Three-Dimensional Speaker Localization: Audio-Refined Visual Scaling Factor Estimation

Xinyuan Qian, *Member, IEEE*, Qi Liu, *Member, IEEE*, Jiadong Wang, and Haizhou Li, *Fellow, IEEE*

***Abstract*—Neither a monocular RGB camera nor a small-size microphone array is capable of accurate three-dimensional (3D) speaker localization. By taking advantage of accurate visual object detection, and audio-visual complementary sensor fusion, we formulate the three-dimensional (3D) speaker localization problem as a visual scaling factor estimation problem. As a result, we effectively reduce the traditional audio-only 3D speaker localization from an exhaustive grid search to a one-dimensional (1D) optimization problem. We propose a multi-modal perception system with two optimization approaches. We show that the proposed methods are effective, accurate, and robust against interference and, as corroborated by indicative empirical results on real dataset, competitive to the conventional uni-modal and the state-of-the-art audio-visual speaker localization approaches.**

***Index Terms*—Multimodal perception, speaker localization, TDoA, audio-visual fusion, dynamic sensor weighting.**

I. Introduction

MULTIMODAL perception is fertile research ground that merits further investigation, and has been extensively used in cognitive science, behavioral science, and neuroscience owing to its capabilities of enabling brains to learn meaningful information from different sensory modalities, including sound, sight etc [1]. Audio and vision, as the major perception peripheral in Human Computer Interaction (HCI) systems, convey significant and complementary information for scene understanding [2]–[5].

Acoustic Sound Source Localization (SSL) with multi-channel microphones [6]–[8], as one of the most profound localization techniques, has spurred on-going interests for many far-field speech applications, such as automatic speech recognition [9] and separation [10]. However, acoustic SSL only works well in relatively clean conditions, and suffers from noise and reverberation distortions. Moreover, it relies on an extensive grid search algorithm to find the 3D position e.g. Steered Response Power (SRP) [11], that is computationally demanding. To address that, Stochastic Region Contraction (SRC) [12], hierarchical search [13], [14] and vectorization [15] are proposed to speed up the scanning, which usually restrict the search to a two-dimensional (2D) space. As pointed out by [16], a static small-size microphone array is incapable of localizing the speaker in 3D domain.

In some specific environments, e.g., indoor healthcare service robots, where line-of-sight acoustic propagation is impaired by obstacles, SSL techniques perform poorly with low localization accuracy. There have been studies where HCI systems with visual information are developed to provide assistance and support for object localization. However, computer vision is sensitive to illumination variation and complex occlusions, which limits its applications.

The recent advances in artificial neural networks (ANN) have created immense opportunities for the development of SSL [17]. A convolutional neural network-based method has been proposed for 2D object detection from monocular images [18]. Despite progress, using single RGB image captured from monocular camera for 3D object localization remains a challenging task. Compared with the solutions such as LiDAR and stereo vision [19], the monocular method is far from satisfactory due to the lack of depth-of-field information [20]. To deal with this problem, ANN-based 3D monocular methods for object localization have attracted increasing attentions owing to their superior results [20]. However, their success relies on the availability of a large amount of human-labelled training data, that is not always available.

Audio-only (AO)- and video-only (VO)-based processing methods are complementary. They have their respective merits and drawbacks. It is attractive for SSL to benefit from the best of both. In this paper, we exploit the audio-visual signals captured from the calibrated platforms to improve the localization accuracy of a visible and audible speaker in a 3D domain.

Let us denote t as the time index, $\mathbf{s}_t^{1:N}$ as the audio signals captured by a N-channel microphone array, indexed by $n \in \mathcal{N} = \{1, \ldots, N\}$, $\mathbf{B}_t$ as the image captured by a standard monocular RGB camera. $\mathbf{\Omega} \in \mathbb{R}^{3\times3}$ and $\mathbf{\Xi} \in \mathbb{R}^{3\times4}$ are represented as the camera intrinsic and extrinsic matrices, respectively. $\mathbf{p}^{1:N} \in \mathbb{R}^{3\times N}$ indicates the 3D location of microphones. We aim to estimate the target 3D location via:

$$\mathbf{o}_t = f(\mathbf{s}_t^{1:N}, \mathbf{B}_t | \mathbf{p}^{1:N}, \mathbf{\Omega}, \mathbf{\Xi}) \tag{1}$$

where $\mathbf{o}_t \in \mathbb{R}^{3\times1}$, and f is a transfer function. For the ease of clarity, the time index t is omitted without loss of generality.

