Exploiting the Complementarity of Audio and Visual Data in Multi-Speaker Tracking *

Yutong Ban, Laurent Girin, Xavier Alameda-Pineda and Radu Horaud
INRIA Grenoble Rhône-Alpes, LJK and GIPSA Lab, Univ. Grenoble Alpes, France.

Abstract

Multi-speaker tracking is a central problem in human-robot interaction. In this context, exploiting auditory and visual information is gratifying and challenging at the same time. Gratifying because the complementary nature of auditory and visual information allows us to be more robust against noise and outliers than unimodal approaches. Challenging because how to properly fuse auditory and visual information for multi-speaker tracking is far from being a solved problem. In this paper we propose a probabilistic generative model that tracks multiple speakers by jointly exploiting auditory and visual features in their own representation spaces. Importantly, the method is robust to missing data and is therefore able to track even when observations from one of the modalities are absent. Quantitative and qualitative results on the AVDIAR dataset are reported.

1. Introduction

We address the problem of automatically tracking several moving persons using both visual and audio data. The scenario that we address is a social scene with people interacting with each other, in particular they speak to each other. The tracking system is dedicated to be embedded in a humanoid robot placed among the humans and also involved in the interaction. The cameras and microphones used to capture the audio-visual (AV) data are typically placed on the robot head, and are at mid-distance from the people (one to five meters). In such a context, multi-speaker tracking aims to identify, localize and follow the different persons involved in the social interaction in order to prepare the robot for higher level tasks such as automatic speech recognition and multi-person dialog.

Both visual and audio data are particularly rich but they are also difficult to process. The visual observations consist, *e.g.* of bounding boxes provided by a person detector. Such a detector provides both good person localization and person appearance when people in the scene are clearly separated, which allow to simultaneously track multiple persons. In practice multi-person visual detection and tracking suffers from occlusions and missing data: people wander around, cross each other, turn their head away from the camera, move in and out of the camera field of view, etc. Hence visual detectors may yield both false positives and missing detections. The nature of auditory data can help to partially solve this issue: audio processing suffers neither from limited field-of-view nor from occlusions. However, natural speech often happens intermittently and simultaneously [1]. In addition, every day indoor environments (domestic or office) suffer from a significant amount of reverberation, and the audio signals recorded by the microphones can be modeled as time-varying convolutive mixtures of speech signals uttered by several persons, with the number and identity of speakers varying over time. This kind of mixture signals are still quite difficult to separate [2], yet this is a prerequisite for automatic speech recognition and dialog handling. Clearly, the knowledge of who is where and when in the scene could help separating the sound sources, for instance by using beamforming techniques [3].

The vast majority of research studies on multi-person tracking exploits visual information only. However, some methods exist for audio-visual tracking. Particle filters and probability hypothesis density (PHD) filters are the most common frameworks for audio-visual tracking. However, the particle generation procedure may lead to a high computational cost. [4, 5] 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, [6] employs 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 AV combination is mostly to increase the sampling efficiency, with audio and visual data required to be available simultaneously. Alternatively, [7]

*Work supported by the European Research Council through the Advanced Grant #340113 *Vision and Hearing in Action* (VHIA).

used a Markov chain Monte Carlo particle filter (MCMC-PF) to increase sampling efficiency. Still in a particle filter tracking framework, [8] 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, along a different line, [9] 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 the critical point. In addition, the tracking state-space is fixed in the sense that the number of tracked people is set in advance, and their identity is not allowed to change as the scene evolves. Furthermore, most of the existing methods, up to our knowledge, are designed for meeting-room settings, *e.g.* using a distributed sensor network. However, seldom are multi-person audio-visual tracking methods designed to work with robots. Commonly, methods designed for HRI applications, *e.g.* [10], use simple particle filter techniques for audio-visual data fusion.

In this paper we propose a novel multi-speaker tracking method inspired from previous research on "instantaneous" audio-visual fusion [11, 12]. A dynamic Bayesian model is investigated to smoothly fuse acoustic and visual information over time from their feature spaces. The visual observations consist of bounding boxes provided by a person detector (head detection in our case) and audio observations consist of binaural features extracted from two-channel audio recordings. We propose an efficient solution based on a variational approximation of the posterior distribution of multi-speaker location, and an associated expectation-maximization (EM) procedure. The solution takes the form of a multi-target audiovisual Kalman filter where both visual and audio information are processed on common grounds. The proposed method can deal with visual clutter and missing data, *e.g.* people wander around, cross each other, turn their head away from the camera, move in and out of the camera field of view, and acoustic clutter and missing data, *e.g.* mixed speech signals with silent intervals, environmental noise, reverberant room. Note that an extended Kalman filter (EKF) has already been used in [13], but it was for single-speaker audio-visual tracking. [14] proposed a method with a Kalman filter for visual tracking, an EKF for audio tracking separately, and a late fusion applied for audio-visual combination. To the best of our knowledge, this is the first work proposing a robust combination of visual and audio information into a single Kalman filter for multi-speaker tracking in realistic scenarios. Furthermore, the computational complexity and on-line nature of the proposed method makes it a good candidate to be embedded in a humanoid robot.

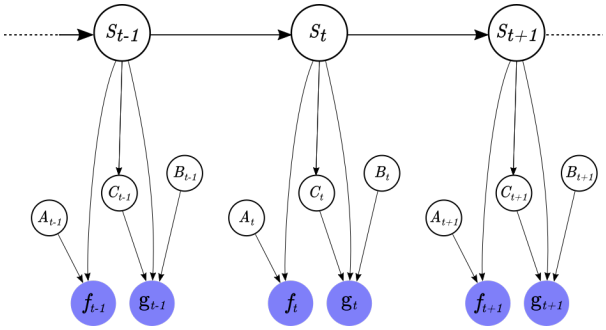


Figure 1. Graphical representation of the proposed probabilistic model.

