Audio-Visual Cross-Attention Network for Robotic Speaker Tracking

Xinyuan Qian, *Member, IEEE*, Zhengdong Wang, Jiadong Wang, Guohui Guan, and Haizhou Li, *Fellow, IEEE*

***Abstract*—Audio-visual signals can be used jointly for robotic perception as they complement each other. Such multi-modal sensory fusion has a clear advantage, especially under noisy acoustic conditions. Speaker localization, as an essential robotic function, was traditionally solved as a signal processing problem that now increasingly finds deep learning solutions. The question is how to fuse audio-visual signals in an effective way. Speaker tracking is not only more desirable, but also potentially more accurate than speaker localization because it explores the speaker's temporal motion dynamics for smoothed trajectory estimation. However, due to the lack of large annotated dataset, speaker tracking is not well studied as speaker localization. In this paper, we study robotic speaker Direction of Arrival (DoA) estimation with a focus on audio-visual fusion and tracking methodology. We propose a Cross-Modal Attentive Fusion (CMAF) mechanism, which explores self-attention to learn intra-modal temporal dependencies, and cross-attention mechanism for inter-modal alignment. We also collect a realistic dataset on a robotic platform to support the study. The experimental results demonstrate that our proposed network outperforms the state-of-the-art audio-visual localization and tracking methods under noisy conditions, with an improved accuracy of 5.82% and 3.62% at SNR $= -20$ dB, respectively.**

Manuscript received 1 August 2021; revised 5 March 2022, 1 July 2022, and 13 October 2022; accepted 17 November 2022. Date of publication 2 December 2022; date of current version 23 December 2022. This work was supported in part by The Science and Engineering Research Council, Agency for Science, Technology and Research (A*STAR), Singapore, through the National Robotics Program under Human-Robot Interaction Phase 1 under Grant No. 192 25 00054, in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen), in part by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen under Grant No. B10120210117-KP02, Grant UDF01002333, and Grant UF02002333, in part by the internal project of Shenzhen Research Institute of Big Data under Grant T00120220002, and in part by the Research Foundation of Guangdong Province under Grant 2019A050505001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Keisuke Kinoshita. *(Corresponding author: Jiadong Wang.)*

Xinyuan Qian is with the Department of Computer Science and Technology, University of Science and Technology Beijing, Beijing 100083, China, also with the Chinese University of Hong Kong, Shenzhen 518172, China, and also with the Shenzhen Research Institute of Big data, Shenzhen 51872, China (e-mail: qianxy@ustb.edu.cn).

Zhengdong Wang and Jiadong Wang are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: e0572572@u.nus.edu; jiadong.wang@u.nus.edu).

Guohui Guan is with the Department of Electrical Engineering and Computer Sciences, Univeristy of California at Berkeley, Berkeley, CA 94702 USA (e-mail: guohuiguan@berkeley.edu).

Haizhou Li is with the Guangdong Provincial Key Laboratory of Big Data Computing, Chinese University of Hong Kong, Shenzhen 518172, China, also with the Shenzhen Research Institute of Big data, Shenzhen 51872, China, also with the Department Electrical and Computer Engineering, National University of Singapore, Singapore 119077, and also with the University of Bremen, 28359 Bremen, Germany (e-mail: haizhouli@cuhk.edu.cn).

Digital Object Identifier 10.1109/TASLP.2022.3226330

***Index Terms*—Speaker tracking, direction-of-arrival, audio-visual fusion, cross-modal attention.**

I. Introduction

SPEECH is one of the most significant mediums of communication between humans and machines. Speaker localization and tracking are helpful in many human-robot interaction (HRI) applications, such as speech enhancement [1] and separation [2], music information processing [3], [4]. They can be estimated via the arrival time or energy level differences between signals from two spatially separated microphones [5], [6]. The Signal Processing (SP)-based Sound Source Localization (SSL) techniques are analytical solutions under certain assumptions about the signal, noise type, and environmental conditions, which may vary in practice. As an alternative, recently, researchers have proposed Deep Learning (DL)-based approaches that build machine learning models to bypass explicit sound propagation modeling and other required priors [7]. Those approaches model the mapping from acoustic features to speaker locations, and have demonstrated significant performance gain over the SP-based methods, unless the training and testing data are of different conditions [8]. Despite much progress, the techniques that solely rely on acoustic signals are always affected by adverse acoustic conditions [9].

Humans use multi-modal cues to explore, capture, and perceive the real world. In addition to audio, vision is another primary stream that conveys significant information [10]. Many studies confirmed the advantages of audio-visual fusion, such as visually indicated sound separation [11], video-infused audio in-painting [12], and embodied navigation [13]. If a visually tracked object emits sound, its location can also be inferred using SSL techniques. Audio and vision offer complementary characteristics [9]. For example, one may achieve improved tracking accuracy by using sound to estimate the speaker trajectories in unseen regions of a camera [14] or by using visual cues to predict a target location during silent periods [15].

Since audio and vision operate in different spaces, i.e., audio in a 3D space and video on a 2D image plane, in most audio-visual localization-based applications, sensor calibration information, which constructs the mapping between different coordinates, is required. Using calibrated sensors, one can align a DoA to specific 2D locations on an image plane [16], [17], or map a target image location to a 3D spatial space [18], [19]. However, such a calibration process is labor-intensive [20] and precise calibration information is hard to come by.

TABLE I
SUMMARY OF THE SIGNIFICANT STATE-OF-THE-ART (SOTA) DL-BASED SPEAKER DOA LOCALIZATION AND TRACKING APPROACHES AND OUR PROPOSED METHODS (INDICATED WITH "PROP.")

Reference	Dataset	Modality	Loc	Track	Input features	Network
1. [21]	simulation [U]	audio	✓		GCC-PHAT	MLP
2. [22]	simulation [U]	audio	✓		ITD, ILD	MLP
3. [23]	real [U]	audio	✓		GCC, cepstrogram	CNN
4. [24]	simulation [U]	audio	✓		STFT	CNN
5. [25]	real [U]	audio	✓		MUSIC eigenvector	MLP
6. [26]	simulation [U]	audio	✓		Phase STFT	CNN
7. [27]	simulation [U]	audio		✓	Magnitude and phase STFT	CRNN
8. [28]	SSLR	audio	✓		GCC-PHAT	MLP
9. [29]	SSLR	audio	✓		Real and imaginary STFT	CNN ResNet
10. [30]	RSL	audio	✓		ITD	CNN
11. [31]	LOCATA	audio		✓	STFT	CRNN
12. [32]	DCASE	audio		✓	GCC-PHAT, log-mel spectrogram	CRNN
13. [9]	SSLR	audio-visual	✓		GCC-PHAT, face detection	MLP
14. prop. AV-CRNN	AVRI	audio-visual		✓	GCC-PHAT, log-mel spectrogram, face detection	CRNN
15. prop. CMAF	AVRI	audio-visual		✓	GCC-PHAT, log-mel spectrogram, face detection	Attention

"LOC" denotes the speaker localization task, while "track" denotes speaker tracking (the letter [U] indicates the data is publically unavailable).

Not to be burdened with such sensor calibration, one solution is using DL techniques to transform the localization task into a data-driven optimization problem, where a model gradually learns the mapping from input signals to speaker location ground truth [28]. It is noted that the success of DL techniques is based on a large amount of training data. However, most existing datasets provide data of audio-only localization [28], [33], [34], or audio-visual localization, but with a short recording duration [35]. The scarcity of datasets hinders the DL-based speaker tracking studies [9].

In this paper, we tackle the speaker location estimation problem with signals captured by multi-modal sensors mounted on a real robot. We are particularly interested in exploiting three unique properties of audio-visual signals for speaker tracking under noisy acoustic conditions: (1) both audio and visual signals are sequential; (2) the speaker locus at either an audio or visual frame is temporally correlated to that at the neighboring frames; and (3) audio and visual signals from the same speaker are highly correlated as far as the speaker locus is concerned. To this end, we propose an audio-visual cross-attention network. We consider our work is of significant importance with the following contributions.

1) We make the first attempt at audio-visual speaker spatial DoA estimation using DL-based techniques and a tracking strategy.
2) We propose a Cross-Modal Attentive Fusion (CMAF) architecture that explores the self-attention mechanism to learn intra-modal temporal dependencies and the cross-attention mechanism for inter-modal alignment.
3) We develop a DL-enabled audio-visual dataset with signals captured by a real robot. The monocular image sequences, multi-channel microphone array signals, and speaker 3D location annotations are provided.
4) We demonstrate that the proposed CMAF outperforms the state-of-the-art uni-modal and multi-modal approaches.[1]

[1]We will release the dataset and the source code.

The rest of the paper is organized as follows. Section II gives a comprehensive review of the related works. Section III formulates the research problem. Section IV first characterizes the audio and video processing, and then elaborates the proposed audio-visual tracking network. Section V summarizes the existing DL-based datasets, followed by a detailed description of our self-collected audio-visual dataset. Experiments are conducted and analyzed in Section VI. Limitations and future works are discussed Section VI-H. Finally, we conclude in Section VII.

II. RELATED WORK

Let us start with a review of the significant SSL approaches. Then, we discuss how speaker tracking is different from the localization task, and how the incorporation of vision can help improve performance. The most relevant DL-based SSL and tracking approaches are summarized in Table I.

