# Alignment of Binocular-Binaural Data Using a Moving Audio-Visual Target

Vasil Khalidov, Florence Forbes, Radu Horaud

# Alignment of Binocular-Binaural Data Using a Moving Audio-Visual Target

Vasil Khalidov *#1, Florence Forbes *2, Radu Horaud *3

* *INRIA Grenoble Rhone-Alpes*
*655 avenue de l'Europe*
*38330 Montbonnot Saint-Martin, France*

# *Idiap Research Institute*
*Centre du Parc, rue Marconi 19, PO Box 592*
*1920 Martigny, Switzerland*

[1] Vasil.Khalidov@idiap.ch [2] Florence.Forbes@inria.fr, [3] Radu.Horaud@inria.fr

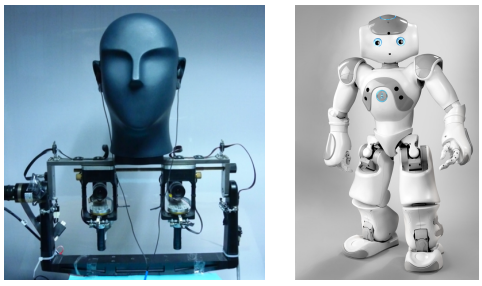
***Abstract*—In this paper we address the problem of aligning visual (V) and auditory (A) data using a sensor that is composed of a camera-pair and a microphone-pair. The original contribution of the paper is a method for AV data aligning through estimation of the 3D positions of the microphones in the visual-centred coordinate frame defined by the stereo camera-pair. We exploit the fact that these two distinct data sets are conditioned by a common set of parameters, namely the (unknown) 3D trajectory of an AV object, and derive an EM-like algorithm that alternates between the estimation of the microphone-pair position and the estimation of the AV object trajectory. The proposed algorithm has a number of built-in features: it can deal with A and V observations that are misaligned in time, it estimates the reliability of the data, it is robust to outliers in both modalities, and it has proven theoretical convergence. We report experiments with both simulated and real data.**

## I. Introduction

Audiovisual (AV) scene analysis has become an increasingly popular research topic during the past years due to many useful applications: human-robot interaction [1], multimodal interfaces [2], audio-visual tracking [3], [4], object localization [5], etc. Various attempts to build computational paradigms for AV scene analysis consider the issue of integration as the cornerstone of the approaches. A popular association principle for the auditory and visual data found in the literature is co-localization [2], [6], [7], meaning that observations from different modalities are fused together as if they were generated from the same spatial source. This leads to an important question: How to align auditory and visual observation spaces, so that the co-localization principle makes sense?

This paper addresses the problem of aligning auditory and visual data gathered with a sensor composed of two cameras and two microphones, e.g, Figure 1. If considered separately, the two cameras are capable of providing dense 3D localization information while the two microphones can be combined to yield partial (azimuthal) sound source localization [8]. In order to *align* the data gathered with these two different sensorial modalities, one has to address the issue of binocular-binaural calibration to guarantee that the two modalities are expressed in the same common coordinate frame (metric alignment) and that they occur simultaneously (temporal alignment). This is a difficult problem that has not been properly addressed in the past.

*MMSP'13, Sept. 30 - Oct. 2, 2013, Pula (Sardinia), Italy.*
*978-1-4799-0125-8/13/$31.00* 



(a) (b)

Fig. 1. Typical binocular-binaural heads include sophisticated devices such as POPEYE shown in (a) and which is composed of an active stereoscopic camera-pair and microphone-pair plugged into the ears a dummy head mounted onto a motor-controlled pan/tilt mechanism, or a camera-pair and a microphone-pair embedded into the motor-controlled head of a consumer robot such as the humanoid robot Nao shown in (b).

There are several difficulties that need to be tackled when aligning auditory (A) and visual (V) data. Firstly, A and V observations belong to two different physical spaces that possess different mathematical properties. Secondly, the A and V observations are not aligned in time and thus it is not obvious how to associate visual events to audio events occurring within a small time interval. Thirdly, the overall sampling rate is not constant, some time intervals contain more observations than others, and thus are more informative. Finally, data from both modalities are strongly affected by noise and outliers, such as visual objects that do not emit sounds, acoustic reverberations, background noise, etc.

### A. Related Work

Almost any audio-visual fusion method requires some kind of spatial alignment, temporal alignment, or both. Whenever an array of microphones and several cameras are used, one can perform independently multiple-microphone localization [9] and multiple-camera calibration [10]. Then the spatial alignment of the two modalities is straightforward and consists in finding the relationship between the microphone-centred and visual-centred coordinate frames such that the two types of sensors refer to the same metric representation. While these methods are well suited for smart-room environments and near-field interaction such as smart kiosks, where a large number of cameras and microphones can be deployed [11], [12], they are not practical in the case of a binaural-binocular *active* robot head. Indeed, they cannot be applied to just two

microphones, they assume stationary sensors and require multiple and perfectly synchronized sound sources. Moreover, the spatial layout of these acoustic sources is constrained by the fact that they must lie within the visual fields of the cameras. We note that there are audio-visual sensor configurations, e.g., one camera and an array of microphones, that do not need full spatial calibration. One can estimate the two-dimensional relationship between the *image position* of a visual feature and an auditory event by *mapping* sounds onto the image plane [2], [6], or by using a rough estimate of microphone locations relative to the camera [3]. Alternatively one can estimate a calibration function that maps the two-dimensional image coordinates of a visual event to the one-dimensional audio angle of arrival in a linear microphone array [13]. In the case of one camera and one microphone, spatial alignment is not possible and methods using this minimal sensor configuration work well only if it is assumed a perfect temporal alignment between the image sequence and the one-dimensional acoustic signal [14]. However, methods using just one camera do not permit to take full advantage of three-dimensional audio-visual event localization which has been proved to be very useful for the detection and localization of multiple speakers [1], [5] or for sound-source separation [15]. Moreover, as already explained, we note that the temporal alignment assumption is not at all realistic.