2. Audio-Visual Probabilistic Model

We start by introducing a few notations and definitions. Unless otherwise specified, upper-case letters denote random variables while lower-case letters denote their realizations, *e.g.* $P(X = x)$. For the sake of conciseness we sometimes write (except when it is necessary to avoid ambiguities) $p(x)$. Vectors are written in slanted bold, *e.g.* $\boldsymbol{X}, \boldsymbol{x}$ while matrices are written in bold, $\mathbf{Y}, \mathbf{y}$. The objective is to use audio and visual data to track speakers over time. Let t denote the frame index. Since we suppose that the data are synchronized, this index is common to the two modalities. Let n be a person index, and let N be the upper bound of the number of persons that can simultaneously be present at any time t. Moreover, let $n = 0$ denote *nobody*. We now introduce two latent variables associated with person n at t, the person's location in the image plane $\boldsymbol{X}_{tn} \in \mathcal{X} \subset \mathbb{R}^2$ and the person's velocity $\boldsymbol{Y}_{tn} \in \mathcal{Y} \subset \mathbb{R}^2$, and let $\boldsymbol{S}_{tn} = (\boldsymbol{X}_{tn}, \boldsymbol{Y}_{tn})$. Moreover, let $\{\boldsymbol{F}_{tm}\}_{m=1}^{M_t}$ and $\{\boldsymbol{G}_{tk}\}_{m=1}^{K}$ denote the sets of visual observations and of audio observations, respectively, available at t. Let $\mathbf{F}_t = (\boldsymbol{F}_{t1} \dots \boldsymbol{F}_{tm} \dots \boldsymbol{F}_{tM_t})$, $\mathbf{G}_t = (\boldsymbol{G}_{t1} \dots \boldsymbol{G}_{tk} \dots \boldsymbol{G}_{tK})$, and $\mathbf{O}_t = (\mathbf{F}_t, \mathbf{G}_t)$. As explained below, the number of visual observations may vary over frames, while the number of audio observations remains fixed.

As already mentioned, the objective is to track multiple persons and to estimate over time the audio status (speaking or silent) of each tracked person. This problem can be cast into the estimation, at each time step t, of the maximum a posteriori (MAP) of the following filtering distribution:

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

Where $\mathbf{o}_{1:t} = (\mathbf{o}_1 \dots \mathbf{o}_t)$ and $\boldsymbol{z}_t = (\boldsymbol{a}_t, \boldsymbol{b}_t, \boldsymbol{c}_t)$ jointly denotes three assignment variables that will be made explicit below. By applying Bayes rule and by assuming that $\mathbf{s}_t$ follows a first-order Markov model, while the visual and audio observations are independent, Figure 1, one can write (1) as:

$$p(\mathbf{s}_t, \boldsymbol{z}_t | \mathbf{o}_{1:t}) \propto p(\mathbf{o}_t | \mathbf{s}_t, \boldsymbol{z}_t,) p(\boldsymbol{z}_t | \mathbf{s}_t) p(\mathbf{s}_t | \mathbf{o}_{1:t-1}), \tag{2}$$

with:

$$p(\mathbf{o}_t|\mathbf{s}_t, \boldsymbol{z}_t) = p(\mathbf{f}_t|\mathbf{x}_t, \boldsymbol{a}_t)p(\mathbf{g}_t|\mathbf{x}_t, \boldsymbol{b}_t, \boldsymbol{c}_t), \quad (3)$$

$$p(\boldsymbol{z}_t|\mathbf{s}_t) = p(\boldsymbol{a}_t)p(\boldsymbol{b}_t)p(\boldsymbol{c}_t|\mathbf{s}_t), \quad (4)$$

$$p(\mathbf{s}_t|\mathbf{o}_{1:t-1}) = \int p(\mathbf{s}_t|\mathbf{s}_{t-1})p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})d\mathbf{s}_{t-1}, \quad (5)$$

where (3) is the joint (audio-visual) observed-data likelihood and the probabilities in (4) can be written as: $p(\boldsymbol{a}_t) = \prod_{m=1}^{M_t} p(A_{tm} = n)$ and $p(\boldsymbol{b}_t) = \prod_{k=1}^{K} p(B_{tk} = n)$. We simplify these notations with $\eta_n = p(A_{tm} = n)$ and $\rho_n = p(B_{tk} = n)$. We also have $p(\boldsymbol{c}_t|\mathbf{s}_t) = \prod_{k=1}^{K} p(C_{tk} = r|\mathbf{s}_t)$ whose expression is given by (14) below.

In these formulas, A, B, and C are assignment variables. The notations $A_{tm} = n$ and $B_{tk} = n$ mean that at t the m-th visual observation and the k-th audio observation, respectively, are assigned to person n. Note that we also allow the assignments $A_{tm} = 0$ and $B_{tk} = 0$ meaning that the corresponding visual and audio observations are assigned to *nobody*. In the visual domain, this corresponds to a bad person detection. In the audio domain it corresponds to an acoustic feature that is not uttered by a person, *e.g.* environmental noise. The remaining assignment variables, $\boldsymbol{C}_t$, are associated with the audio generative model described in more detail in Section 2.2 below.

The predictive distribution (5) at t is computed recursively from the marginal of the transition distribution and from the predictive distribution at $t-1$. The transition distribution is computed with:

$$p(\mathbf{s}_t|\mathbf{s}_{t-1}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{s}_{t-1\,n}, \boldsymbol{\Lambda}_{tn}), \quad (6)$$

where the transition matrix is given by:

$$\mathbf{D} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (7)$$

2.1. The Visual Model

The visual data are recorded with an RGB camera. Let a visual observation consist of a bounding box and of a feature vector describing the RGB region inside this box, hence $\boldsymbol{f}_{tm} = (\boldsymbol{v}_{tm}, \boldsymbol{h}_{tm})$, where $\boldsymbol{v}_{tm} \in \mathcal{V} \subset \mathbb{R}^2$ is the bounding-box center and $\boldsymbol{h}_{tm} \in \mathcal{H} \subset \mathbb{R}^P$ is the feature vector. The bounding-box/feature-vector pairs are the result of applying a person detector to each image. One may indifferently use a full-person detector, an upper-body detector or a face detector. As for computing the feature vector, one may use any of the numerous region descriptors available in the computer vision literature, such as a color histogram, HOG, SIFT, or one among the many CNN-based feature representations that were recently made available.

