SRP-DNN: LEARNING DIRECT-PATH PHASE DIFFERENCE FOR MULTIPLE MOVING SOUND SOURCE LOCALIZATION

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

[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, hongliu@pku.edu.cn, lixiaofei@westlake.edu.cn

ABSTRACT

Multiple moving sound source localization in real-world scenarios remains a challenging issue due to interaction between sources, time-varying trajectories, distorted spatial cues, etc. In this work, we propose to use deep learning techniques to learn competing and time-varying direct-path phase differences for localizing multiple moving sound sources. A causal convolutional recurrent neural network is designed to extract the direct-path phase difference sequence from signals of each microphone pair. To avoid the assignment ambiguity and the problem of uncertain output-dimension encountered when simultaneously predicting multiple targets, the learning target is designed in a weighted sum format, which encodes source activity in the weight and direct-path phase differences in the summed value. The learned direct-path phase differences for all microphone pairs can be directly used to construct the spatial spectrum according to the formulation of steered response power (SRP). This deep neural network (DNN) based SRP method is referred to as SRP-DNN. The locations of sources are estimated by iteratively detecting and removing the dominant source from the spatial spectrum, in which way the interaction between sources is reduced. Experimental results on both simulated and real-world data show the superiority of the proposed method in the presence of noise and reverberation.

Index Terms— Direct-path phase difference, multiple moving sound source localization, direction of arrival, deep neural network.

1. INTRODUCTION

Sound source localization aims to determine the relative position of sound sources with respect to microphone array. As an important characteristics of directional sources, location information is widely used in real-world applications such as human-robot-interaction, and signal processing tasks including speech enhancement and source separation [1]. Recently, more and more works focus on localization in practical noisy and reverberant scenarios, but most of them either localize single moving source which avoids interaction between sources, or localize multiple static sources using long-time microphone signals. The dynamic trajectories of multiple moving sound sources pose new challenges to this task, which needs to timely estimate the locations of competing sources for each required time.

Traditional methods, such as generalized cross correlation (GCC) [2], steered response power (SRP) [3] and multiple signal classification (MUSIC) [4], are widely used for sound source localization. To deal with the multi-source case, these methods are sometimes combined with time-frequency (TF) processing [5–8], where the W-disjoint orthogonal (WDO) assumption [9] is used to simplify the problem of multiple source localization on broadband to that of single source localization in individual TF bins. Recently, deep learning based methods have shown promising localization performance [10–15]. They usually treat the localization task as either a location classification [10–12] or a location/feature regression [13–15] problem. Location classification methods output the posterior probabilities of location classes, but the dimension of output will increase with the number of candidate locations. Regressing the locations or features of multiple sources is difficult, because the assignment between multiple outputs and multiple training targets become ambiguous, and the dimension of regression output ought to vary with the number of sound sources.

Based on the fact that sound sources move continuously over time, many works try to exploit temporal context for moving sound source localization [13, 16–18]. Diaz-Guerra *et al*. performed a three-dimensional (3D) convolutional neural network (CNN) over the sequence of SRP-phase transform (SRP-PHAT) spatial spectrum to track single source [17]. Li *et al*. recursively estimated the direct-path relative transfer function (DP-RTF) using a short memory for online multiple-speaker localization [16]. Despite the progress of these works, it still requires a method to well exploit temporal context to quickly detect multiple moving sources and meanwhile filter out the outlier estimates caused by noise and reverberation.

This paper works on taking full use of the spatial-temporal context information to localize multiple moving sound sources. Considering the localization robustness of direct-path features [14, 15], the sequence of direct-path inter-channel phase differences (IPDs) is predicted by deep neural network (DNN) for each microphone pair. These predicted IPDs of all microphone pairs are used to construct the SRP spatial spectrum. This improved SRP is referred to as SRP-DNN. The contributions of this work are summarised as follows.

Learning direct-path IPD sequence for multiple moving sources: A causal convolutional recurrent neural network (CRNN) is designed to predict the direct-path IPD sequence for each microphone pair. This architecture fully exploits the TF patterns of direct-path IPDs, such as the temporal smoothness due to the continuous movement of sound sources, and the linearity along frequencies due to the linear relation between IPD and time difference of arrival (TDOA). The magnitude and phase spectrograms of dual-channel microphone signals are taken as the network input. As for the network output, it is difficult to well separate the direct-path IPDs of different sources from the overlapped microphone signals. Hence, the learning target is designed in a weighted sum format, where the weight reflects the source activity and the summed value compounds the direct-path IPDs. Importantly, the learned weighted direct-path IPDs can be directly used to construct the spatial spectrum for multiple sources.

