3D Audiovisual Speaker Tracking with Distributed Sensors Configuration

Frank Sanabria-Macias
Department of Electronics
University of Alcalá.
Alcalá de Henares, Madrid, Spain
ORCID: 000-0002-6946-0326

Marta Marron-Romera
Department of Electronics
University of Alcalá.
Alcalá de Henares, Madrid, Spain
ORCID: 0000-0001-7723-2262

Javier Macias-Guarasa
Department of Electronics
University of Alcalá.
Alcalá de Henares, Madrid, Spain
ORCID: 0000-0002-3303-3963

***Abstract*—Smart spaces are environments equipped with a set of sensors with the main objective of understanding humans' behavior within them, their interactions and to improve human-machine interfaces. Audiovisual tracking is used to know people's position in the environment and if they are talking, through the use of cameras and microphones. In this work we present an audiovisual tracking solution with a single camera and microphone array in a distributed configuration. Our idea is to exploit the estimation of azimuth and elevation from audio information to be fused with the position estimation obtained from a Viola and Jones based observation model. The fact that the microphone array is not co-located with the camera will allow to reduce the distance estimation uncertainty from the video model and improve tracking accuracy. The system was evaluated on the AV16.3 database on single speaker sequences, outperforming results of state-of-the-art, under these conditions.**

***Index Terms*—Smart Space, Audiovisual Tracking, Particle Filter**

I. Introduction and Previous Works

Smart spaces are environments equipped with a set of sensors, communication and computing systems, with the main objective of understanding human behaviors within them, their interactions and to improve human-machine interfaces. In this context, one of the core low level information refers to the presence, position, orientation (pose), and voice activity status of users within the space, as these features play a major role in high level interactions between the users and the environment. Audiovisual speaker tracking is a technique that can be applied in this context to locate people in the environment using audio and video information. The main idea is to combine the best features of both sensors to improve the tracking accuracy and robustness.

In the last few years there has been an increase in the number of proposals for multiple speaker tracking in smart spaces using audio and video information [1]–[10]. One relevant research line is focused in a challenging speaker tracking scenario, where there are a variable and unknown number of speakers [1]–[3]. Another line focuses on finding a better visual appearance model to track multiple speakers in indoor environments [5], while others proposals are centered on audiovisual tracking in compact configurations (co-located camera and microphone array) for applications like human-robot interfaces [6]–[9].

In [9] an extensive review to the state of the art in audiovisual speaker tracking is presented. In this review, the different literature proposals were classified according to different aspects like in which space the tracking is done (image plane, ground or three-dimensional (3D)), if the configuration of the sensors is co-located or distributed, the number of sensors used, etc. In recent proposals there appears to be a tendency to reduce the number of sensors, being one camera and one microphone array in [2], [3], [6], [9], [11]. Many of these systems track the speakers in the image plane (2D) [2], [3], [8], [11], [12]. In [6] and [9] the idea of tracking in the 3D space with one camera and one microphone array is presented. One of the problems reported for this task is the low accuracy in the user distance estimation, as it is based on the size of the detected face or person's body. In the audio modality, the problem is even worse, since a small microphone array has larger distance errors than video.

Another characteristic in the audiovisual speaker tracking literature is the common use of color histogram-based likelihood approaches to track faces in the video modality [1]–[3], [11], [13]. However, these likelihoods are not very accurate in determining the object location. These models also suffer from the problem of not being able to correctly locate the mouth for different face poses. Another typical approach is to apply face or person detectors [8], [9], but when the detection fails, the tracker must use color information. A 3D visual only tracking system was recently presented in [14] for face/mouth location with particle filters using a probabilistic Viola and Jones based observation model [15]. This model uses the knowledge learned by the Viola and Jones classifier, that is modified to be used as a likelihood estimator in a particle filter. The previous results in [14], show that this observation model is accurate when tracking faces in the image plane domain.

In this paper we will combine this visual observation model with audio information, to carry out audiovisual face/mouth tracking in a distributed configuration, using a single camera and a single microphone array which are not co-located. Our strategy is to exploit the estimation of azimuth and elevation only using audio information, and not the distance estimation,

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness under projects HEIMDAL-UAH (TIN2016-75982-C2-1-R) and ARTEMISA (TIN2016-80939-R); and by the University of Alcalá under project ACÚFANO (CCG19/IA-024).

which is highly error-prone. The fact that the microphone array is not co-located with the camera will allow to reduce the distance estimation uncertainty from the video model and improve the tracking accuracy.

