AUDIO-VISUAL TRACKING OF MULTIPLE SPEAKERS VIA A PMBM FILTER

Jinzheng Zhao[⋆]*, Peipei Wu*[⋆]*, Xubo Liu*[⋆]*, Yong Xu*[†]*, Lyudmila Mihaylova*[‡]*, Simon Godsill*[§]*, Wenwu Wang*[⋆]

[⋆] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
[†] Tencent AI Lab, Bellevue, WA, USA
[‡] Department of Automatic Control and Systems Engineering, University of Sheffield, UK
[§] Department of Engineering, University of Cambridge, UK

ABSTRACT

Audio-visual tracking of multiple speakers requires to estimate the state (e.g. velocity and location) of each speaker by leveraging the information of both audio and visual modalities. Estimating the number of speakers and their states jointly remains a challenging problem. We propose an Audio-Visual Possion Multi-Bernoulli Mixture Filter (AV-PMBM) that can not only predict the number of speakers but also give accurate estimation of their states. We also propose a novel sound source localization technique based on DOA information and a deep learning based object detector to provide reliable audio measurements for the AV tracker. To our knowledge, this represents the first attempt using PMBM for multi-speaker tracking with audio visual modalities. Experiments on the AV16.3 dataset demonstrate that AV-PMBM achieves state-of-the-art performance in optimal sub-pattern assignment (OSPA).

Index Terms— multiple-speaker tracking, audio-visual fusion, PMBM filter

1. INTRODUCTION

Tracking multiple speakers simultaneously plays a key role in many civilian applications such as speech recognition [1], human-computer interaction [2], speaker diarization [3] and monitoring [4].

Audio and visual signals, as two important modalities, can provide complementary information to improve tracking robustness and accuracy. For example, if speakers are occluded or disappear from the camera field of view, they can be tracked using audio signals; if the audio information is corrupted by background noise and room reverberation, visual data can be used to locate and detect the speakers. Thus information of multiple modalities can work jointly to improve the tracking performance. There are, however, several challenges that need to be addressed in audio-visual multi-speaker tracking, including (1) fusion of the audio-visual measurements to obtain optimal estimates of the states of the speakers, (2) dealing with the unknown and time-varying number of speakers.

To tackle these problems and provide reliable tracking, several algorithms have been proposed. Particle filter has been used in audio-visual multi-speaker tracking [5]. However, it cannot deal with the time-varying number of speakers. Audio-visual SMC-PHD filter [6] has been proposed to estimate the variable number of speakers and their states jointly. Generalized Labeled Multi-Bernoulli (GLMB) filter has also been applied in audio-visual multi-speaker tracking [7], by incorporating the label information in the posterior density.

Based on the conjugacy property that the predicted and updated distribution follows the prior distribution, Possion multi-Bernoulli mixture (PMBM) filter was proposed in [8] [9], which uses a Possion process to represent the states of undetected objects and uses a multi-Bernoulli mixture to represent detected objects. PMBM has shown superior performance compared to other Bernoulli-based filters in terms of accuracy and speed [10]. It has been applied to object tracking using visual information [11] or LiDAR data [12]. To our knowledge, PMBM has not been exploited in the scenarios of audio-visual speaker tracking.

In this paper, two new contributions are made. Firstly, AV-PMBM is proposed. Secondly, we propose a deep learning based detection method for multiple speakers and this technique can provide reliable audio measurements for the AV tracker. Experiments on AV16.3 show that our tracker can estimate the number of speakers and the state of each speaker accurately. It also gives more robust performance when occlusion happens, as compared with the baseline methods.