Assuming that the M_t visual observations available at t are independent and that the appearance of a person is independent his/her position in the image, we obtain the following decomposition of the visual likelihood in (3):

$$p(\mathbf{f}_t|\mathbf{x}_t, \boldsymbol{a}_t) = \prod_{m=1}^{M_t} p(\boldsymbol{v}_{tm}|\mathbf{x}_t, a_{tm})p(\boldsymbol{h}_{tm}|\mathbf{h}, a_{tm}), \quad (8)$$

where the bounding-box centers and the feature vectors are drawn from the following distributions, respectively:

$$p(\boldsymbol{v}_{tm}|\mathbf{x}_t, A_{tm} = n) = \begin{cases} \mathcal{N}(\boldsymbol{v}_{tm}; \boldsymbol{x}_{tn}, \boldsymbol{\Phi}) & \text{if } 1 \leq n \leq N \\ \mathcal{U}(\boldsymbol{v}_{tm}; \text{vol}(\mathcal{V})) & \text{if } n = 0, \end{cases} \quad (9)$$

and

$$p(\boldsymbol{h}_{tm}|\mathbf{h}, A_{tm} = n) = \begin{cases} \mathcal{B}(\boldsymbol{h}_{tm}; \boldsymbol{h}_n) & \text{if } 1 \leq n \leq N \\ \mathcal{U}(\boldsymbol{h}_{tm}; \text{vol}(\mathcal{H})) & \text{if } n = 0, \end{cases} \quad (10)$$

where $\boldsymbol{\Phi} \in \mathbb{R}^2$ is a covariance matrix associated with the error in the bounding-box center, $\mathcal{U}(\cdot; \text{vol}(\cdot))$ is the uniform distribution with $\text{vol}(\cdot)$ being the support volume of the space spanned by the set of observations, $\mathcal{B}(\cdot; \boldsymbol{h}_n)$ is the Bhattacharya distribution, and $\mathbf{h} = (\boldsymbol{h}_1 \dots \boldsymbol{h}_n \dots \boldsymbol{h}_N) \in \mathcal{H} \subset \mathbb{R}^P$ are the feature vectors that describe the appearances of the N persons.

2.2. The Audio Model

Without loss of generality, we assume that the audio data are recorded with a microphone pair. The short-time Fourier transform (STFT) is applied to the left- and right-microphone signals. The effect of applying the STFT is twofold: by sliding a window of a fixed size along the temporal signal and by applying the FFT to the windowed signal one obtains (i) a sequence of audio frames, and for each frame (ii) a complex-valued vector of K Fourier coefficients, where K is the number of frequencies, *e.g.* $K = 512$. By computing the ratio between the left and right Fourier coefficients for each frequency $k \in \{1 \dots K\}$, we obtain a complex number, whose module corresponds to the inter-channel level difference (ILD) and whose argument corresponds to the inter-channel phase difference (IPD). By representing the IPD with the sine and cosine of the argument, we obtain the observations at each frame t: $\mathbf{g}_t = (\boldsymbol{g}_{t1} \dots \boldsymbol{g}_{tk} \dots \boldsymbol{g}_{tK}) \in \mathcal{G} \subset \mathbb{R}^{3\times K}$. By assuming that

these observations are independent, the audio likelihood in (3) can be written as:

$$p(\mathbf{g}_t|\mathbf{x}_t, \boldsymbol{b}_t, \boldsymbol{c}_t) = \prod_{k=1}^{K} p(\boldsymbol{g}_{tk}|\mathbf{x}_t, b_{tk}, \boldsymbol{c}_t) \tag{11}$$

It is well known that both the ILD and IPD contain audio direction information. Nevertheless, because of the presence of reverberation, the ILD-IPD to audio-direction mapping is non-linear. As proposed in [15] we approximate this non-linear mapping by a piecewise affine transformation. The joint probability of the observed and latent variables can be written as:

$$p(\boldsymbol{g}_{tk}, \boldsymbol{x}_{tn}, B_{tk} = n) = \sum_{r=1}^{R} p(\boldsymbol{g}_{tk}|\boldsymbol{x}_{tn}, B_{tk} = n, C_{tk} = r) \times p(\boldsymbol{x}_{tn}|B_{tk} = n, C_{tk} = r)p(C_{tk} = r)p(B_{tk} = n), \tag{12}$$

with: $p(\boldsymbol{g}_{tk}|\boldsymbol{x}_{tn}, B_{tk} = n, C_{tk} = r) = \mathcal{N}(\boldsymbol{g}_{tk}; \mathbf{L}_{kr}\boldsymbol{x}_{tn} + \boldsymbol{l}_{kr}, \boldsymbol{\Sigma}_{kr})$, $p(\boldsymbol{x}_{tn}|B_{tk} = n, C_{tk} = r) = \mathcal{N}(\boldsymbol{x}_{tn}; \boldsymbol{\nu}_r, \boldsymbol{\Omega}_r)$, $p(B_{tk} = n) = \rho_n$, $p(C_{tk} = r) = \pi_r$. In these formulas, the matrix $\mathbf{L}_{kr} \in \mathbb{R}^{3\times 2}$ and the vector $\boldsymbol{l}_{kr} \in \mathbb{R}^{3\times 1}$ characterize the r-th affine transformation that maps the location $\boldsymbol{x}_{tn}$ of audio source n onto its corresponding binaural feature $\boldsymbol{g}_{tk}$, while the covariance matrix $\boldsymbol{\Sigma}_{kr} \in \mathbb{R}^{3\times 3}$ captures both the presence of noise in the data as well as the reconstruction error due to the affine approximation. Consequently, the audio likelihood (11) is provided by:

$$p(\boldsymbol{g}_{tk}|\mathbf{x}_t, B_{tk} = n, C_{tk} = r) = \begin{cases} \mathcal{N}(\boldsymbol{g}_{tk}; \mathbf{L}_{kr}\boldsymbol{x}_{tn} + \boldsymbol{l}_{kr}, \boldsymbol{\Sigma}_{kr}) & \text{if } 1 \leq n \leq N \\ \mathcal{U}(\boldsymbol{g}_{tk}; \text{vol}(\mathcal{G})) & \text{if } n = 0 \end{cases} \tag{13}$$