In [14], the face likelihood model was used in two tracking tasks. In the first one, using a single video camera, tracking is done in the image plane (2D) without estimating depth. In the second task, the tracking is done in 3D, taking advantage of the intersection of the location estimation from the 3 video cameras. So, the main differences of this work as compared with that in [14], are the combined use of audio and video information, and the 3D localization carried out using a single camera, with the aid of the acoustic information. The distance between the mouth and the camera is estimated considering the face size in the image plane (video only modality). Then, the 3D mouth position estimation is refined by intersecting the video estimation with the estimated speaker direction from the microphone array, thus combining both information sources in the audiovisual modality.

II. Tracking with Particle Filters

Bayesian Filters (BFs) are techniques that estimate the posterior probability density function (PDF) of a system along time, given a set of measurements [16]. The estimation uses a mathematical model called the state-space model. In this model, the state ($\mathbf{x}_k$) and observation ($\mathbf{z}_k$) vectors are defined, being k the corresponding time instant. State vectors characterize the properties of the system to be estimated (position, velocity, physical dimensions, etc.), and observation vectors consider the measurements taken from the sensors.

The estimation is done in two steps: prediction and update. In the prediction stage, a prior PDF calculation of the state vector given the previous states ($p(\mathbf{x}_{k-1}|\mathbf{Z}_{k-1})$) is done using a motion model ($p(\mathbf{x}_k|\mathbf{x}_{k-1})$, see equation (1)). In the update stage, a posterior PDF is computed (see equation (2)) by including the observation vector information:

$$p(\mathbf{x}_k|\mathbf{Z}_{k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{Z}_{k-1})d\mathbf{x}_{k-1} \quad (1)$$

$$p(\mathbf{x}_k|\mathbf{Z}_k) = \frac{p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{Z}_{k-1})}{p(\mathbf{z}_k|\mathbf{Z}_{k-1})}, \quad (2)$$

where $\mathbf{Z}_k = \{\mathbf{z}_{k'}, k' = 1, ..., k\}$ and $p(\mathbf{z}_k|\mathbf{x}_k)$ is the likelihood defined in the observation model [16].

Particle filters (PFs) are a particular class of BFs that approximate the states distribution with a set of weighted samples $\{\mathbf{x}_k^i, w_k^i\}$, called particles, as shown in equation (3):

$$p(\mathbf{x}_k|\mathbf{Z}_k) \approx \sum_{i=1}^{N} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i), \quad (3)$$

where N is the number of particles being used.

The estimation stage is then carried out by applying *importance sampling* [16] (IS), which is a statistical technique to estimate the properties of a distribution when its samples are generated by another one.

III. Proposed Method

Our proposal is a PF to track the speaker's mouth position with a microphone array and a video camera (Figure 1 shows the general scheme of the proposed algorithm). The state vector is defined from the 3D position of the particles, and their velocity $\mathbf{x}_k^i = [x^i,\ y^i,\ z^i,\ v_x^i,\ v_y^i,\ v_z^i]$. The particles $\mathbf{x}_k^i$ are initialized in a determined position, being the initial weight $w_{k=0}^i = 1/N$ for each one, with $i = 1, ..., N$. These particles are propagated according to the motion model, in a frame by frame basis, given a predicted particle set $\{\mathbf{x}^i, w^i\}_{k|k-1}$. In the update stage, first, the audio and video data $\mathbf{z}_k = \{\mathbf{z}_k^a, \mathbf{z}_k^v\}$ are used to evaluate the likelihood of each predicted particle $l_i^{av} = f(l_i^a, l_i^v)$ with the observation model, (see section III-D). Where the ^{a},^{v} and ^{av} superscripts refer to the audio, video, and audiovisual modalities respectively. Then, the posterior weights w_k^i are computed by updating the particles' previous weights $w_{k|k-1}^i$ with the likelihood values. The speaker position estimation is obtained as a weighted average of the updated particle set. A resample step eliminates low weight particles and multiply particles with a high weight obtaining a new set of particles $\{x^{i\star}, w^{i\star}\}_k$. This new set is used as the prior distribution in the next time step $k+1$.

A. Motion model

In each iteration, the particles are propagated using the Langevin motion model [17] (commonly used in the acoustic speaker tracking literature [18]), and it has two parameters: a constant speed v and the rate of change β.

