ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413923

SUPERVISED DIRECT-PATH RELATIVE TRANSFER FUNCTION LEARNING FOR BINAURAL SOUND SOURCE LOCALIZATION

Bing Yang[1,2]*, Xiaofei Li*[2]*, Hong Liu*[1]

[1]Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, China
[2]Westlake University & Westlake Institute for Advanced Study, Hangzhou, China

bingyang@sz.pku.edu.cn, lixiaofei@westlake.edu.cn, hongliu@pku.edu.cn

ABSTRACT

Direct-path relative transfer function (DP-RTF) refers to the ratio between the direct-path acoustic transfer functions of two channels. Though DP-RTF fully encodes the sound directional cues and serves as a reliable localization feature, it is often erroneously estimated in the presence of noise and reverberation. This paper proposes a supervised DP-RTF learning method with deep neural networks for robust binaural sound source localization. To exploit the complementarity of single-channel spectrogram and dual-channel difference information, we first recover the direct-path magnitude spectrogram from the contaminated one using a monaural enhancement network, and then predict the DP-RTF from the dual-channel (enhanced-) intensity and phase cues using a binaural enhancement network. In addition, a weighted-matching softmax training loss is designed to promote the predicted DP-RTFs to be concentrated for the same direction and separated for different directions. Finally, the direction of arrival (DOA) of source is estimated by matching the predicted DP-RTF with the ground truths of candidate directions. Experimental results show the superiority of our method for DOA estimation in the environments with various levels of noise and reverberation.

Index Terms— Relative transfer function, binaural sound source localization, direction of arrival, deep neural network.

1. INTRODUCTION

Sound source localization has been investigated intensively in last decades due to its wide application in teleconferencing, robot audition, etc. Many researchers adopt a dual-stage approach which consists of localization feature extraction and feature-to-location mapping [1, 2]. Deep learning has been successfully applied to the localization task recently. Under the dual-stage localization framework, deep neural network (DNN) can be used to either extract localization features [3, 4], or build the mapping from the localization features to source location [5, 6]. Commonly used localization feature includes inter-channel time difference (ITD) [7], inter-channel phase difference (IPD) [8], inter-channel intensity difference (IID), relative transfer function (RTF) [9, 10], etc. The source can be easily localized with aforementioned localization features under a noise-free and anechoic condition. However, in practical acoustic scenes, noise and reverberation often contaminate the direct-path propagated source signal and degrade the accuracy of feature estimation.

Many methods aim to remove the effect of acoustic interferences on the direct-path feature extraction. One common way to reduce the adverse effect is to select the time-frequency (TF) bins dominated by direct sound. Unsupervised TF bin selection methods include coherence test [11], direct path dominance (DPD) test [12], etc. With the development of deep learning techniques, supervised methods [13, 14, 15, 16] are also used to identify single-source-dominant T-F bins. For these methods, the selection error, i.e. miss-detections and false-detections, of TF bins, will lead to localization feature estimation error. Instead of retaining or discarding certain TF bins, some methods dedicate to directly remove the acoustic interferences and retain direct-path information. Knapp *et al*. adopted a cross-power spectrum weighting scheme to improve the robustness of ITD in the presence of noise [7]. Pak *et al*. trained a DNN to enhance the interference-contaminated IPD on the sinusoidal tracks [4]. Li *et al*. identified the direct-path relative transfer function (DP-RTF) of a single speaker, with a convolutive transfer function approximation and an inter-frame spectral subtraction algorithm designed for reflection exclusion and noise removal respectively [17]. Despite decades of research, estimating a robust localization feature under adverse acoustic conditions still remains a challenging problem.

This paper proposes a supervised DP-RTF learning method to preserve the time and intensity difference cues of direct-path signal, and meanwhile suppress the contamination of noise and reverberation. First, to fit the real-value DNN, the complex DP-RTF is changed into a real-value representation which is a concatenation of the IID and the sinusoidal functions of the IPD. Then, the monaural enhancement network (MEnet) and the binaural enhancement network (BEnet) are employed to learn the DP-RTF. Specifically, the MEnet predicts the clean direct-path log magnitude spectrogram from the contaminated one, and then the predicted spectrogram together with the phase components are passed to the BEnet to predict the DP-RTF. As the single-channel spectrogram and dual-channel difference information are complementary and helpful for suppressing the affection of acoustic interferences, the combination of BEnet and MEnet can provide more reliable DP-RTF prediction. To make the DP-RTF estimation more suited for direction of arrival (DOA) estimation, a weighted-matching softmax training loss is designed to enforce the inter-class compactness and inter-class separability. Finally, with the predicted DP-RTF, the DOA can be estimated by a simple but effective matching method. Experiments demonstrate the effectiveness of our method under various acoustic conditions.