2. AUDIO-VISUAL MEASUREMENTS

2.1. Visual Measurements

Previous work [5] [6] employs color histogram as visual measurements, by comparing the similarities of potential areas in an image frame to a reference image. Using deep learning techniques, face detectors can provide more efficient and accurate detections. We use a face detector to provide visual measurements. In [7] and [13], Mxnet [14] is employed to detect human face. However, Mxnet often fails to give a detection result when people are not facing towards the camera. Thus, we use a more robust face detector DSFD [15], which can output coordinates and confidence scores of the bounding boxes

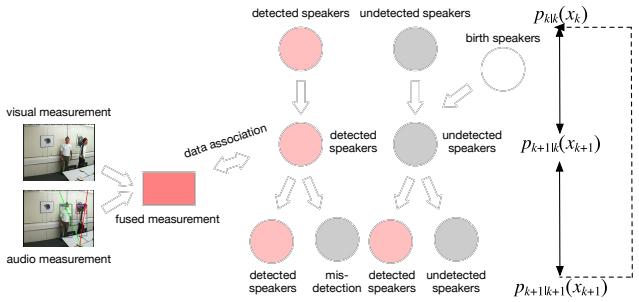

Fig. 1. The model diagram of the proposed AV-PMBM, consisting of an audio-visual fusion module (left) and a tracking module (right).

used to detect faces. More specifically, $\mathbf{b}_{k,i} = [x, y, w, h]^T$ represents the i-th bounding box at time k, where $[x, y]$ is the top left coordinates of the bounding box and $[w, h]$ is the width and height of the bounding box, respectively. Bounding boxes whose confident scores are above a predefined threshold are treated as reliable measurements and are converted to mouth position:

$$\mathbf{o}_{k,i}^{v} = \boldsymbol{W} \cdot \mathbf{b}_{k,i} \tag{1}$$

where $\mathbf{o}_{k,i}^{v}$ is regarded as the i-th visual measurement at time k, $\boldsymbol{W} = [I, \mathrm{diag}(0.5, 0.75)]$ is the coordinates conversion matrix from face bounding boxes to mouth positions, as defined in [13].

2.2. Audio Measurements

Audio information can be used to locate and track people when visual measurements are unavailable or unreliable [7] [13]. Global coherence field (GCF) can be used to locate sound source obtained by summing up the value of Generalized Cross Correlation with Phase Transform (GCC-PHAT) among all microphone pairs. However, the estimated location is not accurate if there are more than one speaker due to the presence of spurious peaks in the GCF map. Although GCC-PHAT deemphasis [16] can provide position estimation in multi-speaker scenarios by re-calculating the GCF map after locating the first speaker, its performance drops significantly as the number of speakers increases. Therefore, we propose a new method that can provide reliable locations for multiple speakers using Direction of Arrival (DOA) lines, i.e. from the center of the microphone array towards the direction of sound sources.

The DOA line provides the angle of a sound source but it can not decide its depth (i.e. the distance of the sound source from the microphone array). Thus we employ an object detector to help estimate the depth of the sound source. As illustrated by Fig. 2, we use the Yolo-v3 object detector [17] pretrained on the COCO dataset [18] to obtain the bounding boxes $\mathbf{a}_{k,i} = [x, y, w, h]^T$ of the speakers, where the notations are the same as those in $\mathbf{b}_{k,i}$. Detected bounding boxes for other classes such as bowls and chairs are removed. First, we estimate the bounding boxes corresponding to the upper part

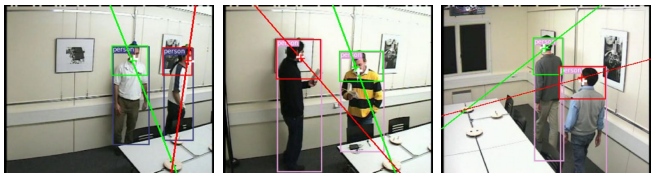

Fig. 2. Mouth positions estimated by DOA combined with object detector. Bounding boxes denoted by *person* are generated by Yolo-v3 detectors. Bounding boxes inside are areas of speakers' upper part of bodies. Lines starting from the microphone array are DOA lines. The ground truths of mouth positions are denoted by white crosses while the estimated mouth positions are marked by green and red crosses.

of the bodies (around the head) of the speakers empirically as $\mathbf{h}_{k,i} = [x, y, w, 0.3 * h]^T$. Then we find the intersection points $\mathbf{s_1} = [x_1, y_1], \mathbf{s_2} = [x_2, y_2]$ between the head bounding boxes and DOA lines. The position of speaker's mouth $\mathbf{o}_k^a$ can be estimated as the middle point between $\mathbf{s_1}$ and $\mathbf{s_2}$. The DOA line is matched to the bounding box if the vertical distance from the mass to this line is the shortest. The unmatched DOA lines are discarded.

