MULTI-TARGET DOA ESTIMATION WITH AN AUDIO-VISUAL FUSION MECHANISM

Xinyuan Qian, Maulik Madhavi, Zexu Pan, Jiadong Wang, Haizhou Li

Department of Electrical and Computer Engineering,
National University of Singapore, Singapore

ABSTRACT

Most of the prior studies in the spatial Direction of Arrival (DoA) domain focus on a single modality. However, humans use auditory and visual senses to detect the presence of sound sources. With this motivation, we propose to use neural networks with audio and visual signals for multi-speaker localization. The use of heterogeneous sensors can provide complementary information to overcome uni-modal challenges, such as noise, reverberation, illumination variations, and occlusions. We attempt to address these issues by introducing an adaptive weighting mechanism for audio-visual fusion. We also propose a novel video simulation method that generates visual features from noisy target 3D annotations that are synchronized with acoustic features. Experimental results confirm that audio-visual fusion consistently improves the performance of speaker DoA estimation, while the adaptive weighting mechanism shows clear benefits.

Index Terms— audio-visual fusion, sound source localization, adaptive weighting mechanism

1. INTRODUCTION

In human-robot interaction, a robot relies on its Sound Source Localization (SSL) mechanism to direct its attention. Traditionally, SSL approaches only use audio signals and attempt as a signal processing problem [1, 2, 3]. However, those approaches are adversely affected by acoustically challenged conditions, such as noise and reverberation scenarios [4]. To address that, several Neural Networks (NN)-based approaches were explored [5, 6, 4, 7] assuming a sufficient amount of data are available. Specifically, location-related Short-Time-Fourier-Transform (STFT) cues are mapped to sound DoA information in [5, 6] while the Generalized Cross Correlation with Phase Transform (GCC-PHAT) cues are used in [4, 7]. Despite the progress, many research problems remain. One of them is multi-speaker localization in real multi-party human-robot interaction scenarios under acoustic challenging conditions [4].

Considering seeing and hearing are the two most essential human cognitive abilities, studies observed that audio and video convey complementary information and may help to overcome uni-modal limitations of a degradation condition for scene analysis [8, 9, 10]. There is a very broad literature of audio-visual approaches for speaker localization over the past decades [11, 12, 13]. However, it was not until recently that the deep learning-based approaches have attracted more attention, thanks to the increasing computational power and rapid development in NN techniques. Nevertheless, most of these methods aim at locating sound sources in visual scenes [14, 15, 16, 17]. Specifically, an attention mechanism is incorporated into the individual sound and vision network to model the audio-visual image correspondence [14] . A visual saliency network is employed in [15], together with an audio representation network, to feature a SSL module for producing an audio-visual saliency map. An attention network is proposed in [16] to learn the visual regions of a sounding event. By fusing audio and visual features using LSTM and bilinear pooling, the audio assisted visual feature extraction is described in [17]. All the research studies use audio as a supplementary modality for visual localization and require the sound sources to be both audible and visible.

Unlike the prior studies, we aim to perform audio-visual speaker localization in the spatial DoA domain where targets can appear either inside (visible) or outside (invisible) the camera's Field-of-View (FoV). We propose two neural network architectures and make the following contributions in this paper: (1) we propose a novel video simulation method to deal with the lack of video data; (2) for the first time, we design a deep learning network for audio-visual multi-speaker DoA estimation, and (3) we adopt an adaptive weighting mechanism in a simple feedforward network to estimate the multi-modal reliability under different conditions.

2. PROPOSED METHOD

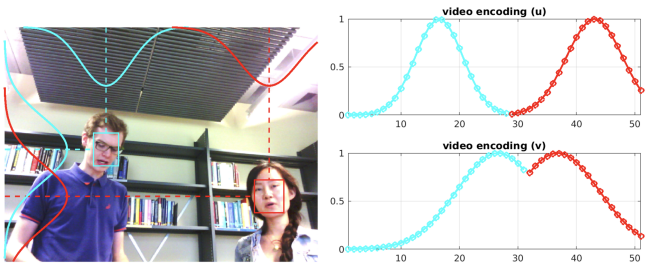
Given a sequence of frame-synchronized audio and video signals captured by a microphone array and a calibrated camera, we aim to estimate the DoA information $\theta = [-180°, 180°)$ for each sound source at each frame. Next, we describe the way we characterize audio and video signals, the video simulation method, and the proposed neural networks.