2. PROBLEM FORMULATION

We consider a single source observed by binaural microphone pair, equipped in the dual ears of a dummy head, in an enclosed environment with additive ambient noise. The signal received by the m-th microphone is denoted as $x_m(t)$ with $t \in [1, T]$ and $m \in \{1, 2\}$. Applying the short-time Fourier transform (STFT) to $x_m(t)$, the mi-

This work is supported by National Natural Science Foundation of China (No. 61673030, U1613209), Science and Technology Plan Project of Shenzhen (No. JCYJ20200109140410340).

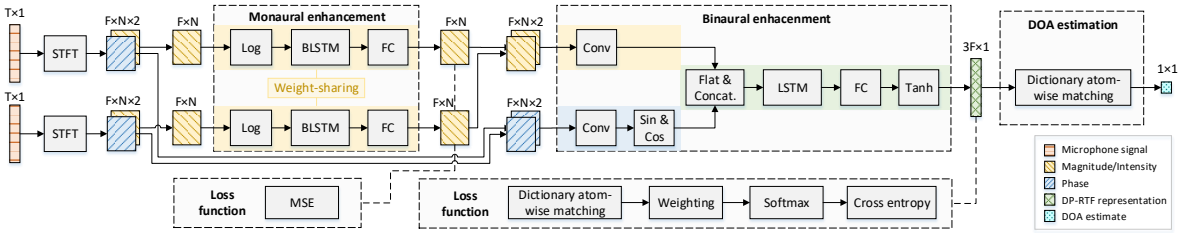

Fig. 1. Framework of the supervised DP-RTF learning for robust DOA estimation.

crophone signal is expressed in the TF domain as

$$X_m(n,f) = H_m(f,\theta)S(n,f) + V_m(n,f), \tag{1}$$

where $n \in [1, N]$ represents the time frame index, $f \in [1, F]$ represents the frequency index, and θ denotes the horizontal DOA of source. N and F refer to the number of utilized frames and frequencies, respectively. Here, $X_m(n,f)$, $S(n,f)$ and $V_m(n,f)$ represent the microphone, source and noise signals in the TF domain, respectively. The acoustic transfer function $H_m(f,\theta)$ involves the direct and reflected propagation paths of the sound source, i.e.,

$$H_m(f,\theta) = H_m^{\mathrm{d}}(f,\theta) + H_m^{\mathrm{r}}(f,\theta), \tag{2}$$

where $H_m^{\mathrm{d}}(f,\theta)$ and $H_m^{\mathrm{r}}(f,\theta)$ denote the acoustic transfer functions of direct-path and reflected propagations, respectively. The direct-path relative transfer function (DP-RTF) [17] is defined as the ratio between the two direct-path acoustic transfer functions, namely

$$R^{\mathrm{d}}(f,\theta) = H_2^{\mathrm{d}}(f,\theta)/H_1^{\mathrm{d}}(f,\theta). \tag{3}$$

Since the difference between the direct-path signals of two channels fully encodes the source location, our goal is to accurately estimate the DP-RTF from the microphone signals $X_m(n,f)$ in the presence of noise and reverberation, so that sound source localization can be performed by directly matching the estimated DP-RTF with the ground truths (lookup) of candidate directions.