Additionally, the space of possible audio-source locations is modeled with a Gaussian mixture model with means $\{\boldsymbol{\nu}_1 \dots \boldsymbol{\nu}_r \dots \boldsymbol{\nu}_R\} \subset \mathbb{R}^{2\times 1}$ and covariances $\{\boldsymbol{\Omega}_1 \dots \boldsymbol{\Omega}_r \dots \boldsymbol{\Omega}_R\} \subset \mathbb{R}^{2\times 2}$. This yields the following expression for the posterior probabilities in (4), $\forall k \in \{1 \dots K\}, \forall r \in \{1 \dots R\}$:

$$p(C_{tk} = r|\boldsymbol{x}_{tn}) = \frac{\pi_r \mathcal{N}(\boldsymbol{x}_{tn}; \boldsymbol{\nu}_r, \boldsymbol{\Omega}_r)}{\sum_{i=1}^{R} \pi_i \mathcal{N}(\boldsymbol{x}_{tn}; \boldsymbol{\nu}_i, \boldsymbol{\Omega}_i)}. \tag{14}$$

Finally, the above model is characterized by the following set of parameters:

$$\boldsymbol{\theta} = \{\mathbf{L}_{kr}, \boldsymbol{l}_{kr}, \boldsymbol{\Sigma}_{kr}, \boldsymbol{\nu}_r, \boldsymbol{\Omega}_r, \pi_r\}_{k=1,r=1}^{k=K,r=R}, \tag{15}$$

which can estimated from a training dataset $\{\boldsymbol{x}_j, \boldsymbol{g}_j\} \in \mathcal{X} \times \mathcal{G}$ and using the methodology proposed in [16].

3. Variational Inference

The maximization (1) is intractable because of the complexity of the posterior distribution. Indeed, the integration (5) does not have an analytical solution. Consequently, the evaluation of the expectations computed with respect to this distribution is also intractable which in turn does not lead to an efficient EM algorithm. We overcome this problem via an approximate solution, namely we assume that the posterior (2) factorizes as:

$$\overline{q}(p(\boldsymbol{z}_t, \mathbf{s}_t|\mathbf{o}_{1:t}) \approx q(\boldsymbol{z}_t, \mathbf{s}_t) = q(\boldsymbol{z}_t) \prod_{n=0}^{N} q(\boldsymbol{s}_{tn}). \tag{16}$$

The optimal solution $\overline{q}(\boldsymbol{z}_t, \mathbf{s}_t)$ is given by [17, 18]:

$$\log \overline{q}(\boldsymbol{s}_{tn}) = \mathbf{E}_{q(\boldsymbol{z}_t)\prod_{m\neq n} q(\boldsymbol{s}_{tm})}[\log p(\boldsymbol{z}_t, \mathbf{s}_t|\mathbf{o}_{1:t})], \tag{17}$$

$$\log \overline{q}(\boldsymbol{z}_t) = \mathbf{E}_{q(\mathbf{s}_t)}[\log p(\boldsymbol{z}_t, \mathbf{s}_t|\mathbf{o}_{1:t})]. \tag{18}$$

We now assume that the variational posterior $\overline{q}(\boldsymbol{s}_{t-1\,n})$ follows a Gaussian distribution parameterized by the mean $\boldsymbol{\mu}_{t-1\,n}$ and the covariance $\boldsymbol{\Gamma}_{t-1\,n}$:

$$\overline{q}(\boldsymbol{s}_{t-1\,n}) = \mathcal{N}(\boldsymbol{s}_{t-1\,n}; \boldsymbol{\mu}_{t-1\,n}, \boldsymbol{\Gamma}_{t-1\,n}). \tag{19}$$

By substituting (19) into (5) and combining it with (6), the predictive distribution (5) becomes:

$$p(\mathbf{s}_t|\mathbf{o}_{1:t-1}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{s}_{tn}; \mathbf{D}\boldsymbol{\mu}_{t-1\,n}, \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^\top + \boldsymbol{\Lambda}_{tn}).$$

We can now compute (17) and (18). We start with substituting the previous equation in (5) and we use the formulas for (3) and (4) derived in the previous section. By grouping the terms in (17) and by identification, this yields the following Gaussian distribution:

$$\overline{q}(\boldsymbol{s}_{tn}) = \mathcal{N}(\boldsymbol{s}_{tn}; \boldsymbol{\mu}_{tn}, \boldsymbol{\Gamma}_{tn}), \tag{20}$$

with:

$$\boldsymbol{\Gamma}_{tn} = \Big[\Big(\boldsymbol{\Lambda}_n + \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^\top\Big)^{-1} + \Big(\sum_{m=1}^{M_t} \alpha_{tmn}\Big)\boldsymbol{\Phi}^{-1} + \sum_{k=1}^{K}\sum_{r=1}^{R} \beta_{tknr}\mathbf{L}_{kr}^\top\boldsymbol{\Sigma}_{kr}^{-1}\mathbf{L}_{kr}\Big]^{-1}, \tag{21}$$

$$\boldsymbol{\mu}_{tn} = \boldsymbol{\Gamma}_{tn}\Big[\Big(\boldsymbol{\Lambda}_n + \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^\top\Big)^{-1}\mathbf{D}\boldsymbol{\mu}_{t-1\,n} + \boldsymbol{\Phi}^{-1}\Big(\sum_{m=1}^{M_t} \alpha_{tmn}\boldsymbol{f}_{tm}\Big) + \sum_{k=1}^{K}\sum_{r=1}^{R} \beta_{tknr}\mathbf{L}_{kr}^\top\boldsymbol{\Sigma}_{kr}^{-1}(\mathbf{g}_{kr} - \boldsymbol{l}_{kr})\Big], \tag{22}$$

where $\alpha_{tmn} = \overline{q}(A_{tm} = n)$ (respectively $\beta_{tknr} = \overline{q}(B_{tk} = n, C_{tk} = r)$) is the posterior probability of assigning the m-th visual observation (respectively the k-th auditory observation) to person n. The formulas for these posteriors are derived from (18) as show below.