A. Speaker Localization

Speaker localization using sound is a well-established area of research. Conventional SP-based speaker localization methods generally belong to four categories: (1) time delay estimation, e.g., Time Difference of Arrival (TDoA) (2) sub-space methods, (3) beamforming methods, and (4) histogram analysis methods. Among the four categories, time delay-based methods attract the most attention. In particular, Generalized Cross Correlation (GCC) estimates the sound location at the maximum correlation between the inter-microphone signals. As only phase information conveys TDoA, Generalized Cross Correlation with Phase Transform (GCC-PHAT) eliminates the amplitude and uses only phase of the cross spectrum for robustness against noise interference [5]. Studies show that speaker localization benefits from the use of a multi-channel microphone array [36]. By aggregating information from multiple microphone pairs, we overcome errors from an individual microphone pair. As an example, Steered Response Power PHAse Transform (SRP-PHAT) [6] promotes this concept by estimating the

speaker DoA at the hypothesis with the maximized accumulated GCC-PHAT value. Despite much progress, SP-based methods remain to be improved under adverse acoustic conditions [28]. Nonetheless, the SP-based features are shown to be effective under controlled acoustic conditions.

Recently, DL-based SSL approaches, also called neural solutions, show superiority over the conventional SP-based approaches for their generalization ability under reverberation and noise conditions [21], [29], [37]. The neural solutions employ a network architecture that learns to map input acoustic features to sound source locations. They differ in terms of feature representation and network architecture. As listed in Table I, SP-based acoustic spatial features are mostly used, e.g., GCC-PHAT [21], [23], [28], Short-Time-Fourier-Transform (STFT) variants [24], [26], [27], [29], [31], [32], eigenvectors of spatial covariance matrix [25], Inter-channel Time Delay (ITD) and Inter-channel Level Differences (ILD) [22]. Several common network architectures are studied that include Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), and CNN ResNet (with residual block). In this paper, we further study neural solutions by exploring a novel multi-modal fusion technique.

B. Speaker Tracking

In speaker localization, we tackle frame-level localization, which does not consider sound source motion dynamics. Unlike speaker localization, speaker tracking not only locates the speaker, but also follows the speaker's motion over consecutive frames. Speaker tracking is required in many real-world robotic applications. As it benefits from sound source motion dynamics, it potentially provides more accurate position information than speaker localization.

Traditional parametric tracking approaches often rely on a recursive Bayesian estimation paradigm, such as Kalman Filter (KF) [38] and Particle Filter (PF) [16], [18], [39]. PF is a stochastic inference strategy that relies on sequential importance sampling to approximate the target posterior distribution using weighted samples. To complement audio cues, vision is shown effective [17], [18], [40]. For example, [18] uses PF for speaker tracking, where audio-visual fusion is implicitly handled during the recursive particle likelihood update stage. [41] adopts the Extended Kalman Filter (EKF) paradigm to achieve dynamic weighting of audio-visual streams according to the instantaneous sensor reliability measure. However, those parametric approaches require prior knowledge of the scene, such as spatial distribution and respective velocity of a sounding object, or manual tuning of specific hyperparameters according to the scene composition and dynamics.

Unlike parametric approaches, DL-based trackers do not require problem-specific parameter settings. For example, the Recurrent Neural Network (RNN) learns to capture long-term information. Others [27], [32] use consecutively stacked CNN and RNN layers, that is referred to as Convolutional Recurrent Neural Network (CRNN). In terms of audio-visual fusion, it is common to track a sound source on an image plane. For example, [42] generates a heat map to visualize the image location of a sounding object. [43] represents an audible and visible object as the trajectory of a potential sound source through space and time. When it comes to the spatial DoA domain, few studies have tackled the audio-visual tracking problem, let alone their fusion mechanism. One of the reasons that hinder the research is the lack of a large annotated dataset. As the first attempt, we tackle this problem with a simple but effective visual simulation method in our prior work [9] by taking advantage of sensor calibration information, where a straightforward audio-visual weighting mechanism is studied. Nonetheless, in [9], we did not investigate the temporal dynamics of speaker motion as the involved sound source is stationary.

It is tempting to combine parametric methods with DL techniques for speaker tracking. For example, [44] proposes Backprop Kalman Filter (BKF), which downsizes the observations to a lower-dimensional space through a nonlinear encoding network. Inspired by [44], a DL-based KF extension for audio-visual speaker tracking was introduced in [45] where dynamic multi-modal stream weights are jointly learned during model optimization. Despite comparable performance with standard KF-based methods, this approach still requires manual parameter tuning.

C. Summary

Most of the previous DL-based studies are conducted on synthetic recordings with a spatially static sounding object, either using audio signals [21], [22], [23], [24], [25], [26], [27], [29], [30], [31], [32], or audio-visual signals [9]. None of them explore the problem of DL-based audio-visual speaker spatial DoA tracking, which is the focus of this paper. Fig. 1 illustrates a general block diagram of our proposed architecture. In addition to multi-modal input data capture, it consists of three stages: (1) the front-end feature processing block, (2) the multi-modal fusion block, and (3) the speaker DoA classifier. We discuss this in further detail next.

III. Problem Formulation

We start by defining a few notations and definitions. Unless otherwise specified, matrices are in bold uppercase, e.g., $\mathbf{X}$, $\mathbf{Y}$, vectors are in bold lowercase, e.g., $\boldsymbol{x}, \boldsymbol{y}$, variables are in lowercase, e.g., x, y, and functions are in calligraphic font, e.g., $\mathcal{F}$. Let us denote the synchronized audio and image sequences captured by an M-channel microphone array and a camera mounted on a robotic platform, as $\boldsymbol{s}^{1:M}$ and $\mathbf{I}$. We aim to estimate DoA of a speaker, denoted as $\theta_t \in [1^\circ, 360^\circ]$, in each frame $t = 1, .., T$. We formulate the problem as a regression problem, which seeks to predict DoA as one of the 360 discrete classes, denoted as $\theta = \{j | j \text{ is an integer, and } 1 \leq j \leq 360\}$.

As DoAs are spatially continuous, instead of adopting the one-hot output coding, we use a Gaussian-like vector [28], denoted as $p_t(\theta)$, to represent the *posterior probability likelihoods* of a speaker presence in the direction of θ_t,

$$p_t(\theta) = \exp\left(-\frac{|\theta - \theta_t|^2}{\sigma_\theta^2}\right) \tag{1}$$

where $p_t(\theta)$ is centered on the ground truth θ_t with a standard deviation σ_θ. To be noted, the prefix $\frac{1}{\sqrt{2\pi\sigma_\theta}}$ of the Gaussian

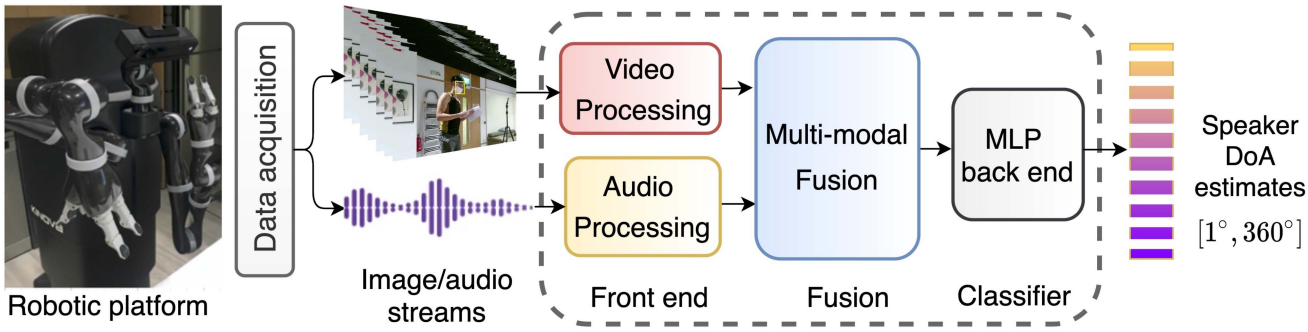


Fig. 1. The proposed network architecture for speaker DoA estimation given the audio and video input streams captured by a robotic platform. Apart from the input multi-modal data, it consists of three stages: (1) audio and video front-end feature processing block (red and yellow rectangles), (2) multi-modal fusion block (blue rectangle), and (3) MLP back-end DoA classifier (grey rectangle). On the right-most part, we use a color bar with the varying color from purple to yellow indicating the DoA variations from 1° to 360°.

distribution is dropped in Eq. (1) as well as in [28], since it is a constant which has no impact on the algorithm.

As will be discussed later, we will adopt a DL technique to learn the mapping from the multi-modal inputs to the speaker DoA *posterior probability*,

$$\hat{p}_t(\theta) = \mathcal{F}(\mathbf{s}^{1:M}, \mathbf{I}; \Omega) \tag{2}$$

where $\mathcal{F}(\cdot)$ is the proposed network with Ω the learnable parameters. In this way, the locus of the speaker is approximated by the DoA value of the highest probability,

$$\hat{\theta}_t = \arg\max_{\forall\theta} \ \hat{p}_t(\theta) \tag{3}$$

We adopt the Mean Square Error (MSE) loss for the posterior probability-based coding in (1),

$$\mathcal{L}_t(\boldsymbol{s}^{1:M}, \mathbf{I}; \Omega) = \sum_{\theta=1}^{360} ||p_t(\theta) - \hat{p}_t(\theta)||_2^2 \tag{4}$$

For brevity, we drop the time index t hereafter.