3. SUPERVISED LEARNING FOR LOCALIZATION

3.1. DP-RTF representation

In theory, the direct-path acoustic transfer function, more specifically the head-related transfer function (HRTF) in the present binaural localization context, can be expressed as $H_m^{\mathrm{d}}(f,\theta) = \alpha_m(f,\theta)e^{-j\omega_f \tau_{m(\theta)}}$, where ω_f denotes the angular frequency of the fth frequency, and $\alpha_m(f,\theta)$ and $\tau_{m(\theta)}$ are the propagation attenuation factor and the time of arrival from the source to the mth microphone, respectively. Substituting it into Eq. (3), the DP-RTF can be rewritten as

$$R^{\mathrm{d}}(f,\theta) = \frac{\alpha_2(f,\theta)}{\alpha_1(f,\theta)} e^{-j\omega_f(\tau_2(\theta)-\tau_1(\theta))}. \tag{4}$$

The DP-RTF is indeed a complex value, which encodes IID and IPD information in its magnitude and argument respectively. However, the complex DP-RTF cannot be directly processed by the real-value DNN. Instead of directly learning complex DP-RTF, we carefully design the DP-RTF representation without information loss to fit the real-value DNN.

The phase-magnitude decomposition is done to map the complex domain to real values without losing location information. The phase of DP-RTF is exactly the IPD, and we denote it as $\Delta P(f,\theta) = \angle R^{\mathrm{d}}(f,\theta)$, where $\angle$ is the phase operator of complex numbers. The IPD is in the range from $-\pi$ to π. It tends to be periodically wrapped with the increasing of frequency or time difference, and discrete when $\omega_f(\tau_2(\theta) - \tau_1(\theta))$ reaches $\pi + 2i\pi$ with an integer i. To avoid the phase wrapping ambiguity, the sinusoidal functions of IPD is used instead, namely $\sin \Delta P(f,\theta)$ and $\cos \Delta P(f,\theta)$, which are continuous in [-1,1] as the location of the sound source varies. Since the DP-RTF is defined as a ratio between two values, the magnitude of DP-RTF is asymmetrical w.r.t. the broadside direction of the two microphones. Instead, we transform the magnitude into log domain, as the IID defines

$$\Delta I(f,\theta) = 20\log_{10}\left|R^{\mathrm{d}}(f,\theta)\right| / \Delta I_{\max}, \tag{5}$$

where $\Delta I_{\max}$ is an empirically-set maximum value of IID used for normalization, which normalizes the IID into the range of [-1,1] to balance the contribution of the IID and IPD information. The DP-RTF representation contains the full-band normalized IID, the sine and cosine of the IPD, namely

$$\begin{aligned}\mathbf{r}(\theta) =& [\Delta I(1,\theta),\ldots,\Delta I(F,\theta), \sin\Delta P(1,\theta),\ldots,\sin\Delta P(F,\theta),\\ & \cos\Delta P(1,\theta),\ldots,\cos\Delta P(F,\theta)]^T \text{ in } \mathbb{R}^{3\mathrm{F}\times 1},\end{aligned} \tag{6}$$

where $(\cdot)^T$ denotes vector transpose. Each element of $\mathbf{r}(\theta)$ is in the range of [-1, 1]. The complex DP-RTF can be reversely recovered using this DP-RTF representation.

To estimate this DP-RTF using the noisy and reverberant microphone signals, we leverage both the monaural spectral pattern learning and the inter-channel difference learning in this work. Fig. 1 shows the framework of the supervised DP-RTF learning for robust DOA estimation. It consists of monaural enhancement network (MEnet) and binaural enhancement network (BEnet). The monaural magnitude spectrogram is first enhanced using a single-channel speech enhancement method, and then the DP-RTF is estimated using the dual-channel enhanced-intensity and phase information.

3.2. Monaural enhancement

Deep learning has been widely used for monaural speech enhancement [18]. Deep monaural speech enhancement is especially effective for the magnitude spectrogram due to the structured magnitude spectral pattern of speech. In contrast, the monaural phase enhancement is much more difficult. Compared to the normal dual-microphone setup, the magnitude/intensity difference plays an especially important role for binaural source localization, since the "head shadow" effect makes the magnitude response of the binaural recordings prominently different from each other. Therefore, in this work we adopt the monaural enhancement technique to promote the binaural magnitude difference estimation. Note that the commonly used enhancement technique is to predict mask and add the mask to localization features [14, 15], while our method aims to directly estimate the clean single-channel log spectrogram.