We now develop (18) and it turns out that the visual and audio assignment variables are conditionally independent, and that the assignment variables for each observation are also independent. Formally this can be written as:

$$\overline{q}(\boldsymbol{z}_t) = \prod_{m=1}^{M_t} \overline{q}(a_{tm}) \prod_{k=1}^{K} \overline{q}(b_{tk}, c_{tk}). \tag{23}$$

In addition, we obtain closed-form formulae for the above distributions. Indeed, for the visual assignment we obtain:

$$\alpha_{tmn} = \overline{q}(A_{tm} = n) = \frac{\tau_{tmn}\eta_n}{\sum_{i=0}^{N} \tau_{tmi}\eta_i},$$

where η_n is the prior probability to assign a visual observation to person n, and τ_{tmn} is defined as:

$$\tau_{tmn} = \begin{cases} \mathcal{N}(\boldsymbol{v}_{tm}; \boldsymbol{\mu}_{tn}, \boldsymbol{\Phi}) e^{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Phi}^{-1}\boldsymbol{\Gamma}_{tn}\right)} \mathcal{B}(\boldsymbol{h}_{tm}; \boldsymbol{h}_n) & n \neq 0 \\ \mathcal{U}(\boldsymbol{v}_{tm}; \mathrm{vol}(\mathcal{V})) \mathcal{U}(\boldsymbol{h}_{tm}; \mathrm{vol}(\mathcal{H})) & n = 0. \end{cases}$$

For the audio posterior assignment we obtain:

$$\beta_{tknr} = \overline{q}(B_{tk} = n, C_{tk} = r) = \frac{\kappa_{tknr}\rho_n\pi_r}{\sum_{i=0}^{N}\sum_{j=1}^{R} \kappa_{tknij}\rho_i\pi_j},$$

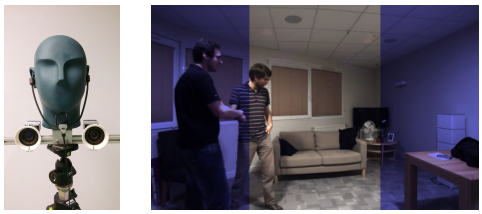
where ρ_n is the prior probability to assign an audio observation to person n, π_r is the prior probability associated with the audio model (14), and κ_{tknr} is given by:

$$\kappa_{tknr} = \begin{cases} \mathcal{N}(\boldsymbol{g}_{tk}; \mathbf{L}_{kr}\boldsymbol{\mu}_{tn} + \boldsymbol{l}_{kr}, \boldsymbol{\Sigma}_{kr}) \\ e^{-\frac{1}{2}\mathrm{tr}\left(\mathbf{L}_{kr}^{\top}\boldsymbol{\Sigma}_{kr}^{-1}\mathbf{L}_{kr}\right)} \mathcal{N}(\hat{\boldsymbol{x}}_{tn}; \boldsymbol{\nu}_r, \boldsymbol{\Omega}_r)) & n \neq 0 \\ \mathcal{U}(\boldsymbol{g}_{tk}; \mathrm{vol}(\mathcal{G})) & n = 0. \end{cases}$$

An approximation is made as follows to have closed-form expression. In (14), $p(\boldsymbol{c}_t|\mathbf{x}_t)$ in (4) is calculated by replacing $\boldsymbol{x}_t$ with $\boldsymbol{x}_{tn}$. Note that we can immediately see that if we need to take the logarithm or the expectation w.r.t. $\boldsymbol{X}_{tn}$ of (14), there is no closed-form solution due to the sum in the denominator. To overcome this problem, we will replace $\boldsymbol{x}_{tn}$ by its prediction from time $t-1$ denoted $\hat{\boldsymbol{x}}_{tn}$. This latter is extracted from $\hat{\boldsymbol{s}}_{tn} = \mathbf{G}\boldsymbol{\mu}_{t-1n}$, where $\boldsymbol{\mu}_{t-1n}$ is the estimated state of source n at time $t-1$.

4. Birth Process

In order to deal with a time-varying number of speakers in the scene, we adopt a birth process that allows to create new tracks. We extend the single modality strategy in [19] to and audio-visual birth process. Indeed, we monitor the posterior distribution of the visual and auditory samples during the past T_{new} frames and perform a statistical test to decide whether a group of observations belong to a new person or they have been correctly assigned to the clutter class. Intuitively, we test the consistency of the auditory and visual observations recently assigned to the clutter class, that is $\mathcal{C} = \{\boldsymbol{f}_{t'k}, \boldsymbol{g}_{t'm}\}_{t'=t-T_{\text{new}}, A_{tk}=0, B_{tm}=0}^{t}$. If the consistency of the new observation sequence satisfies $p(\mathcal{C}) > \tau$, where τ is the probability that these observations are generated by clutter, a new track is created. The consistency is evaluated by calculating the joint probability of the new observation sequence. Consequently $\boldsymbol{\mu}_{tN_t+1}$ is set to the most recent geometric observation in $\mathcal{C}$, and the appearance observations are assigned to the new source.



(a) (b)

Figure 2. (a) The Popeye robot. (b) Experimental settings. Blue areas are *blind areas*, *i.e.* where the visual information is not used.

