INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Sound Event Localization and Detection Based on Multiple DOA Beamforming and Multi-task Learning

Wei Xue, Ying Tong, Chao Zhang, Guohong Ding, Xiaodong He, Bowen Zhou

JD AI Research
{xuewei27,tongying,chao.zhang,dingguohong,xiaodong.he,bowen.zhou}@jd.com

Abstract

The performance of sound event localization and detection (SELD) degrades in source-overlapping cases since features of different sources collapse with each other, and the network tends to fail to learn to separate these features effectively. In this paper, by leveraging the conventional microphone array signal processing to generate comprehensive representations for SELD, we propose a new SELD method based on multiple direction of arrival (DOA) beamforming and multi-task learning. By using multiple beamformers to extract the signals from different DOAs, the sound field is more diversely described, and specialised representations of target source and noises can be obtained. With labelled training data, the steering vector is estimated based on the cross-power spectra (CPS) and the signal presence probability (SPP), which eliminates the need of knowing the array geometry. We design two networks for sound event localization (SED) and sound source localization (SSL) and use a multi-task learning scheme for SED, in which the SSL-related task act as a regularization. Experimental results using the database of DCASE2019 SELD task show that the proposed method achieves the state-of-art performance.

Index Terms: Sound event localization and detection, microphone arrays, beamforming, multi-task learning

1. Introduction

Sound event detection (SED) aims to determine the time period of acoustic events, and has been widely used in applications such as robotics, smart home and surveillance [1–3]. Recently the focus has shifted to not only estimating the temporal information of the acoustic event, but also determining the location of the corresponding sound source. This raises the problem of joint SED and sound source localization (SSL), namely, sound event localization and detection (SELD). A microphone array is usually utilized, such that temporal and spatial samplings are simultaneously performed to describe the acoustic scene.

Conventionally the SED and SSL are separately treated, and there have been enormous researches on each problem. Most of the state-of-art SED systems are now based on deep neural networks (DNNs) [2–6], and the convolutional neural networks (CNN) [5–7] and recurrent neural networks (RNN) [4, 5] are exploited to model the compact representations and the temporal characteristics of the acoustic events, respectively. For SSL, conventional methods are generally based on analysing the cross-correlations between the multichannel signals [8–13], including the generalized cross-correlation (GCC) [8], the multichannel cross-correlation coefficient (MCCC) [14] and the multiple signal classification (MUSIC) [15–17] based approaches, etc. Methods based on the DNNs are also proposed [18–20], which use the cross-correlations as the input feature, and estimate the direction of the arrival (DOA) as a regression or classification problem.

Since both SED and SSL can be achieved by using a DNN, to cope with the SELD problem, DNN-based end-to-end systems can be trained by taking the multichannel temporal spectral features as input, and the SED and SSL results are simultaneously obtained by multi-task learning [21, 22]. Shared bottom hidden layers are used to extract features for both SED and SSL, and then the layers are split into different branches to adapt to the specific task. In [22], the conventional recurrent neural network (CRNN) is used to model the spectral and temporal characteristics of the acoustic events, and the SED and SSL branches are respectively formulated by the feed-forward networks (FNN). A main challenge of SELD is to deal with the overlapping cases in which sources from different DOAs coexist, and the features of different sources collapse with each other. Although the problem can be alleviated by adopting an active source counter [23], or performing data augmentation [24–27], simply performing data-driven supervised training from the overlapped features has limited flexibility to make the network learn to separate these features effectively.

In this paper, the conventional microphone array signal processing is leveraged to generate comprehensive representations for both SED and SSL, and a new method based on multiple direction of arrival (DOA) beamforming and multi-task learning is proposed. The proposed method exploits spectral and spatial features extracted from signals of multiple beams, which orient towards different DOAs. The multiple beams give a diversified description of the acoustic field, and each beam is formed according to the estimated steering vector of each DOA. A steering vector estimation method based on the cross-power spectrum (CPS) and signal presence probability (SPP) is proposed. We design two separate DNNs for SED and SSL, and a multi-task learning scheme is exploited for training the SED network, in which the SSL related tasks act as a regularization. We conduct experiments on both development and evaluation sets of the DCASE2019 SELD task, and the results demonstrate the effectiveness of the proposed method.

The rest of the paper is organized as follows. In Section 2 we describe the problem. Details of the proposed method will be presented in Section 3. We evaluate the proposed method in Section 4 and draw conclusions in Section 5.

