2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP) | 979-8-3503-9796-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISCSLP57327.2022.10037995

Deep Learning Based Audio-Visual Multi-Speaker DOA Estimation Using Permutation-Free Loss Function

Qing Wang[1], *Hang Chen*[1], *Ya Jiang*[1], *Zhe Wang*[1], *Yuyang Wang*[1], *Jun Du*[1*], *Chin-Hui Lee*[2]

[1]University of Science and Technology of China, Hefei, China
[2]Georgia Institute of Technology, Atlanta, GA. USA

jundu@ustc.edu.cn

Abstract

In this paper, we propose a deep learning based multi-speaker direction of arrival (DOA) estimation with audio and visual signals by using permutation-free loss function. We first collect a data set for multi-modal sound source localization (SSL) where both audio and visual signals are recorded in real-life home TV scenarios. Then we propose a novel spatial annotation method to produce the ground truth of DOA for each speaker with the video data by transformation between camera coordinate and pixel coordinate according to the pin-hole camera model. With spatial location information served as another input along with acoustic feature, multi-speaker DOA estimation could be solved as a classification task of active speaker detection. Label permutation problem in multi-speaker related tasks will be addressed since the locations of each speaker are used as input. Experiments conducted on both simulated data and real data show that the proposed audio-visual DOA estimation model outperforms audio-only DOA estimation model by a large margin.

Index Terms: sound source localization, DOA estimation, audio-visual fusion, pin-hole camera model

1. Introduction

Sound source localization (SSL) aims to estimate the position of single or multiple sound sources relative to the position of the recording microphone array. In most cases, we are interested in direction of arrival (DOA) for each sound source, hence most of the SSL methods focus on azimuth and elevation angles estimation. Effective SSL is of great importance in many applications including automatic speech recognition (ASR) [1], tele-conferencing [2], robot audition [3, 4], and hearing aids [5].

Many previous studies about SSL pay more attention to audio modality alone. Conventional SSL methods, such as generalized cross-correlation with phase transform (GCC-PHAT) [6], steered response power with phase transform (SRP-PHAT) [7], estimation of signal parameters via rotational invariance technique (ESPRIT) [8], and multiple signal classification (MUSIC) [9], were based on signal processing techniques and usually performed poorly in noisy and reverberant environments. Deep neural network (DNN)-based SSL methods have been proposed in recent years and proven to outperform conventional SSL methods due to their strong regression capability [10]. Grumiaux [11] provided a thorough survey of the audio SSL literature based on deep learning techniques. The output strategy for DOA estimation can be divided into two categories: classification and regression. Convolutional recurrent neural networks (CRNN) were proposed for DOA estimation of multiple sources by using a classification strategy in [12, 13]. Some other works [14, 15] tried to solve the SSL problem as a regression task by directly estimating either Cartesian coordinates or spherical coordinates. Tang [16] demonstrated that regression model achieved better performance over classification model.

By hearing and seeing, human brain is able to perceive surroundings and extract complementary information. Intelligent devices equipped with audio-visual sensors are supposed to achieve similar goals. Fusion of audio and video modalities has shown promising results in many areas, e.g. acoustic scene classification [17], speech enhancement [18], and active speaker detection [19]. The literature on audio-visual localization is sparse compared to the large number of studies for sound source localization [11]. Most of these works [20, 21, 22] mainly focused on localizing sound sources in video clips rather than estimating DOA of sound sources. In [23], the authors first proposed a deep neural network (DNN) architecture for audio-visual multi-speaker DOA estimation by simulating visual features. Promising results were observed in [23] when at most two speakers existed however the performance of localizing more than two speakers remained unknown. Berghi [24] proposed a teacher-student model to perform active speaker detection and localization with the ‘teacher’ network generating pseudo-labels and the ‘student’ network localizing speakers.

Most of previous works only consider localizing one or two concurrent speakers and the existing audio-visual datasets [25, 26, 24] are of limited size. In this paper, we propose a novel audio-visual DOA estimation approach for multi-speaker scenario based on the MISP2021-AVSR corpus [27], a large-scale audio-visual Chinese conversational corpus which contains 141 hours of audio and video data with at most six concurrent speakers. To avoid expensive and time-consuming cost of manual annotation, we propose to produce the ground truth of DOA for each speaker based on the video data and camera calibration. Then we solve the multi-speaker DOA estimation problem as active speaker detection with the ground truth of DOA served as complementary input to acoustic feature. Label permutation problem in multi-speaker related tasks will be addressed since the locations of each speaker are used as input.