5. Experiments

The method is evaluated on the *moving participants* subset of the AVDIAR dataset [20]. It consists of several sequences recording multiple speakers freely moving and chatting in a natural indoor environment. Speakers may occlude each other, look in different directions, occasionally speak simultaneously, etc. As shown in Fig. 2, the AVDIAR dataset was recorded with the Popeye robot, constituted by an acoustic dummy head with a stereo camera pair and a pair of microphones. The cameras provide a field of view of $97^o \times 80^o$ (horizontal $\times$ vertical), with an image resolution of 1920×1200 pixels at 25 FPS. The in-ear microphone pair is placed in the left and right ears of the dummy hear [21]. The audio signal provided is down-sampled to 16 kHz.

The AVDIAR dataset is specifically chosen for its sensing capabilities. Indeed, the combination of auditory and wide-angle visual information in a robotic head are quite unique and highly appropriate for the purpose of the present study, since it allows to simulate an audio-visual multi-speaker track with speakers going in and out of the field of view. In order to do that, we cut the wide-angle image into 3 parts, representing 30%, 40%, and 30% of the image width, as shown in Fig 2 (b). On one side, during the training phase all three parts are used so that the domain of the

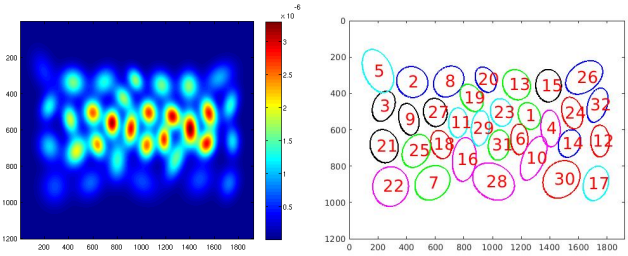
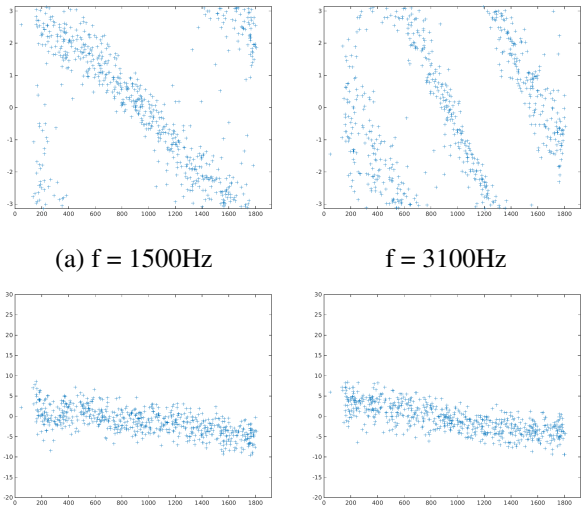


Figure 3. Visualisation of the marginal GMM components in the source position (image) space, for $R = 32$. (a) Audio GMM distribution density (b) Visualisation of location of different component

audio observation model spans the entire field of view of the camera, emulating a factory or built-in calibration. On the other side, during tracking, only the central part of the image is available, thus simulating a standard color camera. In other words, the image content from the blind areas are not used during tracking. The wide-angle image is also used for annotation (and thus evaluation) purposes.

ILD and IPD. Auditory speaker localization is achieved from the interaural level difference (ILD) and the interaural phase difference (IPD), which are the modulus and phase of the ratio between the left and right spectrograms. The spectrograms are computed from a 128 ms segment of the audio signal sampled at 16 kHz. These segments are synchronized with the video (so the segment shift is 40 ms, corresponding to 25 FPS), and from each segment we extract 16 auditory vectors of dimension 64 as explained in the following. Within each segment, we compute eight STFT by sliding a 123 ms window every 5 ms and padded with zeros to reach back the 128 ms. Then these eight 1024-dimension STFTs are averaged, as this is far more robust than computing one 128 ms STFT. The 1024 TF points are finally grouped into $K = 16$ vectors of 64 consecutive TF points.

Training the Audio Generative Model. As described in Section 2, we use a Gaussian mixture regression model to generate audio observations. The model parameters, $\{\mathbf{L}_{kr}, \boldsymbol{l}_{kr}, \boldsymbol{\Sigma}_{kr}, \boldsymbol{\nu}_r, \boldsymbol{\Omega}_r, \pi_r\}$ with $k = \{1 \dots K\}, r = \{1 \dots R\}$, are estimated via an EM procedure using a training dataset $\{\mathbf{g}, \boldsymbol{x}\}$ [16] 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 $\boldsymbol{x}_t$, i.e. $\sum_{r=1}^{R} \pi_j \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\nu}_r, \boldsymbol{\Omega}_r)$, obtained via training, is illustrated in Fig. 3. This figure shows that the field of view is well covered by the Gaussian components of our model. Fig 4 displays the ILD and IPD features for the training dataset as a function of the source horizontal position in the image, and for two example frequencies. We see that the IPD feature exhibits a coarse piece-wise linear profile, due



(a) f = 1500Hz f = 3100Hz

(b) f = 1500Hz f = 3100Hz

Figure 4. IPD (a) (in radians) and ILD (b) (in dB) features used in the training dataset of the audio observation model, plotted as a function of the horizontal source position (in image pixel, arbitrarily indexed from 1 to 1920). These features were calculated from white noise sequences emitted from 800 positions.

to phase wrapping. The noisy aspect of the profile is possibly due to reverberation and justifies a probabilistic observation model.

Voice Activity Detection. Even if the use of white noise is justified during the training of the auditory observation model, the proposed system targets the localization of speakers, and thus it must correctly deal with speech signals. To this end, the voice activity detector (VAD) in [22] is used to detect the time intervals in which speech is present, the others are discarded. Even when speech is present in the audio signal, many time-frequency bins may not carry significant information due to the natural sparseness of the speech signal. The TF bins are thresholded, leading to partially empty auditory observations, which is not a problem for the auditory observation probabilistic model. The threshold is learned for each video sequence by assuming that the first few frames correspond to silence. Overall, when the VAD detects speech activity, 16 partially filled auditory observations are collected, otherwise no auditory information is used.

Visual observations. A face/head detector is used to provide the visual observations. In practice, we extract the bounding box of the head, from the full-body pose estimated with [23].