2. Problem Description

We consider the SELD task in the DCASE2019 challenge as a representative of the problem in this paper, in which a 3-dimensional microphone array is used to capture the signals of potentially multiple sources. Given a development dataset which consists of sound event recordings of different types, DOAs and overlapping patterns, the aim is to jointly estimate the time interval, azimuth and elevation angles of each sound event for a new recording from the same array.

With M microphones, the short-time Fourier transform

http://dx.doi.org/10.21437/Interspeech.2020-2759

(STFT) domain reverberant signal in the m-th microphone is expressed as

$$Y_m(t,f) = \sum_{q=1}^{Q} H_{m,q}(f) S_q(t,f) + V_m(t,f), \quad (1)$$

where t and f are the temporal frame and frequency indexes, respectively, and Q is the maximum number of considered sources. We denote $S_q(t,f)$ as the STFT-domain signal of the q-th source in the time-frequency (TF) bin (t,f), and $H_{m,q}(f)$ as the frequency-domain room impulse response (RIR) from the q-th source to the m-th microphone. $V_m(t,f)$ is the noise signal in the m-th microphone which is assumed to be uncorrelated with the source signals.

3. Proposed Method

3.1. Multiple DOA Beamforming

The microphone array could perform spatial filtering that enhances the target sources and attenuates the inferences, which is beneficial for the SELD task. Although the networks can be expected to learn the spatially-filtered signals from raw multichannel inputs, optimizing and interpreting the learned intermediate representations are not straightforward. It might be possible to first estimate the DOAs of multiple sources, and then conduct SED on the signal from each estimated DOA. However, the error of multi-source DOA estimation could introduce an extra risk to the robust SED.

In this paper, instead of estimating the signal from the DOA of each source, we propose to perform multiple DOA beamforming, which evenly steers the beams towards different DOAs, such that spatially-distribute sources and noise signals can be separated. Along with the multichannel raw observations, the beamformed signals from multiple DOAs provide a richer description of the acoustic environment, and a specialised representation of each source. The signals from source-absent directions can also be regarded as an estimation of the noise field. When using the beamformed signals from multiple DOAs for the DNNs, a better SELD performance can be expected.

In the following of this section, the methods of estimating the steering vector and noise covariance matrix for the beamforming as well as the beamformer design will be introduced.

3.1.1. Steering Vector Estimation

With the labelled training data, the steering vector for each DOA in the training data can be derived without knowing the geometry of the microphone array.

From the signal model (1), without loss of generality, we assume that the q-th source is from azimuth and elevation of (θ, ϕ), then the $M \times 1$ steering vector $\mathbf{A}(\theta,\phi,f)$ for (θ,ϕ) is defined as

$$\mathbf{A}(\theta,\phi,f) = [1, \frac{H_{2,q}(f)}{H_{1,q}(f)}, ..., \frac{H_{M,q}(f)}{H_{1,q}(f)}]^T. \quad (2)$$

Given the multichannel labelled data, for each (θ,ϕ), a new multichannel signal can be obtained by concatenating all the time intervals that consist of only one source and that the source is from (θ,ϕ). Since only one source exists in the multichannel signal, we consistently denote the source index as q without loss of generality. According to (1), the CPS between the m-th and the n-th microphones is expressed by

$$\begin{aligned} R_{mn}(f) &= \mathbb{E}\{Y_m(t,f)Y_n^*(t,f)\} \\ &= H_{m,q}H_{n,q}^* R_{ss}(f) + \mathbb{E}\{V_m(t,f)V_n^*(t,f)\}, \end{aligned} \quad (3)$$

where $R_{ss}(f)$ is the variance of the source signal. If ignoring the noise-related term, the m-th element of the steering vector can be computed based on the CPS as

$$A_m(\theta,\phi,f) = \frac{R_{m1}(f)}{R_{11}(f)} = \frac{H_{m,q}H_{1,q}^* R_{ss}(f)}{H_{1,q}H_{1,q}^* R_{ss}(f)} = \frac{H_{m,q}(f)}{H_{1,q}(f)}. \quad (4)$$

