Calibration of A Binocular-Binaural Sensor Using a Moving Audio-Visual Target

Vasil Khalidov, Florence Forbes, Radu Horaud

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Calibration of A Binocular-Binaural Sensor Using a Moving Audio-Visual Target

Vasil Khalidov, Florence Forbes and Radu Horaud

N° 7865

January 2012

Domaine 4

ISSN 0249-6399 ISRN INRIA/RR--7865--FR+ENG

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE | INRIA | centre de recherche GRENOBLE - RHÔNE-ALPES

Calibration of A Binocular-Binaural Sensor Using a Moving Audio-Visual Target

Vasil Khalidov*, Florence Forbes† and Radu Horaud ‡

Domaine : Perception, cognition, interaction

Équipes-Projets MISTIS & PERCEPTION

Rapport de recherche n° 7865 — January 2012 — 27 pages

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 estimating the 3D positions of the microphones in the visual-centred coordinate frame defined by the stereo camera-pair. Assuming that the latter is calibrated, the problem is twofold: estimate the trajectory of an audio-visual (AV) object that freely moves in the scene and estimate the locations of the two microphones. We explore the geometric and physical properties of the two sensorial modalities within two generative models. These models are then used to project the AV object onto both the visual and auditory observation spaces. We exploit the fact that these two distinct data sets are conditioned by a common set of parameters, namely the (unknown) 3D trajectory of the AV object. We derive an EM-like algorithm that alternates between the estimation of the microphone-pair position and the estimation of 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.

Key-words: Audio-visual fusion, stereoscopic vision, binaural hearing, time-difference of arrival, interaural time difference, microphone localization, maximum likelihood, EM algorithm.

This work was supported by the European project HUMAVIPS, under EU grant FP7-ICT-2009-247525.

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

† MISTIS team

‡ PERCEPTION team

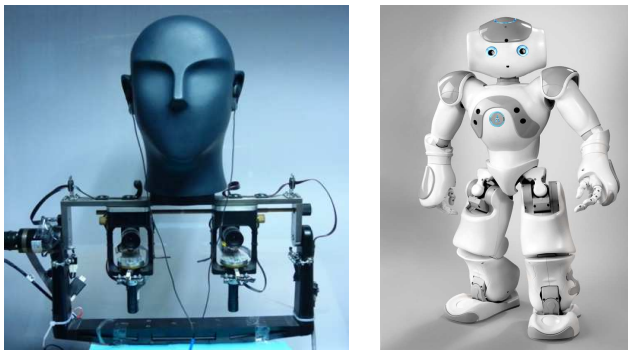
Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Calibration d'un Capteur Binoculaire-Binaural Utilisant Un Objet Audio-Visuel

Résumé : Dans cet article nous abordons le problème de l'alignement visuel (V) et auditif (A) de données en utilisant un capteur qui est composé d'une paire de caméras et une paire de microphones. L'apport original de ce document est une méthode pour estimer les positions 3D des microphones dans le repère défini par la paire stéréo. En supposant que cette dernière est calibrée, le problème est double: estimer la trajectoire d'un objet audio-visuel (AV) qui se déplace librement dans la scène et d'estimer les emplacements des deux microphones. Nous explorons les propriétés géométriques et physiques des deux modalités sensorielles dans deux modèles génératifs. Ces modèles sont ensuite utilisés pour projeter l'objet AV sur les deux espaces d'observation (visuel et auditif). Nous exploitons le fait que ces deux ensembles de données distincts sont conditionnés par un ensemble commun de paramètres, savoir la trajectoire (inconnue) 3D de l'objet AV. Nous dérivons un algorithme de type EM qui alterne entre l'estimation de la position des microphones et l'estimation de la trajectoire de l'objet AV. L'algorithme proposé a un certain nombre de fonctions intégrées: il peut faire face des observations A et V qui sont mal alignées dans le temps, il estime la fiabilité des données, il est robuste aux valeurs aberrantes dans les deux modalités, et sa convergence est prouve d'un point de vue théorique. Nous rapportons des expériences avec des données simulées et réelles.

Mots-clés : Fusion audio-visuelle, vision stéréoscopique, audition binaurale, différence de temps d'arrivée, différence de temps interaurale, localisation de microphone, maximum de vraisemblance, algorithme EM.

Contents



(a) (b)

Figure 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).

