ACCOUNTING FOR ROOM ACOUSTICS IN AUDIO-VISUAL MULTI-SPEAKER TRACKING

Yutong Ban[1,2], *Xiaofei Li*[1], *Xavier Alameda-Pineda*[1,2], *Laurent Girin*[1,2,3] *and Radu Horaud*[1,2]

[1] INRIA Grenoble Rhône-Alpes,[2] Univ. Grenoble Alpes, [3] CNRS, Grenoble-INP, GIPSA-Lab, France

ABSTRACT

Multiple-speaker tracking is a crucial task for many applications. In real-world scenarios, exploiting the complementarity between auditory and visual data enables to track people outside the visual field of view. However, practical methods must be robust to changes in acoustic conditions, e.g. reverberation. We investigate how to combine state-of-the-art audio-source localization techniques with Bayesian multi-person tracking. Our experiments demonstrate that the performance of the proposed system is not affected by changes in the acoustic environment.

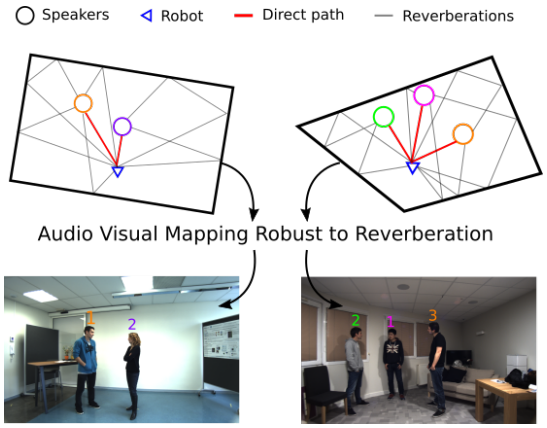
Index Terms— audio-visual fusion, multi-speaker tracking, sound-source localization, dereverberation.

1. INTRODUCTION

Multi-speaker tracking is a crucial task in many applicative scenarios such as human-computer/robot interaction, surveillance and monitoring systems, etc. The vast majority of methods use vision, and therefore they suffer from such limitations as visual occlusions, limited field of view, lighting conditions, etc. Audio processing can help overcoming these limitations due to the complementary nature of the information encoded in the acoustic signals. However, in order to exploit these signals jointly with images, fusion of multi-modal information is required in order to account for audio and visual data corruption, proper to multi-person indoor environments, e.g. a robot interacting with a group of persons holding a conversation.

In this article we address the challenging task of tracking multiple moving speakers with auditory and visual data, with special emphasis on accounting for the room acoustics, i.e. audio-visual tracking robust to reverberation. The use of two complementary modalities is beneficial when the information is correctly processed and fused, however the inference algorithms need to be robust to noise and outliers present in both modalities. Moreover, and in the particular case of audio and vision, the algorithms have to be carefully designed to exploit the inherent nature of the two modalities. Indeed, while the visual observations, e.g. face detection, is almost continuous for speakers looking towards the camera and within the field of view, natural speech often happens intermittently with occasional overlaps between several speech signals [1]. Importantly, including acoustic signals in the overall inference, opens the door to identify which source emitted which part of the speech signal: that is to separate and diarize the sources [2]. From the opposite point of view, it is quite clear that the knowledge of who is where and when in the scene could help separating the sound sources, for instance by using beamforming techniques [3], specially for moving sources [4].

This research has received funding from the ERC Advanced Grant VHIA (#340113).



(a) Meeting room (b) Living room

Fig. 1. Different rooms may have different sound reverberation characteristics. Extracting the direct path of a sound allows learning an audio-visual mapping that is robust to reverberation.

The literature on multi-person audio-visual tracking is sparse compared to the vast number of papers dealing with multi-person visual detection and tracking, e.g. [5, 6, 7, 8, 9]. Methods explicitly devoted to audio-visual tracking use particle filters or a probability hypothesis density (PHD) framework. Generally speaking, one must be careful when using these methods since the particle generation procedure may lead to a high computational cost. [10] and [11] proposed a method using the source direction of arrival (DOA) to determine the propagation of particles and combined it with a mean-shift algorithm to reduce the computational complexity. Similarly, [12] employed the DOA angles of the audio sources to reshape the typical Gaussian noise distribution for particle propagation and to weight the observation model afterwards. The methods presented above are based on audio-guided visual-particle generation, and the goal of audio-visual combination is mostly to increase the sampling efficiency, with the requirement that audio and visual data must be simultaneously available. Alternatively, [13] used a Markov chain Monte Carlo particle filter (MCMC-PF) to increase sampling efficiency. Still in a particle filter tracking framework, [14] proposed to use the maximum global coherence field of the audio signal and image color-histogram matching to adapt the reliability of audio and visual information. Finally, [15] used visual tracking information to assist source separation and beamforming.