In practice, the CPS is estimated by recursive smoothing as:

$$\begin{aligned} &\hat{R}_{mn}(t,f) \\ &= \alpha(t,f)\hat{R}_{mn}(t-1,f) + [1-\alpha(t,f)]Y_m^*(t,f)Y_n(t,f), \end{aligned} \quad (5)$$

where $0 < \alpha(t,f) < 1$ is a smoothing factor, and $\hat{R}_{mn}(f)$ is calculated by averaging the CPS over all frames. In order to reduce the effect of noise in CPS estimation, the smoothing factor is controlled by the SPP $\rho(t,f)$, as

$$\alpha(t,f) = \begin{cases} 1, \text{ if } \rho(t,f) < \kappa_{\text{cps}}; \\ \alpha_y, \text{ otherwise}, \end{cases} \quad (6)$$

where κ_{cps} is a threshold determining whether a TF bin is speech/noise-dominated for CPS estimation, and α_y is the smoothing factor adopted when updating the CPS. It can be seen that the CPS does not update in noise-dominant TF bins when $\rho(t,f) < \kappa_{\text{cps}}$ and $\alpha(t,f) = 1$. The SPP is estimated using the first channel signal based on the method in [28].

3.1.2. Noise Covariance Matrix Estimation

With the SPP, for each utterance, all noise-dominated TF bins are used to estimate the noise covariance matrix $R_{vv}(f)$. In the f-th frequency bin we have

$$R_{vv}(f) = \frac{1}{T}\sum_t \mathbf{y}(t,f)\mathbf{y}^H(t,f)w(t,f), \quad (7)$$

where T is the number of frames of the utterance, $\mathbf{y}(t,f) = [Y_1(t,f), Y_2(t,f), ..., Y_M(t,f)]^T$ is the $M \times 1$ signal vector, $w(t,f)$ equals to one if $\rho(t,f) < \kappa_{\text{ncm}}$ and zero otherwise.

3.1.3. Beamformer Design

We utilize multiple MVDR beamformers, which steer the beams towards P different DOAs, to achieve separation of different source signals. The MVDR beamformer coefficients for each DOA (θ,ϕ) is computed according to the estimated steering vectors and noise covariance matrix.

As is shown in Fig. 1, the DOAs of the beamformers are chosen as symmetric in elevation and equally spaced in azimuth, since sources are usually more distributed over azimuth.

The output signals from the multiple DOA beamforming are used to extract the features for both SSL and SED, which will be elaborated later in the next two sections.

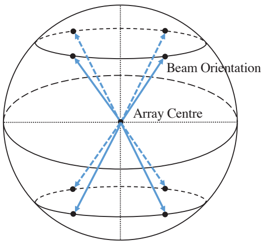

Figure 1: *Beam orientations for multiple DOA beamforming.*

3.2. SSL

3.2.1. Features

The DOA of the sound source is closely related to the phase differences and intensity differences between microphones. The phase difference can be expressed by the phase of the CPS. With M microphones, all pairwise $M(M-1)/2$ CPS are estimated, then in frame t, a CPS phase feature vector is formed by

$$\mathbf{r}(t) = [\mathbf{r}_{21}(t), \mathbf{r}_{31}(t), ..., \mathbf{r}_{M(M-1)}(t)]^T, \tag{8}$$

where $\mathbf{r}_{mn}(t) = [\angle R_{nm}(t,1), ..., \angle R_{nm}(t,F)]$, and F is the number of frequency bins. $\angle R_{nm}(t,f)$ computes the phase angle of $R_{nm}(t,f)$ within the range $(-\pi,\pi]$. Here a small constant smoothing factor 0.2 is adopted in (5) to increase the tracking speed of SSL.

The multichannel raw observations and the signals from multiple DOA beamforming are jointly used to compute the intensity differences. By steering towards different DOAs, the augmented multiple DOA beamforming outputs give a more comprehensive sampling of the intensity distribution of the sound field. In each frame, the intensity of each signal is computed by averaging the power densities over all TF bins within the $[-5,5]$ context window. An $(M+P)\times 1$ frame-wise normalized intensity feature vector is obtained by gathering the intensities of all signals, and normalizing the vector such that the minimum and maximum elements of the vector are 0 and 1, respectively.

3.2.2. Architectures

We formulate the SSL as a classification problem, and each class represents a candidate DOA that has appeared in the training data. Based on the phase and intensity difference features, the network structure for SSL is depicted in Fig. 2 (a). Since the CPS phase features span over the whole frequency range, CNN is adopted to extract a compacted representation. The compacted feature from CNN and the intensity feature vector are then concatenated and sent into the gated recurrent units (GRU) based layers, such that the temporal evolution of the acoustic events are taken into account. Finally linear layers with sigmoid activations in the output layer are used to project the hidden units to the DOA label, under the binary cross-entropy optimization criterion. Both CNN layers and GRU layers are composed of several CNN and GRU blocks with architectures illustrated in the Fig. 2 (c) and (d), respectively.