This work is supported by National Natural Science Foundation of China (No. 62073004), Science and Technology Plan of Shenzhen (No. JCYJ20200109140410340).

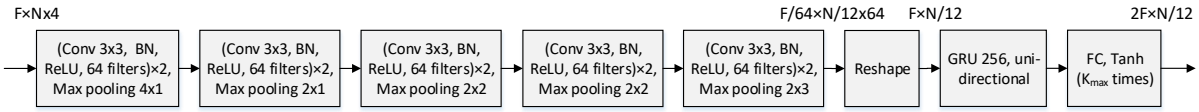

Fig. 1. Causal CRNN architecture to estimate the sequence of direct-path phase difference.

Iterative source detection and localization: One trivial way to perform direction of arrival (DOA) estimation is directly applying the peak detection method [7] on the spatial spectrum , which however is inaccurate especially for the case where multiple peaks are merged due to the interaction between sources. To solve this problem, this work proposes a new method that iteratively detects and removes the dominant source from the overall spatial spectrum, which makes it possible to separate the merged peaks.

2. METHOD

Suppose that there are multiple moving sound sources in the noisy and reverberant environment, the signal captured by the m-th microphone is formulated in the short-time Fourier transform domain as:

$$X_m(n,f) = \sum_{k=1}^{K} H_m(n,f,\boldsymbol{\theta}_k) S_k(n,f) + V_m(n,f), \quad (1)$$

where $m\in[1,M]$, $k\in[1,K]$ $n\in[1,N]$ and $f\in[1,F]$ are the indices of microphones, sound sources, time frames and frequencies, respectively. $\boldsymbol{\theta}_k=[\theta_k^{\text{ele}},\theta_k^{\text{azi}}]^T$ denotes the 2D DOA of the k-th sound source, which consists of elevation $\theta_k^{\text{ele}}\in[0,\pi]$ and azimuth $\theta_k^{\text{azi}}\in[-\pi,\pi)$. Here, $(\cdot)^T$ denotes matrix/vector transpose. $\theta_k^{\text{ele}}=0$ and $\theta_k^{\text{azi}}=0$ are defined along the positive z-axis and the positive x-axis, respectively. $X_m(n,f)$, $S_k(n,f)$ and $V_m(n,f)$ represent the microphone, source and noise signals, respectively. $H_m(n,f,\boldsymbol{\theta}_k)$ denotes the time-varying acoustic transfer function from the k-th source to the m-th microphone.

2.1. Classical SRP-PHAT

The classical SRP-PHAT algorithm [19] estimates the spatial spectrum by averaging the frame-wise GCC-PHAT over frequencies and nonredundant microphone pairs, namely:

$$P(\boldsymbol{\theta},n) = \frac{2}{M(M-1)} \sum_{m=1}^{M-1} \sum_{m'=m+1}^{M} G_{mm'}(\boldsymbol{\theta},n), \quad (2)$$

where $\boldsymbol{\theta}$ denotes the candidate DOA for spatial spectrum construction. The frame-wise GCC-PHAT [20, 21] for one pair of microphone signals is computed as:

$$G_{mm'}(\boldsymbol{\theta},n) = \frac{1}{F} \sum_{f=1}^{F} \Re\left\{ \Psi(n,f) e^{-j\omega_f \tau_{mm'}(\boldsymbol{\theta})} \right\}, \quad (3)$$

with the PHAT cross-power spectrum being:

$$\Psi(n,f) = \frac{X_m(n,f)X_{m'}^*(n,f)}{\left|X_m(n,f)X_{m'}^*(n,f)\right|} = e^{j\angle X_m(n,f)X_{m'}^*(n,f)}, \quad (4)$$

where ω_f denotes the angular frequency of the f-th frequency. $\Re\{\cdot\}$ and $|\cdot|$ denote the real part and magnitude of complex number, respectively. For far-field model where the propagation paths from one sound source to different microphones are regarded to be parallel, the TDOA between signals captured by the m-th and the m'-th microphones is computed as:

$$\tau_{mm'}(\boldsymbol{\theta}) = d_{mm'}(\boldsymbol{\theta})/c = (\mathbf{p}_m - \mathbf{p}_{m'})^T \mathbf{u}(\boldsymbol{\theta})/c, \quad (5)$$

with the Cartesian coordinates of the direction $\boldsymbol{\theta}$ on a unit sphere being:

$$\mathbf{u}(\boldsymbol{\theta}) = [\sin(\theta^{\text{ele}})\cos(\theta^{\text{azi}}), \sin(\theta^{\text{ele}})\sin(\theta^{\text{azi}}), \cos(\theta^{\text{ele}})]^T, \quad (6)$$

where $\mathbf{p}_m$ denotes the location coordinate of the m-th microphone and c is the speed of sound.

