Audio-Visual Speaker Diarization in the Framework of Multi-User Human-Robot Interaction

Timothée Dhaussy, Bassam Jabaian, Fabrice Lefèvre, Radu Horaud

AUDIO-VISUAL SPEAKER DIARIZATION IN THE FRAMEWORK OF MULTI-USER HUMAN-ROBOT INTERACTION

Timothée Dhaussy[1]*, Bassam Jabaian*[1]*, Fabrice Lefèvre*[1]*, Radu Horaud*[2]

[1]LIA-CERI, Avignon University, [2]Inria at Université Grenoble Alpes

ABSTRACT

The speaker diarization task answers the question "who is speaking at a given time?". It represents valuable information for scene analysis in a domain such as robotics. In this paper, we introduce a temporal audio-visual fusion model for multi-users speaker diarization, with low computing requirement, a good robustness and an absence of training phase. The proposed method identifies the dominant speakers and tracks them over time by measuring the spatial coincidence between sound locations and visual presence. The model is generative, parameters are estimated online, and does not require training. Its effectiveness was assessed using two datasets, a public one and one collected in-house with the Pepper humanoid robot.

Index Terms— speaker diarization, multimodal, human-robot interaction

1. INTRODUCTION

Matching speech signal to its emission source is a crucial task to perform an accurate analysis of a scene with different speakers. Commonly named speaker diarization, it consists in assigning audio segments to classes that correspond to speaker identities. This task brings an answer to the question "who spoke when?" [1]. Early work on speaker diarization focused on audio modality [2, 3]. Nowadays, typical audio speaker diarization systems that use audio as input are composed of three components: (1) Speech segmentation, where the audio input is decomposed into short segments, each segment is supposed to have only one speaker, and the noise is filtered out. This component can be seen as a voice activity detection module. (2) Extraction of the audio embedding from the segmented sections through various methods. The most noticeable are MFCCs [4], speaker factors [5] and i-vectors [6]. And (3) Clustering module, where the extracted audio embeddings are clustered into speakers. For this task the number of speakers is determined. It is also possible to pre-process upstream using speech enhancement and denoising technics which lead to significant improvement through deep learning [7]. Speaker representation has been largely improved with the arrival of neural networks and other new methods such as d-vector [8] and x-vector [9]. An interesting alternative consists in the fusion of audio and visual data. The combination of these two modalities provides complementary information and, therefore, audiovisual approaches to speaker diarization are likely to be more robust than audio-only approaches. It can be associated with a face or mouth tracking through spatial coincidence on the image plane [10, 11], on a ground plane [12] or in 3D [13]. Methods for tracking person in 3D require spatially distributed camera networks and microphone arrays which are not tractable in the case of complex (real) scenarios. On the other hand, methods relying on plane or ground image may present a lack of information and suffer much more from occlusion but they offer the advantage of easier data collection and can be utilized in a wider range of scenarios. In addition of mouth position, several methods are based on the synergy between utterances and lip movements through different approaches such as mutual information [2] and deep learning [14].

In this study, we propose a method which models this fusion through the spatial coincidence of visual and sound source localization (SSL) and by combining this concordance model with a dynamic Bayesian formulation that tracks the identity of the active speaker. SSL provides several benefits in multi-user conversations such as the ability to handle overlapping speech segments, eliminating the need for a voice activation module. The proposed method can be applied in various acoustic conditions by leveraging spatial information from SSL and face location. This paper is organized as follows: in Section 2 we present our model for multi-user speaker diarization. Section 3 presents the experimental setup and the evaluation of the proposition.

