IRM, it can weigh the target speech signal well only when additive noise exists in the acoustic environment. However, IRM may not suitable for T-F units full of reverberation, which can be regarded as uncorrelated interference.

Two main issues exist in real-valued T-F mask-guided SSL: (1) most T-F masks are designed for a monoaural signal; and (2) T-F mask definitions usually involve only the spectral amplitude, signal power, or SNR. For the first issue, because a monoaural mask usually does not consider the difference between multichannel signals , the binaural cues may be destroyed during T-F unit selection. The localization information extracted from the binaural cues is incomplete or even promiscuous, which leads to inaccurate sound source results. In terms of the second issue, owing to the deficiency of the phase information, which is more significant than the amplitude information for the SSL task, the weight of T-F units may be measured ambiguously. Consequently, the mis-selected T-F units will also generate an inaccurate localization feature, leading to the degradation of performance.

To solve these confusing issues, several methods [18, 20] based on the simultaneous processing of binaural signals exhibiting the ability to preserve binaural cues have been proposed. A phase-sensitive mask, including a measure of phase [21]; the complex IRM, employed for both magnitude and phase spectra [22] and the dereverberation mask [23] are proposed to yield a better estimation of clean speech. However, these T-F masks are elaborately designed for speech separation. If we apply these T-F masks directly to SSL, they eliminate the influence of the early and late reverberations to some extent, but still circuitously.

Motivated by these studies, we propose a novel complex T-F mask-guided binaural SSL approach. This work mainly makes three contributions. First, different from previous monoaural T-F masks used in speech separation, the proposed mask is customized for binaural SSL. It is dedicated to highlighting the optimal T-F units while resisting the disturbances of noise and reverberation. Second, this complex mask is designed to reserve the direct path component of the head-related transfer function(HRTF) from mixed binaural spectra, the HRTF-reserved mask. Third, the experiments demonstrate that the ITD and ILD, calculated from an estimation of HRTF, can lead to lower localization error compared with features extracted from the received binaural signals.

Section 2 formulates the binaural signal model and related works. The definition of the HRTF-reserved mask and system overview are described in Section 3. Section 4 describes the experimental setup and exhibits experimental results with different acoustic environments. Finally, conclusions are given in Section 5.

2 | RELATED WORK

2.1 | Binaural signal model

The received binaural signals $x^i(n)$ in the time domain can be formulated as:

$$x^i(n) = s(n) \otimes h^i(n) + v^i(n), \forall i = l, r, \tag{1}$$

where $\otimes$ denotes the convolution operation, n is the time index and $s(n)$ and $v^i(n)$ represent the clean sound signal and additive noise signal, respectively. l and r refer to the index of the left and right channels. The $h^i(n)$ denotes the impulse response between the source and ear, consisting of an indoor acoustic property and head-related impulse response [7].

After applying short-time Fourier transform (STFT), the binaural signal in the time domain is transformed to the time-frequency domain, which can be modeled as:

$$X^i(t,f) = S(t,f)H^i(f,\theta) + V^i(t,f), \tag{2}$$

where $X^i(t,f)$, $S(t,f)$ and $V^i(t,f)$ represent the spectra of the received binaural signal, clean speech and noise signal, respectively. $H^i(f,\theta)$ is the frequency-domain version of the binaural room impulse response (BRIR), in which the propagation path contains the direct path, early reflections and late reverberation. θ denotes the corresponding azimuth and t and f denote the index of time frame and frequency bin, respectively.

In conventional approaches, physical localization cues such as ITD (or IPD) and ILD are directly extracted from the received signals for each time frequency pair [24].