2. Proposed Method

In this section, we describe our proposed approach for multi-speaker DOA estimation using audio and video data. Real data is recorded in the home TV scenario with several people sitting and chatting in Chinese. In home TV scenario, people are always sitting, so our study is focused on estimating the azimuth angle only. Firstly, we introduce how to generate DOA labels with video clips. Then we describe the proposed multi-modal DOA (MDOA) and audio-only DOA (ADOA) estimation models for multi-speaker situation.

979-8-3503-9796-3/22/$31.00 ©2022 IEEE

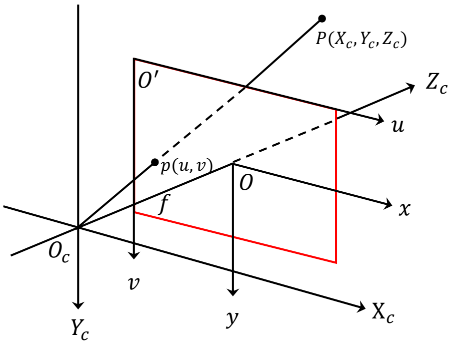


Figure 1: *Transformation from pixel coordinate to camera coordinate. (X_C, Y_C, Z_C): camera coordinate; (x, y): image coordinate; (u, v): pixel coordinate.*

2.1. Spatial Annotation Method

It is expensive and time-consuming to annotate DOA labels manually, therefore we propose to produce the ground truth of DOA with the video data by transformation between camera coordinate and pixel coordinate according to the pin-hole camera model [28]. The linear microphone array is placed on in the horizontal axis x of the camera coordinate system, with its center coinciding with the origin of the coordinate system. Based on this, we transfer the target point in the pixel coordinate to the camera coordinate and use the resulting point as the sound source location.

Several effective techniques were used to detect the head-and-shoulder of target speakers in the video. In particular, we adopted a head-and-shoulder detection model based on yolo-v5[1] and the Deep SORT algorithm [29] to track and match multiple speakers in the same video clips simultaneously. The average missing rate of head-and-shoulder detection is less than 1% at the frame level on MISP2021-AVSR dataset. We average the pixel coordinates of the top-left and bottom-right points of the head-and-shoulder detection bounding box, which is considered to be the location of the mouth.

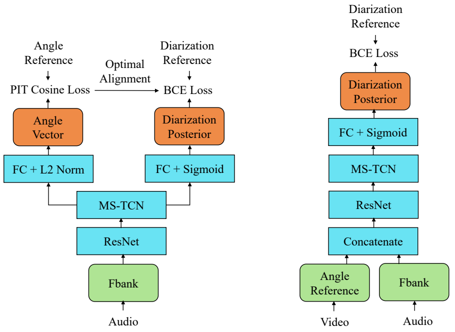
Figure 1 shows the transformation process from pixel coordinate to camera coordinate. O_c, O and $O^{'}$ denote the origins of camera coordinate, image coordinate and pixel coordinate, respectively. Z_c corresponds to the camera's optical axis and f represents the image distance. Let us define $p(u, v)$ as the pixel coordinate of one target speaker in the image plane. Then the corresponding point in the camera coordinate is denoted as $P(X_c, Y_c, Z_c)$. By using the pin-hole camera model [28], point in the image coordinate is expressed as follows:

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \frac{1}{Z_c} \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{pmatrix} \tag{1}$$

The affine transformation between image coordinate and pixel coordinate can be written as:

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{dx} & 0 & u_0 \\ 0 & \frac{1}{dy} & v_0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \tag{2}$$

where dx and dy represent the physical length of each pixel on the horizontal axis x and the vertical axis y, respectively.



(a) ADOA estimation (b) MDOA estimation

Figure 2: *The overall network architectures of ADOA and MDOA estimation models.*

The coordinates of point O in the pixel coordinate system are represented by u_0 and v_0. And $\lambda = \begin{pmatrix} \frac{f}{dx} & 0 & u_0 & 0 \\ 0 & \frac{f}{dy} & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$ is the intrinsic parameters obtained by camera calibration. With pixel coordinates $p(u, v)$ and intrinsic parameters λ, we can get the camera coordinates $P(X_c, Y_c, Z_c)$ as follows:

$$P = \Phi(p, \lambda) \tag{3}$$

where Φ is the transformation function. As we use monocular camera in our study, the value of Z_c is always one. Finally, the angle between the vector O_cP and the axis O_cX_c represents the azimuth angle of one target speaker.