2.1. Audio features

The GCC-PHAT is widely used to calculate the time different of arrival (TDOA) between any two microphones in a microphone array [4, 7]. We adopt it as the audio feature [1] due to its robustness in the noisy and reverberant environment [18] and the fewer tunable parameters than the other counterparts e.g. STFT [5]. Let S_l and S_p be the Fourier transforms of audio sequence at l and p^{th} channels of the microphone array, respectively. We compute the GCC-PHAT features with different delay lags τ as:

$$\text{GCC-PHAT}_{lp}(\tau) = \sum_k \mathcal{R}\left(\frac{S_l[k](S_p[k])^*}{|S_l[k](S_p[k])^*|} e^{j\frac{2\pi k}{N}\tau}\right) \quad (1)$$

This research work is supported by the Neuromorphic Computing project, Programmatic Grant No. A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain), and Human-Robot Interaction Phase 1 (Grant No. 192 25 00054) from the National Research Foundation, Prime Minister's Office, Singapore under the National Robotics Programme.



(a) face detections (b) visual feature encoding

Fig. 1. Visual feature encoding from face detection bounding boxes. The feature resembles the horizontal (top) and vertical (bottom) axis of the image.

where $*$ denotes the complex conjugate operation, $\mathcal{R}$ denotes the real part of complex number and N denotes the FFT length. Here, the delay lag τ between two signals arrived is reflected in the steering vector $e^{j\frac{2\pi k}{N}\tau}$ in Eq. 1.

2.2. Visual features and simulation

With the advent of deep learning, accurate face detection at low computational cost becomes widely available [19]. Let us define $\mathbf{b}_d = (u, v, w, h)_d^\intercal$ as the face detection bounding box d ($d \leq D$), where $^\intercal$ denotes transpose, (u, v) are the horizontal and vertical positions of the top-left point, (w, h) are the width and height, and D is the number of detected faces. The central point of detection is thus computed as:

$$\boldsymbol{\mu}_d = (u + \frac{1}{2}w, v + \frac{1}{2}h)_d^\intercal \tag{2}$$

The visual feature is encoded as the exponential part of the multi-variant Gaussian distribution (in u and v direction) with the standard deviations specified by the detection width and height and achieves the maximum at the central point:

$$\mathcal{V}(\mathbf{x}) = \begin{cases} \max_d \; e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_d)\Sigma_d^{-1}(\mathbf{x}-\boldsymbol{\mu}_d)^\intercal} & D > 0, \\ \mathcal{U}(\mathbf{x}) & otherwise \end{cases} \tag{3}$$

where $\mathbf{x}$ indicates the potential image positions, $\Sigma_d = diag(w_d^2, h_d^2)$ is a diagonal covariance matrix, and $\mathcal{U}(\mathbf{x})$ indicates uniform distribution. The components in $\mathcal{V}(\mathbf{x})$ are re-sampled to the same length of GCC-PHAT.

Audio-visual parallel data are not abundantly available. However, it is possible to obtain the camera's extrinsic and intrinsic calibration parameters ζ^e and ζ^i, the 3D location $\mathbf{p} = (x, y, z)^\intercal$ of a sound source. We propose a novel method to synthesize visual features in synchrony with the audio features by Eq. 3. The overall pipeline of visual feature generation is illustrated in Fig. 2(a) and the process is formulated next.

We first add three-variant Gaussian distributed spatial noise to the target 3D location $\mathbf{p}$ to account for possible face detection error, and transfer the resulting point to the camera coordinates given the extrinsic parameters:

$$\tilde{\mathbf{p}}_c = \Phi(\mathcal{N}(\mathbf{p}, \Sigma_p) \mid \zeta^e) \tag{4}$$

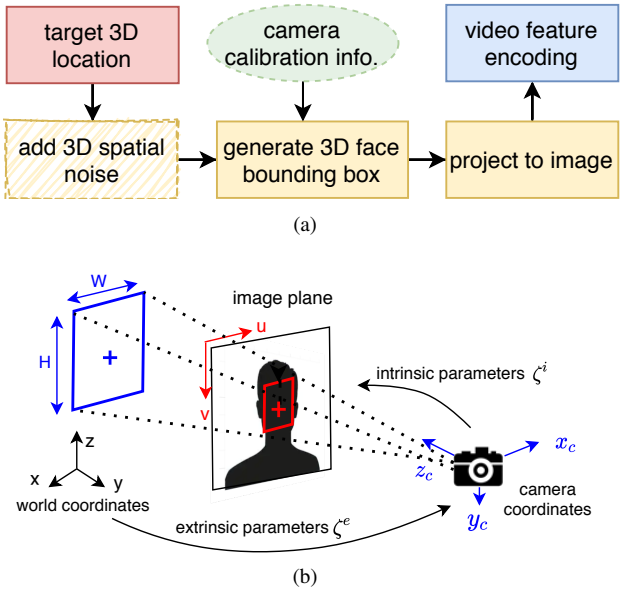


Fig. 2. (a) Pipeline to generate face bounding boxes and visual features and (b) 3D-to-image bounding box projection. (x, y, z): world coordinates; (x_c, y_c, z_c): camera coordinates; (u, v): image coordinates.