IPD and ILD of the binaural signals in Equation (2) can be respectively extracted as:

$$\widehat{\phi}(t,f) = \angle \frac{X_r(t,f)}{X_l(t,f)}, \tag{3}$$

$$\widehat{\lambda}(t,f) = 20\log_{10}\frac{|X_r(t,f)|}{|X_l(t,f)|}, \tag{4}$$

where $\widehat{\phi}(t,f)$ denotes the IPD at the $t-th$ time frame and $f-th$ frequency bin, and $\widehat{\lambda}(t,f)$ denotes the ILD at the $t-th$ time frame and $f-th$ frequency bin.

As mentioned, the additive noise and reverberation component will disturb the localization features in the specific T-F units. The features calculated from the single source–dominated T-F pairs, which involve only direct path propagation, can lead to better performance.

2.2 | Direction of arrival estimation via template matching

We exploit the template matching method [7] to estimate the sound source. First, the offline template establishment is conducted by:

$$\lambda^T(f,\theta) = 20\log_{10}\left|\frac{\mathrm{HRTF}_r(f,\theta)}{\mathrm{HRTF}_l(f,\theta)}\right|, \tag{5}$$

where T represents the template, $\mathrm{HRTF}_r(f,\theta)$ and $\mathrm{HRTF}_l(f,\theta)$ denote the T-F domain (time t is omitted) HRTFs on the right and left ears for azimuth θ, respectively. Similarly, ITD templates $T_p^T(f,\theta)$ can be established by:

$$\phi^T(f,\theta) = \frac{1}{f}\angle\frac{\mathrm{HRTF}_r(f,\theta)}{\mathrm{HRTF}_l(f,\theta)}. \tag{6}$$

Because the azimuths are known in the template establish stage, we can calculate the theoretical ITD.

Second, both the ITD and ILD are considered for SSL. Their estimation can be respectively calculated as:

$$\widehat{\lambda}(f,\theta) = 20\log_{10}\left|\frac{H_{dp}^r(f,\theta)}{H_{dp}^l(f,\theta)}\right|, \tag{7}$$

$$\widehat{\phi}(f,\theta) = \frac{1}{f}\angle\frac{H_{dp}^r(f,\theta)}{H_{dp}^l(f,\theta)}. \tag{8}$$

Finally, the DOA estimation is obtained by minimizing the hybrid distances, which can be formulated as:

$$\widehat{\theta} = \underset{\theta_j}{\operatorname{argmin}}\sum_f \left\{D_T(f,\theta_j)\cdot D_I(f,\theta_j)\right\}, \tag{9}$$

where j is the azimuth index, $D_T(f,\theta_j)$ denotes the distance between ITD estimation and ITD templates, and $D_I(f,\theta_j)$ denotes the distance between ILD estimation and ITD templates.

3 | CONVOLUTIONAL NEURAL NETWORK–BASED HEAD-RELATED TRANSFER FUNCTION–RESERVED MASK ESTIMATION

The schematic diagram of the proposed binaural SSL is illustrated in Figure 1. In the binaural signal simulation phase, the binaural signals are generated by convolving the BRIR with the clean speech signal. Moreover, because different kind of noises with various SNRs always exist in real scenarios, we generated noisy binaural signals by adding a noise signal to the clean binaural signals. During training, the STFT is performed on the received signals. After that, the data block, which is composed of the imaginary and real parts of T-F units, is fed to a convolutional neural networkm (CNN). The ITD and ILD calculated directly by HRTF are regarded as feature templates for the DOA estimation. In the test stage, the HRTFs are estimated by multiplying the T-F units of binaural signals with the HRTF-reserved mask predicted from the trained CNN. With regard to template matching, we measure the distance between the templates and the binaural cues extracted from the estimated HRTFs. The azimuth corresponding to the minimum distance is determined as the final DOA [7].