Considering the better context modeling ability of RNN over feed-forward neural networks and convolutional neural networks, we adopt the bi-directional long short-term memory (BLSTM) to estimate the magnitude spectrogram of clean speech from that of noisy and reverberant speech. To facilitate training and to be consistent with the log-IID defined in (5), a log operation is applied to compress the dynamic range of the magnitude spectrogram, and the mapping from contaminated signal to clean one will be performed in the

log magnitude domain. The BLSTM model consists of two hidden layers and each layer contains 1024 units. A fully connected (FC) layer is used as output layer. The model is trained by minimizing the mean-squared error (MSE) between the estimated and the clean log magnitude. The loss function is formulated as

$$L_{\mathrm{ME}} = \frac{1}{NF}\sum_{n=1}^{N}\sum_{f=1}^{F}|\hat{M}_m(n,f) - M_m(n,f)|^2, \tag{7}$$

with the clean target log magnitude

$$M_m(n,f) = \log_{10}|H_m^{\mathrm{d}}(f,\theta)S(n,f)|, \tag{8}$$

where $\hat{M}_m(n,f)$ is the predicted log magnitude of clean speech. Accounting for the following DP-RTF estimation, we consider the direct sound as the target signal, which means this magnitude enhancement network performs both noise reduction and dereverberation. It is important to note that the two microphones share the same magnitude enhancement network.

3.3. Binaural enhancement

To imitate the process of building DP-RTF representation, the intensity and phase of two channels are fed into two separate processes to learn the independent patterns related to DP-RTF, and then are concatenated before passed to a joint learning process. The enhanced log magnitude spectra of the two microphone channels are stacked along the microphone channel dimension, and then fed into a convolutional layer with 64 1×1 kernels to extract the inter-channel intensity features for each frequency and time frame. Similarly, the phases of the two microphone channels are stacked along the microphone channel dimension, and then a convolutional layer with 64 1×1 kernels is applied to extract the inter-channel phase features for each frequency and time frame. The phase-based features are closely related to IPD cues, which are further activated by sine and cosine functions to obtain a IPD feature in the format of DP-RTF representation.

The intensity-based and phase-based features produced by the convolutional layers are flattened and concatenated along the frequency and feature map dimensions. With multiple-frame features, we use uni-directional long short-term memory (LSTM) model to capture both long-range and short-range temporal context information and output the single-frame full-band DP-RTF $\hat{\mathbf{r}}$. Full bands are taken into account due to the mutual dependence of the localization features across frequencies. The LSTM model consists of two hidden layers and each layer contains 512 units. The output FC layer is activated by a tanh function to fit each element of the predicted feature into the range from -1 to 1.

One straightforward training target for binaural enhancement is the ground truth DP-RTF, and correspondingly the loss would be a normal regression loss, such as MSE. The MSE loss tends to minimize the distance between the predicted DP-RTF and its ground truth, which can be considered as an intra-class distance. To improve the localization robustness, we propose a weighted-matching (WM) softmax loss to not only minimize the intra-class distance, but also maximize the inter-class distance, i.e. the distance between the predicted DP-RTF and the true DP-RTF of other directions. The WM softmax loss consists of a dictionary atom-wise matching, a weighting scheme, a softmax function and a cross-entropy loss (see Fig. 1), which is formulated as

$$L_{\mathrm{BE}} = -\log\left(\frac{e^{-w(g)u(g)}}{\sum_{c=1}^{C} e^{-w(c)u(c)}}\right), \tag{9}$$

where $u(\cdot)$ represents the dictionary atom-wise matching, and $w(\cdot)$ is the weighting scheme.

Table 1. Room configuration for training and test data

Dataset	Room size [m^3]	Distance [m]	RT_{60} [s]	SNR [dB]
Training	7.0×8.0×5.0	1.50: 0.50: 3.00, 3.40	0: 0.17: 0.85	-5: 5: 20
	6.0×6.0×3.5	1.75, 2.25	0: 0.22: 0.88	-5: 5: 20
	4.0×5.5×3.0	0.50, 1.00	0: 0.25: 0.75	-5: 5: 20
	3.8×3.0×2.5	0.75, 1.25	0: 0.30: 0.90	-5: 5: 20
Test	6.0×8.0×3.8 (Large)	0.60, 1.50, 2.40, 3.30	0.2: 0.2: 0.8	5
			0.6	-5: 5: 15
	5.0×7.0×3.0 (Medium)	0.70, 1.40, 2.10	0.2: 0.2: 0.8	5
			0.6	-5: 5: 15
	4.0×4.0×2.7 (Small)	0.80, 1.30	0.2: 0.2: 0.8	5
			0.6	-5: 5: 15