We evaluate the correctness of the spatial annotation algorithm by selecting ten points in the recording room and comparing their oracle angles with the annotated angles. The difference of angles is less than $0.5°$, which demonstrates that our proposed spatial annotation algorithm could accurately generate ground truth of DOAs.

2.2. Neural Network Architecture

We design two models to solve ADOA and MDOA estimation for multi-speaker situation. Figure 2 illustrates the overall network architectures of ADOA and MDOA estimation models.

For audio data, we extract Mel Filter Bank (Fbank) features, which are then fed into the ResNet [30] encoder to learn high-level feature representations. ResNet contains several hidden layers with each hidden layer consisting of several blocks. We conduct ablation experiments to find proper ResNet architecture for DOA estimation. Multi-Scale Temporal Convolution Network (MS-TCN) [31] is adopted to model the temporal structures within the signal by using several 1D temporal convolutions with different kernel sizes. MS-TCN includes four blocks with each block consisting of two sequential modules. Each module has three branches of temporal convolution with different kernel sizes. The kernel sizes are set to 3, 5, and 7, respectively with the channel number equal to 256.

There is a big difference of the inputs and outputs between ADOA estimation and MDOA estimation models. For ADOA estimation, only audio data is used as input and there are two branches of fully-connected (FC) layers in the output. We aim to predict the angle vector represented by *cosine* and *sine* values of the azimuth angle instead of the angle itself. Hence, L2 normalization is used to make sure that the magnitude of the an-

[1] https://github.com/ultralytics/yolov5

gle vector is equal to 1. Cosine loss function is adopted to evaluate the angle error between reference azimuth and predicted azimuth. We use permutation invariant training (PIT) strategy to solve the label alignment problem for multiple sound sources. And sigmoid activation is employed to predict diarization posterior of each speaker with binary cross entropy (BCE) loss calculated with the optimal alignment determined by minimizing the Cosine loss. The total loss function for ADOA estimation model is expressed as:

$$E^{\mathrm{ADOA}} = E^{\mathrm{ADOA}}_{\cos} + E^{\mathrm{ADOA}}_{\mathrm{bce}} \tag{4}$$

where $E_{\cos}$ and E_{bce} are written as:

$$E^{\mathrm{ADOA}}_{\cos} = \frac{1}{N}\sum_{n}\frac{\sum_t\sum_s y^s_{n,t}[1-\cos(\theta^{\phi^*(s)}_{n,t},\hat{\theta}^s_{n,t})]}{A_n} \tag{5}$$

$$E^{\mathrm{ADOA}}_{\mathrm{bce}} = \frac{1}{N\times S}\sum_{n}\frac{\sum_t\sum_s m_{n,t}\mathrm{bce}(y^{\phi^*(s)}_{n,t},\hat{y}^s_{n,t})}{B_n} \tag{6}$$

$$\phi^* = \mathop{\arg\min}_{\phi\in permu(S)}\frac{\sum_t\sum_s y^s_{n,t}[1-\cos(\theta^{\phi^*(s)}_{n,t},\hat{\theta}^s_{n,t})]}{A_n} \tag{7}$$

where N represents the batch size and $n = 1,2,...,N$. The two functions $\cos(\cdot)$ and $\mathrm{bce}(\cdot)$ are the Cosine loss and BCE loss, respectively. The frame index $t = 1,2,...,T$ and the speaker index $s = 1,2,...,S$ with $S = 6$ denoting that the number of output nodes is fixed with 6 in the diarization prediction branch. The angle reference and predicted azimuth angle are represented by $\theta^s_{n,t}$ and $\hat{\theta}^s_{n,t}$ for the n-th sample, t-th frame and s-th speaker, respectively. The diarization reference and predicted diarization posterior are represented by $y^s_{n,t}$ and $\hat{y}^s_{n,t}$, respectively. $A_n = \sum_t\sum_s y^s_{n,t}$ denotes the total activation number for the n-th sample. If the n-th sample and the t-th frame is silent then $m_{n,t} = 0$ otherwise $m_{n,t} = 1$, so $B_n = \sum_t m_{n,t}$ represents the number of non-silent frames in the n-th sample. And $permu(S)$ is a permutation of $1,2,...,S$. Cosine loss is calculated when the speakers are active while BCE loss is calculated for non-silent frames.