3.1 | Head-related transfer function–reserved time-frequency mask

For T-F mask-guided SSL, several T-F masks are available [19]. The complex-valued mask is considered owing to its ability to restore the STFT coefficient. The typical complex IRM that suppresses the contribution of the T-F unit, including the noise signal and reverberation, is defined as:

$$cIRM(t,f) = \frac{S(t,f)H_{dp}(f,\theta)}{X(t,f)}, \tag{10}$$

where $H_{dp}(f,\theta)$ denotes the direct-path HRTF[1] corresponding to azimuth θ at frequency f. The microphone index is omitted for the sake of simplicity. The STFT coefficient of the direct-path signal $X_{dp}(t,f)$ is obtained after multiplying the received signal $X(t,f)$ and the $cIRM(t,f)$. This process can be formulated as:

$$X_{dp}(t,f) = cIRM(t,f)X(t,f) = S(t,f)H_{dp}(f,\theta). \tag{11}$$

To reserve the HRTF from the direct-path speech signal, we define a novel T-F mask as:

[1]In this paper, we relax the HRTF definition and use the term HRTF to describe the frequency response from a target source to binaural microphones.

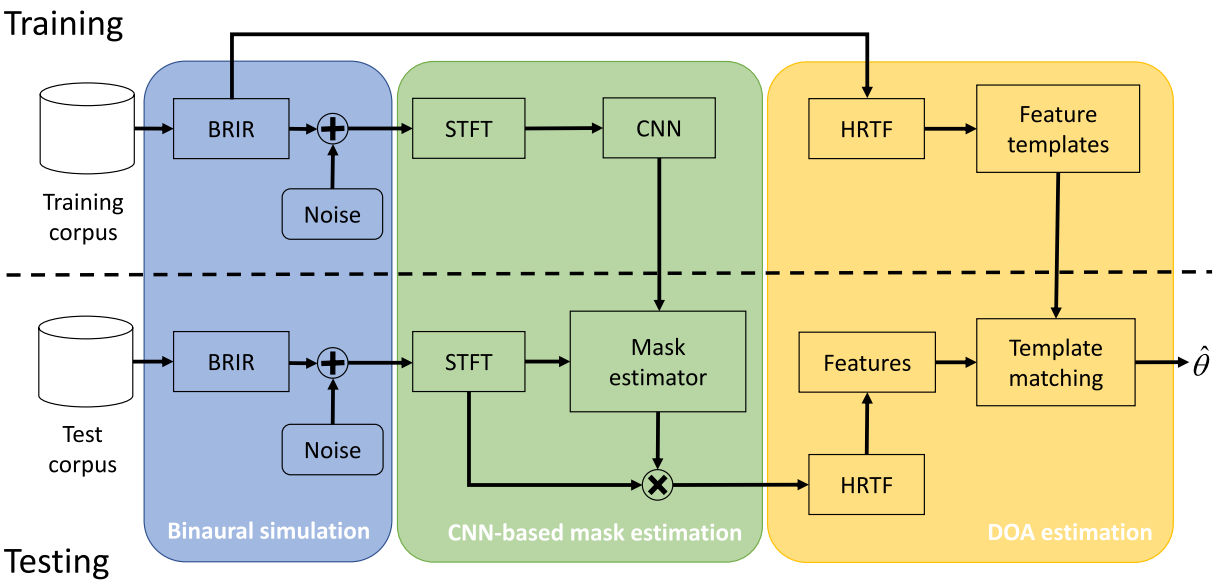

FIGURE 1 Schematic diagram of proposed head-related transfer function–reserved mask-guided sound source localization method. The upper part is the training phase constructing a convolutional neural network–based mask estimator and the binaural feature templates. The lower part is the test phase to estimate the azimuth through template matching and time-frequency masking.

$$cIRM_{re}(t,f) = \frac{1}{S(t,f)}. \tag{12}$$

According to Equations (11) and (12), $cIRM_{re}$ is a de-speech operation. Therefore, the desired HRTF can be obtained by:

$$H_{dp}(f,\theta) = cIRM_{re}(t,f)X_{dp}(t,f). \tag{13}$$

Combining $cIRM(t,f)$ with $cIRM_{re}(t,f)$, a fused mask $cIRM_{fu}(t,f)$ can directly extract the HRTF from the received signal:

$$cIRM_{fu}(t,f) = cIRM(t,f)cIRM_{re}(t,f), \tag{14}$$

$$H_{dp}(f,\theta) = cIRM_{fu}(t,f)X(t,f). \tag{15}$$

Unlike previous manually designed T-F masks, we first emphasize the T-F units containing the direct-path signal. Then, the proposed T-F mask directly eliminates the speech component from the direct-path speech signal while preserving the HRTF of the sound source.

