Received: 18 November 2020 | Accepted: 21 December 2020

DOI: 10.1049/cit2.12024

CAAI Transactions on Intelligence Technology

ORIGINAL RESEARCH PAPER

# Enhancing direct-path relative transfer function using deep neural network for robust sound source localization

Bing Yang[1,2] | Runwei Ding[1] | Yutong Ban[3,4] | Xiaofei Li[2] | Hong Liu[1]

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

[2]Westlake University & Westlake Institute for Advanced Study, Hangzhou, China

[3]SAIIL, Massachusetts General Hospital, Boston, Massachusetts, USA

[4]CSAIL, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA

**Correspondence**

Runwei Ding, Shenzhen Graduate School, Peking University, China.
Email: dingrunwei@pku.edu.cn

**Funding information**

National Natural Science Foundation of China, Grant/Award Numbers: No.61673030, U1613209; National Natural Science Foundation of Shenzhen, Grant/Award Number: No. JCYJ20200109140410340

**Abstract**

This article proposes a deep neural network (DNN)-based direct-path relative transfer function (DP-RTF) enhancement method for robust direction of arrival (DOA) estimation in noisy and reverberant environments. The DP-RTF refers to the ratio between the direct-path acoustic transfer functions of the two microphone channels. First, the complex-value DP-RTF is decomposed into the inter-channel intensity difference, and sinusoidal functions of the inter-channel phase difference in the time-frequency domain. Then, the decomposed DP-RTF features from a series of temporal context frames are utilized to train a DNN model, which maps the DP-RTF features contaminated by noise and reverberation to the clean ones, and meanwhile provides a time-frequency (TF) weight to indicate the reliability of the mapping. The DP-RTF enhancement network can help to enhance the DP-RTF against noise and reverberation. Finally, the DOA of a sound source can be estimated by integrating the weighted matching between the enhanced DP-RTF features and the DP-RTF templates. Experimental results on simulated data show the superiority of the proposed DP-RTF enhancement network for estimating the DOA of the sound source in the environments with various levels of noise and reverberation.

## 1 | INTRODUCTION

Sound source localization has a wide range of applications such as teleconferencing, robot audition, hearing aids, and so forth. With the development of deep learning techniques, lots of data-driven sound source localization works are built in a supervised manner [1]. According to the role of the deep learning model plays, these methods are classified into four categories, namely signal-to-location [2], feature-to-location [3,4], spatial spectrum-to-location [5], and feature-to-feature [6,7]-based methods. Among these methods, the feature-to-feature-based method is simple and effective for improving the performance of sound source localization in noisy and reverberant environments, as it is the data driven and the extracted features can adapt to various acoustic conditions.

The spatial features utilized for localization include the time and the intensity differences between dual-microphone signals. Inter-channel time difference (ITD) is commonly estimated by searching the maximum of the generalized cross-correlation (GCC) function [8]. Inter-channel phase difference (IPD) is another time difference feature and owns an approximate linear property with respect to frequency [9]. Moreover, inter-channel intensity difference (IID) is computed as the energy ratio of the signals captured by two microphones. Relative transfer function (RTF) [10,11] encodes time and intensity information in its argument and magnitude respectively, which is the ratio between the acoustic transfer functions of the two channels. Other high-level localization features include the cross-correlation function (CCF) [3], the eigen vectors of spatial correlation matrix associated with signal subspace [12], and so forth. Overall, the sound source can be easily localized with the 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 localization feature estimation, which further leads to a significant drop on the localization performance.

Many methods aim to remove the effect of acoustic interferences on the direct-path localization feature extraction.

Some research works give high attention values to the direct sound dominant time-frequency (TF) regions using a TF weighting scheme, which can be classified into unsupervised methods [13–16] and supervised methods [17,18]. However, these methods do not refine the value of localization features. Though the IPD enhancement method [6] has been used to fine-tune the localization features using deep neural network (DNN), it only considers the time difference information of current time frame, but intensity difference and temporal context information are also important for localization. We aim to investigate how to make full use of the time and intensity difference information of both the historical and current time frames, in order to recover the clean localization features from the existing contaminated ones, so that the sound source can be robustly localized.

This article designs a direct-path relative transfer function (DP-RTF) enhancement network to preserve the time and intensity difference information of direct-path signal and suppress the contamination of noise and reverberation. The complex DP-RTF is decomposed as the IID and the sinusoidal functions of the IPD, in order to fit the real-value DNN framework and to explicitly present the localization cues. Then the DP-RTF is enhanced using a DNN which nonlinearly maps the contaminated DP-RTF feature of multiple temporal context frames to one single-frame clean feature. The DP-RTF enhancement network can jointly predict the clean DP-RTF and the TF reliability weight, by adopting a weighted mean square error (MSE) with a weight-maximization regularization term. The trained DP-RTF enhancement network can significantly depress the effect of noise and reverberation on the DP-RTF estimation. For each TF bin, the enhanced DP-RTF are matched with the template of candidate directions. The direction of arrival (DOA) of the sound source is determined by integrating the matching functions united with the predicted TF weight from multiple TF bins. Experiments using simulated data demonstrate the effectiveness of our method under noisy and reverberate acoustic conditions. The main contributions of this paper are summarized as follows:

(1) We design a DP-RTF enhancement network to recover the DP-RTF features from the contaminated localization features for robust sound source localization. Different from the method in [6] which recovers the frame-wise time difference information using current time frame, the proposed model considers the short-term temporal context information, and jointly recovers both time and intensity difference information.