All methods presented above are within a sampling framework, in which the trade-off between tracking quality and computational cost is usually one of the critical points. In addition, they were specifically designed to track a fixed number of people, and therefore the state space has fixed dimensionality. Moreover, these methods are evaluated in meeting-room like scenarios, meaning that

several visual and auditory sensors are used at different positions in the room. Little is known on the performance of these methods in ego-centric conditions, that is when all sensors are confined within a small volume (e.g. the head of a robot). The methods for audio-visual multi-speaker tracking designed for robot applications are rather rare. [16] exploit the framework of particle filtering for audio-visual localization of a single speaker with a robotic platform. Very recently, we have proposed a probabilistic framework for audio-visual multi-speaker tracking on a robotic platform [17]. For the sake of reducing the computational complexity the auditory features used in the previous work were straightforward binaural features, which are highly sensitive to the reverberation and other acoustic conditions, as we will demonstrate in the experimental section.

In this paper we propose to incorporate robustness to room acoustics into the audio-visual multi-speaker tracking method of [17]. More precisely, we propose to develop a tracking algorithm that is robust to speech reverberation in such a way that there is no need for retraining whenever the algorithm is tested in a room that is different than the room used for training. Indeed, the audio localization algorithm used in [17] is highly sensitive to changes in the acoustic conditions, typical of robotic scenarios in the wild. Therefore, we investigate how the recently proposed direct-path related transfer function (DP-RTF) [18, 19] features could be exploited within a Bayesian multi-speaker tracking method, as illustrated in Fig 1. Even if the general framework can process any visual localization features, we aim at evaluating the robustness of auditory features. For the sake of a fair comparison with previous works, we will only modify the auditory features and, consequently, the audio observation model, and leave the rest of the model (visual features and probabilistic dynamic model) intact.

The rest of the paper is organized as follows. The next section is devoted to describe the DP-RTF features. Section 3 depicts the audio-visual multi-speaker framework and the new auditory probabilistic observation model. Results on a publicly available dataset are discussed in Section 4.

2. ACOUSTIC FEATURES

We first describe the acoustic features in the case of one speaker for the sake of clarity. The multi-speaker case is discussed afterward.

Single speaker. Given the speech signal $s(l)$, the recorded signals at the microphone array are:

$$u(l) = s(l) \star a(l), \quad v(l) = s(l) \star b(l), \tag{1}$$

where $\star$ denotes convolution and the room impulse responses $a(l)$ and $b(l)$ encode the propagation path of the sound wave from the source point to the microphones, which is composed of the direct-path and the reflections. The direct-path propagation encodes the relative location of the source with respect to the microphone array. Our goal is to extract the direct-path from the microphone signals, which are distorted by the reflections and by ambient noise. In the short-time Fourier transform (STFT) domain, we have:

$$u_{p,k} = s_{p,k} \star a_{p,k}, \quad v_{p,k} = s_{p,k} \star b_{p,k}, \tag{2}$$

where p and k are the indices of temporal frames and frequency bins, $u_{p,k}$, $v_{p,k}$ and $s_{p,k}$ are the STFT of $u(l)$, $v(l)$ and $s(l)$ respectively, $a_{p,k}$ and $b_{p,k}$ with $p = 0, \cdots, Q-1$ represent the convolutive transfer function (CTF) corresponding to the room impulse responses [20, 21]. The first CTF coefficient $a_{0,k}$ can be interpreted as the k-th coefficient of the Fourier transform of the impulse response segment $a(l)|_{l=0}^{L-1}$, where L is the STFT frame length. If L is small, this segment includes only the direct-path impulse response, and thus the DP-RTF is defined as $\frac{b_{0,k}}{a_{0,k}}$.