3. PROPOSED AV-PMBM

The aim of the proposed tracker is to estimate the number of speakers and the state $\mathbf{x} = (x, v_x, y, v_y)$ of each speaker using both audio and visual data, where (x, y) is the 2D coordinates of the speaker's mouth location and (v_x, v_y) is its corresponding velocity. The overall diagram of the proposed AV-PMBM is shown in Fig.1. In each time step, there are detected speakers that have been associated to measurements, and undetected speakers that exist but are not associated to any measurements. The distribution of undetected speakers $\mathbf{x}^u$ are represented by Poisson point process $\mu(\cdot)$ and the detected speakers $\mathbf{x}^d$ are represented by multiple Bernoulli mixture $f(\cdot)$, where $\mathbf{x}^u$ and $\mathbf{x}^d$ are disjoint subsets of the set of existing speakers $\mathbf{x}$ ($\mathbf{x}^u \uplus \mathbf{x}^d = \mathbf{x}$). The PMBM density $p_k(\cdot)$ at time k represents the information of speakers' state, and is calculated as the convolution of $\mu(\cdot)$ and $f(\cdot)$:

$$p_k(\mathbf{x}) = \sum_{\mathbf{x}^u \uplus \mathbf{x}^d = \mathbf{x}} \mu_k\left(\mathbf{x}^u\right) f_k\left(\mathbf{x}^d\right) \tag{2}$$

3.1. Prediction

The calculation of the predicted distribution $p_{k+1|k}\left(\mathbf{x}_{k+1}\right)$ follows the Chapman Kolmogorov equation:

$$p_{k+1|k}\left(\mathbf{x}_{k+1}\right) = \int \pi\left(\mathbf{x}_{k+1} \mid \mathbf{x}_k\right) p_{k|k}\left(\mathbf{x}_k\right) \delta\mathbf{x}_k \tag{3}$$

where $\pi\left(\mathbf{x}_{k+1} \mid \mathbf{x}_k\right)$ is the transition density. In our experiments, we suppose the transition process is velocity constant [5]. Prediction of undetected speakers and detected speakers

are independent. The predicted density of undetected speakers consists of the union of birth intensity and the predicted density of surviving undetected speakers with surviving probability P^{S}. For detected speakers, the prediction can be achieved with a Kalman filter.

3.2. Update

The distribution at time $k+1$ can be calculated with the measurement model $g\left(\mathbf{z}_{k+1} \mid \mathbf{x}_{k+1}\right)$:

$$p_{k+1|k+1}\left(\mathbf{x}_{k+1}\right)=\frac{g\left(\mathbf{z}_{k+1} \mid \mathbf{x}_{k+1}\right) p_{k+1|k}\left(\mathbf{x}_{k+1}\right)}{\int g\left(\mathbf{z}_{k+1} \mid \mathbf{x}'_{k+1}\right) p_{k+1|k}\left(\mathbf{x}'_{k+1}\right) \delta \mathbf{x}'_{k}} \tag{4}$$

The states of speakers can be divided into four types. For undetected speakers, they are detected for the first time or remain undetected. For detected speakers, they are detected again or they are not detected. We define that speakers are detected if they are associated to a measurement.

For undetected speakers that are misdetected again, the density is decreased by $(1-P^{\mathrm{D}})$, where P^{D} is the detection probability. For undetected speakers that are detected for the first time, the states are represented by a new Bernoulli distribution. For detected speakers that are detected again, a Kalman filter is used to update the states of speakers for each associated measurement. If detected speakers are misdetected, their states remain unchanged.