IV. Proposed Methods

Audio-visual signals provide rich spatial and temporal cues to track the speaker in the scene [10]. It was shown that directly concatenating the frame-level multi-modal features is vulnerable to temporal misalignment and uni-modal outliers [46]. We start with the basics of audio and visual characterization. We then formulate CMAF, the cross-modal attentive fusion architecture for speaker tracking.

A. Audio Processing

In acoustic speaker localization, time-delay based methods have achieved remarkable success with widespread applications thanks to their simple and effective computation. In particular, GCC-PHAT, which facilitates the TDoA estimation between any two arbitrary microphones, is more robust to noise and room reverberations [47]. Herein, we use it as the acoustic feature. Let $\mathbf{S}_{n_1}$ and $\mathbf{S}_{n_2}$ denote the STFT of short-time audio signals at a microphone pair $\{(n_1, n_2), \forall n_1 < n_2 \leq N\}$ where N is the total number of microphones. We use $m \leq M$ to index the microphone pair with M the total pair number. Then, the GCC-PHAT feature with time delay τ is computed as,

$$g^m(\tau) = \sum_k \left(\frac{\mathbf{S}_{n_1}(k)\mathbf{S}^*_{n_2}(k)}{|\mathbf{S}_{n_1}(k)\mathbf{S}^*_{n_2}(k)|} e^{i\frac{2\pi k}{N_s}\tau} \right) \tag{5}$$

where i indicates the imaginary unit, $*$ denotes the complex conjugate, k indicates the frequency bin, and N_s is the STFT length. GCC-PHAT is a feature in the time domain that peaks at the actual time delay. It is noted that its performance is unstable for signals under a low Signal-to-Noise Ratio (SNR).

Since frequency-domain DoA estimation methods are more robust than time-domain methods, in particular in the presence of background noise and reverberation [48], we also incorporate log-mel spectrogram by passing STFT through a mel filter bank to provide a compact representation. The mel Spectrogram is supplementary to GCC-PHAT that operates in the frequency domain. We use $\mathbf{M}^n$ to denote the one calculated in the n-indexed microphone ($n \leq N$). Since DL features are automatically deduced and optimally tuned for the desired DoA outcome, incorporating features from both domains brings the greatest benefits.

In summary, we extract two location-related features from the audio processing block: the time-domain GCC-PHAT features, and the frequency-domain log-mel spectrogram features.

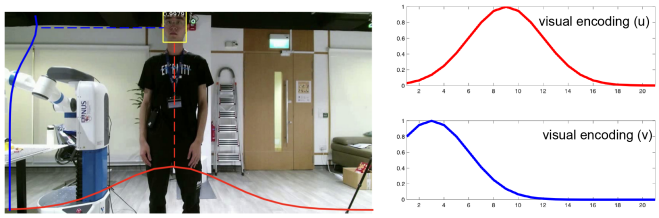
B. Video Processing

The advances in object detection have enabled many real-world applications, such as autonomous driving, robot vision, and surveillance [49]. Human face detection has served as the front-end of many audio-visual tasks, such as speaker tracking [18], diarization [50], and speaker extraction [51].

Object detection results are typically represented by bounding boxes [52]. In this paper, we adopt the tracking-by-detection methodology [53], and use face detection as the front-end for video encoding. Let us denote a detected face bounding box at t as

$$\mathbf{b}_t = (u_t, v_t, w_t, h_t) \tag{6}$$

where (u_t, v_t) indicates the image position of the top-left corner, and (w_t, h_t) the width and height of the bounding box.



(a) (b)

Fig. 2. The encoded visual features from an image face detection bounding box. The red and blue curves follow a Gaussian distribution which centers at the face detection point. The Gaussian standard deviation is proportional to the detection width and height.

We propose to encode the visual locus of a speaker as the concatenation of two Gaussian-like vectors where $\rho_t(u)$ and $\rho_t(v)$ represent the *likelihoods* of a speaker visually present along the image's horizontal axis u and vertical axis v. We use the same formulation as [9],

$$\rho_t(u) = \begin{cases} \exp\left(-\frac{|u-\mu_{u,t}|^2}{\sigma_u^2}\right) & \mathbf{b}_t \neq \emptyset, \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $\mu_{u,t} = u_t + \frac{1}{2}w_t$ is the horizontal center of $\mathbf{b}_t$, and σ_u is the standard deviation. The vertical representation $\rho_t(v)$ has the same format as $\rho_t(u)$ by replacing u with v, $\mu_{u,t}$ with $\mu_{v,t} = v_t + \frac{1}{2}h_t$, and σ_u with σ_v in (7). Notice that when there is no face detection, the visual features are set to all-zero vectors.

In Fig. 2, we show an example of the encoded visual features of a face detection result, where the red and blue curves indicate the posterior probability of the horizontal and vertical positions of the bounding box. The curves peak at the central points of the detected face.

C. Cross-Modal Attentive Fusion (CMAF)

We extract the location-related features from audio and video, respectively. Then, we design a deep module to refine the location-related cues by taking into account the temporal variations among cross-modalities.

Self-attention in neural networks learns which element to focus on in a sequential signal, while cross-attention learns the interaction between audio and visual signals. Self-attention makes use of historical data to make a decision, that is particularly useful for speaker tracking because it considers a receptive field instead of a single frame, and benefits from the temporal correlation properties of audio and visual signals separately. Cross-attention exploits the synchronization information between audio and visual signals, which takes advantage of the audio-visual correlation properties. We believe that by employing both self-attention and cross-attention in the proposed CMAF network, we make full use of multi-modal cues.

We detail the architecture of CMAF in Fig. 3. For brevity, we exclude the input streams and DoA outputs from the figure. We first apply a fully-connected (FC) layer to project the multi-channel information from each of the three individual modalities (i.e., M-channel GCC-PHAT $g^m(\tau)$, N-channel log-mel spectrogram $\mathbf{M}^n$, and two-channel visual features) into the latent representations, denoted as $\mathbf{X}_\rho$, $\mathbf{X}_g$, and $\mathbf{X}_\mathbf{M}$. Then, we adopt six parallel CMAF blocks which enumerate all order combinations of the three latent representations. Finally, the resulting cross-attentive features are concatenated for speaker DoA estimation.

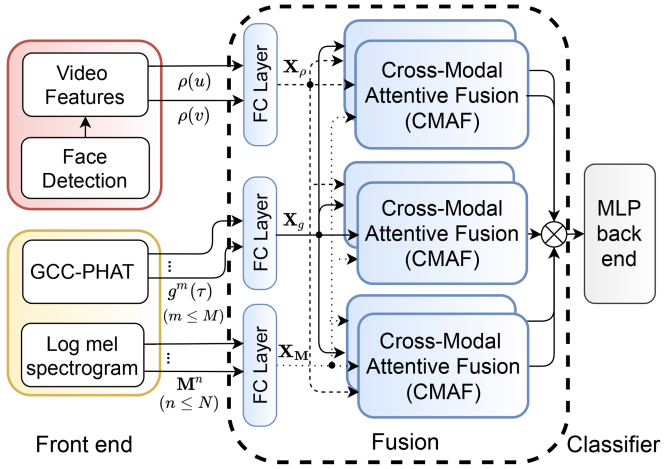


Fig. 3. The network architecture of the proposed CMAF model (we ignore the input streams and DoA estimates). The model first uses separate Fully-Connected (FC) layers to project each of the three encoded multi-channel features (two dimensional face features $\rho(u), \rho(v)$, acoustic GCC-PHATs $g^m(\tau)$ with $m \leq M$, and log-mel spectrogram features $\mathbf{M}^n$ with $n \leq N$) into the latent representations, $\mathbf{X}_\rho$, $\mathbf{X}_g$ and $\mathbf{X}_\mathbf{M}$. Then, six parallel CMAFs are applied to any two modalities. The resulting cross-attentive features are concatenated to the back-end speaker DoA classifier ($\otimes$ indicates the concatenation operation).

Specifically, in Fig. 4, we illustrate one CMAF block that deals with two arbitrary inputs, denoted $\mathbf{X}_\alpha$ and $\mathbf{X}_\beta$. Two attention modules, Multi-head Self-Attention (M-SAtt) and Multi-head Cross-Attention (M-CAtt), are detailed in the left-most and right-most panels.

1) Self-Attention (SAtt): SAtt allows the model to attend to all features at the scale of the input sequences. As described in [54], it maps a query to a set of key-value pairs. For instance, given a sequential input $\mathbf{X}_\alpha$ from modality α, the scaled dot-product self-attention matrix, which represents the energy, is computed as,

$$SAtt(\mathbf{X}_\alpha) = softmax\left(\frac{\mathbf{X}_\alpha \mathbf{W}_{Q_\alpha} \mathbf{W}_{K_\alpha}^\intercal \mathbf{X}_\alpha^\intercal}{\sqrt{d_k}}\right) \mathbf{X}_\alpha \mathbf{W}_{V_\alpha} \tag{8}$$

$$= softmax\left(\frac{\mathbf{Q}_\alpha \mathbf{K}_\alpha^\intercal}{\sqrt{d_k}}\right) \mathbf{V}_\alpha \tag{9}$$

where W_{Q_α}, W_{K_α} and W_{V_α} are the trainable parameters of three individual FC layers which project $\mathbf{X}_\alpha$ into a common space. $\mathbf{Q}_\alpha = \mathbf{X}_\alpha \mathbf{W}_{Q_\alpha} \in \mathbb{R}^{T_\alpha \times d_q}$ is a set of queries, $\mathbf{K}_\alpha = \mathbf{X}_\alpha \mathbf{W}_{K_\alpha} \in \mathbb{R}^{T_\alpha \times d_k}$ is a set of keys, and $\mathbf{V}_\alpha = \mathbf{X}_\alpha \mathbf{W}_{V_\alpha} \in \mathbb{R}^{T_\alpha \times d_v}$ is the corresponding set of values (where $T_\star$ and $d_\star$ denote the sequence length and feature dimension, respectively). The *softmax*$(\cdot)$ is used to normalize the weights.