We notice that: $u_{p,k} \star b_{p,k} = s_{p,k} \star a_{p,k} \star b_{p,k} = v_{p,k} \star a_{p,k}$. Dividing both sides by $a_{0,k}$ and reorganizing the terms in vector form, (2) rewrites:

$$v_{p,k} = \mathbf{x}_{p,k}^\top \, \mathbf{y}_k, \tag{3}$$

where $\mathbf{x}_{p,k} = [u_{p,k}, \ldots, u_{p-Q+1,k}, v_{p-1,k}, \ldots, v_{p-Q+1,k}]^\top$,

$$\mathbf{y}_k = \left[\frac{b_{0,k}}{a_{0,k}}, \ldots, \frac{b_{Q-1,k}}{a_{0,k}}, -\frac{a_{1,k}}{a_{0,k}}, \ldots, -\frac{a_{Q-1,k}}{a_{0,k}}\right]^\top.$$

We see that the DP-RTF appears as the first entry of the reverberation model $\mathbf{y}_k \in \mathbb{C}^{(2Q-1)\times 1}$. This equation is defined for one frame, i.e. the p-th frame. To estimate the vector $\mathbf{y}_k$, we collect $O > 2Q-1$ frames and solve a least square problem. In addition, an inter-frame spectral subtraction method is proposed in [18, 22] to remove the possible additive noise.

Multiple speakers. The single speaker case is relatively easy since all the frames are associated to a time invariant $\mathbf{y}_k$. In [19], neighbor frames are assumed to be associated to the same speaker, and therefore to the same $\mathbf{y}_k$. This assumption relies on the well-known sparsity of speech signals in the STFT domain. In practice, we propose to estimate the DP-RTF of the current frame by stacking only the previous $O \approx 3.5Q$ values. This is a good trade-off between a robust DP-RTF estimation and assuming that vector $\mathbf{y}_k$ is constant over time. At time t, the audio observations, denoted by $\boldsymbol{g}_{tk}$ correspond to the current estimate of the DP-RTF features, which is the first entry of $\mathbf{y}_k$: $\boldsymbol{g}_{tk} = \hat{\mathbf{y}}_k|_t^1$, where $\hat{\mathbf{y}}$ is an estimate of $\mathbf{y}$. We assume that for each time t at frequency k, the $\boldsymbol{g}_{tk}$ is associated to the position of only one speaker in the scene, which is well observed in practice.

When several microphone pairs are available, we can compute the DP-RTF for each microphone pair and then consider all these data as audio observations. Importantly, the relation between different frequency bins of different microphone pairs depends on the relative position of the speakers with respect to each other and to the microphones. Therefore we cannot guarantee that the audio feature at the k-th frequency bin corresponds to the same speaker location for all microphone pairs. Therefore, the features of different microphone pairs must be considered as independent observations. Virtually, if we exploit three microphone pairs, we will have $3K$ audio observations, where K denotes the number of considered frequencies. We are now left with the following challenging problems: (i) how to represent the audio and visual observations in a joint space, (ii) how to automatically assign the different observations to each of the speakers, and (iii) how to fuse them so as to infer the speakers' position.

3. MULTIPLE-SPEAKER AUDIO-VISUAL TRACKING

We are inspired from the multi-speaker tracking framework presented in [17], and in order to be able to evaluate the effect of the DP-RTF features for tracking purposes, we keep the same visual

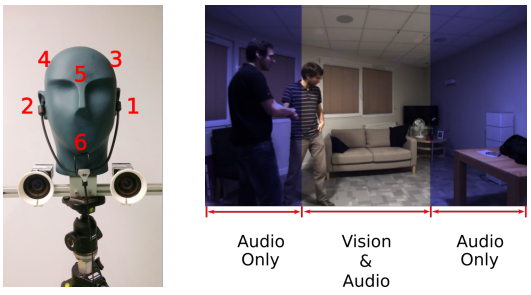


Fig. 2. (a) Robotic head with 6 microphones (red numbers); (b) Experimental simulation of limited field-of-view.