3.3. Multi-task Learning for SED

3.3.1. Features

We design a separate network for SED. The features for the SED network consist of two parts: a) the $(M+P)$-channel Log-Mel features of the multichannel observations as well as the multiple-DOA beamforming outputs; b) the CPS phase features used for the SSL network, which contain the spatial information. Since the Log-Mel features could indicate the level difference between channels, the intensity features used for SSL are excluded. It should be noted that the DOA is not related to the explicit spectral information of the signal, thus the Log-Mel features are not exploited for SSL in the previous section.

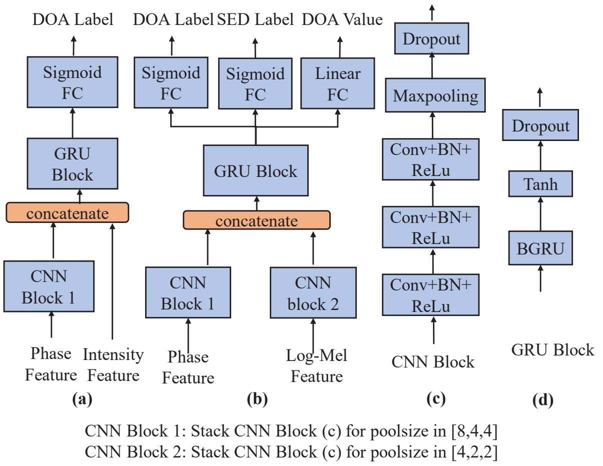

Figure 2: *Network Architectures for SSL and SED. (a) SSL Network; (b) SED Network; (c) CNN block; (d) GRU block.*

3.3.2. Architectures

A multi-task learning based network architecture for SED is shown in Fig. 2 (b). The network has a similar lower-layer structure with the SSL network, except that the CNN is also used to extract hidden representations from the Log-Mel features. Since the spatially distributed sound events could be temporally overlapped, in order to integrate the spatial discrimination capability into the network, a triple-task learning scheme is developed, which forces the network to predict the DOA of different sources, and serves as regularization for the SED target. The SED is solved as a classification problem, and both regression-based and classification-based targets are adopted for SSL, since empirical studies show that keeping both SSL targets outperforms utilizing either one of them. The optimization criteria for DOA regression and SED classification are mean square error and binary cross-entropy, respectively. The target for the regression based SSL is a $2Q\times 1$ vector whose elements are the azimuths and elevations of the Q sources.

We note that when only the SSL network is used for DOA estimation, there is an ambiguity of assigning the DOA to the correct source in overlapping scenarios. The regression based DOA estimation helps to solve the problem by providing an anchor point for each source. In addition, although the SSL targets are included in the SED network, the SSL network is still needed since the spectral features for the SED network could impose a negative effect on SSL.

With the SED and SSL networks, the SELD is solved by a) detecting sound events, b) estimating the DOA of each event by using the SSL network, with the regression based SSL results from the SED network as anchor points in the overlapping cases.

4. Evaluation

4.1. Data and Experimental Setup

The dataset of the DCASE2019 SELD task, which is recorded using a four-channel spherical microphone array, is for evaluation. The dataset consists of a 400-utterance development set and a 100-utterance evaluation set, and the development set is further divided into four 100-utterance cross-validation splits to facilitate training. Each utterance is sampled at 48 kHz, and has a fixed duration of one minute. Eleven types of acoustic events are included, and up to $Q = 2$ events appear simultaneously. The azimuths and the elevations of the sound sources are distributed within the range of $[-180^\circ, 170^\circ]$ and $[-40^\circ, 40^\circ]$, respectively, all with a 10° increment, thus totally $36 \times 9 = 324$ DOAs are included in the dataset.

The STFT is conducted with a frame length of 2048 samples and the hop size is 960 samples, resulting 3000 frames for each one-minute utterance. Only the lower 512 frequency bins are used. Eight target directions are chosen for multiple DOA beamforming, whose two elevations are in $\{-20^\circ, 20^\circ\}$ and four azimuths are in $\{-170^\circ, -80^\circ, 10^\circ, 100^\circ\}$, respectively. The 96-dimension Log-Mel features for SED are extracted from the STFT spectrum. In the implementation, we set $\kappa_{\text{cps}} = 0.6$, $\kappa_{\text{ncm}} = 0.3$ and $\alpha_y = 0.9$. The network configurations of the proposed method is summarized in Fig. 2 (e). For the SED network, the loss functions of SED classification, DOA regression and DOA classification are combined with a weight of $[1, 50, 50]$ for joint optimization. In the training stage, the development data is augmented by speed perturbation with perturbation factors of 0.9 and 1.1 to improve the generalization ability of the trained models.