with noise covariance matrix $\Sigma_p = diag(\sigma_x^2, \sigma_y^2, \sigma_z^2)$ assuming that the additive noises to (x, y, z) are independent, and Φ is the transformation using the pin-hole camera model [20].

Then, we geometrically create the 3D face bounding box whose plane is perpendicular to the camera's optical axis (z_c in Fig. 2(b)), and project to the image plane:

$$\chi = \Psi(\tilde{\mathbf{p}}_c + \mathbf{v} \mid \zeta^i) \tag{5}$$

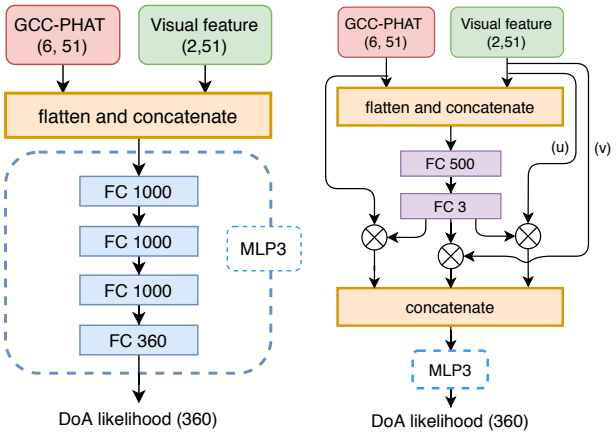
where Ψ is the 3D-to-image projection, $\mathbf{v}$ is the translation vector which equals to $(-\frac{W}{2}, -\frac{H}{2}, 0)^\intercal$ for the top-left point χ^{tl} and $(\frac{W}{2}, \frac{H}{2}, 0)^\intercal$ for the bottom-right point χ^{br}, respectively. W and H are the width and height assumptions of a real human face.

Finally, the simulated face detection bounding box $\mathbf{b}$ is computed as $\mathbf{b} = cat(\chi^{tl}, \chi^{br} - \chi^{tl})$, where cat denotes a concatenation operation to form a column vector.

2.3. Neural network architecture

We propose two NN architectures for audio-visual speaker DoA estimation based on Multilayer Perceptron (MLP), namely MLP Audio-Visual Concatenation (MLP-AVC) and MLP Audio-Visual Adaptive Weighting (MLP-AVAW), which specify different ways of audio-visual feature fusion and classifier design as illustrated in Fig. 3.

MLP-AVC consists of three hidden layers, denoted as MLP3 in Fig. 3(a) by a dotted blue box, each one is a fully-connected layer with ReLU activation [21] and batch normalization [22]. It takes the flattened and concatenated GCC-PHAT and visual features as an input vector. The network is trained to predict the probability of DoA labels, as in [4], using a sigmoid output layer. MLP-AVC adopts an early fusion strategy by concatenating audio and visual features. We hypothesize that such early fusion doesn't learn to pay



(a) MLP-AVC (b) MLP-AVAW

Fig. 3. Proposed NN architectures for 360° DoA estimation (red: audio block; green: video block; blue: standard MLP network; purple: adaptive weighting block; orange: feature reformatting block). The input dimension $(6, 51)$ represents 51 GCC-PHAT coefficients for each of the 6 microphone pairs and $(2, 51)$ represents 51 visual feature encoding for the image horizontal and vertical directions.

Table 1. MAE (°) and ACC (%) of the noisy target 3D locations ($\mathcal{N}(\mathbf{p}, \Sigma_p)$ in Eq. 4) for visual feature generation of the loudspeaker cases. Results are measured on frames accounting into DR.

Train (loudspeaker)			**Test-loudspeaker**		
DR	MAE	ACC	DR	MAE	ACC
11.3 %	6.67	46.0%	9.2 %	6.28	48.9 %

selective attention to uni-modal features, that are crucial in face of missing data or noisy data.