It is difficult to predict the azimuth angle for multiple speakers using only audio signals due to the label permutation problem. We propose to solve the multi-speaker DOA estimation task as speaker diarization by feeding the azimuth angles into the model input as shown in Figure 2(b). The MDOA estimation model learns to predict which speaker is active and the corresponding reference azimuth angle of active speaker is then selected as the prediction. It is much easier to choose an angle from a finite candidate set in MDOA than to make a direct prediction of continuous angle in ADOA. BCE loss function is adopted for MDOA estimation model computed on the active frames as follows:

$$E^{\mathrm{MDOA}}_{\mathrm{bce}} = \frac{\sum_n\sum_t\sum_s m_{n,t}\mathrm{bce}(y^s_{n,t},\hat{y}^s_{n,t})}{S'B} \tag{8}$$

$$S' = \max(S_n) \tag{9}$$

where $B = \sum_n B_n$ denotes the total number of non-silent frames in the current batch and S_n denotes the number of speakers in the n-th sample. The speaker index $s = 1,2,...,S'$ with S' denoting the maximum number of speakers among all samples in the batch.

Table 1: *Overview of the real dataset.*

Dataset	Training	Validation	Testing	Total
Duration (h)	97.5	11.2	7.5	116.2
Room	21	5	5	31
Speaker	200	21	23	244

3. Experimental Results and Analysis

3.1. Experimental Setup

To evaluate the effectiveness of our proposed MDOA estimation model, we conduct our experiments on a simulated dataset and a real dataset. A python package named pyroomacoustics [32] was adopted to simulate data by generating room impulse responses (RIRs). Audio signals are from TIMIT corpus [33] as it consists of a variety of speakers. About 125 hours of data were simulated for the simulated database with 111 hours of data for training and 14 hours of data for testing. The ground truth of DOA for each speaker was calculated with the locations of speakers and microphone array. The simulated audio data is clean without background noise.

Real dataset is recorded in home TV rooms with several people sitting and chatting. Audio data and video data are collected by far-field microphone array and far-field camera, respectively. The linear microphone array is placed 3-5m away from the speaker, consisting of 6 sample-synchronised omnidirectional microphones. And the distance between adjacent microphones is 35 mm. It is a subset of MISP2021-AVSR corpus [27] containing 116 hours of data, half of which is recorded with television on. Details about the real dataset is listed in Table 1. Note that there is no overlap of speakers and rooms among these three subsets, namely training, validation and testing.

In the test stage, the angle prediction is performed every 100 ms. Assume that the test utterance consists of L frames and the angle prediction of the l-th frame is represented by $\hat{\theta}_l$, we select p_l highest peaks of diarization posterior as active speaker prediction with p_l equal to the number of active speakers in the l-th frame. And the corresponding DOAs of the p_l speakers are selected as predicted DOAs. We use $Permutation\ Invariant\ Mean\ Absolute\ Error$ (PIMAE) to evaluate the difference between reference DOAs and predicted DOAs. PIMAE is calculated using Hungarian algorithm [34] to find the least angle distance given a set of ground truth angles and its respective predicted angles:

$$\mathrm{PIMAE}(^\circ) = \frac{\sum_l \mathcal{H}(\theta_l,\hat{\theta}_l)}{\sum_l p_l} \tag{10}$$

where $\mathcal{H}(\cdot)$ represents the Hungarian algorithm; θ_l and $\hat{\theta}_l$ denote the reference and predicted angle lists for the l-th frame. We also adopt $Accuracy$ (ACC) metric [14] to measure the percentage of correct predictions with a spatial localization error allowance of 20°:

$$\mathrm{ACC} = \frac{\sum_l\sum_{j=1}^{p_l}\mathbf{1}_{<20^\circ}}{\sum_l p_l} \tag{11}$$

where $\mathbf{1}$ denotes the indicator function and we use the localization error allowance according to the Sound Event Localization and Detection (SELD) task of Detection and Classification of Acoustic Scenes and Events (DCASE) 2021 Challenge[2].