Four metrics are used for evaluation, which are the F-score and error rate (ER) for SED, and the DOA error (in degree) and frame recall (FR) for SSL. An SELD score is calculated based on the above four metrics to provide an overall evaluation of the SELD performance, as

$$\text{SELD score} = \frac{\text{ER} + \text{DOA Error}/180 + (2 - \text{F-score} - \text{FR})}{4}$$

4.2. Results

The proposed method (Triple-task + BF) is compared with the baseline system [22] and the systems from top two systems of the DCASE2019 challenge SELD task [23, 29]. To examine the effectiveness of multiple-DOA beamforming and multi-task learning, two variations of the proposed method, which respectively use only the multichannel signals for feature extraction (Triple-task), and exploit only the SED classification & DOA regression task for the SED network (Dual-task + BF), are taken for comparison.

The performances of different systems are summarized in Table 1 and Table 2. The results on the development set show that the propose method substantially outperforms the baseline method, and achieves the better SELD score than the top systems of the DCASE2019 challenge. The proposed method performs best on SED, and yields SSL results comparable with the methods in [23]. Compared with the ”Triple-task learning” system, the proposed method achieves a relative reductions of 38.3% on SED ER and 34.5% on DOA error, which clearly shows that the richer description of the sound field provided by multiple-DOA beamforming is helpful to both SED and SSL. In addition, we notice that the triple-task learning scheme performs better than the dual-task learning counterpart, by adopting a more strict regularization for SED. We should note that although separate SSL and SED networks are used, a better SED could improve the SSL performance by increasing the frame recall. By analysing the results on the evaluation set, similar conclusions can be drawn.

Table 1: *Performances on the DCASE 2019 SELD dev set*

System	ER	F-score	DOA	FR	SELD score
DCASE Baseline [22]	34.0%	79.9%	28.5°	85.4%	21.1%
Method in [23]	14.0%	89.3%	5.7°	**95.6%**	8.1%
Method in [29]	13.0%	92.8%	6.7°	90.8%	8.3%
Triple-task	15.4%	87.6%	8.1°	86.4%	11.5%
Dual-task + BF	13.0%	92.2%	7.4°	89.9%	8.8%
Triple-task + BF	**9.5%**	**94.1%**	**5.3°**	93.1%	**6.3%**

Table 2: *Performances on the DCASE 2019 SELD eval set*

System	ER	F-score	DOA	FR	SELD score
DCASE Baseline [22]	28.0%	85.4%	24.6°	85.7%	24.5%
Method in [23]	8.0%	94.7%	**3.7°**	**96.8%**	4.6%
Method in [29]	8.0%	95.5%	5.5°	92.2%	5.8%
Triple-task	10.4%	90.7%	6.1°	89.7%	8.3%
Dual-task + BF	9.0%	91.9%	5.5°	90.9%	7.3%
Triple-task + BF	**5.6%**	**96.9%**	**3.7°**	95.5%	**3.8%**

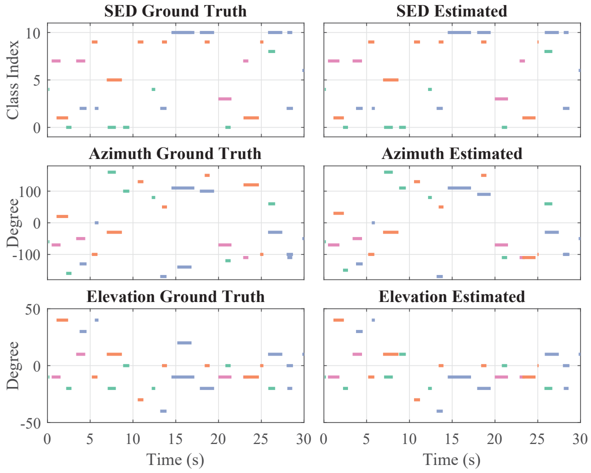

Figure 3: *Illustration of the SELD performance of the proposed system in overlapping conditions.*

Fig. 3 displays the SELD performance of the proposed method on one utterance. We can observe that for almost all cases the proposed system can yield both accurate and stable SED and SSL estimates.