To make full use of self-attention, a multi-head attention mechanism is applied [54], which enables a model to attend to various projections of an input. The resulting multi-head

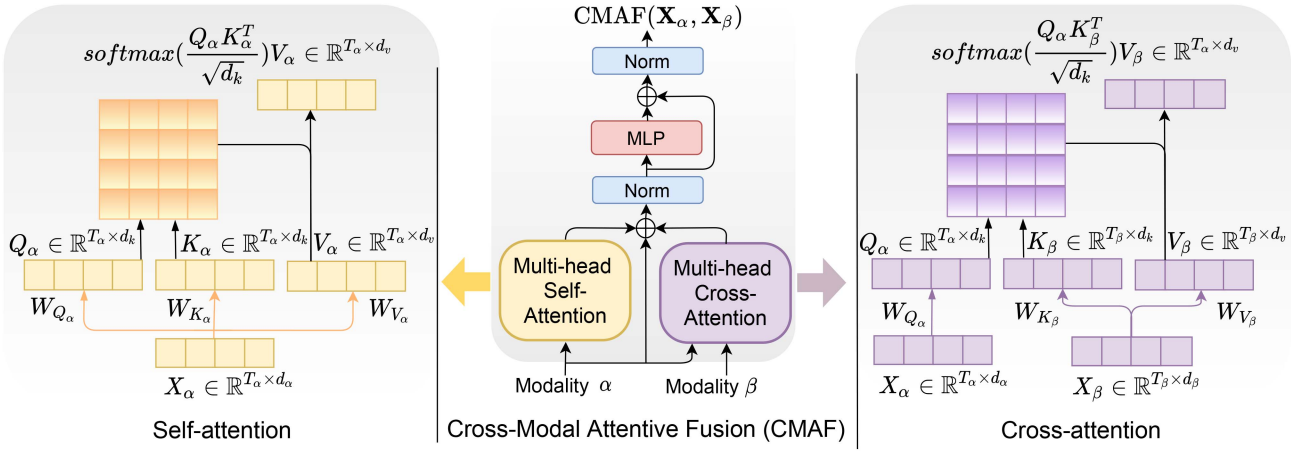


Fig. 4. An individual CMAF block (Fig. 3) given the input streams X_α and X_β from modality α and β. The self-attention and the cross-attention block are displayed on the left-most and the right-most sides, respectively ($\oplus$ indicates the pair-wise addition operation).

self-attention, denoted as M-SAtt, is then formulated as,

$$M - SAtt(\mathbf{X}_\alpha) = cat\left(SAtt_1(\mathbf{X}_\alpha), \ldots, SAtt_H(\mathbf{X}_\alpha)\right)\mathbf{W}_\alpha \tag{10}$$

where *cat* denotes concatenation, H is the total number of heads, $\mathbf{W}_\alpha$ is the set of trainable parameters that have been associated with the concatenated $SAtt(\cdot)$ representation. Moreover, each self-attention module indexed by $h \leq H$ consists of independent trainable parameters $\mathbf{W}_{Q_\alpha,h}$, $\mathbf{W}_{K_\alpha,h}$ and $\mathbf{W}_{V_\alpha,h}$, respectively.

2) Cross-Attention (CAtt): Technically, CAtt queries the feature of one modality to the other modality and vice versa. Since the learned audio and visual features share the same spatial correspondence, we use the CAtt module to achieve collaborative fusion while preserving the intra-modal characteristics. The right-most part of Fig. 4 illustrates the details. With the cross-attention module, each modality keeps updating its features via external information from the other modality.

Given the latent representations $\mathbf{X}_\alpha$ and $\mathbf{X}_\beta$ of modalities α and β, the scaled dot-product cross-attention (i.e., a variant of cross-correlation) achieves latent adaptation across modalities through a formulation similar to (8). This process is formulated as,

$$CAtt(\mathbf{X}_\alpha, \mathbf{X}_\beta) = softmax\left(\frac{\mathbf{X}_\alpha \mathbf{W}_{Q_\alpha} \mathbf{W}_{K_\beta}^\intercal \mathbf{X}_\beta^\intercal}{\sqrt{d_k}}\right)\mathbf{X}_\beta \mathbf{W}_{V_\beta} \tag{11}$$

$$= softmax\left(\frac{\mathbf{Q}_\alpha \mathbf{K}_\beta^\intercal}{\sqrt{d_k}}\right)\mathbf{V}_\beta \tag{12}$$

where $\mathbf{W}_{Q_\alpha}$, $\mathbf{W}_{K_\beta}$ and $\mathbf{W}_{V_\beta}$ are the projection parameters of the query and the key-value pairs, respectively.

Accordingly, the M-CAtt mechanism follows the same principle as the self-attention counterpart (10),

$$M{-}CAtt(\mathbf{X}_\alpha, \mathbf{X}_\beta) = cat\left(CAtt_1, \ldots, CAtt_H\right)\mathbf{W}_{\alpha,\beta} \tag{13}$$

where we abbreviate the cross-attention representation (11) as $CAtt_{1,\ldots,H}$ with the subscript indexing the head number, and $\mathbf{W}_{\alpha,\beta}$ indicates the set of trainable parameters.

3) CMAF: We jointly model the temporal recurrence, co-occurrence, and synchrony of the multi-modal features in the CMAF block through the self-attention and cross-attention modules. The intermediate CMAF features are computed as,

$$\tilde{\mathbf{X}}_{\alpha,\beta} = \mathbf{X}_\alpha + M{-}SAtt(\mathbf{X}_\alpha) + M{-}CAtt(\mathbf{X}_\alpha, \mathbf{X}_\beta) \tag{14}$$

where the skip connections can help preserve the identity information from the input stream $\mathbf{X}_\alpha$.

The final CMAF outcome is formulated as,

$$CMAF(\mathbf{X}_\alpha, \mathbf{X}_\beta) = LN\left(MLP\left(LN\left(\tilde{\mathbf{X}}_{\alpha,\beta}\right)\right)\right) \tag{15}$$

where LN stands for layer normalization, and *MLP* includes two FC layers with a ReLU activation function.

The overall CMAF block models the interactions between each cross-modal pair. With three input multi-modal streams, six parallel CMAFs are adopted (as illustrated in Fig. 3) where all the cross-attentive features are concatenated to the back-end DoA classifier at the last step.

V. Dataset

We now first review the existing datasets which are suitable for DL-based speaker DoA estimation, to motivate the design of a new dataset for Audio-Visual Robotic Interface (AVRI).

A. Existing Datasets

We summarize all datasets in Table II and characterize them in terms of modality, sensing platform, type of sounding source, recording duration, and data annotation.

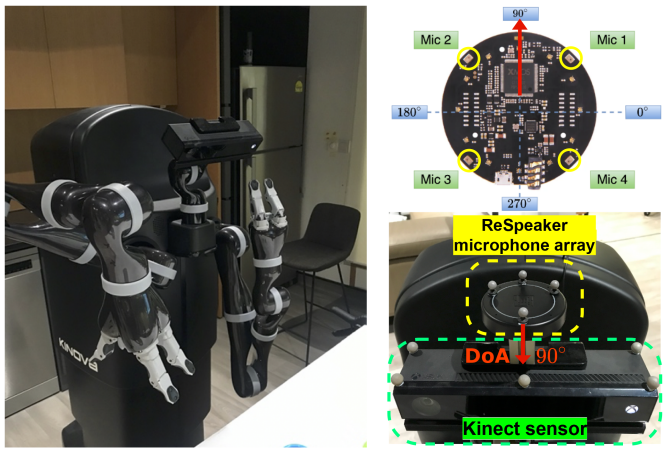
1) The LOCalization And TrAcking (LOCATA) dataset [55] is recorded with different array configurations in a high-reverberant indoor environment. Annotations include microphones and target 3D locations and Voice Activity Detector (VAD) labels.
2) The Realistic Speech Localization (RSL) dataset [34] is recorded in a low noise and nearly reverberation-free environment, using a 4-channel ReSpeaker microphone

TABLE II
SUMMARY OF THE SIGNIFICANT DL-BASED SSL DATASET (L: LOUDSPEAKER; H: HUMAN; NA: INFORMATION NOT AVAILABLE)

Name	Modality	Sensors	Source	Duration	Annotations
(1) LOCATA [55]	audio	DICIT, eigenmike, hearing aids, NAO robot	L/H	NA-	3D location, VAD
(2) RSL [34]	audio	4-channel ReSpeaker array	L	22 hrs	3D location
(3) SSLR [28]	audio-visual	Pepper robot	L/H	25 hrs/4 mins	3D location, VAD
(4) MuMMER [35]	audio-visual	Pepper robot	H	1 hr	2D face location
(5) AVRI (ours)	audio-visual	Kinect, 4-channel ReSpeaker array	H	9 hrs	3D location, VAD

TABLE III
SPECIFICATIONS OF THE AVRI DATASET

Category	Details
Number of speakers	11 (6 female and 5 male)
Number of clips	43
	(w/wo face masks - 22/21)
	(Chinese/English - 22/21)
Shortest clip	6 min 13 s
Longest clip	18 min 58 s
Total duration	8 h 57 min
Range of DoA *i.e.*, azimuth	$[1^\circ, 360^\circ]$
Recording equipments	(a) ReSpeaker USB 4-Mic Array
	(b) Kinect for Windows v2
	(c) OptiTrack system[6]
	(d) KINOVA MOVO robot
	(e) close-talk system[7]



(a) (b)

Fig. 5. The AVRI recording platform consists of: (a) the KINOVA robot and (b) the multi-modal sensory platform: a 4-channel ReSpeaker microphone array is mounted on top of a Kinect sensor (the red arrow directs DoA $= 90^\circ$).

array.[2] A loudspeaker is placed at two different heights (1 m and 1.5 m height from the ground plane) with DoA of a sound source ranging from 1 to 360 degrees (5-degree resolution).