### B. Contributions

The contribution of this paper is a new method for aligning data from a binaural microphone set with data from a stereoscopic camera system through microphone location estimation. The audiovisual calibration setup is shown in Figure 2. The *audiovisual target* used for calibration consists of a loudspeaker that emits a white-noise acoustic signal and a light source. This target is freely moved in front of the binocular-binaural robot head. We exploit geometrical and physical relations between the 3D scene space, where the target moves, and auditory and visual observation spaces, to formulate the problem as coupled estimation of 3D target trajectory and 3D locations of the two microphones. We propose a Gaussian mixture model (GMM) formulation and we derive an EM algorithm that alternates between assigning audio-visual observations to the target (E-step) and estimating the model parameters, namely the locations of the microphones, the trajectory of the target, and the mixture's priors, means and variances (M-step). The proposed method has a number of desirable built-in features: it can deal with auditory and visual observations that are misaligned in time, it estimates the reliability of the data, it is robust to outliers such as reverberations, and it has proven theoretical convergence.

The remainder of the paper is organized as follows. Section II describes the audio-visual alignment model. This leads to a maximum likelihood formulation and to an associated EM algorithm that is described in detail in Section III. Results obtained with simulated and real data are shown in Section IV. Section V concludes the paper along with a short discussion.

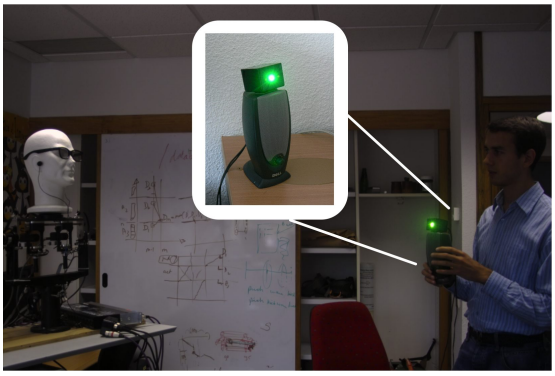


Fig. 2. Audio-visual device used to align the auditory and visual spaces. An LED light bulb is mounted onto a speaker which makes the visual localization more precise. White noise is played throughout the recording to improve the auditory localization.

## II. Observation Space Alignment Through Multimodal Trajectory Matching

Two cameras and two microphones observe an audiovisual target, e.g., Figure 2. This target consists of an auditory source and a visual source and moves along a free and unknown trajectory $\boldsymbol{s}(t) = (x(t), y(t), z(t))$ in the 3D scene $\mathbb{S} \subset \mathbb{R}^3$. The audiovisual target is observed at times $\{t_m^{\mathrm{f}}\}_{m=1}^{M}$ in the visual observation space, and at times $\{t_k^{\mathrm{g}}\}_{k=1}^{K}$ in the auditory observation space. This gives rise to two sets of observations, visual ($\mathbf{F}$) and auditory ($\boldsymbol{G}$):

$$\mathbf{F} = \{\boldsymbol{f}_m\}_{m=1}^{M}, \qquad \boldsymbol{f}_m = \boldsymbol{f}(t_m^{\mathrm{f}}) \in \mathbb{F} \subset \mathbb{R}^3, \tag{1}$$

$$\boldsymbol{G} = \{g_k\}_{k=1}^{K}, \qquad g_k = g(t_k^{\mathrm{g}}) \in \mathbb{G} \subset \mathbb{R}. \tag{2}$$

One important ingredient of our model is that it considers an audiovisual *generative model*, i.e., the transformations $\mathcal{F} : \mathbb{S} \to \mathbb{F}$ and $\mathcal{G} : \mathbb{S} \to \mathbb{G}$ that map a 3D audiovisual object onto the visual and auditory observation spaces. Assuming a pinhole camera model and a rectified stereoscopic pair of images [10], the mapping $\mathcal{F}$ associates a point $\boldsymbol{s} = (x, y, z) \in \mathbb{S}$ from the 3D scene to a *stereoscopic* visual observation $\boldsymbol{f}$:

$$\boldsymbol{f} = (u, v, d) = \mathcal{F}(\boldsymbol{s}) = (x/z, y/z, 1/z)^{\top}, \tag{3}$$

where $(u, v)$ are the 2D coordinates of a left-image point and $d$ is the *horizontal disparity* between two matched points Note that the *projective* mapping defined by (3) is one-to-one and is invertible. Similarly, assuming constant-velocity sound propagation, the auditory mapping $\mathcal{G}$ relates a point $\boldsymbol{s} = (x, y, z) \in \mathbb{S}$ in the 3D scene to an auditory observation $g$, the *interaural time difference* (ITD) of a sound emitted from $\boldsymbol{s}$ and perceived by the left and right microphones:

$$g = \mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}) = c^{-1}\Big(\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\ell}\| - \|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}\|\Big) \tag{4}$$

where $c \approx 343\mathrm{ms}^{-1}$ is the sound speed and $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$ are the 3D positions of the left and right microphones. Unlike the binocular visual model, the binaural mapping $\mathcal{G}$ is not injective: in the 3D space there exists a hyperboloid that is associated to an auditory observation $g$.