Evaluation Protocol. We evaluate the performance of the proposed method using standard MOT (multi-object tracking) metrics: the multi-object tracking accuracy (MOTA), which combines false positives (FP), missed targets = false negative (FN), and identity switches (ID); the false alarm

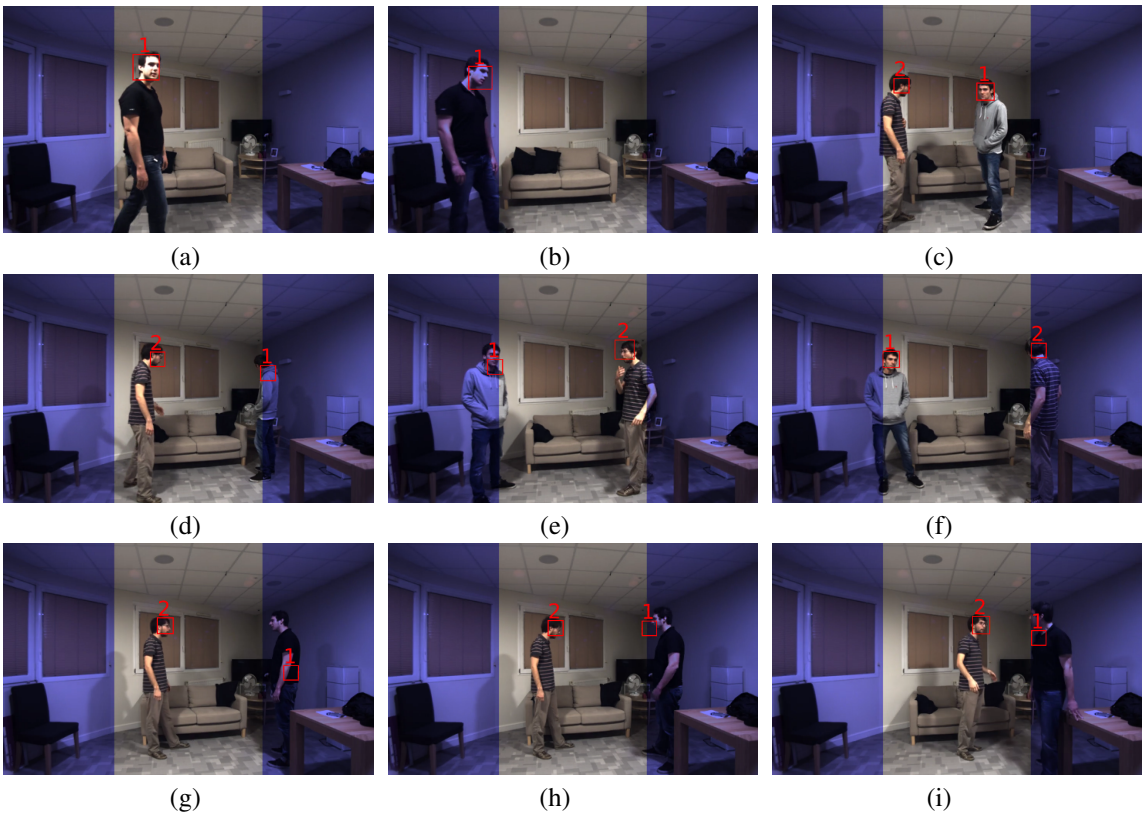


Figure 5. Qualitative result on AVDIAR dataset: (a-b) Seq03-1P-S0M1, (c-f) Seq21-2P-S1M1, (g-i) Seq20-2P-S1M1.

per frame (FAF); the ratio of mostly tracked trajectories (MT); the number of track fragmentations (Frag); the tracking recall (Rcll) and tracking precision (Prcn).

We observed that auditory cues are truly complementary to the visual cues, when the latter are available they are more precise than the former, which are available independently of the position of the speaker. Indeed, visual observations are more precise, but auditory localization is possible outside the field-of-view of the camera. Thus, we expect that tracking with only the acoustic modality would be less accurate than using only the visual modality. Specially since the estimation of the elevation from the two inear dummy head microphones is poorly reliable. We thus compare only the azimuth overlap of the bounding boxes when the person is out of field-of-view. Also, the MOTA score is calculated with an overlap threshold set to 0.9.

Regarding the parameters, the visual observation covariance matrix $\mathbf{\Phi}_m$ is set as a diagonal matrix with values of $\{0.25, 2.5, 2.5, 2.5\}$ times the face-detection bounding box area. This allows the tracker to have more flexibility in elevation than in azimuth. Similarly, the covariance matrix of the dynamical state model $\mathbf{\Lambda}_{tn}$ is set to $\{10^{-2}, 2 \times 10^{-2}, 2 \times 10^{-2}, 2 \times 10^{-2}, 10^{-4}, 10^{-4}\}$ times the corresponding bounding box area.

Results. Qualitative results are illustrated in Fig 5. The proposed multi-speaker tracking algorithm exhibits good performance. Fig 5 (a) and (b) illustrate the results on sequence Seq03-1P-S0M1, which is a single-person sequence. When the speaker appears in the field-of-view, the tracker starts to track the speaker using both audio and visual information. When the speaker stops speaking or goes out of field-of-view, only one modality is present. In such cases, the proposed method continues to track the speaker (see an example in Fig 5 (b)). For the multi-speaker scenarios with a time-varying number of persons, the proposed method gives very promising results (see Fig 5 (c) to (i)). Here also, the tracking is robust to the fact that one of the speakers can go outside the field-of-view or can stop speaking. When only the acoustic modality is present for one speaker, the tracking result is generally less accurate than with the visual information. Especially, the accuracy of elevation estimation drops more significantly than the accuracy of azimuth estimation, as illustrated in Fig 5 (g). Again, the reason for poor elevation estimation is that only one horizontal pair of microphones is used in this experimental settings. However, the tracking in the "blind" regions benefits not only from the audio information but also from the state dynamical model, hence the tracking is particularly robust in Fig 5 (d, f, h, i).