3) The Sound Source localization for Robots (SSLR) [28] is recorded using the humanoid Pepper robot[3] where four microphones and a stereo-vision are mounted on the robot head. It mostly uses a loudspeaker for recording, whereas human recordings only last for 4 minutes.
4) The MultiModal Mall Entertainment Robot (MuMMER) dataset [35] uses the Pepper robot to record the audio-visual streams. However, it only provides 2D face location annotations, which are not suitable for speaker spatial location estimation.

In summary, LOCATA and RSL only include audio signals. Among the audio-visual datasets, SSLR only has 4 minutes of human video, while MuMMER lacks the annotated 3D speaker location. To support our study, we propose a large-scale AVRI dataset, which is elaborated in the following.

B. AVRI Dataset

We collect the AVRI dataset, as summarized in Table III, that features several unique properties: (1) it involves both multi-channel audio signals and images, (2) it is recorded from on-site human speakers, (3) it includes 3D location annotations, and (4) it not only enables DL-based applications, but also supports SP-based techniques. To promote research in DL-based audio-visual speaker localization and tracking, the AVRI dataset and the source code of this work are publicly available together with this paper.

For HRI applications in smart home scenarios, the recording environment makes the most difference. Thus, we design the recording in a real indoor reverberant room equipped with various furniture (e.g., table, sofa, and chairs) where the data are captured from a real robot and affected by ego and background noise. Reverberation time is approximated to $\mathrm{RT}_{60} \approx 0.35$ s according to [56]. The whole recording procedure lasts for two months, while we do not constrain the robot location and room arrangements. This makes the data more realistic in a robotic scenario. Specifically, a Kinect sensor is used to capture RGB image sequences (with a resolution of 960×540 pixels), and audio signals are captured with a four-channel circular ReSpeaker microphone array (i.e., at the sampling frequency of 16 kHz). The setup of the multisensory equipment is illustrated in Fig. 5(b), which is mounted on a KINOVA robot[4] (Fig. 5(a)).

We invite 6 female and 5 male speakers, that is, 11 participants in total. For each speaker, we record 3∼4 clips with each length varying from 6 to 18 minutes. During the recording, the participants freely move around the robot while reading a given script. In particular, 22 recordings were read in Chinese, while 21

[2]https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/
[3]Pepper robot: https://www.softbankrobotics.com/emea/en/pepper
[4]KINOVA robot: https://www.kinovarobotics.com/en

were read in English. Moreover, participants wore a face mask in 22 recordings.

To annotate the speech activities, we adopt a wireless presenter system[5] where a lavalier microphone is clipped on the collar of each participant to acquire the close-talk speech. Moreover, to facilitate speaker spatial localization, we incorporate an OptiTrack system to annotate the 3D locations of speakers and sensors. In Fig. 5(b), the small silvery dots stuck on the sensors are the reflection points (annotators) of the OptiTrack system.

VI. Experiments

As studies show that DL-based localization methods are superior to SP-based methods [9], [21], [26], [32], we focus on comparing our proposed methods with the significant DL-based DoA estimation methods. The experiments are conducted on AVRI dataset to compare CMAF with the competitive baselines. We also provide qualitative analysis and visualization in a comparative study. Furthermore, we would like to show that appropriate audio-visual fusion greatly improves the robustness of speaker tracking in noisy acoustic conditions.

A. Baselines

We compare our proposed CMAF method with four competitive SOTA methods where three of them use audio as input (GCC-MLP [28], STFT-ResNet [29], A-CRNN [32]), and the fourth one uses audio-visual signals (AV-MLP [9]). For AV-MLP and A-CRNN [9], [32], we use their provided source codes, while for GCC-MLP and STFT-ResNet [28], [29], we reproduce their results.

All methods are tested with the same parameter settings for fair comparison. We briefly summarize the methodology of the baseline methods as follows:

1) GCC-MLP [28]: It incorporates GCC-PHAT as the acoustic feature, where an MLP network with three fully-connected hidden layers is used as the DoA classifier.
2) STFT-ResNet [29]: It uses STFT as input, which captures richer temporal and spatial information than GCC-PHAT. A CNN network with residual blocks [57] is used to avoid the problem of vanishing gradients and to learn the DoA feature representation.
3) A-CRNN [32]: It stacks GCC-PHAT and log-mel spectrogram as the inputs to a CRNN network to capture the temporal motion dynamics. Although this work was originally designed for sound event tracking, it is easily extendable to track a speaker.
4) AV-MLP [9]: It uses GCC-PHAT and Gaussian-encoded visual features ((7)) as input. The two modalities are concatenated to form a global audio-visual representation for the back-end MLP-based DoA classifier.

To be noted, since AV-MLP [9], the only comparable DL-based DoA estimation work, is our previous method replies on visual augmentations, we facilitate a new data collection and novel method proposals to encourage more explorations in this field.

[5]https://www.shure.com/en-ASIA/products/wireless-systems/blx_wireless/blx188-cvl-dual-presenter-set

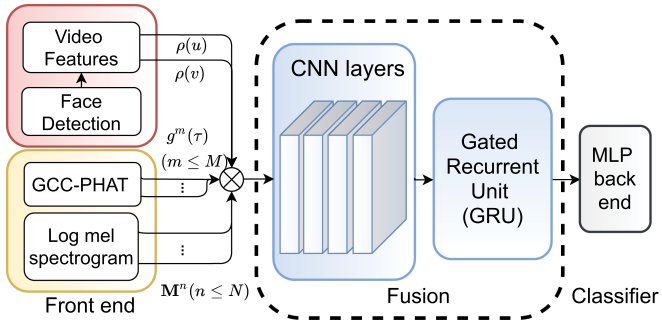


Fig. 6. The network architecture of the contrastive model i.e., AV-CRNN. The encoded multi-modal features are firstly concatenated, then feed into a stack of CNN blocks to learn a high-level representation. Then, a GRU module is adopted to incorporate the temporal information before the final DoA classifier ($\otimes$ indicates the concatenation operation, which aggregates the multi-modal features as the entire input representation).

B. A Contrastive Model: AV-CRNN

Besides the four baselines in Section VI-A, we also implement a contrastive model to examine the contributions of the proposed self-attention and cross-attention mechanism in CMAF. We introduce a CRNN module, i.e., a stacked CNN and GRU architecture, in place of the parallel fusion module in CMAF, referred to as AV-CRNN in Fig. 6, where the dashed box denotes the CRNN architecture.

AV-CRNN has the same front-end feature extractor and the back-end classifier as CMAF, but employs a different multi-modal fusion mechanism without cross attention, that allows us to clearly show the effect of the proposed CMAF. In the CRNN module, since the translation invariant characteristics of CNN are effective in processing multi-modal data [58], we concatenate these features into stacked CNN layers. Each 2D CNN block is followed by an average pooling, a batch normalization, and a ReLU activation, to extract high-level location-related information. To model speaker temporal dynamics that are not captured by CNN layers, we apply a Gated Recurrent Unit (GRU) module where each unit consists of multiple gates to identify the temporal information to store, ignore, and eventually trigger the output. In this way, the recurrent layers can accumulate the evolution of spatial parameters from neighboring time frames to facilitate speaker tracking. Finally, the GRU outcome is taken by the classifier to produce class-wise DoA posterior probabilities ($\hat{p}_t(\theta)$ in (2)).

C. Evaluation Metrics

We evaluate the methods in terms of mean absolute error (MAE) and accuracy (ACC). The symbols '↑' and '↓' indicate the desired direction of improvement.

MAE (°) is calculated as the average difference between the ground truth and the estimated DoA,

$$\text{MAE}\ (\downarrow) = \frac{1}{T}\sum_{t=1}^{T}|\theta_t - \hat{\theta}_t| \tag{16}$$

TABLE IV
COMPARISON OF OUR PROPOSED METHODS WITH THE SIGNIFICANT STATE-OF-THE-ART DL-BASED DOA ESTIMATION METHODS ON THE AVRI DATASET

Model	Modality	Design task	MAE (°) ↓	ACC (%) ↑	# Parameters (M)
STFT-ResNet [29]	audio	localization	17.21	67.63	6.315
GCC-MLP [28]	audio	localization	19.03	66.00	2.604
AV-MLP [9]	audio-visual	localization	17.55	68.78	2.647
A-CRNN [32]	audio	tracking	7.91	79.28	7.713
prop. AV-CRNN	audio-visual	tracking	7.58	79.72	7.715
prop. CMAF	audio-visual	tracking	**7.26**	**80.86**	3.825

The best results are marked in bold. The number of trainable parameters are given in million. "Prop." denotes our proposed methods.