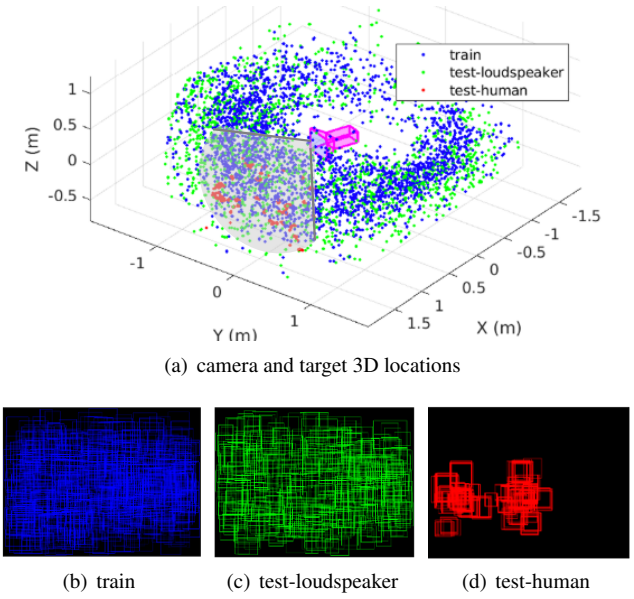
MLP-AVAW introduces an adaptive weighting mechanism, which uses a tiny NN with two fully-connected layers, colored in purple in Fig. 3(b), to learn three adaptive weights for the audio GCC-PHAT feature, video image horizontal and vertical features, respectively. A softmax activation function is applied for weights normalization. We call this as 'adaptive weighting' mechanism as the weights are adapted according to the live input during inference. Finally, the weighted multi-modal features are concatenated for MLP3 to compute DoA.

3. EXPERIMENTS

3.1. Dataset and performance metrics

The existing audio-visual datasets, such as AV16.3 [23], CAV3D [12], and AVASM [24], are either of limited size, or don't provide the spatial ground truth. We, therefore, simulate the synchronized visual features for a SSL dataset of the loudspeaker cases. We choose the recently released SSLR dataset[1] [4], that is recorded in a physical setup from one or two concurrent speakers, and with adequate target 3D annotations. It consists of 4-channel audio recordings at

[1]SSLR dataset: https://www.idiap.ch/dataset/sslr/



(a) camera and target 3D locations

(b) train (c) test-loudspeaker (d) test-human

Fig. 4. (a) Camera and target 3D locations (the gray section indicates the camera's FoV); (b-c) The distribution of the projected face detection bounding boxes (from points in gray region in (a)) on the image plane of different SSLR subsets (blue: train; green: test-loudspeaker); (d) RetinaFace detections [**?**] on test-human.

48 kHz sampling rate, that is organized into three subsets, namely train (loudspeaker), test-human, and test-loudspeaker.

We evaluate the performance of DoA estimates using the same metrics of [4] i.e. *Mean Absolute Error* (MAE) and *Accuracy* (ACC), where MAE is defined as the mean absolute error between the actual and the estimated DoA, while the accuracy allowance of ACC is 5° in the classification prediction.

For the test-human subset, we apply the RetinaFace detector [**?**] to achieve the face bounding boxes. For the train and test-loudspeaker subsets, the visual features are simulated with the method proposed in Sec. 2.2 with a noise covariance matrix $\Sigma_p = diag(0.2, 0.2, 0.2)$. Fig. 4(a) illustrates the ground truth camera (magenta) and target 3D locations for the train (blue), test-loudspeaker (green) and test-human (red) subsets for all frames. Targets in the gray region are inside the camera's FoV, therefore, visible to the camera. We only generate face bounding boxes of visible targets, as visualized in Fig. 4(b-c) and formulated in Eq. 4-5 with the simulated bounding box $\mathbf{b}$. Fig. 4 shows that the face bounding boxes spread well across the FoV with a balanced distribution. We don't generate bounding boxes for speakers that are outside the FoV. As a result, the visual features for the invisible speakers become missing data (the normal distribution in Eq. 3 for visual feature representation) in the audio-visual dataset.

The statistics of simulated visual features are summarized in Tab. 1 where DR represents the percentage of video frames having targets inside the FoV. Low DR means a high percentage of missing visual features. We also report in Tab. 1 the DoA MAE and ACC of the simulated visual features, indicating that the simulated data is of enough difficulty to represent real scenarios.

Table 2. A summary of MAE (°) and ACC (%) of speaker DoA estimation on the SSLR test set (N indicates the number of speakers; the number of audio frames for each subset is given in bracket). We reproduce the results of [4] for comparison.