B. Video observation model

The predicted particles in 3D $(x, y, z)^i$ (representing speaker's mouth positions hypothesis) are projected on the image plane, obtaining $(u, v)^i$. Squares bounding boxes around the mouth points are needed to compute face likelihood. First, bounding box sizes s^i are estimated according to the pin-hole camera model:

$$s^i = \frac{h_r^{3D} \cdot f}{d^i}, \quad (4)$$

where f is the camera focal length, d^i is the distance from the 3D position of the particle i to the camera, and h_r^{3D} is a constant parameter representing the distance (in meters) between the chin and the upper end of the forehead. At this point, 3D particles have been converted to position and size space $(u, v, s)^i$.

The face model used is a modification of the Viola and Jones face detection algorithm [15], which returns a face likelihood value Ω_{exp4} (not binary) using a trained template. As in [14], this model is applied for three different templates, trained for frontal, left, and right poses respectively, given three likelihood values $\Omega_{exp4F}, \Omega_{exp4L}$ and Ω_{exp4R}. The relative position of the bounding box related to the mouth is adapted depending on the face pose [14]. Finally, a global likelihood, l_i^v, is estimated thru the combination of poses likelihoods (5):

$$p(\mathbf{z}_k^v|\mathbf{x}_k^i) \sim l_i^v = \sqrt{\Omega_{i,exp4F}^2 + \Omega_{i,exp4L}^2 + \Omega_{i,exp4R}^2}, \quad (5)$$

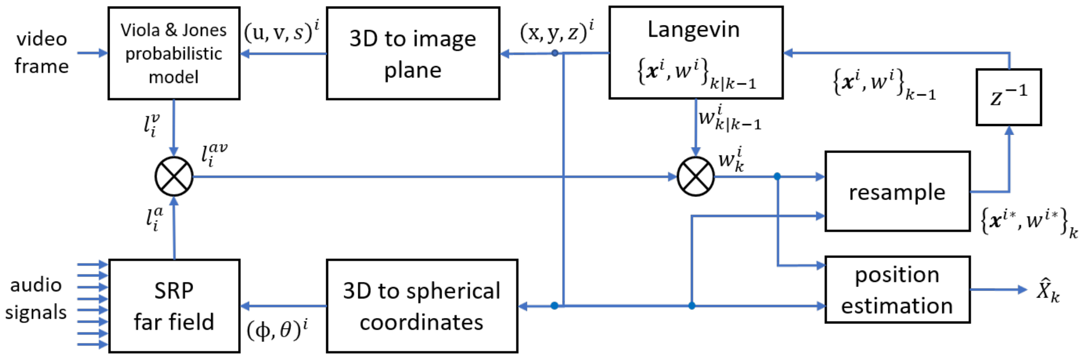

Fig. 1. General scheme of the proposed algorithm.

C. Audio Observation Model

The audio model processing starts by computing the Generalized Cross Correlation with Phase Transform (GCC-PHAT) between each pair (p, q) of microphone signals segments,

$$R^{PHAT}_{s_p,s_q}(\tau) = \sum_{f=0}^{N_f-1} \Psi_{ij}(f) S_p(f) S^*_q(f) e^{j\frac{2\pi f}{N_f}\tau}, \quad (6)$$

where N_f is the number of discrete frequencies, $S_p(f)$ and $S_q(f)$ are the frequency spectra of the signals in the p and q microphones respectively, $\Psi_{pq}(f) = \frac{1}{|S_p(f)S'_q(f)|}$ is the PHAT filter, and τ is the delay.

Then, the SRP-PHAT power of each of the particles is computed as a likelihood value and expressed using the GCC-PHAT terms. In the far-field condition case, the wavefront can be considered flat, so that it is not possible to know the exact distance from the emitting source to the transducers, but only their direction of arrival. In this case, a transformation to particle positions from Cartesian $(x, y, z)^i$ to spherical coordinates is carried out, obtaining the azimuth and elevation $(\phi, \theta)^i$ of each particle, referenced to the center of the microphone array. The SRP-PHAT in the far-field condition can be expressed as:

$$p(\mathbf{z}^a_k|\mathbf{x}^i_k) \sim l^a_i = P(\phi^i, \theta^i) = \sum_{p=1}^{M-1} \sum_{q=p+1}^{M} R^{PHAT}_{s_p,s_q}(\tau^{\phi^i,\theta^i}_{p,q}), \quad (7)$$

where M is the number of microphones in the array, and $\tau^{\phi^i,\theta^i}_{p,q}$ is the theoretical time delay of arrival (TDOA) to the p and q microphones of the sound wave coming from the i^{th} particle direction.