Compared with the original softmax loss, we replace the fully connected layer in softmax loss with a WM strategy. For localization, the candidate localization space can be divided into C discrete directions. The ground truth DP-RTF associated with the cth candidate direction is denoted as $\mathbf{r}(\theta_c)$, and then the DP-RTF dictionary can be formed with ground-truth DP-RTF of the candidate directions $\mathbf{R} = [\mathbf{r}(1), \ldots, \mathbf{r}(C)]$. The dictionary atom-wise matching is defined as $u(c) = \|\hat{\mathbf{r}} - \mathbf{r}(\theta_c)\|^2$, where $||\cdot||$ denotes the Euclidean norm. The matching result can be seen as an indicator of the source presence possibility in the acoustic candidate space. Normally, the smaller of the distance between the predicted DP-RTF and the ground-truth, the closer of the predicted direction and the true direction of the source. However, the predicted DP-RTFs of the true direction and the closer directions tend to be confused during test, due to the mismatch of test and training conditions. To increase the tolerance of the network to this confusion, we weaken the value $u(c)$ when c is close to the true direction during training. The weighting scheme is defined as

$$w(c) = \begin{cases} 1, & c = g \\ \frac{|c-g|}{C-1}, & c \neq g \end{cases}, \tag{10}$$

where g is the index of ground-truth DOA. The true DOA is given a weight of 1. For other directions, the closer to the true direction, the smaller of the weight.

Finally, during test, with the estimated DP-RTF representation $\hat{\mathbf{r}}$, the DOA of the sound source is estimated by taking the direction that minimizes the dictionary matching result, i.e., $\hat{\theta} = \theta_{\arg\min_c \|\hat{\mathbf{r}} - \mathbf{r}(\theta_c)\|^2}$.

4. EXPERIMENTS AND DISCUSSIONS

4.1. Experimental setup

Seven different room configurations are simulated using the image method [19] which is implemented by the Roomsim toolbox [20]. The data generation configurations are summarized in Table 1, among which four room settings are used for training and three for test. All the experiments are carried out for binaural microphones with prominent shadow effect. The speech sound source is located in the same horizontal plane as the two microphones, and the candidate source directions range from -90° to 90° with an interval of 5°. The acoustic impulse response or binaural room impulse response (BRIR) is generated using the Roomsim toolbox and the head-related impulse response of the KEMAR dummy head [21]. We randomly select speech recordings from TIMIT dataset [22], and truncate each to obtain a speech segment with a duration of 0.5 s. These segments are divided into three parts to act as source signals for training, validation and test, respectively. The sensor signals are created by convolving the BRIRs with the source signals. We use the White, Babble and Factory noise files from the NOISEX-92 database [23] as noise signals. Each type of noise signal segments are split as training, validation and test sets, respectively, without overlap between sets. With these noise files, the arbitrary noise field generator

Table 2. Localization accuracy (5° Tol.) [%] of different methods under various rooms and noise type conditions.

Method	Noise	Room			Avg.
		Large	Medium	Small	
DOA-CNN [6]	White	70.76	64.36	62.49	70.00
	Babble	79.32	73.63	71.99	
	Factory	74.01	67.71	65.73	
IPD-EN [4]	White	66.68	58.28	55.64	69.49
	Babble	83.82	77.77	77.69	
	Factory	73.85	67.76	63.91	
Proposed	White	85.52	81.25	80.87	**86.97**
	Babble	93.49	90.80	90.82	
	Factory	88.86	85.80	85.31	

is employed to generate a diffuse noise field [24]. Diffuse noise is scaled and added to each sensor signal according to a given signal-to-noise ratio (SNR), in order to simulate acoustic conditions with various levels of noise. When generating each instance, the source signal, noise signal, RT_{60} and SNR are randomly given within the aforementioned settings. The numbers of instances for training, validation and test are 133200, 26640 and 159840, respectively.