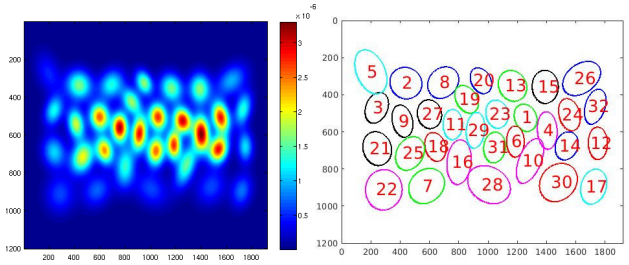


Fig. 3. Visualisation of the marginal GMM ($R = 32$) components in the image space after training. (a) Audio GMM distribution density. (b) Visualisation of location of different components.

pipeline. The main goal is to infer the positions of the speakers, that we encode in a hidden state $\boldsymbol{s}_{tn}$ for speaker n at time t, and $\mathbf{s}_t$ denotes the concatenation of these states. In addition to the position of speaker n, $\boldsymbol{s}_{tn}$ includes the speaker velocity as well as his/her bounding-box size. Each of the visual and audio features must be assigned to each of the speakers, we denote by $\boldsymbol{z}_t$ the set of hidden assignment variables. It consists of three assignment variables: $\boldsymbol{z}_t = (\boldsymbol{a}_t, \boldsymbol{b}_t, \boldsymbol{c}_t)$ (that we detail below). Formally, the tracking problem can be expressed as a MAP problem:

$$\max_{\mathbf{s}_t, \boldsymbol{z}_t} p(\mathbf{s}_t, \boldsymbol{z}_t | \mathbf{o}_{1:t}), \tag{4}$$

where $\mathbf{o}_{1:t} = (\mathbf{o}_1, \ldots, \mathbf{o}_t)$ represents all observations up to time t. $\mathbf{o}_t$ contains both audio observation $\boldsymbol{g}_t$ and visual observation $\boldsymbol{f}_t$. In order to define the probabilistic model, we need to specify the state dynamics, the visual observation model and the audio model.

State dynamics. We assume the speakers' dynamics are first order Markovian following a Gaussian distribution: $p(\boldsymbol{s}_{tn}|\boldsymbol{s}_{t-1,n}) = \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{s}_{t-1,n}, \boldsymbol{\Lambda}_n)$ with transition and uncertainty matrices $\mathbf{D}$ and $\boldsymbol{\Lambda}_n$ respectively.

Visual observation model. For the sake of comparison, we use the very same face detector as in [17]. Each visual observation at time t, $\boldsymbol{f}_{tm}$, consists of the geometric $\boldsymbol{v}_{tm}$ and appearance $\boldsymbol{h}_{tm}$ descriptors. An assignment variable A_{tm} is defined for each of these observations. If assigned to a speaker $1 \leq A_{tm} \leq N$, where N is the maximum number of people may appear in the scenario, these descriptors follow a Gaussian and a Bhattacharya distribution respectively. Otherwise, i.e. $A_{tm} = 0$, the descriptor belongs to a virtual speaker with uniform distribution $\mathcal{U}(\cdot)$. Formally we write:

$$p(\boldsymbol{f}_{tm}|\mathbf{s}_t, A_{tm} = n) = \tag{5}$$
$$\begin{cases} \mathcal{N}(\boldsymbol{v}_{tm}; \boldsymbol{s}_{tn}, \boldsymbol{\Phi})\mathcal{B}(\boldsymbol{h}_{tm}; \boldsymbol{h}_n) & \text{if } 1 \leq n \leq N \\ \mathcal{U}(\{\boldsymbol{v}, \boldsymbol{h}\}_{tm}; \text{vol}(\mathcal{V}, \mathcal{H})) & \text{if } n = 0, \end{cases}$$

where $\boldsymbol{\Phi}$ is the covariance matrix and $\boldsymbol{h}_n$ is the appearance model of the n-th speaker.

Audio observation model. In order to fuse the audio and visual information, we must operate in a common representation space. We opt to exploit a probabilistic generative model to project the audio features onto the visual image plane in a non-linear and principled manner. With the help of a dataset of pairs of image-positions and DP-RTF features, we split the image in R regions within which the visual-audio mapping is approximately linear, and learn these linear transformations. Thus, on top of the observation-to-speaker audio assignment variable, that we denote by B_{tk}, we also need an observation-to-region assignment variable C_{tk}. Similarly to the visual case, when assigned to a speaker (i.e. $1 \leq B_{tk} \leq N$) and to a region (i.e. $1 \leq C_{tk} \leq R$) the observations follow a Gaussian distribution. If assigned to the virtual speaker, they follow a uniform distribution. Synthetically:

$$p(\boldsymbol{g}_{tk}|\mathbf{s}_t, B_{tk} = n, C_{tk} = r) = \tag{6}$$
$$\begin{cases} \mathcal{N}(\boldsymbol{g}_{tk}; \mathbf{L}_{kr}\boldsymbol{s}_{tn} + \boldsymbol{l}_{kr}, \boldsymbol{\Sigma}_{kr}) & \text{if } 1 \leq n \leq N \\ \mathcal{U}(\boldsymbol{g}_{tk}; \text{vol}(\mathcal{G})) & \text{if } n = 0. \end{cases}$$

The parameters $\mathbf{L}_{kr}$, $\boldsymbol{l}_{kr}$ and $\boldsymbol{\Sigma}_{kr}$ are learned during training.

Variational inference. With the proposed probabilistic formulation, the inference problem in (4) cannot be exactly solved without exponentially complex algorithms. Therefore we propose a factorization of the filtering distribution:

$$p(\boldsymbol{z}_t, \mathbf{s}_t|\mathbf{o}_{1:t}) \approx q(\boldsymbol{z}_t) \prod_{n=0}^{N} q(\boldsymbol{s}_{tn}), \tag{7}$$

and solve for the associated variational EM algorithm, see [17]. All the steps of the EM algorithm are in closed-form.

4. EXPERIMENTS

Setup. We use the AVDIAR dataset [23] to evaluate the performance of the algorithm. The dataset consists of several recordings of indoor conversations, where people freely move and chat. We test on two representative sequences: a one-person (1P) sequence and a two-person (2P) sequence. Both sequences contain speakers that enter and leave the field of view. The dataset is recorded with an audio-visual sensor, consisting of a dummy acoustic head equipped with a stereo camera pair and six microphones, e.g. Fig 2-a. The cameras have a field of view of $97^\circ \times 80^\circ$ (horizontal $\times$ vertical), with an image resolution of 1920×1200 pixels and deliver 25 FPS videos. To simulate a natural scenario with people going out of the field of view, we define two "blind" areas, i.e. only audio information is available in these areas, e.g. Fig 2-b. This allows the audio-visual mapping to be learned over the entire field of view. In this way we can consider that the errors due to the audio-visual calibration are negligible with respect to the noise present in the raw data. Regarding the DP-RTF

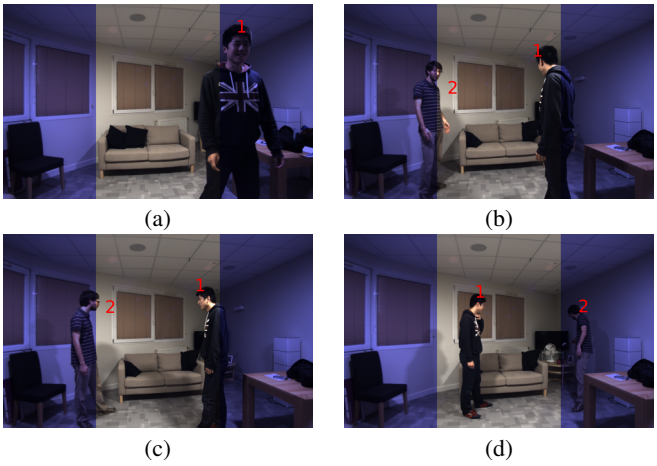


Fig. 4. Examples of results on the AVDIAR dataset: (a) 1P, frame #609. (b, c, d) 2P, frame #137, #507, #1113. Red numbers indicate estimated speakers' position and corresponding identity.

Table 1. Results on the AVDIAR dataset.

Seq.	Method	Rcll(↑)	Prcn(↑)	FAF(↓)	FP(↓)	FN(↓)	IDs(↓)	MOTA(↑)
1P	[12]	68.6	70.1	0.27	191	205	0	39.3
	[17]	76.4	86.9	0.11	75	154	8	63.7
	Prop.	75.5	89.6	0.08	57	160	0	66.7
2P	[12]	56.7	57.1	0.85	1451	1471	92	11.4
	[17]	51.3	63.7	0.59	995	1655	1	22.0
	Prop.	70.5	85.1	0.25	420	1002	2	58.1
Total	[12]	58.6	59.1	0.68	1642	1676	92	15.8
	[17]	55.4	67.7	0.45	1070	1809	9	28.7
	Prop.	71.3	85.8	0.20	477	1162	2	59.5

Table 2. Results with training in a different room.