ACC denotes the ratio of correctly estimated frames,

$$\text{ACC}\ (\uparrow) = \frac{1}{T}\sum_{t=1}^{T}\delta_t,\ \text{where}\ \ \delta_t = \begin{cases} 1 & |\theta - \hat{\theta}_t| \leq \rho \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

where ρ is the accuracy tolerance which is set to 10°.

D. Implementation Details

We split the AVRI dataset (Section V) into non-overlapping 70% and 30% between a training and test set. We apply a face detector to extract the face bounding boxes (i.e., $\mathbf{b_t}$ in (6)), in each frame of the image. We chose [59] considering its high accuracy and robustness for masked faces. A face detection rate of 67.07% is achieved in the whole data set, where 59.43% is for the training set and 84.9% is for the testing set. The frames of no detections are mainly because the speakers move out of the camera's sight. It is noted that different face detection rates between the training and testing set result from subconscious actions of multiple different participants, since we do not constrain their motions. With such face detection rates, visual cues contribute over half of the time for speaker DoA tracking. Whenever no visual cue is available, the audio-visual algorithms rely only on the audio modality.

We set the standard deviation σ_θ of the DoA-based posterior probability (Eq. 1) to 8, the same as [9], [28]. For the audio building block, the STFT is computed over an audio segment of 64 ms (1024 samples) with 50% overlapping. Considering the maximum available TDoA given the inter-distance of a microphone pair, we compute the 21-dimensional GCC-PHAT coefficients with the time delays ranging from -10 to 10 samples. The log-mel spectrogram is computed with 21 mel-scale filters at a frequency range from 20 to 8 kHz. For the visual building block, to be consistent with the audio features, each of the resulting horizontal and vertical visual image features, i.e., $\rho_t(u)$ and $\rho_t(v)$, is of dimension 21 as well. The standard deviation σ_u is empirically set to 3 ((7)), and the same for σ_v. Moreover, since audio and video are of different sampling rates, we resample all features with the same frame rate, i.e., 32 frames per second.

For CMAF, as shown in Fig. 3, we first use separate FC layers to derive the latent representations, $\mathbf{X}_\rho$, $\mathbf{X}_g$, and $\mathbf{X}_\mathbf{M}$ of 128 dimensions. Since we have three input modes, six parallel CMAFs are used. For each involved M-SAtt and M-CAtt blocks, we empirically set the multi-head number and employ $H = 4$ heads for satisfactory performance. For the back-end DoA classifier, we use the same architecture as [28] for all comparable methods, which consists of three FC layers. Except the output layer, each hidden layer is followed by a batch normalization, a ReLu activation function, and a dropout layer. We use Adam Optimizer [60] with a learning rate of 0.001 and a batch size of 32.

E. Comparative Study on AVRI Dataset

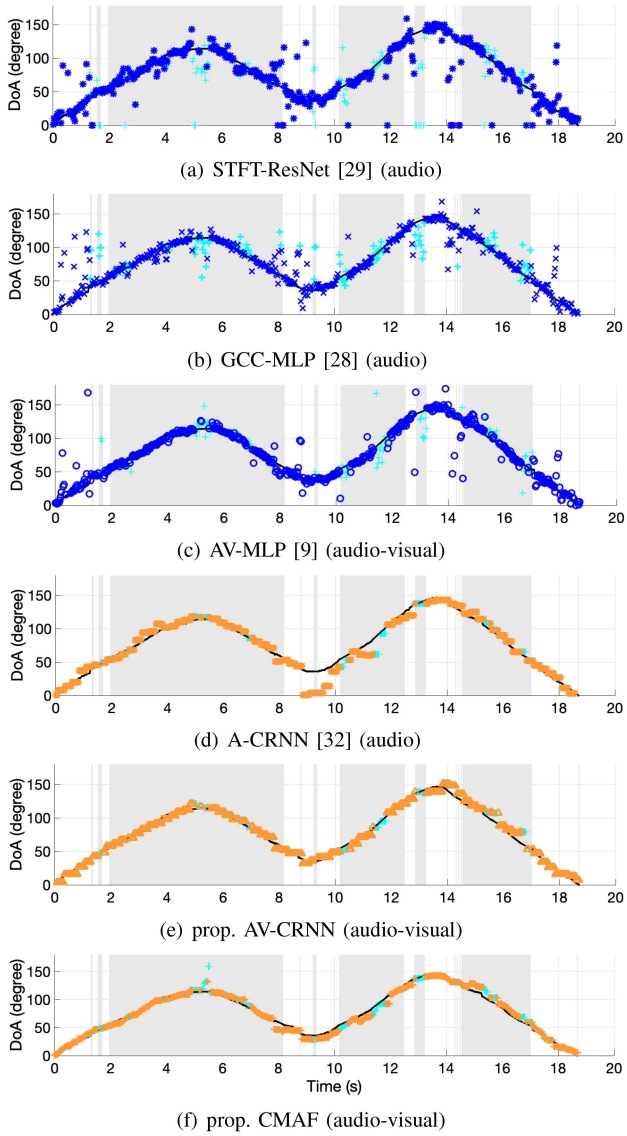
The performance of different methods is summarized in Table IV where the localization and tracking results are grouped separately. We observe that our proposed methods significantly outperform others. We also provide the number of trainable parameters (in million).

For audio-only methods, GCC-MLP simply feeds the GCC-PHAT features to the classifier and achieves the MAE of 19.03° and ACC of 66.00%. STFT-ResNet [29] has slightly better performance than GCC-MLP with the resulting MAE of 17.21° and the ACC of 67.63%. The tracking method A-CRNN [32] uses RNN to temporally filtering the input acoustic features. The reduction of MAE to 7.91° and the increase of ACC to 79.28% emphasize the benefits of consolidating the motion dynamics of the speaker between consecutive frames.

For audio-visual methods, AV-MLP outperforms GCC-MLP with a lower MAE of 17.55° and an improved ACC of 68.78%, which is attributed to visual cues ((7)). When comparing A-CRNN and the group of systems with localization as the design task, it is obvious that temporal filtering (tracking) has a greater impact than visual cues. The results of our contrastive model AV-CRNN corroborate the tracking influence, while compared to A-CRNN, the improved MAE to 7.58° and ACC to 79.72% shows the benefits from vision.

The proposed CMAF model further boosts the performance with the parallel CMAF modules (described in Section IV-C), reducing MAE to 7.26° and improving ACC to 80.86%. In addition to superior performance, CMAF only includes 3.825 million (M) trainable parameters, which are 50% fewer than the CRNN-based tracking methods (7.713 M parameters for A-CRNN, and 7.715 M parameters for AV-CRNN).

To give an intuitive illustration, Fig. 7 displays the speaker trajectory and the DoA estimates of different methods where the horizontal and vertical axes correspond to the time and DoA scale, respectively. We differentiate the *localization* and



(a) STFT-ResNet [29] (audio)

(b) GCC-MLP [28] (audio)

(c) AV-MLP [9] (audio-visual)

(d) A-CRNN [32] (audio)

(e) prop. AV-CRNN (audio-visual)

(f) prop. CMAF (audio-visual)

Fig. 7. Comparison between the speaker ground truth trajectory (the black curve) and the DoA estimates from different methods (the blue curves correspond to the *localization* methods while the orange ones are the *tracking* methods; results of non-speech segments are marked with cyan crosses; frames with face detections are with grey background.). The horizontal axis represents time while the vertical axis represents the DoA variations. The adopted modalities are specified in the bracket. (a) STFT-ResNet [29], (b) GCC-MLP [28], (c) AV-MLP [9], (d) A-CRNN [32], (e) the proposed AV-CRNN method, and (f) the proposed CMAF method.

the *tracking* methods with the color index of blue and orange, respectively. Moreover, results of non-speech segments are marked with cyan crosses while frames with face detections are with grey background. From the figure, we can see that for the localization methods, i.e., STFT-ResNet (Fig. 7(a)), GCC-MLP (Fig. 7(b)) and AV-MLP (Fig. 7(c)), although the vast majority of DoA estimates follow the ground truth (black curve), there are still some spines, resulting from intermediate speech pauses or background noise, randomly distribute over various DoAs. Nevertheless, it is obvious that with the help of vision, Fig. 7(c) produces less spines than Fig. 7(b). Observing the estimated speaker trajectories of the tracking methods, i.e., A-CRNN (Fig. 7(d)), AV-CRNN (Fig. 7(e)) and CMAF (Fig. 7(f)), we find that the tracking mechanism helps remove the DoA outliers (especially in non-speech segments) and thus results in smoother DoA trajectories. Furthermore, when comparing tracking methods with and without visual incorporation (AV-CRNN and CMAF vs. A-CRNN), we can observe the contributions from face detections. In general, our proposed CMAF achieves the most smoothed trajectory over the other competitive methods.

F. Noise Robustness

In real-world applications, audio signals are always corrupted by noise. We would like to test the proposed DL-based model under noisy conditions compared to other models. The Additive White Gaussian Noise (AWGN) is added to the multi-channel audio signals, and the resulting SNR ranges from -20 dB to 10 dB. The results are summarized in Table V.