D. Audio and Video Information Fusion

To fuse the information generated from the audio and video source for each particle, we assume independence between modalities, so that:

$$p(\mathbf{z}^{av}_k|\mathbf{x}^i_k) = p(\mathbf{z}^a_k|\mathbf{x}^i_k) p(\mathbf{z}^v_k|\mathbf{x}^i_k) \quad (8)$$

In practice, audiovisual likelihoods are obtained by computing the product of the likelihoods from both modalities.

$$p(\mathbf{z}^{av}_k|\mathbf{x}^i_k) \sim l^{av}_i = f(l^a_i, l^v_i) = l^v_i \cdot l^a_i \quad (9)$$

E. Weight updates and Position Estimation

The weights for previous step particles are multiplied (updated) by their likelihood values to obtain the new weights.

$$w^i_k = w^i_{k|k-1} \cdot l^{av}_i \quad (10)$$

The position of the speaker $(\hat{\mathbf{X}}_k)$ is then estimated from equation (11):

$$\hat{\mathbf{X}}_k = \sum_{i=1}^{N} w^i_k \mathbf{x}^i_k \quad (11)$$

Finally, the particle set is resampled with the multinomial resampling [19].

IV. Experimental Setup

The proposed tracking system has been tested in two modalities: First only using video information, and second, combining the two information sources in the audiovisual modality.

A. Dataset

The database used for system evaluation is AV16.3 [20], recorded in the *Smart Meeting Room* of the IDIAP research institute, which consists of a $8.2m \times 3.6m \times 2.4m$ rectangular room containing a centrally located $4.8m \times 1.2m$ rectangular table. Two circular microphone arrays of radius $10cm$, each of them composed by 8 microphones, are placed on top of the table. The centers of the two arrays are separated by $80cm$ and the origin of coordinates is located in the middle point between the two arrays. The room is also equipped with 3 video cameras providing different angle views.

The database contains audio and video data taken by the 3 video cameras, and the two circular microphone arrays. The cameras have a frame rate of 25 f.p.s ($40ms$ period) while the audio has been recorded at 16 kHz. The dataset is fully labeled, providing the mouth ground truth location. Synchronization information between the audio and video streams is also available. The experiments were carried out

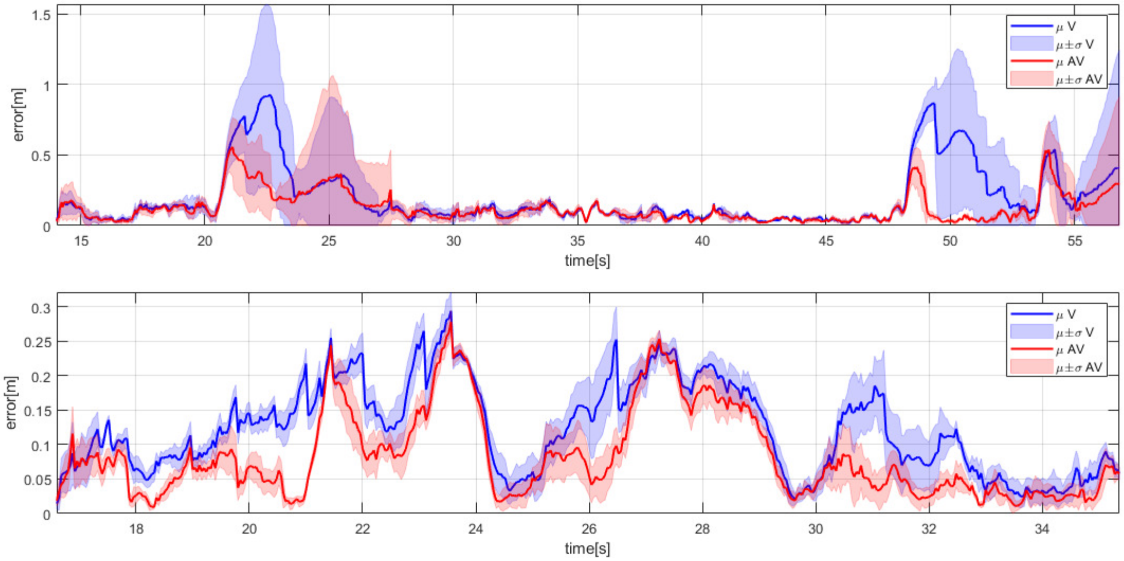

Fig. 2. Mean and standard deviation of error over time in `seq12` and `seq08` (blue: video, red: audiovisual)