The measurement likelihood at time k is denoted by $g\left(\mathbf{z}_{k} \mid \mathbf{x}_{k}\right)$. In audio-visual tracking scenarios, we assume the audio and visual measurements are independent. The joint measurement likelihood is defined as:

$$g\left(\mathbf{z}_{k} \mid \mathbf{x}_{k}\right)=g\left(\mathbf{o}_{k}^{a} \mid \mathbf{x}_{k}\right) \cdot g\left(\mathbf{o}_{k}^{v} \mid \mathbf{x}_{k}\right) \tag{5}$$

Both audio and visual likelihood follow Gaussian distribution centered at the estimated position in Section 2:

$$g\left(\mathbf{o}_{k}^{a} \mid \mathbf{x}_{k}\right) \propto \exp\left[-\left(\mathbf{o}_{k}^{a}-\mathbf{x}_{k}\right)^{T} \Sigma_{a}^{-1}\left(\mathbf{o}_{k}^{a}-\mathbf{x}_{k}\right)\right)\Big] \tag{6}$$

and

$$g\left(\mathbf{o}_{k}^{v} \mid \mathbf{x}_{k}\right) \propto \exp\left[-\left(\mathbf{o}_{k}^{v}-\mathbf{x}_{k}\right)^{T} \Sigma_{v}^{-1}\left(\mathbf{o}_{k}^{v}-\mathbf{x}_{k}\right)\right)\Big] \tag{7}$$

where Σ_a and Σ_v represent the accuracy of audio and visual measurements, respectively. If at some time steps, measurement of one modality is absent, the corresponding likelihood is set to uniform distribution. Audio and video streams can work jointly to provide reliable measurements. Visual information can provide more accurate measurement due to the high-performance face detector, while audio information can help locate speakers if the visual measurement is not available (e.g. when the face detector fails due to speakers not facing towards the cameras).

3.3. Data Association

In AV-PMBM, a multi-Bernoulli Random Finite Set (RFS) represents all potential speakers and the multi-Bernoulli RFS mixture represents estimation of all speakers using multiple data associations. In a multi-Bernoulli RFS mixture, the multi-Bernoulli RFS with the highest data association possibility is selected to derive the states of the speakers.

4. EXPERIMENT

4.1. Dataset

The dataset we use is AV16.3 [19], which is recorded with two 8-microphone arrays with a sampling rate at 16 kHz and three cameras with a sampling rate at 25 fps in an $8.2 \times 3.6 \times 2.4 m^3$ meeting room. People in the room are sitting statically, or standing statically, or walking back and forth while speaking at the same time. Sequences are annotated with 2D ground truth month position of the speakers and in our experiment, we use sequence 18, 19, 24, 25 and 30 of all three cameras and the first microphone array. These sequences are regarded as challenging sequences as occlusions happen and speakers are not facing towards the camera at some moments.

4.2. Evaluation Metric

The metric to evaluate the performance of trackers is Optimal Sub-Pattern Assignment (OSPA) [20]. OSPA is defined on two arbitrary finite sets $M=\left\{m_1, m_2, ..., m_{|M|}\right\}$ and $N=\left\{n_1, n_2, ..., n_{|N|}\right\}$, where $|\cdot|$ is the cardinality of the set:

$$E_{\rho}^{(c)}(M, N)= \left(\frac{1}{|N|}\left(\min_{\pi \in \Pi_{|N|}} \sum_{i=1}^{|M|} d^{(c)}\left(m_i, n_{\pi(i)}\right)^{\rho}+c^{\rho}(|N|-|M|)\right)\right)^{\frac{1}{\rho}} \tag{8}$$

where $c>0$ is the cut-off parameter and $\rho \geq 1$ is the order. $\Pi_{|N|}$ is the set of permutations on $\{1,2,...,|N|\}$. $d^{(c)}\left(m_i, n_{\pi(i)}\right)$ is defined as $\min(c, \|m_i - n_{\pi(i)}\|_2)$. OSPA finds optimal point assignment of points in M and N and calculates the Euclidean distance of the two matched points. Unmatched points left in N will cause cardinality error.