2.2. Direct-Path Phase Difference Learning

For the noise-free and anechoic single-source case, the PHAT cross-power spectrum, i.e. Eq. (4), is actually computing the complex-valued direct-path IPD for the source at $\boldsymbol{\theta}_k$:

$$\Psi^{\boldsymbol{\theta}_k}(n,f) = e^{j\omega_f \tau_{mm'}(\boldsymbol{\theta}_k)}. \quad (7)$$

Accordingly, the value of the frame-wise GCC-PHAT ranges from 0 to 1, and reaches the maximum when $\boldsymbol{\theta}$=$\boldsymbol{\theta}_k$, which is also the case for Eq. (2). However, ambient noise and room reverberation are inevitable in realistic applications. In these cases, using Eq. (4), the direct-path IPD is contaminated by noise, reverberation and interfering sources, which will no doubt lead to less prominent peaks of spatial spectrum and thus DOA estimation error. Therefore, recovering the direct-path IPDs from the noisy and reverberant microphone signals is essential for the SRP-PHAT based localization methods.

In this study, we propose to leverage the strong modeling ability of DNN to learn the direct-path IPDs for each microphone pair. Considering the continuous moving properties of sources over time, the TF patterns of direct-path phase differences are exploited and modelled by a causal CRNN. The convolutional units capture the inter-channel information and its short-term temporal context, and the recurrent units exploit the long-term spatial-temporal context. Overall, the CRNN aims to filter out the contamination of acoustic interferences and recover the direct-path IPDs for multiple moving sources. The causality of this model facilitates the online implementation of sound source localization. The details are described as follows.

Network architecture: Signals recorded by different microphone pairs are treated as independent training samples, and they are processed separately by the proposed network during inference time. The logarithm-magnitude and phase spectrograms of dual-channel signals are taken as the input features of the CRNN. As shown in Fig. 1, the input features are passed to ten causal convolutional modules. Each module consists of a causal convolutional layer followed by a batch normalization (BN) and a rectified linear unit (ReLU) activation function. A max pooling is used to compress the frequency and time dimensions after each two convolutional modules. The output of CNN layers is flatten for the frequency and filter dimensions, and then fed into one-layer uni-directional gated recurrent unit (GRU) with 256 hidden units. A fully connected (FC) layer with an activation of $K_{\max}$ times tanh function is used to output the direct-path phase difference (See details in *Learning target*) for one microphone pair. Here, $K_{\max}$ refers to the maximum possible number of sources. Note that, the frame rate of microphone signals (about 60 frames per second) is normally too high relative to the need of localization frame rate (about 5 frames per second is enough). Therefore, the input frame rate is compressed by a factor of 12 at the network output, and n' is used to denote the frame index of output.

Learning target: To learn the direct-path IPD (or its complex-valued form) for the single-source case, a simple way is to directly regress the real and imaginary parts of Eq. (7), which is expressed in vector form with all frequencies as:

$$\mathbf{r}_{mm'}(\boldsymbol{\theta}_k) = [\cos(\omega_1 \tau_{mm'}(\boldsymbol{\theta}_k)), \sin(\omega_1 \tau_{mm'}(\boldsymbol{\theta}_k)), \ldots, \cos(\omega_F \tau_{mm'}(\boldsymbol{\theta}_k)), \sin(\omega_F \tau_{mm'}(\boldsymbol{\theta}_k))]^T \in \mathbb{R}^{2F\times 1}. \tag{8}$$

For the case of multiple sources, directly regressing multiple vectors will require the dimension of regression output to be variant along with the variance of the number of sound sources. In addition, there will be the assignment ambiguity between the multiple outputs and the multiple training targets, which is similarly encountered by the speech separation task [22,23]. To avoid these problems, we propose to add up the direct-path IPD vectors of multiple sound sources as the training target, which is formulated as:

$$\mathbf{R}_{mm'}(n') = \sum_{k=1}^{K} \beta_k(n') \mathbf{r}_{mm'}(\boldsymbol{\theta}_k). \tag{9}$$

The summation weight $\beta_k(n')$ is defined as the activity probability of the k-th source at the n'-th output frame, which is computed as, over the 12 input frames that corresponding to the n'-th output frame, the proportion of the input frames where the k-th source is active. It ranges from 0 to 1. Correspondingly, the elements of the summed vector are in the range of $[0, K]$. The mean squared error (MSE) between the network output and the learning target is taken as the training loss.