2. PROPOSED METHOD

2.1. Problem definition

First, we introduce the notations and definitions of variables. Scalars are written in italic, vectors in italic bold and matrices are in italic and underlined. Upper-case letters denote random variables while lower-case letters denote their realization. We represent t the time-step index of both visual and audio frames, synchronized with each other. At frame t, there are at most N visual observations, $\underline{X}_t = (\boldsymbol{X}_{t1}, ..., \boldsymbol{X}_{tn}, ..., \boldsymbol{X}_{tN}) \in \mathbb{R}^{2\times N}$, where the random

variable $\boldsymbol{X}_{tn}$ corresponds to the mouth location of person n in image t. Then, a multi-person tracker provides a time series of N image locations, namely $\underline{X}_{1:t} = \{\underline{X}_1, ..., \underline{X}_t\}$ and associated visual-presence binary masks $\underline{V}_{1:t}$, namely variable V_{tn} associated with $\boldsymbol{X}_{tn}$ such that $V_{tn} = 1$ if person n is present in image t and 0 otherwise. $N_t = \sum_n V_{tn}$ represents the number of persons that are observed at frame t. In practical, when $V_{tn} = 0$, $\boldsymbol{X}_{tn} = \boldsymbol{X}_{t-i,n}$ with $t - i$ the most recent timeframe where $V_{t-i,n} = 1$. We also consider a SSL module that provides the azimuth and elevation of the dominant sound source at each audio frame t. The sound-source location can then be mapped onto the image plane, such that an azimuth-elevation pair of observations is transformed into an image location modeled by a random variable $\underline{Y}_t = (\boldsymbol{Y}_{t1}, ..., \boldsymbol{Y}_{tk}, ..., \boldsymbol{Y}_{tK}) \in \mathbb{R}^{2\times K}$ with K audio-visual observations for a visual frame at t and $\underline{Y}_{1:t} = \{\underline{Y}_1, ..., \underline{Y}_t\}$ its time serie. To these audio-visual observations we associate a speech-activity binary masks $\boldsymbol{A}_{1:t} = \{A_1, ..., A_t\}$, such that $A_t = 1$ if there is an active audio source at frame t or 0 otherwise. The objective is to track dominant speaker(s) at time t in associating audio responsibility over time the audio activity (if any) with one of the tracked persons. Audio sources out of the pictures are not taken into account. This is also referred to as audio-visual speaker diarization, addressed below in the framework of temporal graphical models. We introduce a time-series of discrete latent variables, $\boldsymbol{S}_{1:t} = \{S_1, ..., S_t\}$ such that $S_t = n, n \in 1, 2, ..., N$ if person n is both observed and speaks at frame t, and $S_t = 0$ if none of the visible persons speaks at frame t. Notice that $S_t = 0$ represent two different cases: firstly, there is at least one active sound-source at t ($A_t = 1$) but its or their location cannot be associated with one of the visible persons and it can be interpreted as noise, secondly, there is no active sound-source at t, $A_t = 0$. We will also use another latent variable $Z_{t1:K} = Z_{t1}, .., Z_{tK}$ with $Z_{tk} = n$ which represents the attribution of the sound source k to the visual identity n. $Z_{tk} = 0$ means the source k isn't assigned to any person in the image.

2.2. Speaker Diarization Model

The temporal speaker diarization problem can be formulated as finding a maximum-a-posteriori (MAP) solution, namely finding the most probable configuration of the latent state S_t that maximizes the following posterior probability distribution. Also referred to as the filtering distribution it can be express in the following way:

$$\hat{s}_t = \arg\max_{s_t} P(S_t = n|\underline{y}_{1:t}, \underline{x}_{1:t}, \underline{v}_{1:t}, \boldsymbol{a}_{1:t}) \tag{1}$$

Following Bayes formula, the posterior probability (1) can be written and beyond developped as:

$$P(S_t = s_t|u_{1:t}) = \frac{P(u_t|S_t = s_t)G_{s_t}}{\sum_{j=0}^{N} P(u_t|S_t = j)G_j} \tag{2}$$

With $G_n = \sum_{i=0}^{N} P(S_t = n|S_{t-1} = i)P(S_{t-1} = i|u_{1:t-1})$ and $u_t = (\underline{x}_t, \boldsymbol{v}_t, \underline{y}_t, a_t)$.
The evaluation of (2) is recursive and a reasonable number of person simultaneously tracked need to be considered (5-8) in order to keep the calculation tractable. Computation of this equation requires the observed likelihood $P(u_t|S_t = s_t)$ and the transition probabilities $P(S_t = j|S_{t-1} = i)$ explained in the next two subsections.