*Audiovisual calibration* consists in estimating the microphone locations $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$ in the coordinate system in which the mappings $\mathcal{F}$ and $\mathcal{G}$ are defined. Techniques for

stereo calibration are well understood, both from a methodological and practical points of view. Therefore, we assume that the camera-pair is calibrated and the 3D scene points $\boldsymbol{s} \in \mathbb{S}$ are described in a camera-centered frame.

We will assume that both the visual and the auditory observations, $\boldsymbol{f}_m$ and $g_k$, are drawn either from a normal distribution $\mathcal{N}$ around the corresponding predictions generated from a 3D trajectory (*inliers*), or from a uniform distribution $\mathcal{U}$ (*outliers*), e.g., reverberations. An assignment variable is associated with each visual observation $\boldsymbol{A} = \{A_m\}_{m=1}^M$ and with each auditory observation $\boldsymbol{B} = \{B_k\}_{k=1}^K$. The notation $A_m = \textit{inlier}$ means that observation $\boldsymbol{f}_m$ was generated from a trajectory point $\boldsymbol{s}_m$ while $A_m = \textit{outlier}$ means that the observation is an outlier. This yields :

$$P(\boldsymbol{f}_m|\boldsymbol{s}_m) = \mu\mathcal{N}(\boldsymbol{f}_m|\mathcal{F}(\boldsymbol{s}_m), \boldsymbol{\Sigma}) + (1-\mu)\mathcal{U}(V) \quad (5)$$

$$P(g_k|\boldsymbol{s}_k) = \lambda\mathcal{N}(g_k|\mathcal{G}(\boldsymbol{s}_k), \sigma) + (1-\lambda)\mathcal{U}(U) \quad (6)$$

where $\mathcal{N}(\cdot|\mathcal{F}(\boldsymbol{s}_m), \boldsymbol{\Sigma})$ and $\mathcal{N}(\cdot|\mathcal{G}(\boldsymbol{s}_k), \sigma)$ are the 3D and 1D normal distributions respectively. The uniform distributions are parameterized by the visual and auditory support volumes, i.e., $\mathcal{U}(V) = 1/V$ and $\mathcal{U}(U) = 1/U$. The prior probabilities are defined by $\mu = P(A_m = \textit{inlier})$ and by $\lambda = P(B_k = \textit{inlier})$. Both auditory and visual observations $g_k$ and $\boldsymbol{f}_m$ are assumed to be independent and identically distributed for different values of $k$ and $m$.

We impose regularity constraints onto the trajectory $\boldsymbol{s}(t)$:

$$P(\mathbf{s}) \propto \exp\big(-\gamma \sum \|\boldsymbol{s}_{n+1} - \boldsymbol{s}_n\|^2/(t_{n+1}-t_n)\big), \quad (7)$$
$$\mathbf{s} = \{\boldsymbol{s}_n\}_{n=1}^N, \quad \boldsymbol{s}_n = \boldsymbol{s}(t_n) \in \mathbb{S} \subset \mathbb{R}^3$$

where $\gamma > 0$ is a regularization scalar and the time-stamp set $\{t_n\}_{n=1}^N = \{t_m^{\mathrm{f}}\}_{m=1}^M \cup \{t_k^{\mathrm{g}}\}_{k=1}^K$ is taken as an ordered union. Hence, $N \leq M+K$, since the auditory and visual time-stamps $t_m^{\mathrm{f}}$ and $t_k^{\mathrm{g}}$ may coincide for some $m$ and $k$.

The alignment problem is then formulated as the *simultaneous inference* of the unknown 3D trajectory $\boldsymbol{s}(t)$, and the 3D locations of the two microphones $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$. This may well be viewed as the following maximization:

$$\{\mathbf{s}^*, \boldsymbol{\theta}^*, \boldsymbol{\psi}^*\} = \underset{\mathbf{s}\in\mathbb{S}^N, \boldsymbol{\theta}\in\Theta, \boldsymbol{\psi}\in\boldsymbol{\Psi}}{\operatorname{argmax}} \log P(\mathbf{F}, \boldsymbol{G}, \mathbf{s}, \boldsymbol{\theta}\ ;\ \boldsymbol{\psi}), \quad (8)$$

$$\text{with} \quad \log P(\mathbf{F}, \boldsymbol{G}, \mathbf{s}, \boldsymbol{\theta}; \boldsymbol{\psi}) = \log P(\mathbf{F}|\mathbf{s}, \boldsymbol{\theta}; \boldsymbol{\psi}) + \log P(\boldsymbol{G}|\mathbf{s}, \boldsymbol{\theta}; \boldsymbol{\psi}) + \log P(\mathbf{s}) + \log P(\boldsymbol{\theta}) \quad (9)$$

and where $\boldsymbol{\theta} = \{\boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}\}$ are the 3D microphone locations in the cameras' reference frame, $\boldsymbol{\psi} = \{\pi, \lambda, \boldsymbol{\Sigma}, \sigma\}$ are the parameters associated with the mixture distributions (5) and (6), and the trajectory likelihood $P(\mathbf{s})$ is given by (7). Microphone locations $\boldsymbol{\theta}$ are assumed to be uniformly distributed over some compact set $\Theta \subset \mathbb{R}^6$: $P(\boldsymbol{\theta}) = \mathcal{U}(\Theta)$.