using sequences `seq08`, `seq11`, and `seq12`, with single moving speakers, which are the same ones evaluated in [9] and [10], two state-of-the-art proposals that will be used in our performance comparison.

B. System configuration

Audio signals are segmented in windows with 4096 samples ($256ms$), using a window shift of 640 samples ($40ms$), in such a way that each audio segment is associated with one video frame. The number of FFT points is equal to the window size. Each image frame was scaled by a factor of 2 to allow the Viola and Jones templates of $20x20$ pixels to able to ”detect” faces when they are far away from the cameras, although there are frames in which the speaker face size is below $20x20$ pixels (around $15x15$). Next, lens distortion is corrected. The face bounding box size in 3D, h_r^{3D}, is fixed to $0.17m$.

The model parameters ($v = 1ms^{-1}$ and $\beta = 10s^{-1}$) have been selected following [18], in which they are reported to achieve good results. The PF algorithm used was the sequential importance resampling, with 1000 particles. The algorithm was tested with each of the three cameras independently and one of the available microphone arrays (MA1 in AV16.3 [20]). The evaluation metric we used is the Mean Absolute Error (MAE) in 3D, as defined in [9], measured in mm.

V. Results and Discussion

Table I shows the MAE performance metrics (in mm) for each sequence and cameras `C1`, `C2` and `C3`, and for the two modalities (video and audio+video), also including the relative improvement achieved by the audiovisual fusion as compared with the use of video only information ($\Delta_r = \frac{MAE_V - MAE_{A+V}}{MAE_V}$).

From the results, it can be observed that for every combination of sequence and camera, the audiovisual tracking was capable of reducing the errors as compared with a visual only solution. The largest error reduction happens in sequence `seq12` camera 1 (36.1%). The second larger error reduction is show in sequence `08` camera 2 (32%). These two cases with larger reductions are the fist and the third with larger errors in the video modality. In the rest of the cases, the reduction was between 6.8% and 27%, with average improvements around 21%.

TABLE I
Average MAE errors (mm) and relative error reduction by sequence and camera in two modalities, video and audiovisual.

	Video			Audio+Video		
	C1	C2	C3	C1	C2	C3
seq08 Δ_r	70	122	120	66 6.8%	83 32.0%	86 27.9%
seq11 Δ_r	119	98	103	110 7.6%	78 20.6%	89 14.3%
seq12 Δ_r	195	144	142	125 36.1%	125 12.7%	116 18.2%
average Δ_r	128	122	122	100 21.9%	96 21.3%	97 20.3%

A. Analysis of results

Sequence `seq12` presents a speaker with rapid movements and sudden changes of direction, causing the visual tracker estimation not to be able to accurately follow the speaker, thus increasing the tracking error. Figure 2 (top) shows this situation between seconds 21 and 27 and after second 48. In both cases, the audio information allows the tracker to faster recover from the target loss. A similar situation occurs at the end of sequence 11 in a series of loops that the speaker does.

For sequence `seq08`, and using cam 3, there is no target loss, but the visual tracker presents errors above $10cm$ (see bottom graphic in Figure 2) during some time intervals. This error is mainly caused by a wrong speaker distance estimation.

In this case, we can see that the audiovisual solution is capable of keeping the error below that achieved by the only video solution in most of the sequence.

B. Comparison with the state-of-the-art

In Table II we compare our overall average results with those of the AV3T proposal described in [9], and with that of [10], two of the most recent proposals in the literature for audiovisual tracking. In Table II, we also provide the relative error reduction of our proposal and that of [10], as compared with that achieved by AV3T.

From the results shown, it's clear that our visual tracker reduces the error of the AV3T video only counterpart around 70%, and around 39% in the audiovisual modality, being also better than the audiovisual proposal of [10].