(2) We use a weighted MSE with a weight-maximization regularization term to guarantee that the TF-wise DP-RTF features are selectively recovered. Using the predicted TF weight, the enhancement network is more concentrated on the features with high weights, therefore high-weights' features will dominate the DOA estimation.

The rest of this paper is organized as follows. Section 2 introduces the DP-RTF based DOA estimation method. Section 3 details the proposed DNN based DP-RTF enhancement network. Experiments and discussions with simulated data are presented in Section 4, and conclusions are drawn in Section 5.

## 2 | DP-RTF BASED DOA ESTIMATION

In an enclosed environment with additive ambient noise, a single source is observed by a pair of microphones. The signal received by the $m$th microphone is formulated as

$$x_m(t) = h_m(t)^* s(t) + v_m(t), \tag{1}$$

where $m \in [1,2]$ is the microphone index, $s(t)$ denotes the source signal, $v_m(t)$ denotes the received noise signal at the $m$th microphone, and $h_m(t)$ is the acoustic impulse response (AIR) from the source to the $m$th microphone. Here, $*$ denotes the convolution operation. Applying the short-time Fourier transform (STFT) to Equation (1), the signal at the $m$th microphone can be rewritten in the TF domain as

$$X_m(n,f) = H_m(f,\theta)S(n,f) + V_m(n,f), \tag{2}$$

where $n \in [1, N]$ denotes the index of time frame, $f \in [1, F]$ denotes the index of frequency, and $\theta$ denotes the horizontal DOA of source. $X_m(n,f)$, $S(n,f)$ and $V_m(n,f)$ represent the STFT coefficients of $x_m(t)$, $s(t)$, and $v_m(t)$, respectively. The acoustic transfer function (ATF) $H_m(f, \theta)$ is the Fourier transform of $h_m(t)$. The ATF contains the direct and reflected propagation paths of sound source, that is,

$$H_m(f,\theta) = H_m^{\mathrm{d}}(f,\theta) + H_m^{\mathrm{r}}(f,\theta), \tag{3}$$

where $H_m^{\mathrm{d}}(f,\theta)$ and $H_m^{\mathrm{r}}(f,\theta)$ denote the ATFs of direct-path and reflected propagations, respectively. The DP-RTF [19] is defined as the ratio between the two direct-path ATFs, namely

$$R^{\mathrm{d}}(f,\theta) = \frac{H_2^{\mathrm{d}}(f,\theta)}{H_1^{\mathrm{d}}(f,\theta)}. \tag{4}$$

Under the anechoic and noise-free condition, $H_m^{\mathrm{r}}(f,\theta)$ and $V_m(n,f)$ are equal to zero. According to Equations (2) and (3), the microphone signal is simplified as

$$X_m(n,f) = H_m^{\mathrm{d}}(f,\theta)S(n,f). \tag{5}$$

Using this simplification, the DP-RTF can be estimated by

$$\widehat{R}^{\mathrm{d}}(n,f) = \frac{X_2(n,f)}{X_1(n,f)}. \tag{6}$$

The estimated DP-RTF is decomposed and rewritten in a vector form, namely

$$\widehat{\mathbf{r}}(n,f) = \left[\frac{20\log_{10}\left|\widehat{R}^{\mathrm{d}}(n,f)\right|}{\Delta I_{\max}}, \sin\angle\widehat{R}^{\mathrm{d}}(n,f), \cos\angle\widehat{R}^{\mathrm{d}}(n,f)\right]^{T}. \tag{7}$$

The DOA of the sound source is estimated by matching the estimated DP-RTF vectors from reliable TF bins with the template, namely

$$\widehat{\theta} = \underset{\theta\in\mathbb{S}}{\operatorname{argmin}} \sum_{n=1}^{N}\sum_{f=1}^{F} \left\|\widehat{w}(n,f)\left(\widehat{\mathbf{r}}(n,f) - \mathbf{r}(f,\theta)\right)\right\|^2, \tag{8}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\mathbb{S}$ is the set of candidate directions, and $\widehat{w}(n,f)$ denotes the TF weight that indicates the reliability of $\widehat{\mathbf{r}}(n,f)$. Here, the ground-truth DP-RTF vector $\mathbf{r}(f,\theta)$ is defined in Equation (13) (see Section 3.1), and $\mathbf{r}(f,\theta)$ of all candidate directions are used as the matching template.

## 3 | DP-RTF ENHANCEMENT NETWORK

The above-mentioned framework provides a solution to the single sound source localization on the anechoic and noise-free assumption (see Figure 1a). However, noise and reverberation are inevitable in real-world scenarios. The DP-RTF estimation using Equation (6) is a biased approximation under noisy and reverberant conditions, which can introduce an obvious deviation of DOA estimation. Hence, we add a DP-RTF enhancement network to the above-mentioned framework, in order to recover the clean DP-RTF feature from the contaminated ones to suit the reverberant and noisy environments (see Figure 1b).

### 3.1 | Input and target

As the complex DP-RTF estimates cannot be directly processed by the real-value DNN, to fit the real-value DNN, the input and target complex DP-RTF values are transformed into a real-value form.