		Loudspeaker				**Human**				**Overall**	
		N=1 (178k)		**N=2 (29k)**		**N=1 (788)**		**N=2 (141)**			
		MAE	ACC	MAE	ACC	MAE	ACC	MAE	ACC	MAE	ACC
audio	SRP-PHAT [2]	19.00	82.0	36.95	50.0	2.62	93.0	20.90	56.0	21.44	78.0
	MLP-GCC [4]	4.06	94.9	8.10	71.5	4.75	95.1	5.98	75.5	4.63	91.6
audio-visual	MLP-AVC	3.87	94.8	7.80	71.9	**1.84**	97.1	3.89	81.9	4.42	91.7
	MLP-AVAW	**3.73**	**95.0**	**7.28**	**73.6**	2.04	**98.0**	**3.49**	**86.5**	**4.22**	**92.0**

Table 3. A summary of MAE (°) and ACC (%) of speaker DoA estimation on the SSLR test set with different SNRs and face detection swap percentage (FDSP)s. The results are obtained using the MLP-AVAW network architecture.

		MLP-GCC [4]		**Face Detection Swap Percentage**									
				0%		**10%**		**30%**		**50%**		**70%**	
		MAE	**ACC**	**MAE**	**ACC**	**MAE**	**ACC**	**MAE**	**ACC**	**MAE**	**ACC**	**MAE**	**ACC**
audio	**-10**	52.74	24.2	49.69	26.59	49.98	26.5	50.51	26.2	50.87	26.0	51.21	25.8
	0	26.19	54.8	22.19	57.3	22.37	57.2	22.73	57.0	23.01	56.7	23.22	56.6
	10	10.64	78.8	9.34	79.2	9.41	79.1	9.52	79.0	9.61	78.9	9.67	78.8
	20	6.02	89.0	5.68	89.0	5.70	89.0	5.75	89.0	5.79	88.9	5.82	88.9
SNR (dB)	**Clean**	4.63	91.6	4.22	92.0	4.24	92.0	4.28	91.9	4.31	91.9	4.32	91.9

3.2. Parameter settings

The GCC-PHAT is computed for every 170 ms segments with delay lags $\tau \in [-25, 25]$, resulting in 51 coefficients for each microphone pair as in [4]. With 6 microphone pairs, each pair contributing 51 GCC-PHAT coefficients, we obtain 306 GCC-PHAT coefficients. For visual features, the human face width and height are assumed to have $W = 0.14\ m, H = 0.18\ m$, respectively as such in [12]. We adjust the size of the horizontal and vertical visual feature encoding to 51 to match that of GCC-PHAT coefficients.

We use the Adam optimizer [25]. All models are trained for 10 epochs with a batch size of 256 samples and a learning rate of 0.001. Since multi-speaker localization is not a single-label classification problem, we use Mean Square Error (MSE) instead of cross-entropy as the loss function.

3.3. Results

Tab. 2 provides the experimental results on the SSLR test set. Results are separately reported for different subsets and the speaker number (assumed to be known). The best result for each column is in the bold font. We compare the results of MLP-AVC and MLP-AVAW with two audio baseline methods: the traditional Steered Response Power PHAse Transform (SRP-PHAT) method [2] and the state-of-the-art MLP-GCC method [4]. As speakers are not always visible, we don't provide the video-only baseline to avoid unfair comparison. Furthermore, Tab. 1 suggests that it is challenging to expect visual features alone to outperform the audio DoA estimation.

Tab. 2 shows that, by both early fusion of audio-visual features. In particular, MLP-AVC reduces MAE from $4.63°$ (MLP-GCC) to $4.42°$, which confirms the audio-visual fusion benefits. For the test-human subset, speakers are mostly inside the camera's FoV (the red points locate in the gray region in Fig. 4(a)) and DR of the RetinaFace detector [**?**] achieves 100 %, which is much higher than DR in test-loudspeaker (9.2 %). Thus, the MAE degradation in test-human (from $4.75°$ to $1.84°$ and from $5.98°$ to $3.89°$) is more significant than in test-loudspeaker (from $4.06°$ to $3.87°$ and from $8.10°$ to $7.80°$). Besides, further improvements are introduced by the adaptive weighting mechanism in MLP-AVAW, which achieves the best results in most cases with the overall MAE at $4.22°$ and ACC at 92.0%.