## III. Simultaneous Microphone Localization and Trajectory Reconstruction

Formally, (8) is an observed-data log-likelihood. It is well known that direct optimization of this log-likelihood function is intractable because of high dimensionality of the task. Therefore, we adopt a maximum-likelihood with missing data formulation. Hence, (8) is replaced with the *expected complete-data log-likelihood* maximization within the EM algorithm:

$$Q(\boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{s}^{(q)}, \boldsymbol{\theta}^{(q)}, \boldsymbol{\psi}^{(q)}) = \log P(\boldsymbol{\theta}) - \sum_{m=1}^M \alpha_m^{(q)} \left(\|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_m)\|_{\boldsymbol{\Sigma}}^2 + \log|\boldsymbol{\Sigma}| - \log\tfrac{\mu}{1-\mu}\right) - \sum_{k=1}^K \beta_k^{(q)} \left((g_k - \mathcal{G}(\boldsymbol{s}_k;\boldsymbol{\theta}))^2 + 2\log\sigma - \log\tfrac{\lambda}{1-\lambda}\right) + M\log(1-\mu) + K\log(1-\lambda) - \gamma\sum_{n=1}^{N-1} \tfrac{\|\boldsymbol{s}_{n+1}-\boldsymbol{s}_n\|^2}{t_{n+1}-t_n} \quad (10)$$

where the posterior probabilities $\alpha_m = P(A_m = \textit{inlier}|\boldsymbol{f}_m)$ and $\beta_k = P(B_k = \textit{inlier}|g_k)$ are given by the standard formulae:

$$\alpha_m^{(q)} = \frac{\mu^{(q)}\mathcal{N}(\boldsymbol{f}_m|\mathcal{F}(\boldsymbol{s}_m^{(q)}), \boldsymbol{\Sigma}^{(q)})}{\mu^{(q)}\mathcal{N}(\boldsymbol{f}_m|\mathcal{F}(\boldsymbol{s}_m^{(q)}), \boldsymbol{\Sigma}^{(q)}) + (1-\mu^{(q)})\mathcal{U}(V)}, \quad (11)$$

$$\beta_k^{(q)} = \frac{\lambda^{(q)}\mathcal{N}(g_k|\mathcal{G}(\boldsymbol{s}_k^{(q)};\boldsymbol{\theta}^{(q)}), \sigma^{(q)})}{\lambda^{(q)}\mathcal{N}(g_k|\mathcal{G}(\boldsymbol{s}_k^{(q)};\boldsymbol{\theta}^{(q)}), \sigma^{(q)}) + (1-\lambda^{(q)})\mathcal{U}(U).} \quad (12)$$

### A. The proposed EM algorithm

The optimization of (10) can be carried out by an EM algorithm. While the E-step of the algortithm is a standard one, i.e., update the current posteriors (11) and (12), the M-step is more difficult to achieve because of the presence of the visual and auditory mappings $\mathcal{F}$ and $\mathcal{G}$, defined by (3) and (4). Hence, the M-step of the algorithm should be further decomposed into a number of conditional maximization steps:

**CM-1**. Using the current estimates of the mixtures' parameters and the current trajectory of the audiovisual target, the microphone locations are estimated with:

$$\boldsymbol{\theta}^{(q+1)} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \left(\sum_{k=1}^K \beta_k^{(q)} \left(g_k - \mathcal{G}(\boldsymbol{s}_k^{(q)};\boldsymbol{\theta})\right)^2 - \log P(\boldsymbol{\theta})\right) \quad (13)$$

**CM-2**. Each 3D trajectory point $\boldsymbol{s}_n$ is estimated with:

$$\boldsymbol{s}_n^{(q+1)} = \underset{\boldsymbol{s}_n\in\mathbb{S}}{\operatorname{argmin}} \left(\gamma\left(\tfrac{\|\boldsymbol{s}_{n+1}^{(q)}-\boldsymbol{s}_n\|^2}{t_{n+1}-t_n} + \tfrac{\|\boldsymbol{s}_n-\boldsymbol{s}_{n-1}^{(q)}\|^2}{t_n-t_{n-1}}\right) + \delta_n^{\mathrm{f}}\alpha_m^{(q)}\|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n)\|_{\boldsymbol{\Sigma}^{(q)}}^2 + \delta_n^{\mathrm{g}}\beta_k^{(q)}\tfrac{(g_k - \mathcal{G}(\boldsymbol{s}_n;\,\boldsymbol{\theta}^{(q+1)}))^2}{\sigma^{(q)}}\right) \quad (14)$$

where $\delta_n^{\mathrm{f}}$ (or $\delta_n^{\mathrm{g}}$) is equal to 1 if there exists $m$ (or $k$) such that $\boldsymbol{f}_m$ (or $g_k$) is observed at $t_n$.