The theoretical direct-path ATF, namely the head-related transfer function (HRTF) in the binaural localization case, is formulated as

$$H_m^{\mathrm{d}}(f,\theta) = \alpha_m(f,\theta)e^{-j\omega_f\tau_{m(\theta)}}, \tag{9}$$

where $\omega_f$ is the angular frequency of the $f$th frequency. $\alpha_m(f)$ and $\tau_{m(\theta)}$ denote the propagation attenuation factor and the time of arrival from the source to the $m$th microphone, respectively. Substituting Equation (9) into Equation (4), the DP-RTF is rewritten as

$$R^{\mathrm{d}}(f,\theta) = \frac{\alpha_2(f,\theta)}{\alpha_1(f,\theta)}e^{-j\omega_f(\tau_2(\theta)-\tau_1(\theta))}. \tag{10}$$

It can be seen that the DP-RTF encodes IID and IPD information in its magnitude and argument respectively. We extract these two localization cues using a disjoint decomposition, formally written as

$$\Delta I(f,\theta) = 20\log_{10}\left|R^{\mathrm{d}}(f,\theta)\right|, \tag{11}$$

$$\Delta P(f,\theta) = \angle R^{\mathrm{d}}(f,\theta), \tag{12}$$

where $\angle$ is the phase operator of complex numbers. With these two real-value localization cues, the complex DP-RTF value can be recovered.

However, IPD is presented in the range of $[-\pi, \pi]$, and may be periodically wrapped with the increasing of frequency or time difference. Hence, the mean squared IPD error cannot be directly used to reflect the DOA difference, and the DOA estimation using such IPD will fail as well due to the phase wrapping ambiguity. To avoid this ambiguity, the sinusoidal

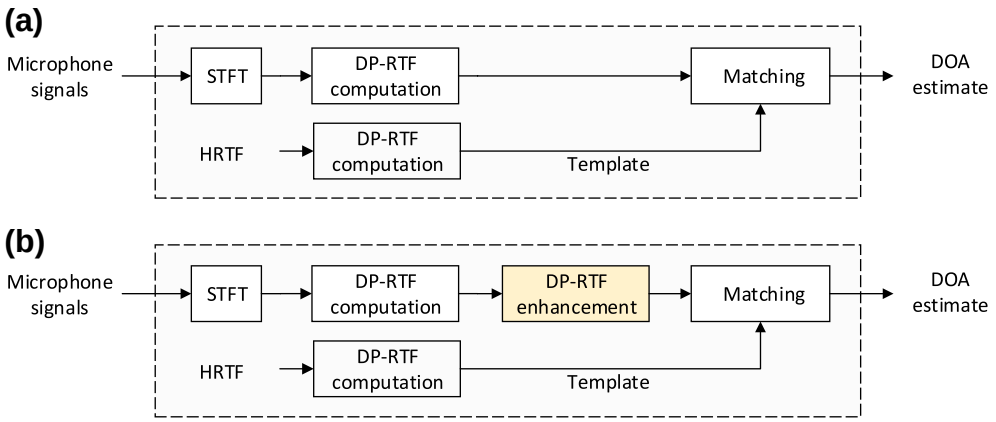


**FIGURE 1** Pipeline for the DP-RTF based DOA estimation method. (a) Without DP-RTF enhancement network. (b) With DP-RTF enhancement network

functions of IPD is used instead. Accordingly, the complex DP-RTF is decomposed into IID, the sine and cosine functions of IPD. The three decomposed parts are concatenated to form the DP-RTF vector ground truth associated with the direction $\theta$, that is,

$$\mathbf{r}(f,\theta)=\left[\frac{\Delta I(f,\theta)}{\Delta I_{\max}},\sin\Delta P(f,\theta),\cos\Delta P(f,\theta)\right]^{T}, \quad (13)$$

where $(\cdot)^T$ denotes vector transpose, and $\Delta I_{\max}$ is an empirically set maximum value of IID used for normalization. As the original IID has a relatively wider range than the sinusoidal functions, the IID is scaled to [−1,1] to balance the contribution of the IID and IPD information. The dimension of $\mathbf{r}(f,\theta)$ is $3\times 1$, and each element is in the range from −1 to 1. By using this transformation, we didn't lose any information of the sound source location. Therefore, the DP-RTF feature can be recovered according to the IID and IPD information contained in RTF vector.

Accordingly, for frame $n$, the input vector is a concatenation of the contaminated DP-RTF vectors from a series of (context) time frames and frequency bands, namely

$$\begin{aligned}\mathbf{C}_{\mathrm{I}}(n)=\quad & [\widehat{\mathbf{r}}(n-C+1,1)^{T},\ldots,\widehat{\mathbf{r}}(n-C+1,F)^{T},\\ & \ldots\\ & \widehat{\mathbf{r}}(n,1)^{T},\ldots,\widehat{\mathbf{r}}(n,F)^{T}]^{T},\end{aligned} \quad (14)$$

and the learning target vector is a concatenation of the clean DP-RTF vectors from multiple frequency bands, namely

$$\mathbf{C}_{\mathrm{T}}^{\mathrm{RTF}}(n)=[\ \mathbf{r}(1,\theta)^{T},\ldots,\mathbf{r}(F,\theta)^{T}\ ]^{T}, \quad (15)$$

where $C$ denotes the number of context time frames and $F$ refers to the number of utilized frequencies. The dimensions of $\mathbf{C}_{\mathrm{I}}(n)$ and $\mathbf{C}_{\mathrm{T}}^{\mathrm{RTF}}(n)$ are $3CF\times 1$ and $3F\times 1$, respectively. By concatenating the DP-RTF feature vectors of $CF$ TF bins into a long input feature vector, the acoustic context information along the time and frequency axes can be captured. Full bands are taken into account due to the mutual dependencies of localization cues on different frequencies.