This learning target is reasonable in the sense that taking the inner product between this summed vector $\mathbf{R}_{mm'}(n')$ (or its prediction) and the direct-path IPD vectors of candidate DOAs (denoted as $\mathbf{r}_{mm'}(\boldsymbol{\theta})$) is equivalent to the weighted summation of the direct-path spatial spectra (for one microphone pair) of multiple sources. This can be interpreted by substituting Eq. (9) into the inner product, namely:

$$\begin{aligned}\mathbf{R}^T_{mm'}(n')\mathbf{r}_{mm'}(\boldsymbol{\theta}) &= \sum_{k=1}^{K} \beta_k(n') \mathbf{r}^T_{mm'}(\boldsymbol{\theta}_k)\mathbf{r}_{mm'}(\boldsymbol{\theta}) \\ &= \sum_{k=1}^{K} \beta_k(n') F G^{\boldsymbol{\theta}_k}_{mm'}(\boldsymbol{\theta}, n'),\end{aligned} \tag{10}$$

with the direct-path spatial spectrum for the k-th source being:

$$G^{\boldsymbol{\theta}_k}_{mm'}(\boldsymbol{\theta}, n') = \frac{1}{F} \sum_{f=1}^{F} \Re\left\{ \Psi^{\boldsymbol{\theta}_k}(n', f) e^{-j\omega_f \tau_{mm'}(\boldsymbol{\theta})} \right\}. \tag{11}$$

2.3. DOA Estimation of Multiple Moving Sources

Using the prediction of the summed direct-path IPD vector, denoted as $\hat{\mathbf{R}}_{mm'}(n')$, the overall spatial spectrum of SRP-DNN is estimated with all microphone pairs as:

$$P'(\boldsymbol{\theta}, n') = \frac{2}{M(M-1)F} \sum_{m=1}^{M-1} \sum_{m'=m+1}^{M} \hat{\mathbf{R}}^T_{mm'}(n') \mathbf{r}_{mm'}(\boldsymbol{\theta}). \tag{12}$$

It is an estimation of the following theoretical value:

$$\sum_{k=1}^{K} \beta_k(n') \frac{2}{M(M-1)} \sum_{m=1}^{M-1} \sum_{m'=m+1}^{M} G^{\boldsymbol{\theta}_k}_{mm'}(\boldsymbol{\theta}, n').$$

Since $\sum_{m=1}^{M-1} \sum_{m'=m+1}^{M} G^{\boldsymbol{\theta}_k}_{mm'}(\boldsymbol{\theta}, n')$ exhibits a high peak at $\boldsymbol{\theta}_k$ and has a very small value at other directions, the value of $P'(\boldsymbol{\theta}_k, n')$ is dominantly contributed by the source at $\boldsymbol{\theta}_k$, and is thus approximately equal to $\beta_k(n')$.

Multiple sources can be localized by the peak detection method [7] which searches the significant peaks of the estimated spatial spectrum that are larger than a predefined threshold, as is done for regular SRP-PHAT. However, the peaks of sources may be merged due to the interaction between sources. To solve this problem, this work proposes to estimate the DOAs of active sources by iteratively detecting and removing the dominant source from the overall spatial spectrum, which is summarized in Algorithms 1. Since the iterative method works independently for each time step, the frame index n' is omitted for simplicity. The DOA of the dominant source $\hat{\boldsymbol{\theta}}^{\mathrm{d}}$ can be easily estimated as the candidate direction having the largest value of $P'(\boldsymbol{\theta})$. According to Eq. (9), the contribution of the dominant source can be removed by subtracting $\hat{\beta}^{\mathrm{d}} \times \mathbf{r}_{mm'}(\hat{\boldsymbol{\theta}}^{\mathrm{d}})$ from $\hat{\mathbf{R}}_{mm'}$, where $\hat{\beta}^{\mathrm{d}}$ is approximated by $P'(\hat{\boldsymbol{\theta}}^{\mathrm{d}})$. The new dominant source can be detected after the contribution of the previous one is removed. The iteration will stop when there is no notable source remained, namely $\hat{\beta}^{\mathrm{d}} < \beta_{\mathrm{TH}}$. Here, β_{TH} is a predefined threshold.