4.3. Experimental Results

AV-PMBM is compared with audio-visual adaptive particle filter [5] (AV-A-PF) and sparse mean-shift smc phd filter [6] (MS-SMC-PHD) in terms of OSPA. Every tracker is tested 10 times and the average results are calculated. For OSPA, we choose $\rho=2$ and $c=5$. The detection probability P^{D} is set to 0.9 and the surviving probability P^{S} is set to 0.6. The birth intensity is set as a Gaussian mixture. At the first frame, AV-PMBM is initialized with the measurements. OSPA results and tracking results are shown in Table 1 and Fig. 3, respectively.

Table 1. OSPA results ([5] and [6] are reproduced with the hyper-parameter c equal to 5).

Sequence	Seq 18			Seq 19			Seq 24			Seq 25			Seq 30			**AvgE**	**AvgC**
Camera	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3		
AV-A-PF [5]	4.91	4.85	5.00	4.75	4.94	4.97	4.87	4.86	4.81	4.97	4.76	4.80	4.90	4.82	4.89	4.87	-
MS-SMC-PHD [6]	-	-	-	-	-	-	4.92	4.89	4.94	5.00	4.77	4.98	4.93	4.97	4.98	4.93	0.80
Proposed	3.44	3.23	3.76	2.92	2.89	2.63	4.11	3.34	3.60	3.06	2.18	2.44	3.33	1.67	2.84	**3.03**	**0.06**

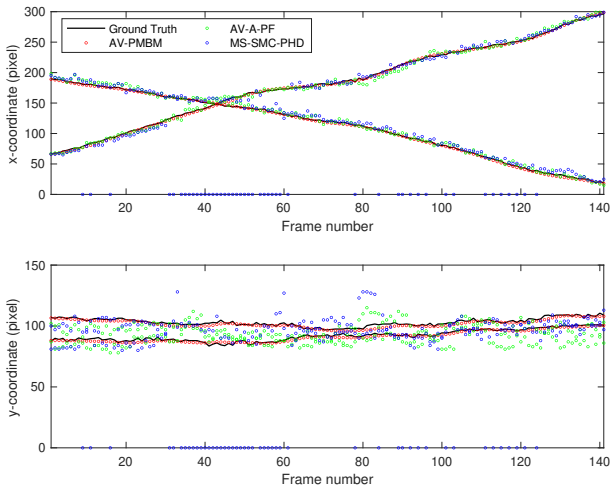

Fig. 3. Tracking results of AV-PMBM, AV-A-PF and MS-SMC-PHD on sequence 25, camera 2 of AV16.3 dataset.

For OSPA, we calculate cardinality error and localization error, respectively. As AV-A-PF is aware of the number of speakers as a prior, its cardinality error is not available. From Table 1, it is clear that AV-PMBM has more accurate tracking performance compared with other methods. In some sequences such as seq25-cam2 and seq30-cam2, it has a lower error. Its overall performance surpasses AV-A-PF though it does not have priors of the numbers of speakers. Fig. 3 shows the visualization of the trackers in an AV sequence. Estimation of AV-PMBM is more close to the ground truth. From frame 30 to 50, speakers are walking towards each other and occlusion happens. MS-SMC-PHD does not work well under this condition (The blue points at the $y = 0$ shows that the tracker fails to track the speakers). AV-PMBM can deal with the occlusion as audio and visual measurements are still reliable and the filter can make correct data associations.

The average cardinality error (AvgC) and the average localization error (AvgE) are shown at the last two columns of Table 1. We can see that AV-PMBM outperforms MS-SMC-PHD by a large margin in terms of cardinality estimation owing to more accurate measurements and reliable data association. We also show the OSPA cardinality error of seq30-cam1 in Fig. 4. It is clear that the cardinality estimation of AV-PMBM is correct at most times. From frame 78 to frame 92, occlusion

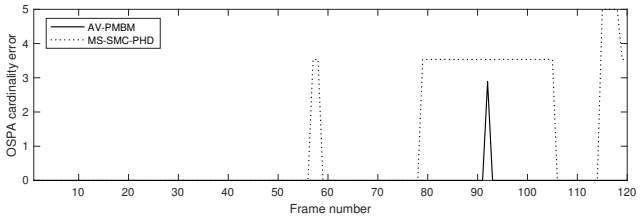

Fig. 4. OSPA cardinality error of AV-PMBM and MS-SMC-PHD on sequence 30, camera 1 of AV16.3 dataset.

happens in sequence. Cardinality estimation by MS-SMC-PHD is erroneous in a large time interval even after occlusion disappears (from frame 92 to frame 106). AV-PMBM shows robust performance to occlusion with only one summit appears at frame 92.