## 3.2 | Network architecture

Considering the complex noise and reverberation generating and mixing process, the mapping from contaminated DP-RTFs to clean DP-RTFs is indeed a complicated non-linear operation. Since neural networks own many levels of non-linearity, we design a DNN model to approximate this highly non-linear relationship. The architecture of the designed DNN model for DP-RTF enhancement is illustrated in Figure 2. The DNN model employs $C$-frame contaminated DP-RTF vector to predict one-frame clean DP-RTF vector and TF reliability weight. The input vector $\mathbf{C}_{\mathrm{I}}(n)$ is first fed into three fully connected (FC) layers to obtain the latent features. Each of the FC layers is with 2048 units and activated by a rectified linear unit (ReLU). Then the latent features are pass to two FC layers, respectively. Accordingly, the output contains two parts, one for DP-RTF vector and the other for TF weight, which are activated by a tan h unit and a sigmoid unit, respectively. The output for DP-RTF vector is with a dimension of $3F\times 1$, namely

$$\mathbf{C}_{\mathrm{O}}^{\mathrm{RTF}}(n)=[\ \widetilde{\mathbf{r}}(n,1)^{T},\ldots,\widetilde{\mathbf{r}}(n,F)^{T}\ ]^{T}. \quad (16)$$

The output for TF weight is with a dimension of $F\times 1$, namely

$$\mathbf{C}_{\mathrm{O}}^{\mathrm{W}}(n)=[\ \widetilde{w}(n,1),\ldots,\widetilde{w}(n,F)\ ]^{T}, \quad (17)$$

where $\widetilde{\mathbf{r}}(n,f)$ and $\widetilde{w}(n,f)$ are DP-RTF vector and TF weight predicted by DNN, respectively.

For network training, one commonly used loss function is the MSE between the output and the target. This loss function treats all frequencies equivalently, which is not suitable for the present DP-RTF enhancement problem. The TF bins where source signal is silent or greatly contaminated by noise or reverberation provide unreliable localization cues, and hence should be disregarded in an indirect manner. To tackle this problem, we add an adaptive weighting scheme to the loss function, and thus the enhanced DP-RTF vector and the TF reliability weight can be jointly learned. The loss function is then defined as

$$\begin{aligned}L(n)=\quad & \frac{1}{3F}\sum_{f=1}^{F}\|\widetilde{w}(n,f)(\widetilde{\mathbf{r}}(n,f)-\mathbf{r}(f,\theta))\|^{2}+\\ & \lambda\frac{1}{F}\sum_{f=1}^{F}|1-\widetilde{w}(n,f)|,\end{aligned} \quad (18)$$

where $\mathbf{r}(f,\theta)$ is the ground-truth DP-RTF vector which is free of the affection of acoustic interferences, and hence it is utilized as the training target of $\widetilde{\mathbf{r}}(f,\theta)$. Here, $\lambda$ denotes the regularization factor. The regularization part is set to avoid trivial solutions, that is zero weights, and to guarantee a sufficient number of TF bins to be employed for training. Using this weighted MSE, the DP-RTF learning disregards those TF bins that cannot provide accurate DP-RTF estimates.

In the test stage, the trained DNN predicts a single-frame full-band estimate of $\widetilde{\mathbf{r}}(n,f)$ and $\widetilde{w}(n,f)$. The DOA of the sound source is estimated using the enhanced DP-RTFs and TF weights of $N$ time frames (and $F$ frequency bands for each frame), according to Equation (8).

# 4 | EXPERIMENTS AND DISCUSSIONS

In this section, the performance of the proposed method is measured. We first give the details of the experimental setup, and then show the experimental results and discussions.

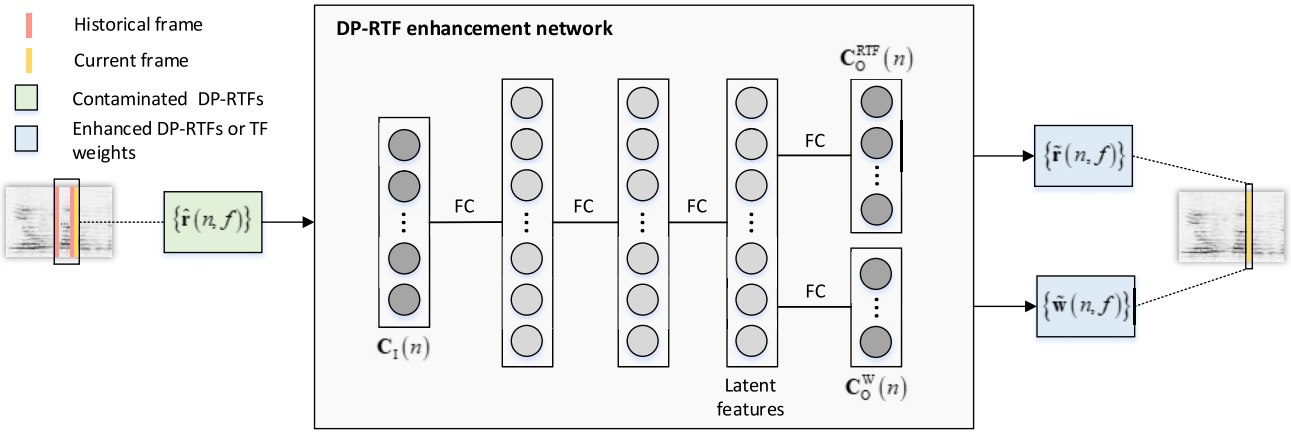


**FIGURE 2** Architecture for the DP-RTF enhancement network. The input is the DP-RTF vector contaminated by noise and reverberation, and the output is the enhanced DP-RTF vector and the TF weight