With regard to neural network training, mask estimation networks are trained using two strategies: separately (*cIRM* and $cIRM_{re}$) or in an integrated way ($cIRM_{fu}$).

3.2 | Convolutional neural network–based mask estimation

Different from monoaural T-F masking, the STFT coefficient of both left and right channels are stacked together as the input data, as shown in Figure 2. On the one hand, both phase and magnitude components, the complete information of the single-channel signal in the T-F domain, are simultaneously considered in a straightforward way. On the other hand, the full spatial information implied in binaural signals can be well-preserved. The input matrix is formulated as:

$$\mathbf{x} = \left[\mathbf{x}^{\mathbf{l}}_{\mathbf{real}}, \mathbf{x}^{\mathbf{l}}_{\mathbf{img}}, \mathbf{x}^{\mathbf{r}}_{\mathbf{real}}, \mathbf{x}^{\mathbf{r}}_{\mathbf{img}}\right] \tag{16}$$

$$\mathbf{x}^{\mathbf{i}}_{\mathbf{k}} = \begin{bmatrix} X^i_k(1,1) & X^i_k(1,2) & \cdots & X^i_k(1,F) \\ X^i_k(2,1) & X^i_k(2,2) & \cdots & X^i_k(2,F) \\ \vdots & \vdots & \ddots & \vdots \\ X^i_k(T,1) & X^i_k(T,2) & \cdots & X^i_k(T,F) \end{bmatrix}, \quad \forall\, k = real, img, \tag{17}$$

where *real* refers to the real part of the complex-value, whereas *img* refers to the imagination component and T and F denote the number of time frames and frequency units, respectively. Binaural signals with a sampling rate of 44.1 kHz are processed by STFT with a Hanning window. The window length is 20 ms (882 samples) with a hop length of 10 ms. Seven frames with frequency bins ranging from 80 to 8000 Hz (159 samples) are packed into a data block. The size of the input data is $7 \times 159 \times 4$ (frame × frequency × channel).

The architecture of the T-F mask estimation network is depicted in Figure 2. We employ a simple CNN with four two-dimensional convolutional layers to predict the T-F mask [18]. The kernel size of each layer is 3×3 with the stride keeping output the same size as input data. The number of filters of each layer is four, corresponding to the imaginary and real parts of the left and right channels. Because the network processes the binaural signals simultaneously, the direct-path propagation from the sound source to microphones can be captured and the binaural cues between T-F pairs can also be preserved. As a