5. Conclusions

We propose an SELD method which leveraged the conventional microphone array signal processing techniques. Multiple-DOA beamforming is used to achieve signal separation and provides a diversified description of the sound field. Based on the CPS and SPP, the steering vector for each DOA is computed and is used to design beamformers for multiple DOAs. A triple-task learning scheme is used, which uses both the regression and classification based SSL criterion to regularize the SED network. The effectiveness of the proposed method is demonstrated by the experiments on the dataset of DCASE2019 challenge SELD task.

6. References

[1] M. Crocco, M. Cristani, A. Trucco, and V. Murino, “Audio surveillance: A systematic review,” *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, p. 52, 2016.

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

[3] C. Grobler, C. P. Kruger, B. J. Silva, and G. P. Hancke, “Sound based localization and identification in industrial environments,” in *IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2017, pp. 6119–6124.

[4] G. Parascandolo, H. Huttunen, and T. Virtanen, “Recurrent neural networks for polyphonic sound event detection in real life recordings,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6440–6444.

[5] S. Adavanne, P. Pertilä, and T. Virtanen, “Sound event detection using spatial features and convolutional recurrent neural network,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 771–775.

[6] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, “Convolutional gated recurrent neural network incorporating spatial features for audio tagging,” in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 3461–3466.

[7] H. Zhang, I. McLoughlin, and Y. Song, “Robust sound event recognition using convolutional neural networks,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 559–563.

[8] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[9] Y. Huang and J. Benesty, “Adaptive multichannel time delay estimation based on blind system identification for acoustic source localization,” in *Adaptive Signal Processing*. Springer Berlin Heidelberg, 2003, pp. 227–247.

[10] J. Tugnait, “Time delay estimation with unknown spatially correlated Gaussian noise,” *IEEE Trans. Signal Process.*, vol. 41, no. 2, pp. 549–558, 1993.

[11] J. Chen, J. Benesty, and Y. Huang, “Time delay estimation in room acoustic environments: an overview,” *EURASIP J. on Applied Signal Processing*, vol. Special issue on advances in multimicrophone speech processing, pp. 1–19, 2006.

[12] W. Xue, S. Liang, and W. Liu, “Interference robust DOA estimation of human speech by exploiting historical information and temporal correlation.” in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, 2013, pp. 2895–2899.

[13] W. Xue, W. Liu, and S. Liang, “Noise robust direction of arrival estimation for speech source with weighted bispectrum spatial correlation matrix,” *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 837–851, 2015.

[14] J. Benesty, J. Chen, and Y. Huang, “Time-delay estimation via linear interpolation and cross correlation,” *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 509–519, Sep. 2004.

[15] R. O. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[16] J. Dmochowski, J. Benesty, and S. Affes, “Broadband MUSIC: Opportunities and challenges for multiple source localization,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007, pp. 18–21.

[17] Y. Zhang and B. P. Ng, “MUSIC-like DOA estimation without estimating the number of sources,” *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1668–1676, Mar. 2010.

[18] K. Youssef, S. Argentieri, and J.-L. Zarader, “A learning-based approach to robust binaural sound localization,” in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 2927–2932.

[19] N. Ma, G. J. Brown, and T. May, “Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] R. Takeda and K. Komatani, “Sound source localization based on deep neural networks with directional activate function exploiting phase information,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 405–409.

[21] T. Hirvonen, “Classification of spatial audio location and content using convolutional neural networks,” in *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[22] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2018.

[23] S. Kapka and M. Lewandowski, “Sound source detection, localization and classification using consecutive ensemble of CRNN models,” DCASE Challenge, Tech. Rep., 2019.

[24] J. Zhang, W. Ding, and L. He, “Data augmentation and prior knowledge-based regularization for sound event localization and detection,” DCASE Challenge, Tech. Rep., 2019.

[25] L. Mazzon, M. Yasuda, Y. Koizumi1, and N. Harada, “Sound event localization and detection using foa domain spatial augmentation,” DCASE Challenge, Tech. Rep., 2019.

[26] K. Noh, C. Jeong-Hwan, J. Dongyeop, and C. Joon-Hyuk, “Three-stage approach for sound event localization and detection,” DCASE Challenge, Tech. Rep., 2019.

[27] S. P. Chytas and G. Potamianos, “Hierarchical detection of sound events and their localization using convolutional neural networks with adaptive thresholds,” DCASE Challenge, Tech. Rep., 2019.

[28] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[29] Y. Cao, T. Iqbal, Q. Kong, M. B. Galindo, W. Wang, and M. D. Plumbley, “Two-stage sound event localization and detection using intensity vector and generalized cross-correlation,” DCASE Challenge, Tech. Rep., 2019.