Table 1. Quatitative evaluation on AVDIAR dataset

Sequence	Method	Rcll(↑)	Prcn(↑)	FAR(↓)	MT(↑)	FP(↓)	FN(↓)	IDs(↓)	MOTA(↑)
Seq03-1P-S0M1	AV-A-PF	60	60.1	0.36	0	204	205	0	20.1
	Proposed	96.7	98	0.02	1	10	17	0	94.7
Seq20-2P-S1M1	AV-A-PF	47.4	52.3	0.82	0	1070	1300	64	1.5
	Proposed	82.0	86.1	0.25	1	327	445	9	68.4
Seq21-2P-S1M1	AV-A-PF	58.8	58.4	0.79	0	887	874	17	16.1
	Proposed	79.8	79.2	0.40	1	445	429	6	58.5
Overall	AV-A-PF	53.4	55.8	0.72	0	2161	2379	81	9.4
	Proposed	82.5	84.3	0.26	3	782	891	15	66.9

As illustrated in the supplementary videos[1], the proposed method always provides a consistent tracking result. Still, some failure cases remain. One of them is identity switch, which happened once during the sequence Seq20-2P-S1M1. That is because the visual feature (color histogram) is not discriminative enough and from our knowledge, the capacity to re-identify a speaker by audio information is still limited. This can be improved by simply plug-in some more powerful visual features.

Three sequences with varying scenario complexity are selected to benchmark the method. Seq03-1P-S0M1 is a single-speaker moving scenario and the speaker is not speaking all the time. Seq20-2P-S1M1 and Seq21-2P-S1M1 are 2 sequences with 2 moving speakers. The speakers are in a conversation scenario and they take speech turns. Each of the sequences contains 575, 1301, 1126 video frames separately.

We keep the same experimental setting to quantitatively compare our method with [6]. As this baseline method needs DOA line projected on the image, [16] is utilised to provide DOA. The DOA line appears as a vertical line on the image, since only 2 microphones are utilised in the experiment. The quantitative results are shown in Table 1. We can see from these figures that when the complexity of the scenarios increases, the performance decreases. Closer inspection of the table shows that the proposed method gets significantly outperforms the baseline method since it gives a more robust and continuous tracking result. In addition, when the tracked person is out of the field-of-view, the track obtained by the benchmark method [6] remains still inside the field-of-view, and the tracked person is lost. As Table 1 shows, the proposed method has a much higher tracking recall and precision and much lower false positive, false negative and identity switches. Overall the 3 sequences, the proposed method obtains a MOTA score of 66.9, which is a very large improvement over the benchmark method (9.4).

[1] https://team.inria.fr/perception/research/variational_av_tracking/

6. Conclusions

We proposed a novel method for multi-speaker tracking, which incorporates both audio and visual information via a probabilistic generative model. The method provides robust tracking results, even though both audio and visual data are not continuously available. The results show that the proposed method significantly outperforms the baseline in our experiments. Although we did not evaluate and discuss this point in the present paper, the proposed method is also promising from a computational cost point of view, because the use of a variational EM solution to the model is potentially competitive with respect to sampling based methods. The proposed graphical model can be easily extended to solve some other audio-visual related problem and incorporate additional information, e.g. an automatic speaker recognition module. In particular, future work will be dedicated to extend the proposed multi-speaker tracking model to jointly provide speaker diarization and source separation.

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 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, Louisiana, USA, 2017. 1

[2] 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*, 24(8):1408–1423, 2016. 1

[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, 25(4):692–730, 2017. 1

[4] 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*, 18(12):2417–2431, 2016. 1

[5] 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*, pages 344–353, 2017. 1

[6] Volkan Kılıç, Mark Barnard, Wenwu Wang, and Josef Kittler. Audio assisted robust visual tracking with adaptive particle filtering. *IEEE Transactions on Multimedia*, 17(2):186–200, 2015. 1, 8

[7] 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*, 15(2):601–616, 2007. 1

[8] 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*, pages 2896–2900, New-Orleans, Louisiana, 2017. 2

[9] 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*, 6(5):466–477, 2012. 2

[10] Niclas Schult, Thomas Reineking, Thorsten Kluss, and Christoph Zetzsche. Information-driven active audio-visual source localization. *PloS one*, 10(9), 2015. 2

[11] X. Alameda-Pineda and R. Horaud. Vision-guided robot hearing. *The International Journal of Robotics Research*, 34(4-5):437–456, April 2015. 2

[12] I. D. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud. EM algorithms for weighted-data clustering with application to audio-visual scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016. 2

[13] Tobias Gehrig, Kai Nickel, Hazim Kemal Ekenel, Ulrich Klee, and John McDonough. Kalman filters for audio-video source localization. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 118–121, 2005. 2

[14] Eleonora D'Arca, Neil M Robertson, and James Hopgood. Person tracking via audio and video fusion. In *The Ninth IET Data Fusion and Target Tracking Conference: Algorithms and Applications*, London, UK, 2012. 2

[15] A. Deleforge, F. Forbes, and R. Horaud. High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5):893–911, 2015. 4

[16] A. Deleforge, R. Horaud, Y. Y. Schechner, and L. Girin. Co-localization of audio sources in images using binaural features and locally-linear regression. *IEEE Transactions on Audio, Speech and Language Processing*, 23(4):718–731, 2015. 4, 6, 8

[17] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 4

[18] V. Smidl and A. Quinn. *The Variational Bayes Method in Signal Processing*. Springer-Verlag, Berlin, 2006. 4

[19] Yutong Ban, Sileye Ba, Xavier Alameda-Pineda, and Radu Horaud. Tracking multiple persons based on a variational bayesian model. In *European Conference on Computer Vision*, pages 52–67, Amsterdam, Netherlands, 2016. 5

[20] 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*, 39, 2017. 5

[21] Popeye robot. team.inria.fr/perception/avdiar. 5

[22] Javier Ramirez, José C Segura, Carmen Benitez, Angel De La Torre, and Antonio Rubio. Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, 42(3):271–287, 2004. 6

[23] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016. 6