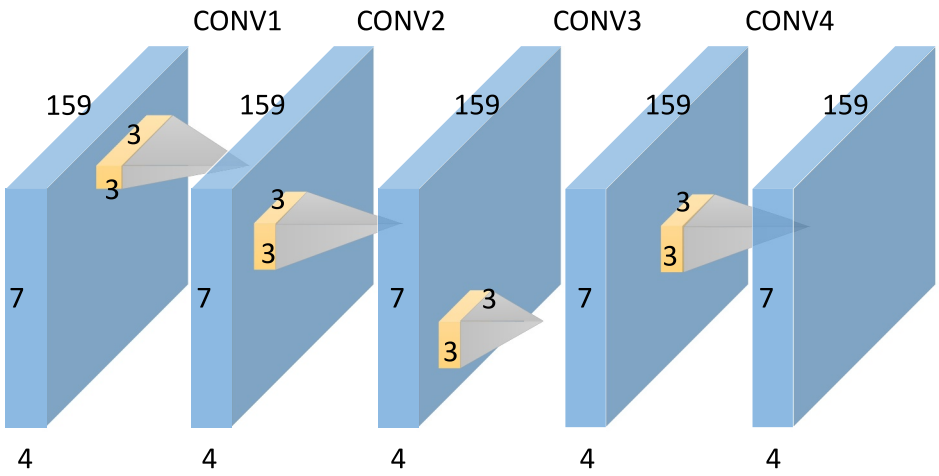

FIGURE 2 Architecture of time-frequency masking network. The shape of input composed of time, frequency and channel (7, 159 and 4) is the same as each layer output

result, the accurate localization feature of the direct-path component can be extracted directly from the $\widehat{H}_{dp}(f,\theta)$.

According to the definition of the HRTF-reserved mask in Equations (10), (12) and (14), the network output is similar to the input matrix:

$$\mathbf{y} = \left[\mathbf{y}^{\mathbf{l}}_{\mathbf{real}}, \mathbf{y}^{\mathbf{l}}_{\mathbf{img}}, \mathbf{y}^{\mathbf{r}}_{\mathbf{real}}, \mathbf{y}^{\mathbf{r}}_{\mathbf{img}}\right] \tag{18}$$

$$\mathbf{y}^{\mathbf{i}}_{\mathbf{k}} = \begin{bmatrix} y^i_k(1,1) & y^i_k(1,2) & \cdots & y^i_k(1,F) \\ y^i_k(2,1) & y^i_k(2,2) & \cdots & y^i_k(2,F) \\ \vdots & \vdots & \ddots & \vdots \\ y^i_k(T,1) & y^i_k(T,2) & \cdots & y^i_k(T,F) \end{bmatrix}, \tag{19}$$

where the y^i_k could be *cIRM*, $cIRM_{re}$ or $cIRM_{fu}$. Each CNN corresponding to a specific T-F mask is trained separately and independently.

The training configurations for three training targets are the same. All networks are conducted with the Pytorch with one NVIDIA GeForce Titan XP GPU. The batch size is set to 16. An Adam optimizer [25] is used to optimize the network parameters by minimizing the mean absolute error(MAE). The initial learning rate is set to 0.003. Then it is divided by 3 every 10 epochs until the performance of validation set no longer improves.

4 | EXPERIMENTS AND ANALYSIS

4.1 | Experimental setup

For the binaural simulation, the HRTFs from the CIPIC HRTF database [26], audio signals from the TIMIT database [27] and noise signal from the Noisex-92 database [28] are exploited to synthesize the received signals.

There are 45 different subjects in the CIPIC HRTF database. Each has 25 azimuths and 50 elevations. The sources are placed at 0 degrees elevation and all azimuths range from [−80 degrees, −60 degrees, −55 degrees, −45 degrees: 5 : 45 degrees, 55 degrees, 65 degrees, 80 degrees], in which 0 degrees is located at the middle front of the head. Subject 21 (i.e., Kemar head) is selected to simulate an acoustic environment for the T-F mask estimation in the following experiments. All subjects are used to build ITD and ILD offline templates, as proposed in Pang et al. [7].

In the experiments, 25, seven, and seven utterances are selected randomly from the TIMIT to generate training, validation and a test set for T-F masking, respectively. Furthermore, to simulate the noisy acoustic environment, babble noise with various SNR, [0:10:30] dB for a training and validation set and [–5:10:25] dB for a test set are added to the noise-free binaural speech signals. The spectrum of the babble noise signal is similar to that of the speech signal and the SNRs are unmatched between the training, validation set and test set, which can increase the credibility of the following experiments of the method.

To evaluate the robustness of the proposed binaural localization method, room impulse responses are also simulated through the Roomsim toolbox [29]. Table 1 shows the room configurations including room size (W, L and H denote the width, length and height of the room, respectively), the distance between the head and source (R), and the head location and reverberation time (RT_{60}). Figure 3 illustrates the simulated acoustic environments for room1 in Table 1.