**CM-3**. The mixtures' parameters $\boldsymbol{\psi} = \{\pi, \lambda, \boldsymbol{\Sigma}, \sigma\}$ are computed using the standard formulae for the priors and the following expressions for the covariances:

$$\boldsymbol{\Sigma}^{(q+1)} = \frac{\sum_{m=1}^M \alpha_m^{(q)}(\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_m^{(q+1)}))(\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_m^{(q+1)}))^\top}{\sum_{m=1}^M \alpha_m^{(q)}} \quad (15)$$

$$\sigma^{2(q+1)} = \frac{\sum_{k=1}^K \beta_k^{(q)}\left(g_k - \mathcal{G}(\boldsymbol{s}_k^{(q+1)};\boldsymbol{\theta}^{(q+1)})\right)^2}{\sum_{k=1}^K \beta_k^{(q)}} \quad (16)$$

We note that the mean values $\mathcal{F}(\boldsymbol{s}_m^{(q+1)})$ and $\mathcal{G}(\boldsymbol{s}_k^{(q+1)};\boldsymbol{\theta}^{(q+1)})$ used in (15) and (16) correspond to the same calculated 3D trajectory $\mathbf{s}^{(q+1)}$ mapped into the visual and auditory observation spaces $\mathbb{F}$ and $\mathbb{G}$.

### B. Initialization

Initialization of an EM algorithm has a significant impact on its performance. A good choice for the starting values $\boldsymbol{\theta}^{(0)}$, $\mathbf{s}^{(0)}$ and $\boldsymbol{\psi}^{(0)}$ reduces the number of iterations and hence the overall elapsed time to find the optimal values. Proper initialization also helps the algorithm to find good estimates for the parameters. The initialization procedure that we propose exploits the properties of the generative mappings $\mathcal{F}$ and $\mathcal{G}$ in the following way.

The initial trajectory $\mathbf{s}^{(0)}$ is found using visual observations only, based on standard stereo triangulation. This provides estimates of the trajectory $\{\boldsymbol{s}_m^{(0)}\}_{m=1}^M$ at times $\{t_m\}_{m=1}^M$. Next, the trajectory is interpolated in order to obtain estimates $\{\boldsymbol{s}_k^{(0)}\}_{k=1}^K$ at times $\{t_k\}_{k=1}^K$.

The initial microphone locations $\boldsymbol{\theta}^{(0)} = \{\boldsymbol{s}_{\mathrm{M}_\ell}^{(0)}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}^{(0)}\}$ are calculated as follows. Parts of the initial trajectory $\mathbf{s}^{(0)}$ that correspond to observations $g_k = 0$ are taken to estimate the plane $\mathcal{M} = \{\boldsymbol{s} \mid \mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}) = 0\}$. By definition of $\mathcal{G}$, this plane is orthogonal to the line segment joining the two microphones and goes through its midpoint. Therefore, we initialize the midpoint $\boldsymbol{s}_\mathrm{M} = (\boldsymbol{s}_{\mathrm{M}_\ell} + \boldsymbol{s}_{\mathrm{M}_\mathrm{r}})/2$ on $\mathcal{M}$ and choose $\boldsymbol{s}_{\mathrm{M}_\ell}^{(0)}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}^{(0)}$ to be symmetric with respect to $\mathcal{M}$ and such that $\boldsymbol{\theta}^{(0)} = \{\boldsymbol{s}_{\mathrm{M}_\ell}^{(0)}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}^{(0)}\}$ lies within the compact support $\Theta$. The distance between the two microphones can be roughly estimated by the maximum and minimum ITD values. These are observed when the sound source lies on the line connecting the two microphones.

The parameters $\boldsymbol{\psi}^{(0)}$ associated with the two mixtures (priors and covariances) are chosen according to the prior knowledge on noise levels for the AV sensor.

The two uniform distributions in (5) and (6) are defined based on setup specifications. The size of the auditory domain is defined by the maximum observed ITD values, while the visual domain size depends on the parameters associated with the stereoscopic calibration and on the observed scene limits.

### C. The Optimization Procedure

To infer the microphone locations $\boldsymbol{\theta} = (\boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}})$ and the 3D trajectory $\mathbf{s}$ we must solve the optimization problems (13) and (14). The minimization of (13) does not admit a closed-form solution, so we use the constrainted simultaneous perturbation stochastic approximation (SPSA) algorithm [16], which turned out to be more efficient for this optimization task than gradient descent, quasi-Newton and Newton-Raphson methods, especially for data with high noise levels.

The 3D points $\{\mathbf{s}_n^{(q+1)}\}_{n=1}^N$ are estimated as the minimizers of (14) (one minimization for each point) using the newly estimated microphone locations $\boldsymbol{\theta}^{(q+1)}$. A closed-form solution for $\mathbf{s}^{(q)}$ does not exist, so we perform coordinate-wise optimization of the trajectory. In practice, it is sufficient to update only a certain amount of points at iteration $(q)$ that give highest values of (14). This way the algorithm can be significantly speeded up.

| Scenario | $\sigma$ | $\boldsymbol{\Sigma}$ | Rounded |
|---|---|---|---|
| Noise 1 | 0.05 | $\mathrm{diag}(10^{-6}, 10^{-6}, 10^{-14})$ | no |
| Noise 1, R | 0.05 | $\mathrm{diag}(10^{-6}, 10^{-6}, 10^{-14})$ | yes |
| Noise 2 | 0.1 | $\mathrm{diag}(10^{-4}, 10^{-4}, 10^{-11})$ | no |
| Noise 2, R | 0.1 | $\mathrm{diag}(10^{-4}, 10^{-4}, 10^{-11})$ | yes |
| Noise 3 | 0.5 | $\mathrm{diag}(10^{-2}, 10^{-2}, 10^{-8})$ | no |
| Noise 3, R | 0.5 | $\mathrm{diag}(10^{-2}, 10^{-2}, 10^{-8})$ | yes |

TABLE I
SIMULATED DATA SETS AND THE CORRESPONDING AUDITORY ($\sigma$) AND VISUAL ($\boldsymbol{\Sigma}$) (CO-)VARIANCE VALUES. A VERSION WITH DISCRETIZED (ROUNDED) AUDITORY OBSERVATIONS IS CONSIDERED IN EACH CASE.