Seq.	Method	Rcll(↑)	Prcn(↑)	FAR(↓)	FP(↓)	FN(↓)	IDs(↓)	MOTA(↑)
1P	[17]	61.3	69.8	0.25	173	252	2	34.5
	Prop.	74.2	88.2	0.09	65	168	0	64.3
2P	[17]	44.9	56.5	0.69	1173	1874	1	10.4
	Prop.	70.2	85.2	0.24	413	1014	2	58.0
Total	[17]	48.0	59.2	0.56	1341	2108	5	14.8
	Prop.	70.8	85.7	0.20	478	1182	2	59.0

features, they are extracted from three microphone pairs (1-2, 3-4, 5-6), using $Q = 10$ and a STFT window of 16 ms with 50% overlap (16-kHz audio sampling). K is set to 64 which represents 4 kHz. The number of regions R in audio observation model is set to 32. We compared the result with particle filtering based method in [12] and with our previous work in [17].

Training. As described in Section 3, we use a Gaussian mixture regression model to generate audio observations. The model's parameters, $\{\mathbf{L}_{kr}, \boldsymbol{l}_{kr}, \boldsymbol{\Sigma}_{kr}\}$ need to be learned for all $k \in \{1, \ldots, K\}$, $r \in \{1, \ldots, R\}$. They are estimated via an EM procedure using a training dataset $\{\mathbf{g}, \mathbf{s}\}$ [17] on the following grounds. 1 s-long white noise signals (to ensure energy in all frequency bins) are emitted by a loudspeaker from 800 different known positions covering the camera field of view. The number of Gaussian components is set to $R = 32$. The distribution of the marginal Gaussian mixture model in source location $\boldsymbol{s}_t$ obtained via training is illustrated in Figure 3. This figure shows that the field of view is well covered by the Gaussian components of our model. We indistinctively refer to this procedure as either training or audio-visual calibration.

Evaluation Metrics. We evaluate the performance of the proposed method using standard MOT (multi-object tracking) metrics [24]: MOT accuracy (MOTA), which combines false positives (FP), missed targets = false negative (FN), and identity switches (ID); the false alarm per frame (FAF); the tracking recall (Rcll) and tracking precision (Prcn). Since audio localisation has unlimited field of view, but is less accurate than visual tracking, we evaluate only the azimuth in the "blind area". MOTA is calculated with an overlap threshold of 0.9.

Results. Table 1 reports the results on the two sequences for the three methods. Firstly, we observe that the results for 1P are better than for 2P, which is expected since 2P is more complex. Also, we notice that the proposed method outperforms the two baseline methods in both sequences. The difference is obvious in the more complex 2P sequence.

Qualitative results are illustrated in Fig 4. We can observe that the tracking results are always around the speakers, even if Fig 4 (b) and (c) are considered as failure cases for quantitative evaluation, since the overlap between the track and the face ground truth is zero.[1]

To further evaluate the robustness of the proposed approach, we trained the audio observation model in a different room than the test room. Results are reported in Table 2. We can clearly see that training in a different room has a very negative effect when the audio features are not robust to the acoustic conditions, i.e. [17]. In contrast, by properly exploiting the DP-RTF features and including them in the tracking framework in a principled manner with a probabilistic observation model, the proposed system is almost unaffected by the changes in the acoustic environment. In short, we propose an audio-visual multi-speaker tracking system that does not require room-specific data to provide highly accurate results.

5. CONCLUSION

We proposed an AV multi-speaker tracking algorithm based on a variational EM algorithm with very reasonable computation cost. The use of DP-RTF audio features makes this system robust to changes in the acoustic conditions, as illustrated by comparison with the use of conventional audio features and another state-of-the-art AV multi-person tracking algorithm.

6. REFERENCES

[1] Dionyssos Kounades-Bastian, Laurent Girin, Xavier Alameda-Pineda, Sharon Gannot, and Radu Horaud, "An EM algorithm for joint source separation and diarization of multichannel convolutive speech mixtures," in *IEEE ICASSP*, New Orleans, Louisiana, USA, 2017.