In total, there are 4500, 1576 and 936 utterances in the training set, validation set and test set, respectively.

4.2 | Results

To evaluate the performance of the proposed T-F mask, we directly measure the MAE of template matching with

TABLE 1 Room configuration for training and test dataset

Dataset	Room	W (m)	L (m)	H (m)	R (m)	Center of head (m)	RT_{60} (s)
Training, validation	Room1	10	6	3	1	(5, 3, 1.5)	0.2, 0.4, 0.6
	Room2	6	5	4.5	1	(3, 2.5, 1.2)	0.2, 0.4, 0.6
	Room3	6	4	3	1	(2, 2, 1.2)	0.2, 0.4, 0.6
Test	Room4	5.5	8	4	1	(3.5, 1.5, 1.2)	0.2, 0.4, 0.6, 0.8

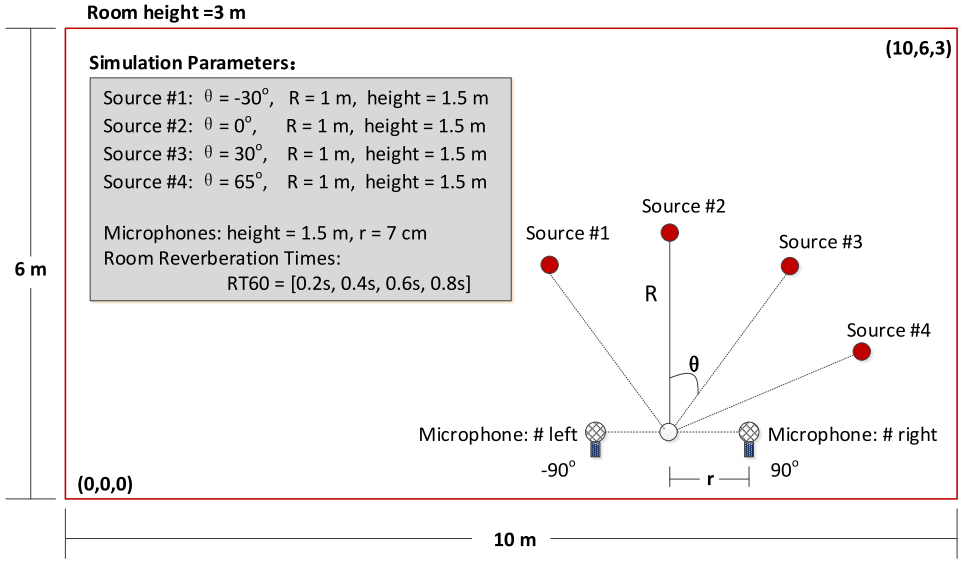

FIGURE 3 Simulated scene and parameters of acoustic environments for room1

TABLE 2 The MAE of DOA estimation (degrees) for models trained in multiconditional environment

	Signal-to-noise ratio (dB)				RT_{60} (s)				
	-5	5	15	25	0.2	0.4	0.6	0.8	Average
TM [7]	41.45	39.72	28.72	14.19	15.80	24.46	29.47	33.53	28.42
TM-complex IRM	**11.60**	**11.54**	11.32	9.93	**8.27**	12.12	15.59	14.87	11.90
TM-HRTFM(sepa)	13.71	12.22	10.36	9.44	10.57	11.10	13.29	15.21	11.98
TM-HRTFM(fuse)	13.00	11.85	**9.85**	**8.07**	9.20	**10.54**	**12.63**	**14.41**	**11.19**

Abbreviations: DOA, direction of arrival; HRTFM, Fuse, networt trained in integrated or fuse way; head-related transfer function–reserved mask; IRM, ideal ratio mask; MAE, mean absolute error; RT, reverberation time; Sepa, networt trained in independant way; SNR, signal-to-noise ratio; TM, template matching.