On the basis of intra-array time-delay features and visual face detections, we propose an Audio-Visual Speaker Localization

Manuscript received February 15, 2021; revised June 15, 2021; accepted June 24, 2021. Date of publication June 28, 2021; date of current version July 26, 2021. This work was supported by the Singapore Government.s Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain) under Programmatic Grant I2001E0053 and by A*STAR, Singapore, through the National Robotics Program under Grant 192 25 00054. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Peter Jax. *(Corresponding author: Qi Liu.)*

The authors are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: eleqian@nus.edu.sg; elelqi@nus.edu.sg; jiadong.wang@nus.edu.sg; haizhou.li@nus.edu.sg).

Digital Object Identifier 10.1109/LSP.2021.3092959

(AVSL) system. As the standout amongst topics in computer vision, face detection delivers impressive 2D accuracy in a 3D domain. By incorporating 2D visual sensing, we re-formulate the audio-only 3D localization problem to an audio-visual 1D visual scaling factor estimation problem. Two objective functions are proposed, where a directional unit vector derived from the face detection points towards the speaker position to guide the audio searching space. This work makes notable contributions in two areas:

- We propose a spatial setup using a monocular camera and a small-size microphone array to exploit audio-visual complementary and interactive characteristics for 3D speaker localization.
- We propose and implement the algorithm with multi-modal objective functions to accurately estimate the depth of speaker location through the audio-refined video scaling factor estimation.

II. Acoustic Sound Source Localization

The time-delay based method [21], e.g., Time Difference of Arrival (TDoA) estimation [22] is one of the most popular SSL techniques, which exploits the fact that the time-of-arrival of a signal is proportional to the source-microphone distance. Given a generic 3D location $\mathbf{o}$, the actual TDoA to a microphone pair, indexed by $m \in \mathcal{M} = \{(m_1, m_2), \forall m_1 < m_2 \leq N\}$, is computed as [11]:

$$\tau_m = \frac{||\mathbf{o} - \mathbf{p}^{m_1}|| - ||\mathbf{o} - \mathbf{p}^{m_2}||}{c} f_a \tag{2}$$

where $||\cdot||$ stands for the ℓ_2 norm, $\mathbf{p}^{m_1}$ and $\mathbf{p}^{m_2}$ are the 3D position vectors. $f_a \in \mathbb{R}_+$ and $c \in \mathbb{R}_+$ hold the sampling frequency (kHz) and the sound speed (m/s), respectively.

To perform robust TDoA estimation in noisy and reverberant environment, Generalized Cross Correlation with Phase Transform (GCC-PHAT) [21], [23] is widely used. Let $\mathbf{S}_{m_1}$ and $\mathbf{S}_{m_2} \in \mathbb{C}$ be the Fourier transforms of audio streams at the m^{th}-indexed microphone pair. We compute the GCC-PHAT features with different time delay τ as:

$$g^m(\tau) = \sum_k \mathcal{R}\left(\frac{\mathbf{S}_{m_1}[k]\mathbf{S}_{m_2}[k]^*}{|\mathbf{S}_{m_1}[k]\mathbf{S}_{m_2}[k]^*|} e^{2\pi j \frac{k}{K}\tau}\right) \tag{3}$$

where j is the imaginary unit, k is the frequency bin index, * and $|\cdot|$ stand for the complex conjugate and the absolute value operations, respectively. $\mathcal{R}$ denotes the real part of complex number, and K is the Fast Fourier Transform (FFT) length. Therefore, the sound source TDoA can be estimated via searching for the GCC-PHAT peaks:

$$\hat{\tau}_m = \text{argmax}_{\tau \in \Gamma_m} \, \text{g}^{\text{m}}(\tau) \tag{4}$$

where $\hat{\tau}_m \in \mathbb{Z}$, $\Gamma_m = [-\tau_m^{max}, \tau_m^{max}]$ is the set of allowable time delay with τ_m^{max} depending on the inter-microphone distance $\tau_m^{max} = \frac{1}{c}||\mathbf{p}^{m_1} - \mathbf{p}^{m_2}||f_a$.