## IV. EXPERIMENTS

**Simulated data**. To get ITD values and associated visual disparities at various depths and angles and imitate the natural limits to the visual field of view, we simulated a spiral 3D trajectory of an audiovisual object:

$$\boldsymbol{s}(t) = (30t\cos(3t), 30t\sin(3t), 100t)^\top,\ t \in [5\pi, 9\pi]. \quad (17)$$

Microphones were set to be located at $\boldsymbol{s}^*_{\mathrm{M}_\ell} = (-85, 120, 10)^\top$ and $\boldsymbol{s}^*_{\mathrm{M}_\mathrm{r}} = (75, 110, -15)^\top$ with respect to a camera-centred coordinate frame. The coordinates are given in millimeters, so the inter-microphone distance was about 160mm.

The observations in visual and auditory spaces were produced according to the generative models (5) and (6). Detector failure levels $1-\pi_*$ and $1-\lambda_*$ are both taken to be 0.05. Detector noise is taken normally distributed with covariance matrix $\boldsymbol{\Sigma}$ and variance $\sigma$ for visual and auditory data respectively. Different settings were considered depending on the amount of noise and its nature, they are summarized in Table I. Since auditory observations (ITDs) are sometimes available only in the discretized space of time shifts, we included data sets with rounded auditory observations for each case.

We assume the auditory and visual data to be acquired at 75Hz and 25Hz respectively. This results in about $M = 3000$ video and $K = 9000$ audio *non synchronized* observations. An example of the generated data in auditory and visual domains is shown in the top part of Figure 3.

We ran 100 optimization iterations of the alignment algorithm. The regularization constant $\gamma$ was set to $10^{-3}$ to favour smooth trajectories in the 3D scene space and filter out all abrupt changes that are due to noise. To increase the algorithm speed we performed trajectory optimization (14) for the 100 worst nodes. This did not have any impact on the convergence, though reduced a lot the computational time.

In order to compare the proposed calibration method to some baseline, we considered a "naive" algorithm that reconstructs the 3D trajectory directly from visual observations and estimates microphone locations using (13). Thus no regularization is performed and no explicit component to model noise is added, which results in a faster optimization that performs rough data alignment. We ran 100 optimization iterations for the "naive" algorithm.

A summary on estimated microphone localization and 3D trajectory reconstruction errors for the two algorithms is given in Table II. Very high precision is observed in case of noiseless

| Algorithm | Scenario | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Noiseless | Noise 1 | Noise 1, R | Noise 2 | Noise 2, R | Noise 3 | Noise 3, R | Real 1 | Real 2 | Real 3 |
| Naive | 0 | 0.13 | 0.23 | 0.21 | 0.3 | 3.46 | 3.52 | 0.32 | 0.6 | 0.41 |
| Proposed | 0 | **0.05** | **0.14** | **0.13** | **0.21** | **3.32** | **3.4** | **0.27** | **0.28** | **0.33** |

TABLE III
TRAJECTORY MISALIGNMENT MEASURES FOR THE PROPOSED AND "NAIVE" CALIBRATION ALGORITHMS.