## 4.1 | Experimental setup

### 4.1.1 | Simulated data

We simulate different room configurations using the image method [20] which is implemented by the Roomsim toolbox [21]. Four acoustic configurations are generated for training and three for test, as shown in Table 1. All experiments are carried out with binaural microphones whose shadow effect is significant. As illustrated in Figure 3, the speech sound source is located in same horizontal plane as the binaural microphones, and the candidate source directions ranges from −90° to 90° with an interval of 5°. The binaural room impulse response (BRIR) is generated using the Roomsim toolbox [21] and the head-related impulse response of the KEMAR dummy head [22]. Speech recordings from TIMIT dataset[1] are truncated and form speech segments with 0.5 s duration. These segments are used as source signals, which are divided into training, validation and test sets, respectively. White, Babble and Factory noise files from the NOISEX-92 database [24] are employed as noise signals. Each type of noise signal segments is divided into training, validation and test sets, respectively. Diffuse noise field [23] is generated using these noise segments. The sensor signals are created by first convolving the source signals with the BRIRs, and then adding the scaled diffuse noise to the convoluted signals according to a given signal-to-noise ratio (SNR).

### 4.1.2 | Parameter settings and evaluation metrics

The sampling rate of the binaural signals used for localization is 16 kHz. The binaural signals are enframed by a window of 32 ms with a frame shift of 16 ms. The frequency ranges from 0 to 4 kHz is used for localization ($F = 128$). The maximum value of IID $\Delta I_{\max}$ is set to 20. The DNN model is trained using the Adam optimizer. The learning rate is set to 0.001. The accuracy of DOA estimation is accessed by the mean absolute error (MAE) and the localization accuracy. The MAE is defined as the average error between the estimated and the ground-truth DOAs over different test instances. The localization accuracy considers a prediction to be correct if the difference between the DOA estimate and the true DOA is less than or equal to 5°.

[1] https://catalog.ldc.upenn.edu/ldc93s1

## 4.2 | Experimental result

### 4.2.1 | Influence of the IID information

To investigate the influence of exploiting IID, we compare the MAE of using only the enhanced IPD and using both the enhanced IID and IPD under different sizes of rooms. The experiments are carried in rooms with different levels of reverberation and noise. The $RT_{60}$ is 0.2, 0.4, 0.6, 0.8 s with SNR being set to 5 dB. The SNR is −5, 0, 10, 15 dB with $RT_{60}$ being set to 0.6 s. The experiment results present in Table 2 are an average of these acoustic conditions. It can be seen that the IPD + IID method performs better than the IPD method in Room 5 and Room 6 while slightly worse in Room 7. Overall, with the IID information, the MAE of DOA estimation can be reduced, which demonstrates the effectiveness of incorporating the IID information in the present framework for binaural sound source localization.

### 4.2.2 | Influence of the temporal context information

We set the number of context time frames $C$ to different values and give the MAE of DOA estimation under different sizes of rooms in Table 3. The experiment data is the same as that used in Section 4.2.1. It can be seen that the MAE of DOA estimation decreases when $C$ varies from 1 to 7 and increases when $C$ is larger than 7. Hence, $C$ is set to 7 in the following experiments, as it is found to be the optimal value under different acoustic conditions.

**TABLE 1** Room configuration for training and test data

| Dataset | Training | | | | Test | | |
|---|---|---|---|---|---|---|---|
| **Room Label** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| Room size ($m^3$) | 7.0 × 8.0 × 5.0 | 6.0 × 6.0 × 3.5 | 4.0 × 5.5 × 3.0 | 3.8 × 3.0 × 2.5 | 6.0 × 8.0 × 3.8 | 5.0 × 7.0 × 3.0 | 4.0 × 4.0 × 2.7 |
| Array centre (m) | (3.00, 3.50, 1.70) | (3.50, 3.00, 1.65) | (2.50, 2.50, 1.40) | (1.20, 1.45, 1.55) | (2.00, 4.00, 1.65) | (2.50, 3.00, 1.50) | (1.80, 1.70, 1.60) |
| Distance (m) | 1.50: 0.50: 3.00, 3.40 | 1.75, 2.25 | 0.50, 1.00 | 0.75, 1.25 | 0.60: 0.90: 3.30 | 0.70, 1.40, 2.10 | 0.80, 1.30 |
| $RT_{60}$ (s) | 0: 0.17: 0.85 | 0: 0.22: 0.88 | 0: 0.25: 0.75 | 0: 0.3: 0.9 | 0.2: 0.2: 0.8 | 0.2: 0.2: 0.8 | 0.2: 0.2: 0.8 |
| SNR (dB) | −5: 5: 20 | −5: 5: 20 | −5: 5: 20 | −5: 5: 20 | −5: 5: 20 | −5: 5: 20 | −5: 5: 20 |