The binaural signals used for localization are with a sampling rate of 16 kHz. They are enframed by a window of 32 ms with a frame shift of 16 ms. The frequency ranges from 0 to 4kHz is used for localization, and correspondingly the number of used frequencies F is 128. The maximum IID value $\Delta I_{\max}$ is set to 20. During training, we train the MEnet first, and then train the BEnet with the MEnet frozen. The model is trained using the Adam optimizer, with a learning rate of 0.0001. We evaluate the performance of DOA estimation using two types of localization accuracy: (i) 5° error tolerance (5° Tol.), considers a prediction is correct if the difference between the DOA estimate and the true DOA is not larger than 5°; (ii) 0° error tolerance (0° Tol.), considers a prediction is correct if it is exactly the true DOA, which is actually a classification accuracy.

4.2. Experimental results

We compare the proposed method with two state-of-the-art methods, which are referred to as DOA-CNN [6] and IPD-EN [4], respectively. The architecture of the DOA-CNN method is with one convolutional layer and three FC layers. The input vector is the phase of the STFT coefficient of single-frame sensor signals. This model outputs the posterior probability of each time frame, and the DOA is determined by taking the average of the posterior probability of multiple frames. The network used in IPD-EN contains four FC layers. It utilizes the sine and cosine of single-frame full-band IPD as localization feature. The DNN architecture maps the contaminated localization features to corresponding clean ones. The DOA estimation in [4] is designed for regular microphone array. To make it work on the binaural data, we modified it with the feature matching technique. The two methods are trained using the same data as our method. Table 2 shows the localization accuracy of all the three methods under various rooms and noise type conditions. Each test signal segment is with a duration of 0.5 s. It is observed that the DOA-CNN and IPD-EN methods achieve comparable performance. IPD-EN performs better for the Babble noise while DOA-CNN works better for the white noise. Compared with DOA-CNN and IPD-EN which localize sources with phase features only, the proposed method aims to pursue a robust localization method using the DP-RTF that encodes both phase and intensity information. As the proposed method takes full use of the monaural spectrogram and binaural difference information to remove the distortion caused by acoustic interferences, it outperforms DOA-CNN and IPD-EN for all the test conditions, which demonstrates the superiority of the proposed method.

To further evaluate the effectiveness of each component of the proposed method, ablation experiments are conducted in the medium room with different levels of noise and reverberation. Table 3 shows the localization results of monaural and binaural models performed solely and jointly. MEnet-only takes the MEnet output to compute IID, and uses the contaminated phase difference. BEnet-only takes contaminated intensities and phases of two microphone channels as network input. It can be seen that the combination of MEnet and BEnet achieves the best performance, which indicates that they are complementary and both contributing to the good performance. The MEnet-only method performs poorly, which means the noisy localization features are largely contaminated by noise and reverberation even with enhanced IID. The MEnet+BEnet method performs better than BEnet-only, especially under the conditions with high-level noise and reverberation. This improvement lies in that the MEnet is able to provide a less contaminated IID to the BEnet.

Table 3. Ablation study for DP-RTF learning (5° Tol.) [%].

Method	SNR [dB] (RT_{60} = 0.6 s)				RT_{60} [s] (SNR = 5 dB)			
	15	10	0	-5	0.2	0.4	0.6	0.8
MEnet-only	63.29	54.64	33.18	24.59	70.02	52.25	42.36	38.38
BEnet-only	94.92	91.76	75.11	58.35	96.43	91.22	85.51	81.80
MEnet + BEnet	94.31	92.03	78.93	63.77	96.40	91.64	87.27	83.26

Table 4. Comparison of different training loss functions.

Method	Tol.	SNR [dB] (RT_{60} = 0.6 s)				RT_{60} [s] (SNR = 5 dB)			
		15	10	0	-5	0.2	0.4	0.6	0.8
MSE	5°	93.38	90.17	74.28	58.20	95.63	89.80	84.44	81.07
	0°	71.32	65.68	46.55	32.58	78.26	66.40	57.96	53.09
Softmax	5°	91.32	88.56	74.62	59.86	95.45	89.43	83.32	79.55
	0°	71.32	66.46	47.25	35.86	80.26	67.46	59.10	55.00
WM softmax	5°	94.31	92.03	78.93	63.77	96.40	91.64	87.27	83.26
	0°	75.45	70.51	52.01	38.68	82.73	71.20	63.75	58.47