From Table V, we can see that the tracking methods using either audio (A-CRNN) or audio-visual (AV-CRNN and CMAF) exhibit superiority over the localization methods (STFT-ResNet, GCC-MLP, and AV-MLP). When SNR ≥ 0 dB, they maintain a great performance of MAE $< 10^{\circ}$ and ACC $> 70\%$. As the SNR degrades, the contribution of visual signals increases. For example, AV-MLP has a slightly higher localization accuracy (59.02%) than GCC-MLP (54.26%) at SNR $= 10$ dB, while it shows a greater improvement in localization accuracy (40.40%) than GCC-MLP (27.92%) at SNR $= -20$ dB.

The same view applies to the two tracking methods (AV-CRNN and A-CRNN). The tracking accuracy improves from 76.26% to 78.18% for SNR $= 10$ dB, and from 33.06% to 42.60% for SNR $= -20$ dB. To be mentioned, the audio-visual localization method AV-MLP (ACC $= 40.40\%$) outperforms the audio tracking method A-CRNN (ACC $= 33.06\%$) at SNR $= -20$ dB, which elaborates that video has a higher impact than the tracking mechanism at high noise interference. In summary, incorporating visual influence and temporal information helps improve the system's robustness. From Table V, the proposed CMAF always achieves the best results where the improved accuracy of 5.82% and 3.62% over AV-MLP and AV-CRNN at SNR $= -20$ dB are observed.

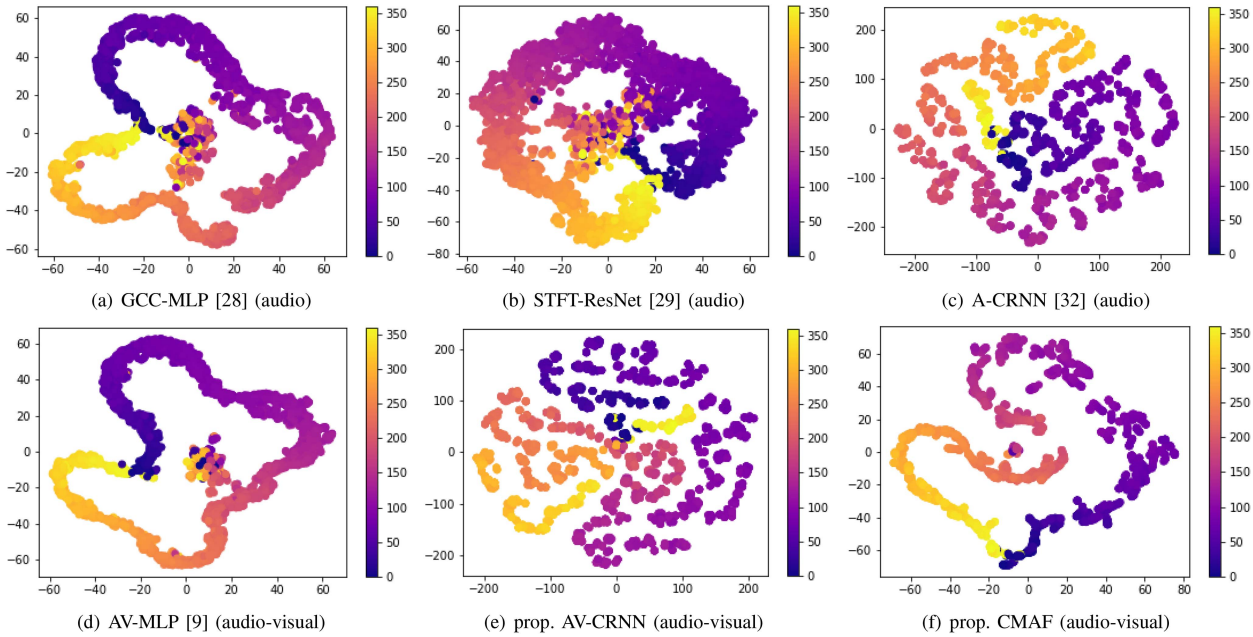
G. Feature Visualization

One of the reasons that DL-based methods are superior to SP-based methods is the encapsulation of feature extraction stages into a learning framework. Thus, we illustrate in Fig. 8 the t-SNE visualization [61] of feature representations which are extracted from the penultimate layer of the speaker DoA classifier. The gradient color varying from purple to yellow corresponds to the DoA range from 1° to 360°. It should be noted that since the azimuth is cyclic, the DoA estimates near 1° and 360° are spatially close. Moreover, since we treat speaker localization as a regression problem on discretized DoA labels, an ideal feature distribution should contain high inter-class variance and low intra-class variance.

TABLE V
COMPARISON OF OUR PROPOSED METHODS WITH THE BASELINES ON THE AVRI DATASET UNDER DIFFERENT SNR CONDITIONS

Model	Modality	Track	-20 dB		-10 dB		0 dB		10 dB	
			MAE (°)↓	ACC (%)↑	MAE (°)↓	ACC (%)↑	MAE (°)↓	ACC (%)↑	MAE (°)↓	ACC (%)↑
STFT-ResNet [29]	audio		41.05	26.78	35.94	30.06	28.85	40.17	21.97	56.49
GCC-MLP [28]	audio		42.76	27.92	37.08	31.37	30.43	41.38	24.67	54.26
AV-MLP [9]	audio-visual		39.98	40.40	34.44	43.35	28.28	50.68	24.23	59.02
A-CRNN [32]	audio	✓	30.77	33.06	13.24	56.57	9.92	71.29	8.87	76.26
prop. AV-CRNN	audio-visual	✓	29.59	42.60	12.91	59.42	9.61	72.12	8.27	78.18
prop. CMAF	audio-visual	✓	**24.17**	**46.22**	**11.51**	**64.52**	**9.03**	**74.85**	**8.00**	**78.81**

The best results are marked in bold.



(a) GCC-MLP [28] (audio) (b) STFT-ResNet [29] (audio) (c) A-CRNN [32] (audio)

(d) AV-MLP [9] (audio-visual) (e) prop. AV-CRNN (audio-visual) (f) prop. CMAF (audio-visual)

Fig. 8. The t-SNE visualization [61] on feature representations extracted from the penultimate layer of the DoA classifier, given the same inputs to different methods in the test set. The gradient color varying from purple to yellow corresponds to the DoA range from 1° to 360°. The top row indicates the audio-only baselines [28], [29], [32] and the bottom row indicates the audio-visual baseline [9], and two of our proposed methods i.e., AV-CRNN and CMAF.

Fig. 8(a) shows the feature representation of the GCC-MLP network [28]. It is observed that given the concatenated multi-channel GCC-PHAT as inputs, in most cases, the network can successfully distinguish different DoA classes. However, the features of distinct classes are gathered around the origin area, which is not informative for the classifier. Fig. 8(b) visualizes the features from the STFT-ResNet [29] where we observe the same limitation as shown in Fig. 8(a), that the network fails on the distinct-class features located around the origin. Fig. 8(c) uses the CRNN module to incorporate a tracking mechanism. By considering the feature variations among neighboring frames, the model can successfully disambiguate the gathered inter-class features around the origin.

Fig. 8(d) shows the use of audio-visual cues for DoA classification. We observe that the features are better clustered than Fig. 8(a), while there are still some distinct class features distributed around the origin. Fig. 8(e) corresponds to the contrastive AV-CRNN network, which considers the temporal dependency among audio-visual features. The feature distribution is similar to Fig. 8(c). For our proposed CMAF in Fig. 8(f), features are cohesive for the same class, while they are divergent for distinct classes. Moreover, it is observed that the DoA features of neighboring classes are spatially close, leading to smooth temporal transition between adjacent DoAs, which is beneficial to the tracking task.

In summary, the t-SNE visualizations in Fig. 8 corroborate the numerical results in Table IV: First, visual cues contribute to improving the audio-only SSL performance. Second, incorporating CRNN-based tracking helps to overcome the inter-class ambiguity. Finally, our proposed CMAF illustrates the best feature visualization with clustered intra-class features and distinct inter-class features. The distances between the different features are proportional to their spatial DoA difference.

H. Limitation and Future Work

The collected AVRI dataset is device-specific. Our proposed method, which is trained on AVRI, herein limits at the specific recording setup i.e., a Kinect with a 4-channel ReSpeaker microphone array mounted on top of it (Fig. 5). When with

a different sensor setup, recordings need to be re-collected for model retraining. Nevertheless, theoretically, as long as the same sensor setup is used, our method can adapt to different room conditions and robotic platforms.

Future works include the DL-based multi-speaker localization and tracking, peculiarly, how to disentangle the identity-specific features from the multi-modal inputs. For observation-to-track assignments, one may be inspired by the permutation invariant training (PIT) mechanism [62] utilized in speech separation as a potential direction. Different room conditions and deployed robotic platforms should be investigated as well. Moreover, how to handle the situation with non-tracked point noise source needs to be explored. Last but not least, deploying an end-to-end network is another promising way for real-world robotic applications.

VII. Conclusion

Multi-modal processing endows the robot with a higher capability of scene understanding. Despite the success of deep learning, most existing multi-modal localization works still rely on SP techniques, where the research development is penalized by the lack of dataset, complex coordinates transformation among heterogeneous sensors, and very few open-sourced algorithms. We consider our work to be of significant importance in the field of audio-visual speaker location estimation. Specifically, we contributed a newly annotated multi-modal dataset that enables DL technique exploration. What's more, we proposed a CMAF framework as the first attempt at DL-based audio-visual speaker DoA localization and tracking. The introduced CMAF module incorporates self-attention and cross-attention to jointly explore the intra- and inter-modality relations for more accurate and robust speaker tracking. The experimental results demonstrate the superiority of CMAF over the other methods. We will make the dataset and the source code publicly available.