[2] https://dcase.community/challenge2021/task-sound-event-localization-and-detection

Table 2: *Experimental results on real dataset for ADOA and MDOA estimation models with different parameters of ResNet encoder.*

Model	ADOA				MDOA			
	Validation		Testing		Validation		Testing	
	PIMAE (°)	ACC	PIMAE (°)	ACC	PIMAE (°)	ACC	PIMAE (°)	ACC
Channel32_Block1	21.04	0.50	22.50	0.45	15.40	0.59	16.88	0.55
Channel32_Block2	21.79	0.52	22.15	0.51	16.13	0.58	16.75	0.54
Channel64_Block2	21.04	0.55	24.51	0.51	15.98	0.59	16.38	0.57
Channel128_Block2	20.80	0.53	23.83	0.45	16.69	0.58	16.37	0.57
Channel128_Block3	27.40	0.47	30.99	0.38	15.10	0.62	16.96	0.56

We extract 64-dimensional Fbank features for audio data. For ADOA model, the angle vector and diarization posterior in the output layer are set with the shape of $(6, 2)$ and (6), respectively. If there are less than six people in one sample, the ground truth of angle and diarization will be padded with zero. For MDOA model, the shape of the diarization posterior in the output layer is determined by S', the maximum number of speakers among all samples in the batch, with the angle reference padded with zero when less than S' speakers are active in one sample.

3.2. Experimental Results

Table 2 lists the experimental results for ADOA and MDOA estimation models on real dataset with different parameters of ResNet encoder. Television background noises exist in audio signals as interferences. ResNet contains four hidden layers, and the number of channels in each layer is doubled progressively. The term "Channel32" denotes that the number of channels in the first hidden layer is set to 32 while the term "Block1" denotes that the number of blocks in each hidden layer is set to 1. As shown in Table 2, the MDOA estimation method outperforms the ADOA estimation method in all parameter configurations, which demonstrates that with spatial locations served as complementary information, it is more accurate to predict active speakers and their DOAs. The performance of ADOA estimation model varies greatly with different encoder parameters while the MDOA estimation model achieves similar and stable performances when using different number of channels and blocks, which proves the robustness of the MDOA estimation method. Take the third row of "Channel64_Block2" for example, the PIMAE decreases from 24.51° to 16.38° and the ACC increases from 0.51 to 0.57, achieving a relative 33% decrease in PIMAE and a relative 12% increase in ACC. Rather than make a direct prediction of continuous angle, it is easier to choose the correct azimuth angle from a candidate set with video data providing spatial information.

We list the experimental results on simulated dataset in Table 3. The parameter configuration used for ResNet encoder is "Channel64_Block2". Much better results are achieved on simulated dataset than real dataset. This is because that audio signals in Simulated Dataset are not corrupted with noise. For simulated data, the MDOA estimation model achieves 5.77° for PIMAE and 0.90 for ACC, yielding a relative 65% decrease in PIMAE and a relative 27% increase in ACC compared with ADOA estimation model.

Figure 3 shows a visualization of the prediction results for a test sample. The MDOA model outperforms the ADOA model in both diarization and angle predictions. At about 10 seconds, there are three concurrent speakers, which is difficult to directly predict the continuous angles and large localization distance is got by the ADOA model. However, accurate azimuth angles are predicted by the MDOA model with useful spatial location information of the speakers. From about 10 to 20 seconds, only one speaker (with index 2) is talking and the MDOA model can correctly predict the active speaker and the corresponding angle. Whereas the ADOA model makes a wrong prediction about active speaker at some segments with not consistent angles.

Table 3: *Experimental results on simulated dataset for ADOA and MDOA estimation models.*

Model	ADOA		MDOA	
	PIMAE (°)	ACC	PIMAE (°)	ACC
Channel64_Block2	16.67	0.71	5.77	0.90

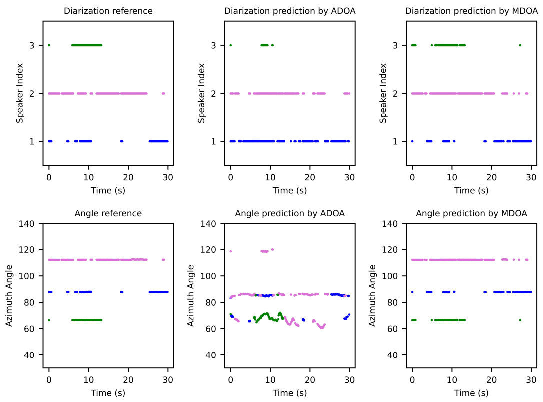


Figure 3: *Prediction results for a test sample.*

4. Conclusion

In this paper, we propose a deep learning based approach for multi-speaker DOA estimation using permutation-free loss function with audio-visual data. A novel spatial annotation method is adopted to generate the ground truth of DOA for each speaker with the video data according to the pin-hole camera model. By using spatial location information as complementary input, multi-speaker DOA estimation could be solved as a classification task of active speaker detection with permutation-free loss function, which provides a new perspective on multi-modal sound source localization. Experiments on real and simulated datasets demonstrate the superior performance of our proposed model compared to audio-only DOA estimation model in terms of both PIMAE and ACC metrics.