III. Audio-Visual Speaker Localization

A monocular RGB camera and a small-size microphone array are utilized to jointly localize a speaker in 3D space. After back-projecting a target image hypothesis to a ray of the camera, the intersection between the ray and the microphone array's receptive area facilitates speaker localization.

Let $\mathbf{f} = (u, v, w, h)^\intercal$ be the face detection bounding box where $^\intercal$ denotes the transpose operator, (u, v) indicate the horizontal and vertical locations of the top-left point of the bounding box in the image coordinate, and (w, h) are the width and height of the bounding box. The sound source location (speaking lips) is extracted via:

$$\boldsymbol{\mu} = \boldsymbol{\Psi}\mathbf{f} \tag{5}$$

in which $\boldsymbol{\Psi} \in \mathbb{R}^{2\times4}$ is a pre-defined extraction matrix. The mapping from a 3D point $\mathbf{o} \in \mathbb{R}^{3\times1}$ in the physical world to an image point $\boldsymbol{\mu} \in \mathbb{R}^{2\times1}$ is described by the camera projective transformation. Given a pinhole camera model, the perspective projection is formulated as [24]:

$$\kappa\boldsymbol{\mu} = \boldsymbol{\Omega}\boldsymbol{\Xi}\vec{\mathbf{o}} \tag{6}$$

where κ is the visual scaling factor depending on the object-camera distance, $\boldsymbol{\mu} = [\boldsymbol{\mu}^\intercal, 1]^\intercal \in \mathbb{R}^{3\times1}$ and $\vec{\mathbf{o}} = [\mathbf{o}^\intercal, 1]^\intercal \in \mathbb{R}^{4\times1}$ are the representations of target image and 3D locations in homogeneous coordinates, respectively.

From 6, the desired 3D location can be formulated as:

$$\vec{\mathbf{o}} = \mathcal{O}(\kappa) \triangleq \kappa \begin{pmatrix} \boldsymbol{\Phi} \\ \mathbf{0}\ 1 \end{pmatrix}^{-1} \vec{\mu} \tag{7}$$

where $\mathbf{0} \in \mathbb{R}^{1\times3}$ is an all-zero vector, $\boldsymbol{\Phi} \in \mathbb{R}^{3\times4}$ is the camera projection matrix, and $(\cdot)^{-1}$ indicates matrix inversion. From Eq. 7, we observe that, for $\kappa \to \infty$, the perspective projections of the 3D points $\{\vec{\mathbf{o}} = \mathcal{O}(\kappa), \forall\kappa \in \mathbb{R}_+\}$ all converge to the same image position $\boldsymbol{\mu}$. Thus, it is inaccurate to use the monocular images alone to estimate the 3D speaker location in a world coordinate, due to the projection ambiguity. The projection ambiguity can be redeemed given the prior of the detected object size [24]. However, the resulting κ is erroneous due to different head sizes and the orientations to the camera, which may adversely affect the estimation of the 3D point $\vec{\mathbf{o}}$.

Let's now formulate the 3D speaker localization problem as a 1D scaling factor estimation problem:

$$\hat{\mathbf{o}} = \mathcal{O}(\hat{\kappa})_{h^{-1}} \tag{8}$$

where the subscript h^{-1} indicates the inverse of homogeneous transformation to get rid of the last element of $\vec{\mathbf{o}}$.

We propose two objective functions, namely Multi-channel Cost Function (MCF) and Global Likelihood Function (GLF), and adopt the grid searching method to select the optimal visual scaling factor κ from the pre-defined hypotheses. In particular, MCF considers the localization error resulting from each individual microphone pair while GLF treats the microphone array as an entire unit and considers the accumulated confidence. Details are described in the next subsections.