The MAE values of method that performs best under the corresponding acoustic conditions are represented in bold.

different types of masks or without a mask. For the DOA estimation stage in the third step of the framework, we use joint ITD and ILD template matching [7], denoted as TM in Table 2. Moreover, HRTFM denotes the proposed HRTF-reserved masks including integrated training (fuse) in Equation (14) and independent training (separation, sepa) of two masks in Equations (12) and (10). The comparison of TM-HRTFM with TM can be considered ablation experiments. The bold number in Table 2 means that the corresponding method performs the best under specific acoustic environment.

The comparison of the azimuth localization performance is shown in Table 2. The HRTF-reserved mask reduces localization error compared with the SSL method without a mask or with complex IRM. This demonstrates that the design of the HRTFM is able to realize a de-speech operation and reserve the HRTF corresponding to the direct-path speech signal. In addition, the proposed mask leads to more effective and accurate binaural SSL results compared with no masks or with the complex IRM.

The performance of TM-HRTFM (sepa) is worse than that of TM-HRTFM (fuse). This is because exploiting a single

network to estimate the fused mask directly can avoid introducing a non-negligible accumulated error.

The TM-HRTF (fuse) does not perform as well as TM-complex IRM in the low-SNR environment, whereas TM-HRTF (fuse) performs better in high-reverberant environments. The main reason is that background noise with no specific direction obscures the directional information of the sound source whereas HRTFM is more suitable for reverberation acoustic environments that contain sufficient directional information.

5 | CONCLUSION

A T-F mask is proposed to reserve the HRTF directly from received signals and lead to the extraction of robust binaural cues in the presence of reverberation and noise. A simple CNN is exploited to estimate the HRTF-reserved mask. The HRTF containing the direct-path propagation from the source to the head is available after multiplying the mask with binaural signals in the T-F domain. Then the ITD and ILD can be extracted efficiently from the HRTFs. Thus, experimental results demonstrate that the performance of template matching–based, probability model-based and other two-stage SSL methods can be promoted, especially in a reverberant environment.

Compared with previous handcrafted monoaural T-F masks, the proposed T-F mask is robustly superior to binaural SSL. The contribution of T-F units dominated by the direct-path speech signal can be precisely evaluated. Furthermore, the direct path component of the HRTF is able to be reserved from a mixture of binaural spectra. Thus, SSL guided by the proposed mask adapts to unknown and adverse acoustic environments, especially in the presence of reverberation.

ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (Nos. 61673030 and U1613209) and the National Natural Science Foundation of Shenzhen (No. JCYJ20190808182209321).