Acknowledgment

We thank (1) the research engineer Yunfan Lu for his help in device setup; (2) the NUS-Advanced Robotics Centre for their support in providing the robotic platform, and finally (3) all the volunteers participated data annotation and collection.

Author Contribution Statement

The authors confirm contribution to the paper as follows: (1) idea and design: Xinyuan Qian, Jiadong Wang, Haizhou Li; (2) data collection: Guohui Guan, Zhengdong Wang; (3) analysis and interpretation of results: Xinyuan Qian, Zhengdong Wang, Jiadong Wang; (4) draft manuscript preparation: Xinyuan Qian; (5) manuscript revision: Xinyuan Qian, Haizhou Li, Jiadong Wang. All authors reviewed the results and approved the final version of the manuscript.

References

[1] R. Li, F. Zhao, D. Pan, and L. Dong, "Speech enhancement based on binaural sound source localization and cosh measure wiener filtering," *Circuits, Systems, Signal Process.*, vol. 41, no. 1, pp. 395–424, 2022.

[2] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[3] X. Gao, C. Gupta, and H. Li, "Genre-conditioned acoustic models for automatic lyrics transcription of polyphonic music," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 791–795.

[4] X. Gao, C. Gupta, and H. Li, "Automatic lyrics transcription of polyphonic music with lyrics-chord multi-task learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2280–2294, 2022.

[5] 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.

[6] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput. Speech Lang.*, vol. 11, no. 2, pp. 91–126, Apr. 1997.

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

[8] E. Vargas, J. R. Hopgood, K. Brown, and K. Subr, "On improved training of CNN for acoustic source localisation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 29, pp. 720–732, 2021.

[9] X. Qian, M. Madhavi, Z. Pan, J. Wang, and H. Li, "Multi-target DOA estimation with an audio-visual fusion mechanism," in *Proc. IEEE Int. Conf. Audio, Speech Signal Process.*, 2021, pp. 4280–4284.

[10] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proc. IEEE*, vol. 103, no. 9, pp. 1635–1653, Sep. 2015.

[11] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–53.

[12] H. Zhou, Z. Liu, X. Xu, P. Luo, and X. Wang, "Vision-infused deep audio inpainting," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 283–292.

[13] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum, "Look, listen, and act: Towards audio-visual embodied navigation," in *Proc. Int. Conf. Robot. Automat.*, 2020, pp. 9701–9707.

[14] F. Talantzis, A. Pnevmatikakis, and A. G. Constantinides, "Audio-visual active speaker tracking in cluttered indoors environments," *IEEE Trans. Syst., Man, Cybern.*, vol. 38, no. 3, pp. 799–807, Jun. 2008.

[15] I. D. Gebru, S. Ba, G. Evangelidis, and R. Horaud, "Audio-visual speech-turn detection and tracking," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, Liberec, Czech Republic, 2015, pp. 143–151.

[16] V. Kilic, M. Barnard, W. Wang, A. Hilton, and J. Kittler, "Mean-shift and sparse sampling based SMC-PHD filtering for audio informed visual speaker tracking," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2417–2431, Dec. 2016.

[17] Y. Liu, V. Kılıç, J. Guan, and W. Wang, "Audio–visual particle flow SMC-PHD filtering for multi-speaker tracking," *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 934–948, Apr. 2020.

[18] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Multi-speaker tracking from an audio-visual sensing device," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2576–2588, Oct. 2019.

[19] H. Liu, Y. Li, and B. Yang, "3D audio-visual speaker tracking with a two-layer particle filter," in *Proc. IEEE Int. Conf. Image Process.*, Taipei, Taiwan, 2019, pp. 1955–1959.

[20] O. Lanz, A. Brutti, A. Xompero, X. Qian, M. Omologo, and A. Cavallaro, "Accurate target annotation in 3D from multimodal streams," in *Proc. IEEE Int. Conf. Audio, Speech Signal Process.*, Brighton, U.K., 2019, pp. 3931–3935.

[21] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Audio, Speech Signal Process.*, 2015, pp. 2814–2818.

[22] R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, and S. Goetze, "On sound source localization of speech signals using deep neural networks," in *Proc. Deutsche Jahrestagung für Akustik*, 2015, pp. 1510–1513.

[23] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *Proc. IEEE Int. Conf. Audio, Speech Signal Process.*, 2018, pp. 2386–2390.

[24] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Proc. 138th Audio Eng. Soc. Conv.*, 2015, pp. 328–334.

[25] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2016, pp. 603–609.

[26] S. Chakrabarty and E. A. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 8–21, Mar. 2019.

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

[28] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. Int. Conf. Robot. Automat.*, 2018, pp. 74–79.

[29] W. He, P. Motlicek, and J.-M. Odobez, "Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training," in *Proc. IEEE Int. Conf. Audio, Speech Signal Process.*, 2019, pp. 770–774.

[30] Z. Pan, M. Zhang, J. Wu, J. Wang, and H. Li, "Multi-tones' phase coding (MTPC) of interaural time difference by spiking neural network," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 29, pp. 2656–2670, Jul. 2020.

[31] 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. 770–774.

[32] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Proc. Workshop Detection Classification Acoust. Scenes Events*, May 2019, pp. 402–406.

[33] C. Evers and P. A. Naylor, "Acoustic SLAM," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 26, no. 9, pp. 1484–1498, Sep. 2018.

[34] R. Sheelvant, B. Sharma, M. C. Madhavi, R. K. Das, S. Prasanna, and H. Li, "RSL2019: A realistic speech localization corpus," in *Proc. IEEE 22nd Conf. Oriental COCOSDA Int. Committee Co-Ordination Standardisation Speech Databases Assessment Techn.*, 2019, pp. 1–6.

[35] O. Canévet, W. He, P. Motlicek, and J.-M. Odobez, "The MuMMER data set for robot perception in multi-party HRI scenarios," in *Proc. IEEE Int. Conf. Robot Hum. Interactive Commun.*, 2020, pp. 1294–1300.

[36] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, Berlin, Germany: Springer, 2001, pp. 157–180.

[37] J. Wang, X. Qian, Z. Pan, M. Zhang, and H. Li, "GCC-PHAT with speech-oriented attention for robotic sound source localization," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. pp. 5876–5883.

[38] T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New York, NY, USA, 2005, pp. 118–121.

[39] V. Kılıç, M. Barnard, W. Wang, and J. Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Trans. Multimedia*, vol. 17, no. 2, pp. 186–200, Dec. 2015.

[40] Y. Ban, X. Alameda-Pineda, L. Girin, and R. Horaud, "Variational bayesian inference for audio-visual tracking of multiple speakers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1761–1776, May 2021.

[41] C. Schymura and D. Kolossa, "Audiovisual speaker tracking using nonlinear dynamical systems with dynamic stream weights," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 28, pp. 1065–1078, 2020.

[42] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 631–648.

[43] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 208–224.

[44] T. Haarnoja, A. Ajay, S. Levine, and P. Abbeel, "Backprop KF: Learning discriminative deterministic state estimators," in *Proc. Int. Conf. Neural Inf. Proc. Syst.*, 2016, pp. 4383–4391.

[45] C. Schymura et al., "A dynamic stream weight backprop Kalman filter for audiovisual speaker tracking," in *Proc. IEEE Int. Conf. Audio, Speech Signal Process.*, 2020, pp. 581–585.

[46] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 6292–6300.

[47] D. Florencio, C. Zhang, and Z. Zhang, "Why does PHAT work well in low noise reverberant environment," in *Proc. IEEE Int. Conf. Audio, Speech Signal Process.*, 2008, pp. 2565–2568.

[48] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 24, no. 11, pp. 2171–2186, Nov. 2016.

[49] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*. [Online]. Available: http://arxiv.org/abs/1905.05055

[50] R. Tao, Z. Pan, R. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 3927–3935.

[51] Z. Pan, R. Tao, C. Xu, and H. Li, "Muse: Multi-modal target speaker extraction with visual cues," in *Proc. IEEE Int. Conf. Audio, Speech Signal Process.*, 2021, pp. 6678–6682.

[52] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 941–951.

[53] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *Proc. Int. Conf. Comput. Vis.*, Kyoto, Japan, 2009, pp. 1515–1522.

[54] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.

[55] C. Evers et al., "The LOCATA challenge: Acoustic source localization and tracking," in *Proc. LOCATA Challenge Workshop - A Satell. Event IWAENC*, Tokyo, Japan, 2018, pp. 410–414.

[56] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Amer.*, vol. 37, no. 6, pp. 1187–1188, 1965.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[58] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Proc. Syst.*, Lake Tahoe, Nevada, USA, 2012, pp. 1097–1105.

[59] X. Peng, H. Zhuang, G.-B. Huang, H. Li, and Z. Lin, "Robust real-time face tracking for people wearing face masks," in *Proc. Int. Conf. Control, Automat., Robot. Vis.*, 2020, pp. 779–783.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 375–388.

[61] L. V. D. Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

[62] M. Kolbæk, D. Yu, Z.-H. Tan, and J. 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, Oct. 2017.