5. CONCLUSION

We presented the AV-PMBM tracker using both audio and visual data. We also proposed a novel sound source localization algorithm by leveraging DOA information assisted by a deep learning based object detector. Experimental results show that AV-PMBM can give more accurate estimation of the number of speakers and each speaker's state than two recent baseline methods, especially when occlusion happens. In our paper, the fusion of audio-visual information is achieved on the decision level. In the future, we will also examine audio-visual fusion in the feature level.

6. ACKNOWLEDGEMENT

This research is sponsored in part by Tencent AI Lab Rhino-Bird Gift Fund and a PhD studentship from the Doctoral College of University of Surrey. The research is also funded in part by the US Army Research Laboratory and the UK MOD University Defence Research Collaboration (UDRC) in Signal Processing under the SIGNetS project. It is accomplished under Cooperative Agreement Number W911NF-20-2-0225. The views and conclusions contained in this document are of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, the MOD, the U.S. Government or the U.K. Government. The U.S. Government and U.K. Government are authorised to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

7. REFERENCES

[1] G. Potamianos, C. Neti, and S. Deligne, "Joint audio-visual speech processing for recognition and enhancement," in *AVSP 2003-International Conference on Audio-Visual Speech Processing*, 2003.

[2] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 882–894, 2010.

[3] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1086–1099, 2017.

[4] A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, and S. Pankanti, "Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking," *IEEE signal processing magazine*, vol. 22, no. 2, pp. 38–51, 2005.

[5] V. Kılıç, M. Barnard, W. Wang, and J. Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Transactions on Multimedia*, vol. 17, no. 2, pp. 186–200, 2014.

[6] V. Kılıç, 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 Transactions on Multimedia*, vol. 18, no. 12, pp. 2417–2431, 2016.

[7] S. Lin and X. Qian, "Audio-visual multi-speaker tracking based on the glmb framework." in *INTERSPEECH*, 2020, pp. 3082–3086.

[8] J. L. Williams, "Marginal multi-bernoulli filters: Rfs derivation of mht, jipda, and association-based member," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 51, no. 3, pp. 1664–1687, 2015.

[9] Á. F. García-Fernández, J. L. Williams, K. Granström, and L. Svensson, "Poisson multi-bernoulli mixture filter: direct derivation and implementation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 4, pp. 1883–1901, 2018.

[10] Y. Xia, K. Granstrcom, L. Svensson, and Á. F. García-Fernández, "Performance evaluation of multi-bernoulli conjugate priors for multi-target filtering," in *2017 20th International Conference on Information Fusion (Fusion)*. IEEE, 2017, pp. 1–8.

[11] S. Scheidegger, J. Benjaminsson, E. Rosenberg, A. Krishnan, and K. Granström, "Mono-camera 3d multi-object tracking using deep learning detections and pmbm filtering," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 433–440.

[12] S. Pang and H. Radha, "Multi-object tracking using poisson multi-bernoulli mixture filtering for autonomous vehicles," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7963–7967.

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

[14] X. Wu, R. He, Z. Sun *et al.*, "A lightened cnn for deep face representation," *arXiv preprint arXiv:1511.02683*, vol. 4, no. 8, 2015.

[15] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "Dsfd: dual shot face detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5060–5069.

[16] A. Brutti, M. Omologo, and P. Svaizer, "Localization of multiple speakers based on a two step acoustic map analysis," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4349–4352.

[17] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[19] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "Av16. 3: An audio-visual corpus for speaker localization and tracking," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2004, pp. 182–195.

[20] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE transactions on Signal Processing*, vol. 56, no. 8, pp. 3447–3457, 2008.