REFERENCES

1. Wang, D., Brown, G.J.: Computational Auditory Scene Analysis: Principles, Algorithms, and Applications, Wiley-IEEE Press, Piscataway, N.J (2006)
2. Lombard, A., et al.: Multidimensional localization of multiple sound sources using averaged directivity patterns of blind source separation systems. In: IEEE International conference on acoustics, speech and signal processing, pp. 233–236 (2009)
3. Visser, E.: Frequency domain passive broadband speaker localization using a permutation-free blind source separation algorithm. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 673–676 (2007)
4. Farmani, M., et al.: Informed sound source localization using relative transfer functions for hearing aid applications. IEEE/ACM Trans. Audio Speech Lang. Process. 25(3), 611–623 (2017)
5. Kim, S.M., Kim, H.K.: Direction-of-arrival based snr estimation for dual-microphone speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 22(12), 2207–2217 (2014)
6. Ying, D., et al.: Window-dominant signal subspace methods for multiple short-term speech source localization. IEEE/ACM Trans. Audio Speech Lang. Process. 25(4), 731–744 (2017)
7. Pang, C., et al.: Binaural sound localization based on reverberation weighting and generalized parametric mapping. IEEE/ACM Trans. Audio Speech Lang. Process. 25(8), 1618–1632 (2017)
8. Girija Ramesan, K., Suresh, P., Ghosh, P.K.: *Subband weighting for binaural speech source localization*. In: Proocedings of Interspeech, pp. 861–865 (2018)
9. Woodruff, J., Wang, D.: Binaural localization of multiple sources in reverberant and noisy environments. IEEE Trans. Audio Speech Lang. Process. 20(5), 1503–1512 (2012)
10. Ma, N., May, T., Brown, G.J.: Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. IEEE Trans. Audio Speech Lang. Process. 25(12), 2444–2453 (2017)
11. May, T., van de Par, S., Kohlrausch, A.: A probabilistic model for robust localization based on a binaural auditory front-end. IEEE Trans. Audio Speech Lang. Process. 19(1), 1–13 (2011)
12. Faller, C., Merimaa, J.: Source localization in complex listening situations: selection of binaural cues based on interaural coherence. J. Acoust. Soc. Am. 116, 3075–3089 (2004)
13. Mohan, S., et al.: Localization of multiple acoustic sources with small arrays using a coherence test. J. Acoust. Soc. Am. 123(4), 2136–2147 (2008)
14. Wang, Z., Zhang, X., Wang, D.: Robust speaker localization guided by deep learning-based time-frequency masking. IEEE Trans. Audio Speech Lang. Process. 27(1), 178–188 (2019)
15. Wang, Z.Q., Zhang, X., Wang, D.: *Robust tdoa estimation based on time-frequency masking and deep neural networks*. In: Proceedings of Interspeech, pp. 322–326 (2018)
16. Pak, J., Shin, J.W.: Sound localization based on phase difference enhancement using deep neural networks. IEEE Trans. Audio Speech Lang. Process. 27(8), 1335–1345 (2019)
17. Liu, H., et al.: Robust interaural time difference estimation based on convolutional neural network. In: IEEE International Conference on Robotics and Biomimetics, pp. 352–357 (2019)
18. Liu, H., Wu, L., Yang, B.: Synergistic optimization based binaural time-frequency masking for speech source localization. In: IEEE International Conference on Robotics and Biomimetics, pp. 2363–2369 (2019)
19. Wang, D., Chen, J.: Supervised speech separation based on deep learning: an overview. IEEE Trans. Audio Speech Lang. Process. 26(10), 1702–1726 (2018)
20. Sun, X., et al.: A deep learning based binaural speech enhancement approach with spatial cues preservation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5766–5770 (2019)
21. Erdogan, H., et al.: Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 708–712 (2015)
22. Williamson, D.S., Wang, Y., Wang, D.: Complex ratio masking for monaural speech separation. IEEE Trans. Audio Speech Lang. Process. 24(3), 483–492 (2016)
23. Sun, Y., et al.: Two-stage monaural source separation in reverberant room environments using deep neural networks. IEEE Trans. Audio Speech Lang. Process. 27(1), 125–139 (2019)
24. Algazi, V.R., Avendano, C., Duda, R.O.: Elevation localization and head-related transfer function analysis at low frequencies. J. Acoust. Soc. Am. 109(3), 1110–1122 (2001)

25. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations, pp. 1–13 (2014)
26. Algazi, V.R., et al.: The cipic hrtf database. In: IEEE Workshop on the applications of signal processing to audio and acoustics, pp. 99–102 (2001)
27. Garofolo, J., et al.: Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1 (NASA STI/Recon Technical Report N93), (1993)
28. Varga, A., Steeneken, H.J. 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. 12(3), 247–251 (1993)
29. Campbell, D., Palomaki, K., Brown, G.: A MATLAB simulation of shoebox room acoustics for use in research and teaching. Comput. Inf. Syst. J. 3, 48–51 (2005)

How to cite this article: Liu, H., et al.: Head-related transfer function–reserved time-frequency masking for robust binaural sound source localization. CAAI Trans. Intell. Technol. 7(1), 26–33 (2022). https://doi.org/10.1049/cit2.12010