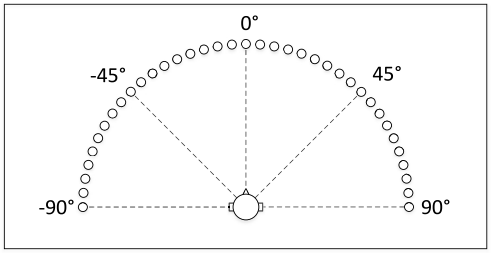


**FIGURE 3** Illustration for the candidate directions of sound sources

**TABLE 2** Influence of the IID information under different rooms (C = 1)

| | MAE (degrees) | | | |
|---|---|---|---|---|
| **Method** | **Room 5** | **Room 6** | **Room 7** | **AVG.** |
| IPD [6] | 14.28 | 18.47 | **20.23** | 17.66 |
| **DP-RTF(IPD + IID)** | **13.93** | **17.97** | 20.64 | **17.51** |

**TABLE 3** Influence of the temporal context information under different rooms

| | MAE (degrees) | | | |
|---|---|---|---|---|
| **Temporal Context** | **Room 5** | **Room 6** | **Room 7** | **AVG.** |
| C = 1 | 13.93 | 17.97 | 20.64 | 17.51 |
| C = 3 | 9.18 | 11.99 | 13.45 | 11.54 |
| C = 5 | 8.43 | 10.78 | 11.87 | 10.36 |
| ***C* = 7** | **7.35** | **9.40** | **10.38** | **9.04** |
| C = 9 | 7.66 | 9.95 | 11.16 | 9.59 |
| C = 11 | 7.64 | 9.88 | 11.20 | 9.57 |
| C = 13 | 7.86 | 10.02 | 10.64 | 9.51 |
| C = 15 | 8.42 | 10.64 | 11.99 | 10.35 |

### 4.2.3 | Influence of the TF weighting scheme

The DOA estimation results without and with the TF weighting scheme are present in Table 4. The experiment data is the same as that used in Section 4.2.1. For the DOA estimation without TF weighting scheme, the TF weight $\widehat{w}(n,f)$ is set to one. For the loss with weighting scheme, the regularization factor $\lambda$ is set to 0.01 and 0.5, respectively. The result show that compared with the DOA estimation without the TF weighting scheme, using this weight achieves a smaller MAE, which verifies the effectiveness of the weighting scheme. Besides, with a larger regularization factor, the MAE of DOA estimation become higher. This is because that a stronger punishment of regularization term can result in that the TF weight is closer to one, which will achieve a similar performance to that without TF weighting scheme. In the following experiments, $\lambda$ is set to 0.01.

**TABLE 4** Influence of the TF weighting scheme under different rooms

| | MAE (degrees) | | | |
|---|---|---|---|---|
| **Method** | **Room 5** | **Room 6** | **Room 7** | **AVG.** |
| w/o weighting | 7.35 | 9.40 | 10.38 | 9.04 |
| **w / weighting ($\lambda$ = 0.01)** | **7.05** | **8.98** | **9.99** | **8.67** |
| w/ weighting ($\lambda$ = 0.5) | 7.04 | 9.01 | 10.2 | 8.75 |

### 4.2.4 | Robustness evaluation

To illustrate the effectiveness of the proposed DP-RTF enhancement network, we estimate the DP-RTF without and with the enhancement network, respectively, and plot the phase and amplitude of the DP-RTF estimate as a function of frequency bins in Figure 4. Each presented phase or amplitude corresponds one DP-RTF estimate in a certain TF bin. We use 31-frame DP-RTF estimates for evaluation. For the method without the enhancement network, the TF weight is set to one. It can be observed that the phase and amplitude of DP-RTF estimated without the enhancement network is scattered, while that provided with the enhancement network is clustered around the ground-truth lines. The proposed enhancement network provides more accurate DP-RTF estimates, which show the ability to preserve the direct-path localization cues and meanwhile reduce the effect of noise and reverberation.

The proposed method is compared with other three methods, IPD-NN [6], RTF and RTF-CT [13]. The IPD-NN method uses a four-layer DNN to map the contaminated IPD features to the corresponding clean ones. The RTF method