2.3. EM Audio-Visual Observation Model

The main feature of the proposed model is its ability to robustly associate the SSL at time t with a person. The expectation-maximization for Gaussian mixture model infers the posterior probability that a person utters speech from audio and visual observations that are mapped onto the same mathematical space. We distinguish two cases. The first one If there is no audio activity at time $t(A_t = 0)$, the posterior can be evaluated with the following formula, where c is a small positive scalar, e.g., $c = 0.2$:

$$P(S_t = n|\boldsymbol{y}_t, \underline{x}_t, \boldsymbol{v}_t, A_t = 0; c) = \begin{cases} c/N_t & \text{if } 1 \leq n \leq N \\ 1 - c & \text{if } n = 0. \end{cases} \tag{3}$$

If a sound-source is active at time t, $(A_t = 1)$, we assign it to a visual identity n such that $Z_{tk} = n$ plays the role of an assignment variable in a mixture model. Its location y_{tk} is assumed to be drawn from the following Gaussian/uniform mixture:

$$P(\boldsymbol{y}_{tk}|\underline{x}_t, \boldsymbol{v}_t, A_t = 1; \theta_t) = \sum_{n=1}^{N} p_{tn} v_{tn} \mathcal{N}(\boldsymbol{y}_{tk}|\boldsymbol{x}_{tn}, \underline{\Sigma}_{tn}) + p_{t0}\mathcal{U}(\beta) \tag{4}$$

where $\theta_t = (\{p_{tn}\}_{n=0}^{N}, \{\underline{\Sigma}_{tn}\}_{n=0}^{N}, \beta)$ denotes the set of model parameters, namely the prior, $\sum_{n=1}^{N} v_{tn}p_{tn} + p_{t0} = 1$, the 2×2 covariance matrices $\underline{\Sigma}_{tn}$, and parameter β that characterizes the outlier component of the mixture, namely a uniform distribution. The parameter set θ_t can be estimated via the EM algorithm for Gaussian mixtures.

The algorithm begin with E-step that evaluates the posterior probabilities r_{tkn} using current parameters values θ_t, Z_t is our assignment variable, $Z_{tk} = n$ means $\boldsymbol{y}_{tk}$ is generated by component n. We first compute $r_{tkn} \forall n, 1 \leq n \leq N$ which correspond that a sound source is associated with a visible person:

$$r_{tkn} = \frac{p_{tn} v_{tn} \mathcal{N}(\boldsymbol{y}_{tk}|\boldsymbol{x}_{tn}, \underline{\Sigma}_{tn})}{\sum_{i=1}^{N} p_{ti} v_{ti} \mathcal{N}(\boldsymbol{y}_{tk}|\boldsymbol{x}_{ti}, \underline{\Sigma}_{ti}) + p_{t0}\mathcal{U}(\beta)} \quad \forall n, 1 \leq n \leq N \tag{5}$$

We can also write the probability that a sound source is not associated with a visible person $n = 0$, either because

it corresponds to a sound emitted by a non visible person or emitted by another type of source, i.e., the posterior of the uniform component of the mixture:

$$r_{tk0} = \frac{p_{t0}\mathcal{U}(\beta)}{\sum_{i=1}^{N} p_{ti}v_{ti}\mathcal{N}(\boldsymbol{y}_{tk}|\boldsymbol{x}_{ti}, \underline{\Sigma}_{ti}) + p_{t0}\mathcal{U}(\beta)} \quad (6)$$

M step re-estimates the parameters using the current responsibilities.

$$\underline{\Sigma}_{tn}^{new} = \frac{1}{R_{tn}} \sum_{k=1}^{K} r_{tkn}(\boldsymbol{y}_{tk} - \boldsymbol{x}_{tn})(\boldsymbol{y}_{tk} - \boldsymbol{x}_{tn})^T + \varepsilon I \quad (7)$$

$$p_{tn}^{new} = \frac{R_{tn}}{K} \quad (8)$$

with $\varepsilon > 0$ is a scalar acting as a parameter to prevent empty clusters, and I is the 2×2 identity matrix and where we have defined:

$$R_{tn} = \sum_{k=1}^{K} r_{tkn} \quad (9)$$

The algorithm can be easily initialized by setting all the priors equal to $1/N + 1$ and by setting all the variances equal to a positive scalar. Because the component means are fixed, the algorithm converges in only a few iterations.
We have N faces and K sources, which represent a combinatorial problem at each iteration. For each possible association at t, we have to consider all the possible cases for the next step. These computations between the faces and the sources will explode in the course of time. To address this issue and keep the audio-visual model tractable, instead of computing all combinations, we factorize all sources in one dominant source $y_{tn_t^*}$. It first requires to choose the person with highest speaking probability represented by the prior:

$$n_t^* = \max_n p_{tn} \quad (10)$$

Therefore, the mean source $\boldsymbol{y}_{tn_t^*}$ is the sound source location that is considered the most probable based on $\boldsymbol{x}_{tn_t^*}$:

$$\boldsymbol{y}_{tn_t^*} = \frac{\sum_{k=1}^{K} r_{tkn_t^*}\boldsymbol{y}_{tk}}{R_{tn_t^*}} \quad (11)$$

$P(S_t = n|\boldsymbol{y}_{tn_t^*}, \underline{x}_t, \boldsymbol{v}_t, a_t)$ can be calculated given $\boldsymbol{y}_{tn_t^*}$, $\forall n, 1 \leq n \leq N$:

$$P(S_t = n|u_{tn_t^*}) = \frac{\pi_{tn}v_{tn}\mathcal{N}(\boldsymbol{y}_{tn_t^*}|\boldsymbol{x}_{tn}, \underline{\Sigma}_{tn})}{\sum_{i=1}^{N} \pi_{ti}v_{ti}\mathcal{N}(\boldsymbol{y}_{tn_t^*}|\boldsymbol{x}_{ti}, \underline{\Sigma}_{ti}) + \pi_{t0}\mathcal{U}(\beta_t)} \quad (12)$$

and for $n = 0$

$$P(S_t = 0|u_{tn_t^*}) = \frac{\pi_{t0}\mathcal{U}(\beta_t)}{\sum_{i=1}^{N} \pi_{ti}v_{ti}\mathcal{N}(\boldsymbol{y}_{tn_t^*}|\boldsymbol{x}_{ti}, \underline{\Sigma}_{ti}) + \pi_{t0}\mathcal{U}(\beta_t)} \quad (13)$$

with $u_{tn^*} = (\boldsymbol{y}_{tn_t^*}, \underline{x}_t, \boldsymbol{v}_t, a_t)$.
Finally, by noting that the observed-data likelihood $P(u_{tn_t^*})$ does not depend on S_t and by assuming a uniform distribution over the priors of visible person n ($v_{tn} = 1$), i.e., $\pi_{t0} = \pi_{tn} = 1/(N_t + 1)$, we obtain the following observation model:

$$\begin{aligned} P(u_{tn_t^*}|S_t = n) &= P(S_t = n|u_{tn_t^*})P(u_{tn_t^*})/P(S_t = n) \\ &= P(S_t = n|u_{tn_t^*})P(u_{tn_t^*})/\pi_{tn} \\ &\propto P(S_t = n|u_{tn_t^*}). \end{aligned} \quad (14)$$

This enables to replace the observed likelihood (left hand side of (14)) with the posterior (right hand side of (14)) in (2).

2.4. State Transition Model Audio-Visual

The state transition probabilities, $p(S_t = j|S_{t-1} = i)$, provide the temporal modality for tracking speech turns along timestep. $p(S_t = j|S_{t-1} = i)$ is computed through several cases based on the presence/absence of persons and on their speaking status (for convenience and without loss of generality we set $v_{t0} = 1$):

$$P(S_t = j|S_{t-1} = i) = \begin{cases} p_s & \text{if } i = j \text{ and } v_{t-1,i} = v_{ti} = 1 \\ (1 - p_s)/N_t & \text{if } i \neq j \text{ and } v_{t-1,i} = v_{tj} = 1 \\ 0 & \text{if } v_{t-1,i} = v_{t-1,j} = 1 \text{ and } v_{tj} = 0 \\ 1/N_t & \text{if } v_{t-1,i} = 1, v_{ti} = 0 \text{ and } v_{tj} = 1 \\ 1/N & \text{if } v_{t-1,i} = 0 \text{ and } v_{ti} = 0 \end{cases} \quad (15)$$