A. MCF-Optimization

The speaker locations estimated from all microphone pairs and the camera share the same spatial space. To leverage such spatial correspondence, we build an objective function at κ

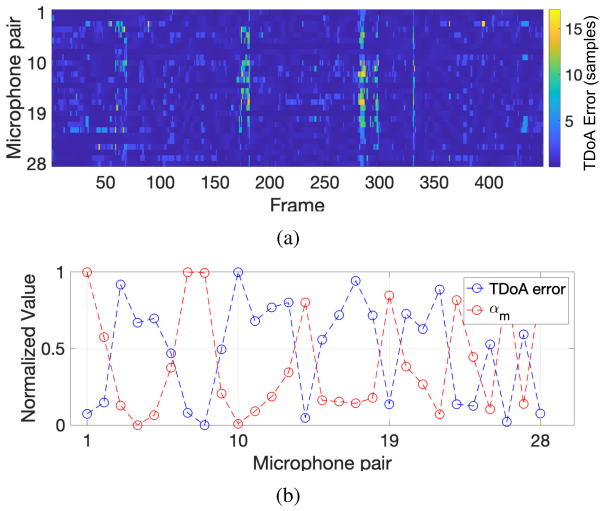

Fig. 1. (a) The absolute TDoA error (in sample) across time frames and microphone pairs, where yellow or blue denotes greater or smaller error, respectively; and (b) The variations of the TDoA error and the proposed adaptive parameter α_m (in Eq. 10) at different microphone pairs averaged over time.

indicating the difference between audio and video estimates:

$$\mathcal{B}_m(\kappa) = \overbrace{||\mathcal{O}(\kappa)_{h^{-1}} - \mathbf{p}^{m_1}|| - ||\mathcal{O}(\kappa)_{h^{-1}} - \mathbf{p}^{m_2}||}^{\text{video}} - \overbrace{\frac{\hat{\tau}_m c}{f_a}}^{\text{audio}} \tag{9}$$

where the first under-bracket item correlates to the theoretical TDoA of the visual 3D location estimate (see Eq. 8) while the second item is derived from the audio GCC-PHAT estimate.

The time delay estimation from the microphone pairs is affected by many factors, such as audio content, distance between the microphone pairs, speaker-microphone distance, and speaking head orientation. Fig. 1(a) indicates the TDoA estimation error varies across each microphone pair (y-axis, totally 28 pairs for a 8-channel array) at different time frames (x-axis, large error at frame 280 results from non-direct speaker head orientation to the array). To leverage different pair-level importance, we design an adaptive weighting parameter $\alpha_m \in [0, 1)$, which updates at every frame, formulated as:

$$\alpha_m = \frac{\tilde{g}^m \delta(\tilde{g}^m \geq \theta)}{\max\left(\sum_m \tilde{g}^m \delta(\tilde{g}^m \geq \theta), \vartheta_\alpha\right)} \tag{10}$$

where $\tilde{g}^m = g^m(\hat{\tau}_m)$, $\delta(\cdot)$ is an indicator function that outputs one if the inequality holds, and ϑ_α is a speech threshold chosen to be 1.4. The parameter $\theta = median(\{\tilde{g}^m, m \in \mathcal{M}\})$ models the median of GCC-PHAT peaks, to dynamically choose half of the pairs with greater reliability while eliminating the others. Herein, we re-use the GCC-PHAT peaks as the reliability measure, considering no additional computation. Fig. 1(b) shows the inverse proportional relationship between the proposed α_m and the TDoA error. The microphone pair with a smaller TDoA error is assigned to a larger α_m due to higher reliability.

The performance of GCC-PHAT significantly degrades when there is interference (i.e., environmental noise and reverberation), which penalizes the utilization of audio-visual spatial correspondence. Note that face detection is not affected by such interference. We incorporate the visual contribution to solve the scaling factor,

$$\hat{\kappa} = \arg\min_{\kappa} \overbrace{\sum_{m\in\mathcal{M}} \alpha_m \left(\mathcal{B}_m(\kappa)\right)^2}^{\text{microphone pairs}} + \overbrace{\alpha_v ||\kappa - \kappa_0||^2}^{\text{camera}}$$
$$\text{s.t.} \sum_{m\in\mathcal{M}} \alpha_m + \alpha_v = 1 \tag{11}$$

where the first under-bracket item correlates to the accumulated difference between multi-modal TDoA estimates at individual microphone pairs (Eq. 9-10), while the second item represents the distance to the visual initial scaling factor κ_0. α_v is the visual dynamic parameter that is adjusted according to audio reliability, which makes video contribute more under adverse acoustic conditions (small α_m) and less otherwise.