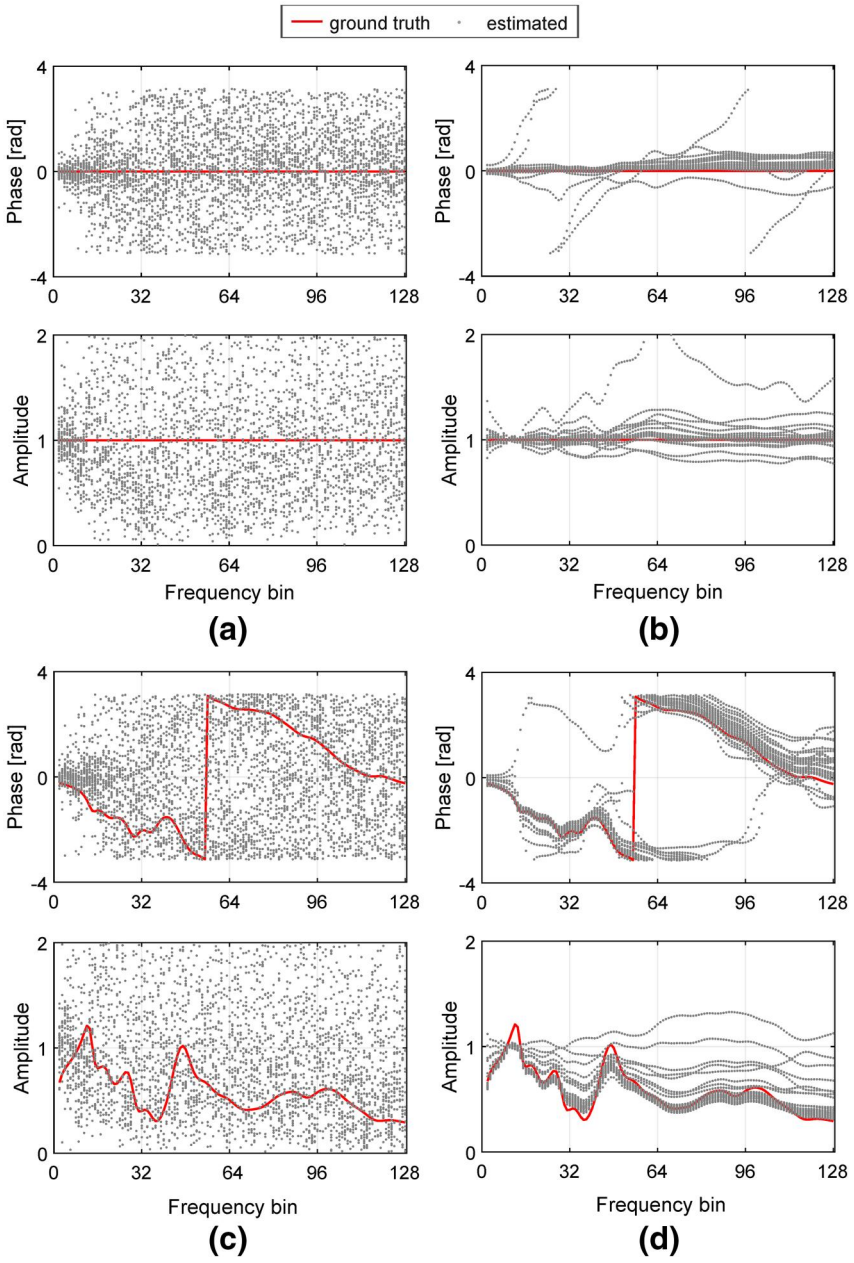


**FIGURE 4** phase and amplitude of the DP-RTF as a function of frequency bins under a typical acoustic condition that $RT_{60}$ = 600 ms and SNR = 5 dB (babble noise) in Room 6. The DP-RTF is estimated without DP-RTF enhancement network in (a) (c), and with DP-RTF enhancement network in (b) (d). The sound source is located at 0° in (a) (b), and 30° in (c) (d). The distance from the sound source to the centre of the microphone array is 2.1 m

means directly using the contaminated DP-RTF features for DOA estimation (namely the pipeline in Figure 1a). It uses a TF weight equalling to one. The RTF-CT method also follows the pipeline in Figure 1a but sets the TF weight by coherence test which is used to select direct path dominated TF bins. For fair comparison, all the comparison methods estimate the DOA by finding the optimal matching between the enhanced feature and the template features of all candidate directions, following the principle of the proposed method. The comparison between these methods is carried out in the environments with different levels of noise and reverberation. Tables 5 and 6 respectively show the MAEs of the four methods under Room 6 with various SNR and $RT_{60}$ conditions. Each test signal segment used for DOA estimation is with a duration of 0.5 s. It can be seen that the proposed method and IPD-NN outperform RTF and RTF-CT in all cases, which demonstrates the superiority of DNN based methods for enhancing the localization feature. The proposed method performs better than the IPD-NN method. This is due to that the proposed method incorporates the IID and temporal context information to feature estimation, which is helpful for improving the robustness of DOA estimation. Compared with the RTF method, both RTF-CT and our method add the DP-RTF enhancement process, but our method achieves a lower MAE than the RTF-CT. It can conclude that employing all data to enhance the DP-RTF is more beneficial than only employing the data selected by the coherence test. Besides, the RTF-CT method applies a hard selection on TF bins, while the proposed method applies a better TF weighting scheme, that is a soft weight.

**TABLE 5** Performance of different methods under various noise conditions (Room 6, $RT_{60}$ = 0.6 s)

| | MAE (degrees) | | | | | ACC (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | **15 dB** | **10 dB** | **0 dB** | **−5 dB** | **AVG.** | **15 dB** | **10 dB** | **0 dB** | **−5 dB** | **AVG.** |
| IPD-NN [6] | 6.61 | 10.61 | 29.27 | 45.61 | 23.03 | 85.08 | 78.35 | 50.93 | 33.05 | 61.85 |
| RTF | 23.69 | 28.49 | 38.42 | 41.14 | 32.94 | 51.17 | 42.73 | 25.96 | 20.63 | 35.12 |
| RTF-CT [13] | 20.96 | 26.30 | 37.12 | 40.32 | 31.18 | 57.70 | 48.02 | 28.14 | 21.88 | 38.94 |
| Proposed ($C$ = 1, w/o weighting) | 5.78 | 9.78 | 28.90 | 43.70 | 22.04 | 87.67 | 79.37 | 50.48 | 32.99 | 62.63 |
| Proposed ($C$ = 7, w/o weighting) | 3.47 | 4.98 | 14.33 | 25.07 | 11.96 | 91.85 | 88.30 | 70.75 | 53.59 | 76.12 |
| **Proposed** ($C$ = 7, w/ weighting) | **3.38** | **4.72** | **13.68** | **23.82** | **11.40** | **92.15** | **88.99** | **71.43** | **54.32** | **76.72** |