The first case of (15) defines the self-transition probability, p_s, e.g., $p_s = 0.8$, of person i present at both $t - 1$ and t. The second case defines the transition probability from person i present at $t - 1$ to another person j present at t. The third case simply forbids transitions from person i present at $t - 1$ to person j present at $t - 1$ but not present at t. The fourth case represents the transition probability from person i present at $t - 1$ but not present at t, to a person j present at t. The fifth case defines the transition probability from person i not present at $t - 1$ to person j that is not present at t. One may easily verify that $\sum_{j=1}^{N} p(S_t = j|S_{t-1} = i) = 1$.

3. EXPERIMENTS

3.1. Data

The proposed method is evaluated on two corpora. The first considered corpus is CAV3D [15]. It contains 20 sequences with duration ranges from 15s to 80s. An evaluation is conducted on a subset SOT composed of 9 sequences with a single speaker and a subset SOT2 composed of 6 sequences with

a single active speaker and a second interfering person (not speaking).

The second corpus is recorded by ourselves on Pepper, a humanoid robot from Softbank Robotics. It contains dialogs between two or three persons. A total of 9 different subjects participated in this experience: 2 women, 7 men. Participants were asked to speak one at a time and try to avoid overlapping. We brought variations to dialogs by asking participants to randomly move in and out the scene, face the robot or look at each others. Different positions in the room were used to obtain different acoustic configurations. In a three-person dialogue, the participant positioned in the middle was requested to remain silent, to act as a distractor. The total duration is around eleven minutes and dominant speaker is carefully labeled in each frame. Windows with a bounding box that is either undetected or inaccurately computed are put aside. It was observed that the Pepper SSL occasionally experiences issues with activation, and in the absence of calibration of the Pepper SSL module, windows containing speech but no SSL detection are discarded too.

3.2. Technical Specifications

For this experiment, speaker diarization model is implemented for CAV3D and the Pepper corpus with some differences. SSLs are extracted with [16], and interpolation is performed from 3D coordinate, given by SSL, to speaker mouth localization using SOT subset. Pepper SSL module retrieves the direction of the emitting source (azimuth and elevation angles) from the TDOAs measured on the different microphone pairs. The angles provided by the sound source localization engine match the real position of the source with an average accuracy of 10 degrees. Transition from angle to image plane localization is made by interpolation. We record SSL from a loudspeaker at different positions in the image and perform a regression to get a mapping angle into image position. Sound sources located out of the image are filtered. We calibrate and fine-tune the parameters of the micro configuration using the first SOT sequence for CAV3D and a training sequence for the Pepper corpus in preparation for the testing phase. Thus we set $\underline{\Sigma} = Diag[300, 800]$, $\beta = 10^7$, $\varepsilon = 100$ for CAV3D and $\underline{\Sigma} = Diag[300, 500]$, $\beta = 300000$, $\varepsilon = 200$ for the Pepper corpus. The remaining parameters are shared between both experiences, $c = 0.2$, $p_s = 0.8$.

3.3. Results

The diarization performance is evaluated by Diarization Error Rate (DER), the lower the better. It contains three terms: Missing Detection (MS), False Alarm (FA), and Speaker Error (SPKE).

To evaluate the acoustic conditions we investigate CAV3D dataset with and without an oracle VAD. The use of an oracle significantly lowers the DER as it gives valuable information to reduce the number of FA. Results on CAV3D are promising, the model losses only 4.28% of DER between SOT and SOT2 without oracle VAD. With a DER of 19.27% on the Pepper corpus we can assume that our model fulfills its diarization goal in a standard robotic case. This method shows interesting results on SOT2 with a SPKE of 0%. The prediction only matches the right person when it detects a speaker. But it substantially decreases on the Pepper dataset. It comes from more complex scenarios and may also be related to Pepper micro quality.