Table 4 shows the comparison of proposed training loss with the MSE loss and the softmax loss. When the proposed WM softmax loss is replaced with a regular softmax loss, the output of the proposed network will not be the DP-RTF feature anymore. Instead, the network will output the class of source direction, which is the same as many deep-classification-based sound source localization methods, such as [5, 25]. With 5° error tolerance, the MSE loss slightly outperforms the softmax loss in most cases, but the softmax loss performs relatively better with 0° error tolerance. The proposed loss achieves the best performance under all acoustic conditions. This indicates the proposed DOA estimation scheme, i.e. DP-RTF feature matching-weighted classification, is better than both the scheme of feature enhancement with MSE loss, and the scheme of classification-based DOA estimation. This superiority is brought by the fact that the proposed loss is able to not only cluster the same direction by minimizing the intra-class distance, but also better discriminate one direction and its neighbor directions by enlarging the inter-class distances.

5. CONCLUSION

This paper proposes to learn DP-RTF with DNN for binaural sound source localization under adverse acoustic conditions. Two complementary enhancement networks are employed, namely MEnet and BEnet. The combination of MEnet and BEnet makes full use of the single-channel spectrogram and dual-channel difference information to remove the distortion on the direct-path signals caused by noise and reverberation. In addition, a WM softmax training loss is used to minimize the intra-class distance and meanwhile enlarge the inter-class separation, which makes the predicted DP-RTF more suited for DOA estimation. Experiments with binaural microphones verify the robustness of our method for DOA estimation especially in scenarios with high level of noise and reverberation.

6. REFERENCES

[1] Ronen Talmon, Israel Cohen, and Sharon Gannot, "Supervised source localization using diffusion kernels," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 245–248.

[2] Hong Liu, Bing Yang, and Cheng Pang, "Multiple sound source localization based on TDOA clustering and multi-path matching pursuit," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 3241–3245.

[3] Duowei Tang, Maja Taseska, and Toon van Waterschoot, "Supervised contrastive embeddings for binaural source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 358–362.

[4] Junhyeong Pak and Jong Won Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 8, pp. 1335–1345, 2019.

[5] Ning Ma, Tobias May, and Guy J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 12, pp. 2444–2453, 2017.

[6] Soumitro Chakrabarty and Emanuel A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[7] Charles H. Knapp and G. Clifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[8] Wenyi Zhang and Bhaskar D. Rao, "A two microphone-based approach for source localization of multiple speech sources," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 18, no. 8, pp. 1913–1928, 2010.

[9] Sebastian Braun, Wei Zhou, and Emanuel A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015, pp. 1–5.

[10] Ziteng Wang, Junfeng Li, Yonghong Yan, and Emmanuel Vincent, "Semi-supervised learning with deep neural networks for relative transfer function inverse regression," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 191–195.

[11] Satish Mohan, Michael E. Lockwood, Michael L. Kramer, and Douglas L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2136–2147, 2008.

[12] Or Nadiri and Boaz Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 22, no. 10, pp. 1494–1505, 2014.

[13] Pasi Pertila and Emre Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 6125–6129.

[14] Zhong-Qiu Wang, Xueliang Zhang, and Deliang Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 1, pp. 178–188, 2019.

[15] Wolfgang Mack, Ullas Bharadwaj, Soumitro Chakrabarty, and Emanuel A. P. Habets, "Signal-aware broadband DOA estimation using attention mechanisms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4930–4934.

[16] Wangyou Zhang, Ying Zhou, and Yanmin Qian, "Robust DOA estimation based on convolutional neural network and time-frequency masking," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 2703–2707.

[17] Xiaofei Li, Laurent Girin, Radu Horaud, and Sharon Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 11, pp. 2171–2186, 2016.

[18] Deliang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 26, no. 10, pp. 1702–1726, 2018.

[19] Jont B. Allen and Daivid 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.

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

[21] William Grant Gardner and Keith D. Martin, "HRTF measurements of a KEMAR," *Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.

[22] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue, "TIMIT acoustic-phonetic continuous speech corpus," 1993.

[23] Andrew Varga and Herman 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 Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[24] Emanuel A. P. Habets, Israel Cohen, and Sharon Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, 2008.

[25] Paolo Vecchiotti, Ning Ma, Stefano Squartini, and Guy J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 451–455.