B. GLF-Optimization

Despite the incorporation of the dynamic weights in MCF-optimization, the errors resulting from inaccurate TDoA estimates may accumulate at microphone pairs. To overcome this, we propose to estimate the visual scaling factor by maximizing the following likelihood function:

$$\hat{\kappa} = \arg\max_{\kappa} \overbrace{\beta_a \sum_{m\in\mathcal{M}} g^m(\tau_m(\kappa))}^{\text{microphone array}} + \overbrace{\beta_v \exp\left(-||\kappa - \kappa_0||\right)}^{\text{camera}}$$
$$\text{s.t.} \sum \beta_a + \beta_v = 1 \tag{12}$$

where $\tau_m(\kappa)$ is the visual theoretical TDoA computed at $\hat{\mathbf{o}}$ (Eq. 6), and $\beta_a \in [0, 1)$ is an adaptive parameter defined as:

$$\beta_a = \frac{\sum_m \tilde{g}^m}{\max(\sum_m \tilde{g}^m, \vartheta_\beta)} \tag{13}$$

where $\vartheta_\beta = 2.8$ is a speech threshold.

Unlike the adaptive parameters in Eq. 10 which apply at each microphone pair, in Eq. 13, we treat the microphone array as a unit sensor which is allocated with a single weight.

IV. Experiments

We evaluate the proposed 3D speaker localization framework on AV16.3 [25], a real-world audio-visual dataset, and benchmark against the competitive SSL systems, AO [11], VO [24], AVz [26] and 2LPF [27]. On the basis of sensor calibration and multi-modal synchronization information, taken from AV16.3, we seek to estimate the speaker's 3D location.

AV16.3 is recorded by two 8-channel circular microphone arrays with 20-cm diameter (placed on a desk) and three cameras (mounted on different room walls). Images are captured at 25 fps with 360×288 pixels while audio signals are sampled at $f_a =$ 16 kHz. The same as [26], [27], we use the speech segments in the single-speaker sequences (i.e., seq08, 11 and 12) where only the first array and individual cameras are involved in our evaluations. Herein, we report the results in terms of Mean Absolute Error (MAE), which measures the distance between our location estimate and the actual speaker 3D location, and Acurracy (ACC) to measure the percentage of correct estimations with 0.2 m as the 3D error allowance.

TABLE I
SSL RESULTS IN MAE (M) ON AV16.3. THE OVERALL ACC IS LISTED IN THE LAST ROW (SEQ: SEQUENCE; NA: INFORMATION UNAVAILABLE; TEMPORAL FILTERING IS INVOLVED IN [27]⋆)

		Uni-modal		Multi-modal			
Seq	**Camera**	**AO** [10]	**VO** [23]	**AVz** [25]	**2LPF** [26]⋆	**MCF**	**GLF**
08	1	0.36	0.25	0.22	0.10	0.09	0.07
	2	0.38	0.40	0.21	0.08	0.13	0.11
	3	0.38	0.40	0.23	0.06	0.15	0.13
11	1	0.60	0.24	0.33	0.26	0.18	0.13
	2	0.60	0.18	0.30	0.08	0.11	0.10
	3	0.61	0.21	0.32	0.07	0.13	0.11
12	1	0.57	0.28	0.43	0.20	0.13	0.11
	2	0.57	0.36	0.36	0.11	0.13	0.12
	3	0.56	0.42	0.37	0.12	0.18	0.18
MAE(m)		0.52	0.31	0.31	0.12	0.14	0.12
ACC%		(20.02)	(40.27)	(53.81)	NA	(79.57)	(83.31)

A. Parameter Settings

The audio FFT window is set at 2^{12} points (viz. 256 ms) which is shifted at the same frame rate as that of the video frame for audio-visual alignment. The sound speed is $c = 342$ m/s. There are a total of $M = 28$ microphone pairs for the 8-channel array. Additionally, the mouth extraction matrix is set to $\Psi = (1\,0\,\frac{1}{2}\,0; 0\,1\,0\,\frac{3}{4})$. The speaker visual direction is extracted from the detected faces according to [18]. The initial scaling factor κ_0 is computed with the assumption of 3D face size $(W, H) = (0.14\ m, 0.18\ m)$, and the scaling factor range becomes $\mathcal{K} = \{\kappa_0 + i\Delta_\kappa\}_{i=-25}^{25}$ with the incremental step $\Delta_\kappa = 0.02\ m$.