Algorithm 1: Iterative source detection and localization.

Input: The predicted direct-path IPD features $\{\hat{\mathbf{R}}_{mm'}\}$.
Output: The DOA estimates $\{\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_k\}$.

```
for k ← 1 to K_max do
    Estimate the spatial spectrum P'(θ) with {R̂_mm'};
    Estimate the DOA of the dominant source:
    θ̂^d ← arg max_θ P'(θ);
    Remove the contribution of the dominant source: β̂^d ← P'(θ̂^d),
    and for all microphone pairs
    R̂_mm' ← R̂_mm' − β̂^d × r_mm'(θ̂^d);
    if β̂^d < β_TH then
        The dominant source at θ̂^d is inactive, k ← k − 1, break;
    end
    θ̂_k ← θ̂^d;
end
return {θ̂_1, ..., θ̂_k}.
```

3. EXPERIMENTS AND DISCUSSIONS

3.1. Experimental Setup

Multi-conditional microphone signals are generated for network training. Following the data generation procedure presented in [17], the size of simulated rooms are randomly selected in the range from 3×3×2.5 m to 10×8×6 m, and the reverberation time is randomly set in the range from 0.2 s to 1.3 s. A 12-microphone array is randomly placed inside the room, and the array geometry is set to be the same as that mounted on the NAO robot head in the localization and tracking (LOCATA) challenge dataset [24]. The sound source moves along a random sinusoidal continuous trajectory. According to these room settings, room impulse responses (RIRs) are generated using the image method [25] implemented by the gpuRIR [26]. Speech recordings are randomly selected from the LibriSpeech corpus [27]. The microphone signals are created by filtering speech recordings by the RIRs, and then adding Gaussian noise with a SNR from 5 dB to 30 dB. In order to increase the diversity of training acoustic conditions, each training sample is generated on-the-fly as a random combination of data settings regarding source trajectories, microphone positions, source signals, noise signals, reverberation times, SNRs, etc. The evaluation is performed on both the simulated dataset and the LOCATA dataset. The real-world data provided by the LOCATA dataset is recorded in a computing laboratory with a size of 7.1×9.8×3 m. The reverberation time is 0.55 s. We consider the development and evaluation sets of tasks 3 and 5 with a single moving source, and also that of tasks 4 and 6 with two moving sources for performance evaluation.

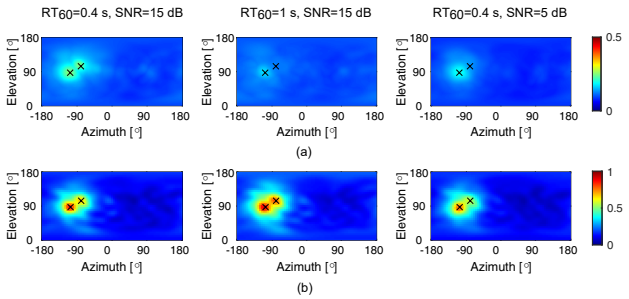

Fig. 2. Illustration of spatial spectra of (a) SRP-PHAT [19] and (b) SRP-DNN for two static sound sources present in the simulated rooms with different levels of noise and reverberation. The black crosses indicate the actual DOAs of sound sources.

The sampling rate of microphone signals is 16 kHz. The window length and the frame shift are 32 ms and 16 ms, respectively. The number of frequencies F is 256. The resolution of candidate azimuths and elevations are both 5°. During training, the length of microphone signals is set to 20 s, and correspondingly the number of time frames N is 1249 which is pooled to be 104 by the network. The model is trained using the Adam optimizer. The size of mini-batch is set to 66 (equals the number of microphone pairs). The performance for the voice-active periods is evaluated with three metrics. The source is considered to be successfully localized if the azimuth error is not larger than 30°. The mean absolute error (MAE) is computed by averaging the absolute azimuth (or elevation) error of all successfully localized sources and time frames. Miss detection (MD) refers to source active but not detected, and false alarm (FA) means source detected but not active. The MD rate (MDR) and the FA rate (FAR) are computed as the percentage of MDs and FAs out of the active sources of all time frames, respectively.