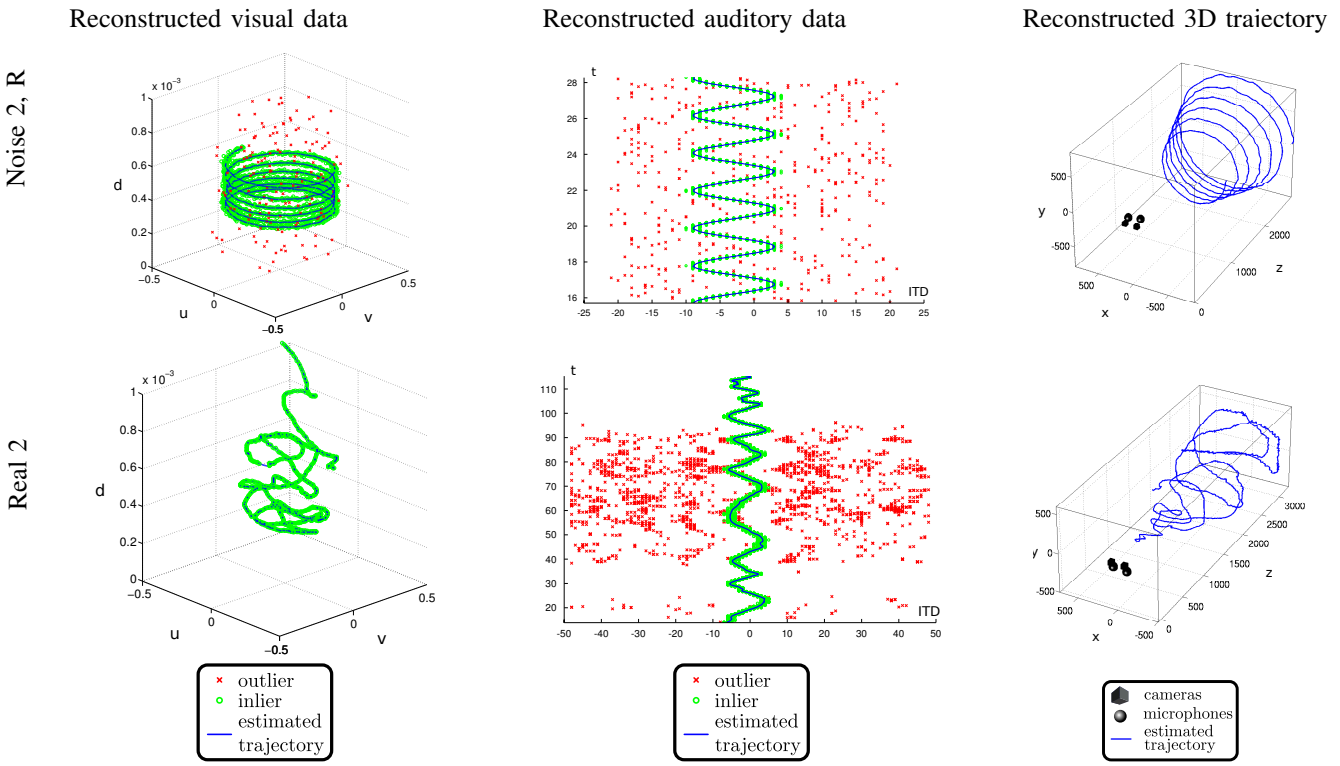



Fig. 3. Auditory and visual spaces alignment results for "Noise 2, R" simulated data experiment (top) and "Real 2" real data experiment (bottom). First two columns show results of classification of visual and auditory observations respectively into inliers (○) and outliers (×). Third column shows camera and microphone locations in the 3D scene and the estimated audio-visual device trajectory. The same trajectory is shown mapped into observations spaces in the first two columns.

| Scenario | Proposed | | | | Naive | | | |
|---|---|---|---|---|---|---|---|---|
| | L | R | A | M | L | R | A | M |
| Noiseless | 1.3 | 1.3 | 0 | 5.9 | **0.7** | **0.8** | 0 | 0 |
| Noise 1 | **19.2** | **19.6** | **2.28** | **27.91** | 223.2 | 224.1 | 87.8 | 7899.9 |
| Noise 1, R | **40.0** | **40.4** | **2.73** | **31.04** | 228 | 230.8 | 96 | 8328.8 |
| Noise 2 | **57.7** | **57.8** | **12.77** | **35.2** | 226.6 | 230.3 | 112.5 | 7830.1 |
| Noise 2, R | **32.6** | **32.7** | **12.65** | **32.27** | 248.2 | 251.8 | 8.3 | 7973.2 |
| Noise 3 | 248.6 | 250.6 | **215.22** | **406.73** | **239.3** | **242.7** | 575.3 | 12013.1 |
| Noise 3, R | **212.8** | **211.5** | **118.11** | **346.98** | 222.8 | 224.6 | 556 | 11192.1 |

TABLE II
ESTIMATED LEFT (L) AND RIGHT (R) MICROPHONE LOCATION ERRORS AND AVERAGE (A) AND MAXIMUM (M) DISTANCES BETWEEN POINTS OF THE ESTIMATED AND GROUND TRUTH TRAJECTORY FOR THE PROPOSED AND "NAIVE" ALGORITHMS IN SIMULATED DATA EXPERIMENTS.

data for both algorithms. Errors are slightly smaller for the "naive" version since no regularization is performed. However, as the amount of noise increases, microphone localization and trajectory reconstruction errors increase rapidly for "naive" approach, while still being reasonable for the proposed approach. Indeed, regularization component and explicit noise modelling in our model result in noise removal and better parameters estimation. We note that noise in visual and auditory modalities has different effect. While visual noise mostly affects 3D trajectory estimation, auditory noise influences the quality of microphone localization, which directly follows from problem formulation in terms of generative models (3) and (4).

Microphone localization errors might seem to be high in noisy conditions. However, this doesn't affect the alignment of auditory and visual observations and can be explained by looking at geometrical and physical properties of the problem. In far field conditions (distances more than 1m from microphones), sensitivity to small displacements (a few cm) of a microphone pair parallel to the plane $\mathcal{M} = \{\boldsymbol{s} \,|\mathcal{G}(\boldsymbol{s}; \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}) = 0\}$ is low in the presence of auditory noise, which can be seen from (4). The sensitivity is lower along the directions that are poorly covered by the observations. In our case natural limits (floor, ceiling) imply less scatter along vertical direction. But even if a microphone pair is not localized precisely, it gives the same quality of observations alignment in far field conditions. For qualitative evaluation we present calibration results for "Noise 2, R" scenario in Figure 3. Though the reported microphone localization errors are about 3 cm, perfect match of visual and auditory observations is achieved at different depths and directions and average trajectory reconstruction error is about 1cm.

For quantitative evaluation of audio-visual alignment, we computed average squared distances from the estimated trajec-

tory points mapped to the auditory observations space $\mathcal{G}(\boldsymbol{s}_k)$ to the corresponding *inlier* observations $g_k$ for both algorithms. The results are presented in Table III. The proposed method clearly outperforms the "naive" one: the reported misalignment measures are small even for noisy data, which shows the benefit of removing outliers and regularizing the trajectory.