B. Experimental Results

We summarize the results for individual audio-visual sequences in TABLE I. It is observed that MCF and GLF consistently outperform the uni-modal methods. The localization error decreases from 0.52 m (AO) to 0.14 m (MCF) while the accuracy increases from 20.02% (AO) to 79.51% (MCF).

It is to be noted that 2LPF [27] employs additional particle filtering which indicates the upper-bound of localization. Nonetheless, MCF and GLF still achieve comparable results without the temporal smoothing. We further observe that GLF outperforms MCF with the MAE decreasing from 0.14 m to 0.12 m and ACC improving from 79.57% to 83.31%. This suggests that maximizing the accumulated likelihood at the microphone array unit is superior to minimizing the error at individual microphone pairs.

We also evaluate the proposed framework in the presence of additive white Gaussian noise at Signal-to-noise Ratio (SNR) ranging from -20 dB to 40 dB. Fig. 2 summarizes the average MAE of the comparative methods, namely, AO [11], AVz [26], VO [24], and the proposed MCF and GLF. We observe that MCF and GLF significantly outperform AO [11] and AVz [26] under all SNR conditions. AVz applied acoustic localization on the video-suggested speaker height plane with the performance degrades at low SNR due to the dominant impact of audio. By comparing with VO, it is clear to see the superiority of our proposal at high SNR (≥ 10 dB). With degrading SNR (≤ 0 dB), MCF and GLF are upper-bounded by the visual performance.

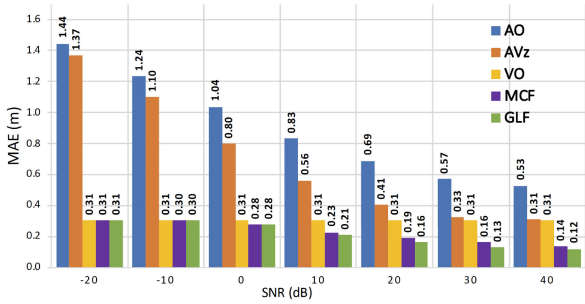

Fig. 2. Comparison of the 3D localization results of different methods in MAE under various SNR (dB).

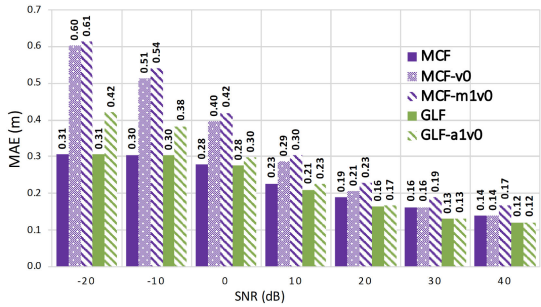

Fig. 3. Ablation study of our proposals under different SNRs (dB). MCF-v0: no video weights ($\alpha_v = 0$); MCF-m1v0: no audio-visual weights ($\alpha_m = 1/M, \alpha_v = 0$) and GLF-a1v0: no audio-visual weights ($\beta_a = 1, \beta_v = 0$).

An ablation study is conducted to understand the contribution of the proposed sensor weighting mechanism at various SNR (see Fig. 3). It is observed that the visual signal doesn't contribute much at a high SNR (40 dB). However, as the SNR decreases, the visual signal's contribution increases. We denote MCF-v0 as the method without visual contribution.

By turning off the adaptive microphone pair weighting mechanism, i.e., keeping all microphone pairs with equal contribution $\alpha_m = 1/M$, we denote the method as MCF-m1v0. We observe that MCF-v0 has a lower MAE than MCF-m1v0, that further validates the rationality of the proposed microphone pair weighting mechanism. With $\beta_v = 0$ (no visual contribution), we denote the method as GLF-a1v0. We observe that the contribution from visual signal increases as the SNR decreases. All results point to the fact that visual signal makes more significant contribution at lower SNR.