Experiences	MS	FA	SPKE	DER
SOT (no sequence 6)	3.29	6.63	/	9.92
SOT (no sequence 6) oracle	1.24	0	/	1.24
SOT2	6.22	8.02	0	14.2
SOT2 oracle	1.14	0	0	1.14
Pepper corpus	5.35	0.1	13.82	19.27

Table 1. The performances (%) of our model for different experience set, A_t is set with oracle VAD derived from diarization labels, or with presence or absence of SSL

Those results can be compared to those of the audio-visual speaker diarization state of the art (SOTA): WST [14] and the audio only speaker diarization SOTA VBx [17] on the AMI corpus [18]. The AMI corpus is a collection of meetings which shows similarities with the Pepper corpus. WTS yields to 21.3% and 21.1% of DER on the two AMI subsets ES and IS and VBx to 38.65% for the whole AMI corpus. Both are computed without an oracle VAD. We denote similar results taking variation of DER between datasets into account. The theoretical complexity of the algorithm is $O(n^2)$. The audio-visual observation model represents 99% of the running time. For 10 sound sources being detected, with 5 considered persons, the running time is $0,0631$ seconds, out of a total running time of $0,0635$ seconds.

4. CONCLUSION

We proposed a model for temporal speaker-diarization based on principled mathematical and algorithmic concepts coupled with two types of perception, SSL and plane image. This model shows good results with a capacity of adaptation to different acoustic conditions without training phase. Thus this diarization method is not biased towards a particular training dataset, hence it is applicable to a large number of practical human-robot interaction scenarios. The VAD function and Robustness are carried by the uniform component of the mixture, which collects sound source locations that are far from the Gaussian components, which are centered around the faces. However to get the audio-image fusion, an interpolation needs to be made between SSL and image plan for every micro configuration, thus removing this micro configuration dependency is a challenge for future work.

5. REFERENCES

[1] S.E. Tranter and D.A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] John Hershey and Javier Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller, Eds. 1999, vol. 12, MIT Press.

[3] D.A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, 2005, vol. 5, pp. v/953–v/956 Vol. 5.

[4] Patrick Kenny, Douglas Reynolds, and Fabio Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–1070, 2010.

[5] Fabio Castaldo, Daniele Colibro, Emanuele Dalmasso, Pietro Laface, and Claudio Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4133–4136.

[6] Stephen H. Shum, Najim Dehak, Réda Dehak, and James R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[7] Tian Gao, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Densely connected progressive learning for lstm-based speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5054–5058.

[8] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopez Moreno, "Speaker diarization with lstm," *arXiv.org*, 2017.

[9] Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, B.J. Borgstrom, Fred Richardson, Suwon Shon, François Grondin, R. Dehak, Paola Garcia, Daniel Povey, Pedro Torres-Carrasquillo, Sanjeev Khudanpur, and Najim Dehak, "State-of-the-art speaker recognition for telephone and video speech: The jhu-mit submission for nist sre18," in *Interspeech*, 09 2019, pp. 1488–1492.

[10] Israel D. Gebru, Silèye Ba, Georgios Evangelidis, and Radu Horaud, "Audio-visual speech-turn detection and tracking," in *Proceedings of the 12th International Conference on Latent Variable Analysis and Signal Separation - Volume 9237*, Berlin, Heidelberg, 2015, LVA/ICA 2015, p. 143–151, Springer-Verlag.

[11] Israel D. Gebru, Sileye Ba, Xiaofei Li, and Radu 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, may 2018.

[12] Yiming Wang and Andrea Cavallaro, "Prioritized target tracking with active collaborative cameras," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2016, pp. 131–137.

[13] Zichao Xiong, Hongqing Liu, Yi Zhou, and Zhen Luo, "Multi-speaker tracking by fusing audio and video information," in *2021 IEEE Statistical Signal Processing Workshop (SSP)*, 2021, pp. 321–325.

[14] Joon Son Chung, Bong-Jin Lee, and Icksang Han, "Who said that?: Audio-visual speaker diarisation of real-world meetings," 2019.

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

[16] Francois Grondin and Francois Michaud, "Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations," 2018.

[17] Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget, "Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks," 2020.

[18] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Iain Mccowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner, "The ami meeting corpus: A pre-announcement," *Lecture Notes in Computer Science*, 07 2005.