**Real data**. The real data experiments were carried out using the audiovisual head-like device shown on Figure 1(a). This device comprises pair of microphones and a pair of stereoscopic cameras with motor-controlled pan, tilt, and vergence movements, Figure 2. It should be emphasized that the acquisitions were made in a normal office room with *no* special arrangements to remove fan noise or reverberations. The ITD values were calculated using the method described in [17]. The 3D visual observations were obtained using standard 3D reconstruction techniques [10] based on matched features in the left and right images.

Three different configurations were considered for narrow (Real 1), medium (Real 2) and large (Real 3) fields of view. This was done by fixing the camera vergence angles on the head-like device. We present input data and the alignment results for "Real 2" setting in Figure 3 (bottom). Since cameras are well-calibrated, noise level in visual observations is low. However, auditory observations obtained in a physically unconstrained environment are significantly contaminated by noise. Nevertheless, our framework succeeds in extracting smooth trajectories based on observations classified as *inliers* by the proposed EM algorithm and rejecting the *outliers*.

Misalignment measures calculated as for the simulated data, are given in Table III. Again, the proposed method clearly outperforms "naive" approach, showing the advantage of explicit noise modelling and considering generative mappings with trajectory regularization.

## V. Discussion

Observation space alignment is a challenging task that is encountered when dealing with integration of multimodal data. Absence of synchronization between the input signals, lack of precision, various types of noise and artifacts, require special methods to be developed and special emitters to be used to produce the data with increased precision and reduced noise.

We presented a framework to align auditory and visual observation spaces, for a device comprising two cameras and two microphones, based on trajectory matching. Our approach uses physically-based generative mappings that relate the unobserved 3D space to the observed spaces and represents the problem as a coordinate system transformation estimation task. This formulation leads to a non-linear optimization problem, that is solved using an EM algorithm. An efficient initialization procedure is proposed, which is based on the geometric properties of the audio and visual generative mappings. This allows to significantly accelerate the optimization.

The performance was evaluated on both simulated and real data against a simple calibration method ("naive" approach). The proposed method achieves accurate alignment of auditory and visual modalities at different depths and directions and clearly outperforms the simple calibration technique, showing the advantage of explicit noise modelling and considering generative mappings with trajectory regularization.

The proposed AV calibration parametrization through microphone locations has several benefits for *active* devices, such as robot heads. Unlike methods that directly map auditory observations to 2D image locations, it operates in 3D and relates 3D locations to auditory observations even for the invisible parts of the scene. AV calibration is kept when performing controlled camera rotations (vergence). This approach may be used for self-calibration of a robot head using controlled head/camera motions in the presence of an AV device.

**Acknowledgments**. This work has been supported by the HUMAVIPS project, grant agreement no. FP7-ICT-247525, funded by the EC 7th Framework Programme.

## References

[1] X. Alameda-Pineda, V. Khalidov, R. P. Horaud, and F. Forbes, "Finding audio-visual events in informal social gatherings," in *Proc. of the 13th Int. Conf. on Multimodal Interfaces*, November 2011, pp. 247–254.

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

[3] J. Vermaak, M. Ganget, A. Blake, and P. Pérez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," in *Proc. of the 8th Int. Conf. on Computer Vision*. IEEE, 2001, pp. 741–746.

[4] P. Perez, J. Vermaak, and A. Blake, "Data fusion for visual tracking with particles," *Proceedings of IEEE*, vol. 92, no. 3, pp. 495–513, 2004.

[5] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud, and R. Horaud, "Detecion and localization of 3D audio-visual objects using unsupervised clustering," in *Proc. of ICMI*, 2008.

[6] T. Hospedales and S. Vijayakumar, "Structure inference for Bayesian multisensory scene understanding," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2140–2157, 2008.

[7] V. Khalidov, F. Forbes, and R. Horaud, "Conjugate mixture models for clustering multimodal data," *Neural Computation*, vol. 23, no. 2, pp. 517–557, 2011.

[8] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Koerner, "A probabilistic model for binaural sound localization," *IEEE Trans. on Systems, Man, and Cybernetics–Part B*, vol. 36, no. 5, pp. 982–994, 2006.

[9] M. Pollefeys and D. Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data," in *IEEE Int. Conf. on Acoustic, Speech, and Signal Processing*, 2008, pp. 2445 –2448.

[10] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, UK: Cambridge University Press, 2003.

[11] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proc. of IEEE Conf. on Acoustics, Speech, and Signal Processing*, 2004, pp. 881–884.

[12] E. Ettinger and Y. Freund, "Coordinate-free calibration of an acoustically driven camera pointing system," in *Second ACM/IEEE International Conference on Distributed Smart Cameras*, sept. 2008, pp. 1 –9.

[13] T. Kuhnapfel, T. Tan, S. Venkatesh, and E. Lehmann, "Calibration of audio-video sensors for multi-modal event indexing," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 2, 2007, pp. 741–744.

[14] Z. Barzelay and Y. Schechner, "Onsets coincidence for cross-modal analysis," *IEEE Trans. on Multimedia*, vol. 12, no. 2, pp. 108–120, 2010.

[15] A. Deleforge and R. P. Horaud, "The cocktail party robot: Sound source separation and localisation with an active binaural head," in *IEEE/ACM Int. Conf. on Human Robot Interaction*, Boston, Mass, March 2012.

[16] J. Spall, *Introduction to Stochastic Searchand Optimization: Estimation, Simulation and Control*. Wiley, 2003.

[17] H. Christensen, N. Ma, S. Wrigley, and J. Barker, "Integrating pitch and localisation cues at a speech fragment level," in *Proc. of Interspeech*, 2007, pp. 2769–2772.