TABLE II
AVERAGE MAE ERRORS (MM) COMPARISON WITH [9], [10]

	AV3T [9]	[10]	Our proposal
`Video`	41	–	12.4
Δ_r		–	69.8%
`Audio+Video`	16	12	9.7
Δ_r		25.0%	39.4%

VI. CONCLUSIONS AND FUTURE WORK

In this work, a new audiovisual tracking proposal for distributed sensor configuration is presented. Evaluations were done with a single camera and one microphone array conditions, for a single speaker tracking task in distributed non-co-located configurations.

Video tracking using face appearance-based observation model is shown to be better than color models and face detection solutions from the state of the art. The audiovisual combination is capable of improving 3D visual only speaker tracking in distributed configurations using a single camera and one microphone array, reducing the depth error that may be significantly high in such configurations. This improvement is especially remarkable in cases where the visual tracking loses the target, where the proposed combination makes tracking much more robust.

As future work, we plan to evaluate this model in multiple speaker scenarios. Also, further investigations will be done in order to improve the dynamic model and to optimize the particle filter parameters to reduce situations where the tracker still loses the target in complex speaker pose configurations.

REFERENCES

[1] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Audio constrained particle filter based visual tracking," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 3627–3631.

[2] ——, "Audio assisted robust visual tracking with adaptive particle filtering," *Multimedia, IEEE Transactions on*, vol. 17, no. 2, pp. 186–200, Feb 2015.

[3] V. Kilic, 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, Dec 2016.

[4] M. Barnard, W. Wang, J. Kittler, S. M. Naqvi, and J. Chambers, "Audio-visual face detection for tracking in a meeting room environment," in *Information Fusion (FUSION), 2013 16th International Conference on*. IEEE, 2013, pp. 1222–1227.

[5] M. Barnard, P. Koniusz, W. Wang, J. Kittler, S. M. Naqvi, and J. Chambers, "Robust multi-speaker tracking via dictionary learning and identity modeling," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 864–880, April 2014.

[6] X. Qian, A. Brutti, M. Omologo, and A. Cavallaro, "3d audio-visual speaker tracking with an adaptive particle filter," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2896–2900.

[7] I. D. Gebru, C. Evers, P. A. Naylor, and R. Horaud, "Audio-visual tracking by density approximation in a sequential bayesian filtering framework," in *Hands-free Speech Communications and Microphone Arrays (HSCMA), 2017*. IEEE, 2017, pp. 71–75.

[8] Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Exploiting the complementarity of audio and visual data in multi-speaker tracking," in *ICCV Workshop on Computer Vision for Audio-Visual Media*, 2017.

[9] 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, pp. 2576–2588, 2019.

[10] H. Liu, Y. Li, and B. Yang, "3d audio-visual speaker tracking with a two-layer particle filter," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1955–1959.

[11] I. D. Gebru, S. Ba, G. Evangelidis, and R. Horaud, "Audio-visual speech-turn detection and tracking," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 143–151.

[12] I. 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*, 2017.

[13] Y. Ban, X. Alameda-Pineda, L. Girin, and R. Horaud, "Variational bayesian inference for audio-visual tracking of multiple speakers," *CoRR*, vol. abs/1809.10961, 2018. [Online]. Available: http://arxiv.org/abs/1809.10961

[14] F. Sanabria-Macias, M. M. Romera, J. Macias-Guarasa, D. Pizarro, J. N. Turnes, and E. J. Marañón-Reyes, "Face tracking with a probabilistic viola and jones face detector," in *IECON 2019 - 45th annual conference of the ieee industrial electronics society*, 2019.

[15] F. Sanabria-Macías, E. Marañón-Reyes, P. Soto-Vega, M. Marrón-Romera, J. Macias-Guarasa, and D. Pizarro-Perez, "Face likelihood functions for visual tracking in intelligent spaces," in *IECON 2013-39th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2013, pp. 7825–7830.

[16] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 174–188, 2002.

[17] K. Wu and A. W. Khong, "Acoustic source tracking in reverberant environment using regional steered response power measurement," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*. IEEE, 2013, pp. 1–6.

[18] X. Zhong, "Bayesian framework for multiple acoustic source tracking," Ph.D. dissertation, The University of Edinburgh, 2010.

[19] J. D. Hol, T. B. Schon, and F. Gustafsson, "On resampling algorithms for particle filters," in *2006 IEEE nonlinear statistical signal processing workshop*. IEEE, 2006, pp. 79–82.

[20] 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.