**TABLE 6** Performance of different methods under various reverberation conditions (Room 6, SNR = 5 dB)

| | MAE (degrees) | | | | | ACC (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | **0.2 s** | **0.4 s** | **0.6 s** | **0.8 s** | **AVG.** | **0.2 s** | **0.4 s** | **0.6 s** | **0.8 s** | **AVG.** |
| IPD-NN [6] | 4.79 | 11.43 | 17.64 | 21.79 | 13.91 | 89.73 | 76.58 | 66.44 | 60.68 | 73.36 |
| RTF | 22.91 | 30.87 | 34.30 | 35.66 | 30.94 | 55.21 | 40.03 | 33.36 | 30.26 | 39.72 |
| RTF-CT [13] | 20.20 | 28.98 | 32.35 | 34.02 | 28.89 | 60.42 | 44.05 | 36.71 | 33.42 | 43.65 |
| Proposed ($C$ = 1, w/o weighting) | 6.00 | 11.73 | 17.23 | 20.63 | 13.89 | 89.13 | 77.84 | 67.61 | 61.23 | 73.95 |
| Proposed ($C$ = 7, w/o weighting) | 2.85 | 5.92 | 8.78 | 9.82 | 6.84 | 95.03 | 87.54 | 81.55 | 78.20 | 85.58 |
| **Proposed** ($C$ = 7, w/ weighting) | **2.86** | **5.76** | **8.25** | **9.36** | **6.56** | **94.89** | **88.47** | **82.34** | **78.81** | **86.13** |

## 5 | CONCLUSION

This article proposes a DP-RTF enhancement network for sound source localization under adverse acoustic conditions. Considering the complex non-linear process of noise and reverberation generating and mixing, we utilize a DNN to model the non-linear regression that discriminates the clean DP-RTFs from the contaminated ones. For training, a novel loss function composed of a weighted MSE loss for DP-RTF and a TF-weight regularization term are proposed to account for the fact that only parts of TF bins contain reliable localization information due to the TF sparsity of the (speech) source signal. Experiments with binaural microphones verify the robustness of our method for DOA estimation especially in scenarios with high level of noise and reverberation. In this work, we focus on the concept of DP-RTF enhancement and TF weight estimation, and a generic DNN model is adopted, which can further be revised with a more advanced network structure, such as the recurrent neural network, as a future work.

## ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (No. 61673030, U1613209), Science and Technology Plan Project of Shenzhen (No. JCYJ20200109 140410340).

## ORCID

*Bing Yang* https://orcid.org/0000-0002-8978-2322

## REFERENCES

1. Talmon, R., Cohen, I., Gannot, S.: Supervised source localization using diffusion kernels. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 245–248. (2011)
2. Vecchiotti, P. et al.: End-to-end binaural sound localisation from the raw waveform. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 451–455. (2019)
3. 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)
4. Chakrabarty, S., Habets, E.A.P.: Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. IEEE J. Sel. Top. Sig. Process. 13(1), 8–21 (2019)
5. Nguyen, T.N.T. et al.: Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network. IEEE Trans. Audio Speech Lang. Process. 28, 2626–2637 (2020)
6. 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)
7. Tang, D., Taseska, M., van Waterschoot, T.: Supervised contrastive embeddings for binaural source localization. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 358–362. (2019)
8. Knapp, C.H., Carter, G.C.: The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. 24(4), 320–327 (1976)
9. Zhang, W., Rao, B.D.: A two microphone-based approach for source localization of multiple speech sources. IEEE Trans. Audio Speech Lang. Process. 18(8), 1913–1928 (2010)
10. Braun, S., Zhou, W., Habets, E.A.P.: 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), pp. 1–5. (2015)

11. Wang, Z. et al.: Semi-supervised learning with deep neural networks for relative transfer function inverse regression. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 191–195. (2018)
12. Yang, B., Liu, H., Pang, C.: Multiple sound source counting and localization based on spatial principal eigenvector. In: Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 1924–1928. (2017)
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. Pavlidi, D. et al.: Real-time multiple sound source localization and counting using a circular microphone array. IEEE Trans. Audio Speech Lang. Process. 21(10), 2193–2206 (2013)
15. Nadiri, O., Rafaely, B.: Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test. IEEE Trans. Audio Speech Lang. Process. 22(10), 1494–1505 (2014)
16. Madmoni, L., Rafaely, B.: Direction of arrival estimation for reverberant speech based on enhanced decomposition of the direct sound. IEEE Trans. Audio Speech Lang. Process. 13(1), 131–142 (2019)
17. Pertila, P., Cakir, E.: Robust direction estimation with convolutional neural networks based steered response power. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6125–6129. (2017)
18. Wang, Z.Q., 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)
19. Li, X. et al.: Estimation of the direct-path relative transfer function for supervised sound-source localization. IEEE Trans. Audio Speech Lang. Process. 24(11), 2171–2186 (2016)
20. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. 65(4), 943–950 (1979)
21. Campbell, D.R., Palomaki, K.J., Brown, G.J.: A MATLAB simulation of shoebox room acoustics for use in research and teaching. Comput. Inform. Syst. J. 9(3), 48–51 (2005)
22. Gardner, W.G., Martin, K.D.: HRTF measurements of a KEMAR. J. Acoust. Soc. Am. 97(6), 3907–3908 (1995)
23. Habets, E.A.P., Gannot, S.: Generating sensor signals in isotropic noise fields. J. Acoust. Soc. Am. 122(6), 3464–3470 (2007)


**How to cite this article:** Yang B, Ding R, Ban Y, Li X, Liu H. Enhancing direct-path relative transfer function using deep neural network for robust sound source localization. *CAAI Trans. Intell. Technol.* 2021;1–9. https://doi.org/10.1049/cit2.12024