3.2. Experimental Results

Since the quality of spatial spectrum is important for the localization performance, we first visualize the spatial spectra of the proposed SRP-DNN method and one SRP-PHAT method [19]. Fig. 2 shows the spatial spectra obtained in the simulated rooms with different levels of noise and reverberation, where two static sources are present. To compute the spatial spectrum, SRP-PHAT estimates the PHAT-weighted cross-power spectrum using Eq. (4), while SRP-DNN predicts the direct-path IPDs in Eq. (7) via DNN. Under the condition that RT_{60}=0.4 s and SNR=15 dB, both methods exhibit sharp and distinct peaks around the actual DOAs. When the acoustic condition becomes worse, the local peaks of the proposed method are preserved, while the peaks of SRP-PHAT become flat and indistinctive. The robustness of SRP-DNN is mainly attributed to the fact that the direct-path IPDs are well recovered by the CRNN and meanwhile the contamination of noise and reverberation is largely reduced.

Three baseline approaches are compared with the proposed SRP-DNN method on the LOCATA dataset, including Cross3D [17], CTF-DPRTF [16] and SRP-PHAT [19]. Cross3D first computes SRP-PHAT and then tracks single sound source by performing 3D CNN on the sequence of the SRP-PHAT spatial spectrum. CTF-DPRTF estimates the azimuths of multiple moving sources with the predicted DP-RTFs. For the proposed SRP-DNN method, we present the results with either peak detection (PD) or iterative source detection and localization (IDL). For the single-source case, the number of source is assumed to be known with K=1. For the multi-source case, the number of source is assumed to be unknown. $K_{\max}$ and β_{TH} are separately set to 2 and 0.2 for SRP-DNN. The setting of β_{TH} is based on some preliminary experiments to seek a good trade-off between MDR and FAR.

Table 1. Performance of different methods on the LOCATA dataset

Method	Single-source (task 3, 5)		Multi-source (task 4, 6)		
	MDR/FAR [%]	MAE (az.|el.) [°]	MDR [%]	FAR [%]	MAE (az.|el.) [°]
Cross3D [17]	0.9	4.9 | 3.3	-	-	- | -
CTF-DPRTF [16]	2.4	3.8 | -	17.6	5.8	4.8 | -
SRP-PHAT [19]	0.8	**2.5** | **2.5**	25.5	12.1	**2.3** | 4.0
SRP-DNN+PD	**0.1**	**2.5** | 2.7	12.5	7.9	2.8 | **3.7**
SRP-DNN+IDL (prop.)	**0.1**	**2.5** | 2.7	**7.4**	**4.0**	2.8 | **3.7**

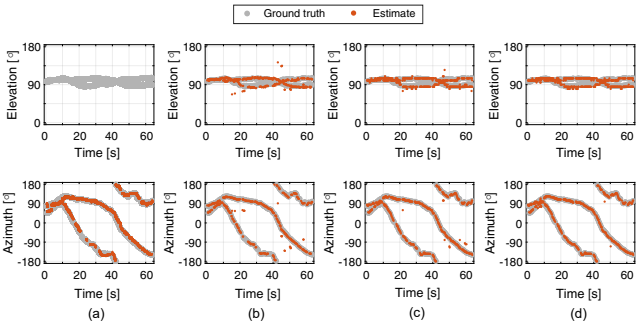

Fig. 3. Illustration of DOA (trajectory) estimation of (a) CTF-DPRTF [16], (b) SRP-PHAT [19], (c) SRP-DNN+PD and (d) SRP-DNN+IDL for one recording from the LOCATA challenge dataset. Two moving sound sources are present in this environment.

The localization results are shown in Table 1. For the single-source case with known source number, the MDR and FAR are equal, and PD and IDL work in the same way. SRP-DNN outperforms the other methods, and its MDR/FAR is extremely low. This indicates that SRP-DNN is able to largely reduce the influence of noise and reverberation, as which may cause spurious peaks. For the multi-source case, SRP-DNN achieves similar MAE and much smaller MDR and FAR compared to SRP-PHAT. This verifies that the proposed training target, i.e. Eq. (9), can well model/preserve the direct-path IPD information of multiple sources, moreover the proposed CRNN model is able to well extract the target vector from microphone signals. Relative to the trivial peak detection method, the proposed IDL method further improves the performance by disentangling the interaction of multiple sources. SRP-DNN+IDL also performs better than CTF-DPRTF on all metrics. Fig. 3 presents an example of localizing two moving sources. It can be seen that SRP-DNN+IDL provides relatively less erroneous DOA estimates (outliers) and more correct estimates, which is generally consistent with the lower rate of FA and MD presented in Table 1.