5. References

[1] H.-Y. Lee, J.-W. Cho, M. Kim, and H.-M. Park, "DNN-based feature enhancement using DOA-constrained ICA for robust speech recognition," *IEEE Signal Processing Letters*, vol. 23, no. 8, pp. 1091–1095, 2016.

[2] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[3] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. IROS*, vol. 2. IEEE, 2003, pp. 1228–1233.

[4] F. Keyrouz, "Advanced binaural sound localization in 3-D for humanoid robots," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 9, pp. 2098–2107, 2014.

[5] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 611–623, 2017.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[7] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Proc. ICASSP*, vol. 1. IEEE, 2007, pp. I–121.

[8] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 37, no. 7, pp. 984–995, 1989.

[9] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[10] Z.-M. Liu, C. Zhang, and S. Y. Philip, "Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections," *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 12, pp. 7315–7327, 2018.

[11] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A review of sound source localization with deep learning methods," *arXiv preprint arXiv:2109.03465*, 2021.

[12] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. EUSIPCO*. IEEE, 2018, pp. 1462–1466.

[13] L. Perotin, R. Serizel, E. Vincent, and A. Guerin, "CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[14] W. He, P. Motlicek, and J.-M. Odobez, "Neural network adaptation and data augmentation for multi-speaker direction-of-arrival estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1303–1317, 2021.

[15] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection," *arXiv preprint arXiv:2101.02919*, 2021.

[16] Z. Tang, J. D. Kanu, K. Hogan, and D. Manocha, "Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks," in *Proc. Interspeech*, 2019, pp. 654–658.

[17] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *Proc. ICASSP*. IEEE, 2021, pp. 626–630.

[18] H. Chen, J. Du, Y. Hu, L.-R. Dai, B.-C. Yin, and C.-H. Lee, "Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement," *Neural Networks*, 2021.

[19] O. Köpüklü, M. Taseska, and G. Rigoll, "How to design a three-stage architecture for audio-visual active speaker detection in the wild," *arXiv preprint arXiv:2106.03932*, 2021.

[20] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proc. ECCV*, 2018, pp. 247–263.

[21] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon, "Learning to localize sound source in visual scenes," in *Proc. CVPR*, 2018, pp. 4358–4366.

[22] V. Sanguineti, P. Morerio, A. Del Bue, and V. Murino, "Audio-visual localization by synthetic acoustic image generation," in *Proc. AAAI*, vol. 35, no. 3, 2021, pp. 2523–2531.

[23] X. Qian, M. Madhavi, Z. Pan, J. Wang, and H. Li, "Multi-target DoA estimation with an audio-visual fusion mechanism," in *Proc. ICASSP*. IEEE, 2021, pp. 4280–4284.

[24] D. Berghi, A. Hilton, and P. J. Jackson, "Visually supervised speaker detection and localization via microphone array," *arXiv preprint arXiv:2203.03291*, 2022.

[25] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "Av16. 3: An audio-visual corpus for speaker localization and tracking," in *Proc. MLMI*. Springer, 2004, pp. 182–195.

[26] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. ICRA*. IEEE, 2018, pp. 74–79.

[27] H. Chen, H. Zhou, J. Du, C.-H. Lee, J. Chen, S. Watanabe, S. M. Siniscalchi, O. Scharenborg, D.-Y. Liu, B.-C. Yin, J. Pan, J.-Q. Gao, and C. Liu, "The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results," in *Proc. ICASSP*, 2022.

[28] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[29] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. ICIP*. IEEE, 2017, pp. 3645–3649.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[31] B. Martinez, P.-C. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *Proc. ICASSP*, 2020, pp. 6319–6323.

[32] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: a python package for audio room simulations and array processing algorithms. corr abs/1710.04196 (2017)."

[33] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[34] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.