V. CONCLUSION

We propose an audio-visual speaker 3D localization framework through estimating the monocular visual scaling factor with assistance of a microphone array. By doing so, we limit the computationally expensive grid searching region from 3D to the ray of the camera. We study a dynamic weighting mechanism between audio and vision, and implement two optimization approaches on real-world dataset. Experiments have validated the proposed visual scaling factor estimation scheme, the audio-visual dynamic weighting mechanism, and the superiority of the proposed methods over the state-of-the-art counterparts, especially under adverse conditions.

REFERENCES

[1] A. Jaimes and N. Sebe, "Multimodal human computer interaction: A survey," in *Proc. Int. Workshop Hum.-Comput. Interact.*, Las Vegas, NV, USA, 2005, pp. 1–15.

[2] 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.

[3] 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.

[4] Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Exploiting the complementarity of audio and visual data in multi-speaker tracking," in *Proc. IEEE Int. Conf. Comput. Vision*, Venice, Italy, 2017, pp. 446–454.

[5] Y. Liu, Q. Hu, Y. Zou, and W. Wang, "Labelled non-zero particle flow for SMC-PHD filtering," in *Proc. Int. Conf. Audio, Speech, Signal Process.*, Brighton, U.K., 2019, pp. 5197–5201.

[6] D. Salvati, C. Drioli, and G. L. Foresti, "Sound source and microphone localization from acoustic impulse responses," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1459–1463, Oct. 2016.

[7] Y. Ban, X. Alameda-Pineda, C. Evers, and R. Horaud, "Tracking multiple audio sources with the von Mises distribution and variational EM," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 798–802, Jun. 2019.

[8] J. Wang, X. Qian, Z. Pan, M. Zhang, and H. Li, "GCC-PHAT with speech-oriented attention for robotic sound source localization," in *Proc. Int. Conf. Robot. Automat.*, 2021, pp. 74–79.

[9] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proc. IEEE*, vol. 109, no. 2, pp. 124–148, Feb. 2021.

[10] L. Wang, J. D. Reiss, and A. Cavallaro, "Over-determined source separation and localization using distributed microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1573–1588, Sep. 2016.

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

[12] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Proc. Int. Conf. Audio, Speech, Signal Process.*, vol. 1, 2007, pp. 1–121.

[13] M. B. Çöteli, O. Olgun, and H. Hacíhabiboglu, "Multiple sound source localisation with steered response power density and hierarchical grid refinement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2215–2229, Nov. 2018.

[14] L. O. Nunes *et al.*, "A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5171–5183, Oct. 2014.

[15] B. Lee and T. Kalker, "A vectorized method for computationally efficient SRP-PHAT sound source localization," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2010, pp. 1–5.

[16] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2193–2206, Oct. 2013.

[17] 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.

[18] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. Int. Conf. Comput. Vision. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 1522–1530.

[19] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun, "3D object proposals using stereo imagery for accurate object class detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1259–1272, May 2017.

[20] X. Zhou, Y. Peng, C. Long, F. Ren, and C. Shi, "MoNet3D: Towards accurate monocular 3D object localization in real time," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11503–11512.

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

[22] H. Cao, Y. T. Chan, and H. C. So, "Maximum likelihood TDOA estimation from compressed sensing samples without reconstruction," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 564–568, May 2017.

[23] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 71–74, Jan. 2011.

[24] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Uni. Press, 2003.

[25] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Proc. Int. Workshop Mach. Learn. Multimodal Interaction*, Jun. 2004, pp. 182–195.

[26] X. Qian, A. Xompero, A. Cavallaro, A. Brutti, O. Lanz, and M. Omologo, "3D mouth tracking from a compact microphone array co-located with a camera," in *Proc. Int. Conf. Audio, Speech, Signal Process.*, Calgary, AB, Canada, 2018, pp. 3071–3075.

[27] 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.*, 2019, pp. 1955–1959.