4. CONCLUSION

This work proposes to learn competing and time-varying direct-path phase differences for robust multiple moving sound source localization. The designed causal CRNN fully exploits the TF patterns to learn the direct-path features which encodes both direct-path phase difference and source activity. Using the predicted direct-path features, the SRP-DNN spatial spectrum shows more clear peaks around actual DOAs of sources even in the adverse noisy and reverberate scenarios. By iteratively detecting and localizing the dominant source, the merged peaks of spatial spectrum can be separated, and accordingly the interaction between sources is reduced. Experimental results show the advantage of the proposed method over other methods for the azimuth and elevation estimation of multiple moving sound sources in both simulated and real-world environments.

5. REFERENCES

[1] Deliang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

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

[3] Dongsuk Yook, Taewoo Lee, and Youngkyu Cho, "Fast sound source localization using two-level search space clustering," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 20–26, 2016.

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

[5] Despoina Pavlidi, Anthony Griffin, Matthieu Puigt, and Athanasios Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2193–2206, 2013.

[6] Lin Wang, Tsz-Kin Hon, Joshua D. Reiss, and Andrea Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 6, pp. 1079–1093, 2016.

[7] Dongwen Ying, Ruohua Zhou, Junfeng Li, and Yonghong Yan, "Window-dominant signal subspace methods for multiple short-term speech source localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 731–744, 2017.

[8] Bing Yang, Hong Liu, Cheng Pang, and Xiaofei Li, "Multiple sound source counting and localization based on TF-wise spatial spectrum clustering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1241–1255, 2019.

[9] Ozgur Yilmaz and Scott Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

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

[11] 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 Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2444–2453, 2017.

[12] Thi Ngoc Tho Nguyen, Woon-Seng Gan, Rishabh Ranjan, and Douglas L. Jones, "Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2626–2637, 2020.

[13] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Selected Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, 2019.

[14] Bing Yang, Xiaofei Li, and Hong Liu, "Supervised direct-path relative transfer function learning for binaural sound source localization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 825–829.

[15] Bing Yang, Hong Liu, and Xiaofei Li, "Learning deep direct-path relative transfer function for binaural sound source localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3491–3503, 2021.

[16] Xiaofei Li, Yutong Ban, Laurent Girin, Xavier Alameda-Pineda, and Radu Horaud, "Online localization and tracking of multiple moving speakers in reverberant environments," *IEEE J. Selected Topics Signal Process.*, vol. 13, no. 1, pp. 88–103, 2019.

[17] David Diaz-Guerra, Antonio Miguel, and Jose R. Beltran, "Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 300–311, 2021.

[18] Alexander Bohlender, Ann Spriet, Wouter Tirry, and Nilesh Madhu, "Exploiting temporal context in CNN based multisource DOA estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1594–1608, 2021.

[19] Romain Lebarbenchon, Ewen Camberlein, Diego di Carlo, Clement Gaultier, Antoine Deleforge, and Nancy Bertin, "Evaluation of an open-source implementation of the SRP-PHAT algorithm within the 2018 LOCATA challenge," in *Proc. LOCATA Challenge Workshop - Satell. Event Int. Workshop Acoust. Signal Enhancement*, 2018.

[20] Zhong-Qiu Wang, Xueliang Zhang, and Deliang Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 178–188, 2019.

[21] Maximo Cobos, Fabio Antonacci, Luca Comanducci, and Augusto Sarti, "Frequency-sliding generalized cross-correlation: A sub-band time delay estimation approach," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1270–1281, 2021.

[22] Dang Yu, Marten Kalbcek, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 241–245.

[23] Morten Kolbk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[24] Heinrich W. Lollmann, Christine Evers, Alexander Schmidt, Heinrich Mellmann, Hendrik Barfuss, Patrick A. Naylor, and Walter Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop*, 2018, pp. 410–414.

[25] Jont B. Allen and Daivid A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[26] David Diaz-Guerra, Antonio Miguel, and Jose R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tools Appl.*, 2020.

[27] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.