Next, we further evaluate the noise robustness of the proposed networks. For audio, we apply additive white Gaussian noise of SNRs varying from $-10\ dB$ to $20\ dB$ on the original SSLR audio signals. For video, we randomly swap up to 70% face detections to the other frames to generate false positives and false negatives. Tab. 3 lists the overall MAE and ACC of MLP-AVAW in comparison with those under clean audio condition. We also provide the MLP-GCC results in the first two columns indicating the audio-only performance without swapping the face detection. From the results, we can see that fusing visual features always brings benefits. Additionally, audio is of more importance than video since with the degradation of SNR, both MAE and ACC are getting worse as Face Detection Swap Percentage (FDSP) increases, the performance degradation is also obvious but not so significant. Even at FDSP=70%, the proposed network still outperforms the MLP-GCC. The performance gains by MLP-AVAW suggest that visual features provide additional information in degraded acoustic conditions.

4. CONCLUSIONS

This paper presented two neural network architectures for multi-speaker DoA estimation using audio-visual signals. The comprehensive evaluation results confirm the benefits of audio-visual fusion and the adaptive weighting mechanism. Besides, we proposed a technique to synthesize visual features from geometric information about the sound sources to deal with lack of annotated audio-visual data. Future work will include exploring network models that can generalize with limited training data.

5. REFERENCES

[1] Charles Knapp and Glifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug 1976.

[2] Michael S Brandstein and Harvey F Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, 1997, pp. 375–378.

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

[4] Weipeng He, Petr Motlicek, and Jean-Marc Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. of Int. Conf. on Robotics and Automation*, 2018, pp. 74–79.

[5] Soumitro Chakrabarty and Emanuël AP Habets, "Multi-speaker DoA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, Mar 2019.

[6] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. of European Signal Processing Conf.*, 2018, pp. 1462–1466.

[7] Zihan Pan, Malu Zhang, Jibin Wu, and Haizhou Li, "Multi-tones' phase coding (mtpc) of interaural time difference by spiking neural network," *arXiv preprint arXiv:2007.03274*, 2020.

[8] Aggelos K Katsaggelos, Sara Bahaadini, and Rafael Molina, "Audiovisual fusion: Challenges and new approaches," *Proc. of the IEEE*, vol. 103, no. 9, pp. 1635–1653, Aug 2015.

[9] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, Nov 2010.

[10] Shankar T Shivappa, Mohan Manubhai Trivedi, and Bhaskar D Rao, "Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey," *Proc. of the IEEE*, vol. 98, no. 10, pp. 1692–1715, Aug 2010.

[11] Matthew J. Beal, Nebojsa Jojic, and Hagai. Attias, "A graphical model for audiovisual object tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 828–836, Jun 2003.

[12] Xinyuan Qian, Alessio Brutti, Oswald Lanz, Maurizio Omologo, and Andrea Cavallaro, "Multi-speaker tracking from an audio-visual sensing device," *IEEE Trans. on Multimedia*, vol. 21, no. 10, pp. 2576–2588, Oct 2019.

[13] Yutong Ban, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud, "Variational Bayesian inference for audio-visual tracking of multiple speakers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Nov 2019.

[14] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon, "Learning to localize sound source in visual scenes," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 4358–4366.

[15] Antigoni Tsiami, Petros Koutras, and Petros Maragos, "STAViS: spatio-temporal audiovisual saliency network," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 4766–4776.

[16] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu, "Audio-visual event localization in unconstrained videos," in *Proc. of European Conf. on Computer Vision*, 2018, pp. 247–263.

[17] Janani Ramaswamy and Sukhendu Das, "See the sound, hear the pixels," in *The IEEE Winter Conf. on Applications of Computer Vision*, 2020, pp. 2970–2979.

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

[19] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye, "Object detection in 20 years: A survey," *CoRR*, vol. abs/1905.05055, 2019.

[20] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2003.

[21] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. of Int. Conf. on Machine Learning*, 2010, pp. 807–814.

[22] Ioffe Sergey and Szegedy Christian, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," vol. 37, pp. 448–456, Jul 2015.

[23] Guillaume Lathoud, Jean-Marc Odobez, and Daniel Gatica-Perez, "AV16. 3: an audio-visual corpus for speaker localization and tracking," in *Machine Learning for Multimodal Interaction*, pp. 182–195. Springer, Martigny, Switzerland, Jun 2004.

[24] Antoine Deleforge, Radu Horaud, Yoav Y Schechner, and Laurent Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 23, no. 4, pp. 718–731, Apr 2015.

[25] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.