[2] Dionyssos Kounades-Bastian, Laurent Girin, Xavier Alameda-Pineda, Radu Horaud, and Sharon Gannot, "Exploting the intermittency of speech for joint separation and diarization," in *IEEE WASPAA*, New Paltz, USA, 2017.

[1] Result videos are available at `https://team.inria.fr/perception/research/tracking_rtf/`.

[3] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[4] Dionyssos Kounades-Bastian, Laurent Girin, Xavier Alameda-Pineda, Sharon Gannot, and Radu Horaud, "A variational EM algorithm for the separation of time-varying convolutive audio mixtures," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1408–1423, 2016.

[5] Sileye Ba, Xavier Alameda-Pineda, Alessio Xompero, and Radu Horaud, "An on-line variational bayesian model for multi-person tracking from cluttered scenes," *Computer Vision and Image Understanding*, vol. 153, pp. 64–76, 2016.

[6] Yutong Ban, Sileye Ba, Xavier Alameda-Pineda, and Radu Horaud, "Tracking multiple persons based on a variational bayesian model," in *ECCV Workshops*, 2016, pp. 52–67.

[7] Yutong Ban, Xavier Alameda-Pineda, Fabien Badeig, Sileye Ba, , and Radu Horaud, "Tracking a varying number of people with a visually-controlled robotic head," in *Intelligent Robots and Systems*, Vancouver,Canada, 2017.

[8] Xavier Alameda-Pineda and Radu Horaud, "Vision-guided robot hearing," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 437–456, 2015.

[9] Israel D. Gebru, Xavier Alameda-Pineda, Florence Forbes, and Radu Horaud, "EM algorithms for weighted-data clustering with application to audio-visual scene analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 12, pp. 2402–2415, 2016.

[10] Mark Barnard, Wenwu Wang, Adrian Hilton, Josef Kittler, et al., "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.

[11] Yang Liu, Wenwu Wang, Jonathon Chambers, Volkan Kilic, and Adrian Hilton, "Particle flow SMC-PHD filter for audio-visual multi-speaker tracking," in *International Conference on Latent Variable Analysis and Signal Separation*, 2017, pp. 344–353.

[12] Volkan Kılıç, Mark Barnard, Wenwu Wang, and Josef Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Transactions on Multimedia*, vol. 17, no. 2, pp. 186–200, 2015.

[13] D. Gatica-Perez, G. Lathoud, J-M. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 601–616, 2007.

[14] Xinyuan Qian, Alessio Brutti, Maurizio Omologo, and Andrea Cavallaro, "3d audio-visual speaker tracking with an adaptive particle filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, New-Orleans, Louisiana, 2017, pp. 2896–2900.

[15] S Mohsen Naqvi, W Wang, M Salman Khan, M Barnard, and JA Chambers, "Multimodal (audio–visual) source separation exploiting multi-speaker tracking, robust beamforming and time-frequency masking," *IET Signal Processing*, vol. 6, no. 5, pp. 466–477, 2012.

[16] Niclas Schult, Thomas Reineking, Thorsten Kluss, and Christoph Zetzsche, "Information-driven active audio-visual source localization," *PloS one*, vol. 10, no. 9, 2015.

[17] Yutong Ban, Laurent Girin, Xavier Alameda-Pineda, and Radu Horaud, "Exploiting the complementarity of audio and visual data in multi-speaker tracking," in *ICCV Workshop on Computer Vision for Audio-Visual Media*, 2017.

[18] Xiaofei Li, Laurent Girin, Radu Horaud, and Sharon Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 2171–2186, 2016.

[19] Xiaofei Li, Laurent Girin, Radu Horaud, and Sharon Gannot, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1997–2012, 2017.

[20] Yekutiel Avargel and Israel Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[21] Ronen Talmon, Israel Cohen, and Sharon Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[22] Xiaofei Li, Laurent Girin, Radu Horaud, and Sharon Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *IEEE ICASSP*, 2015, pp. 320–324.

[23] Israel D. Gebru, Silèye Ba, Xiaofei Li, and Radu Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, 2017.

[24] Anton Milan, Laura Leal-Taixe, Ian Reid, Stefan Roth, and Konrad Schindler, "Mot16: A benchmark for multi-object tracking," 2016.