1 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], avatar-based interfaces [2, 3], human activity recognition [4, 5], video surveillance [6, 7], sound-source separation [8, 9], multimodal interfaces [10], audio-visual tracking [11, 12], object localization [13], etc. A recent survey of audiovisual fusion methods for human-computer interaction can be found in [14]. Recent neurophysiological, imaging, and cognitive studies showed that multisensory integration is one fundamental brick that helps humans to learn and to understand their complex environment and to disambiguate incomplete single-modality data [15, 16]. 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 [17, 18, 10, 19, 20, 21, 22], meaning that observations from different modalities are fused together as if they were generated from the same spatial source. This leads to a very important question: How to align the 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 [23]. In order to *align* the data

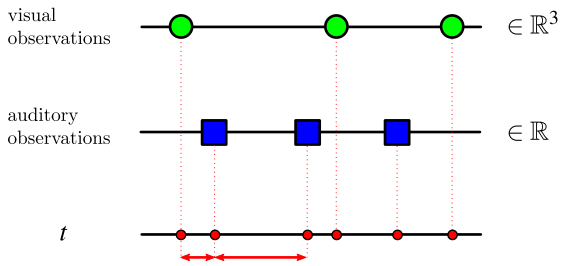


Figure 2: Major difficulties encountered when aligning auditory and visual data: *i)* A and V observations belong to different physical spaces; *ii)* A and V observations are not aligned in time; *iii)* overall sampling rate is not constant; *iv)* A and V data is strongly affected by various kinds of noise and outliers.

gathered with these two different sensorial modalities, one has to address the issue of 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 an extremely difficult problem that has not been properly addressed in the past. Of course, there are calibration methods used in smart rooms, but these methods rely on the use of a large number of microphones which provide precise sound source localization. We do not want to make any a priori assumptions about the geometric and physical (acoustic) properties of the environment, as is often the case with specially arranged spaces. AV objects, e.g., people that emit sounds, can appear at arbitrary 3D positions, projecting to the same location in the image but corresponding to different locations in the auditory space. Moreover, the binocular/binaural sensors that we deal with are active and small camera rotations can lead to significant misalignment between the visual and the auditory data. Therefore, there is a need for a new method to align auditory and visual data gathered with sensors such as the ones shown in Figure 1.

There are several difficulties that need to be tackled when aligning auditory and visual data. First, the auditory (A) and visual (V) observations belong to two different physical spaces that possess different mathematical properties. Second, 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. Third, 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 various kinds of noise and outliers, such as visual objects that do not emit sounds, acoustic reverberations, background noise, etc. The schematic representation of the auditory and visual data streams with a summary of the mentioned difficulties is shown on Figure 2.

1.1 Related Work

Almost any audio-visual fusion method requires some kind of spatial alignment, temporal alignment, or both. Whenever an array of microphones is being used, one can estimate the microphone locations from the time-difference of arrival (TDOA) and matrix factorization [24, 25, 26, 27]. These methods require several synchronized acoustic sources, an anechoic environment, and they can only be applied if the number of microphones is large. Other approaches assume some prior knowledge about the source or microphone locations. For example, [28] proposes a technique to localize a microphone pair with a known distance between the microphones and [29] proposes a method to incrementally localize a microphone relative to a well-calibrated microphone array.

For the purpose of audio-visual calibration, these methods can be used in several ways, such as extending them to a camera-microphone network of sensors [30]. Alternatively, one can perform independently multiple-microphone localization and multiple-camera calibration [31]. 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 [18, 21, 22, 32], 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 they 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 actually 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 [17, 10, 19, 32], or by using a rough estimate of the locations of the microphones relative to the camera [11]. 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 [7]. 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 [33, 8]. 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 [13, 20, 1] or for sound-source separation [34]. Moreover, as already explained, we note that the temporal alignment assumption is not at all realistic.

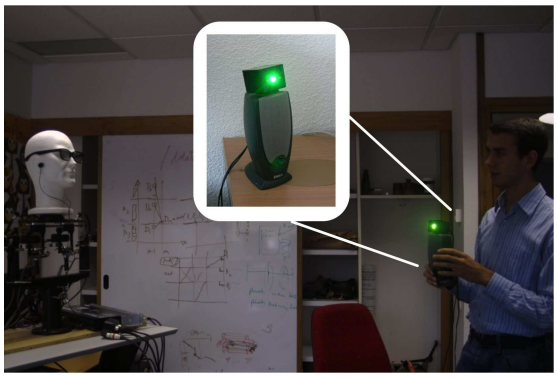


Figure 3: Audio-visual device used to align the auditory and visual spaces. An LED light bulb is mounted onto a speaker which makes the visual localization more precise. White noise is played throughout the recording to improve the auditory localization.

1.2 Contributions

The contribution of this paper is a new method for estimating the position of a binaural set of microphones relative to a stereoscopic camera system. The audiovisual calibration setup is shown in Figure 3. The *audiovisual target* used for calibration consists of a speaker that emits a white-noise acoustic signal and a light source. This target is freely moved in front of the binocular-binaural robot head. Assuming that the stereoscopic camera-pair has been previously calibrated, the task of audio-visual calibration is twofold: (1) estimate the 3D trajectory of the audiovisual target and (2) estimate the 3D locations of the two microphones. We explore the geometric and physical properties of the two sensorial modalities within two coupled generative models. These models are used to map the audiovisual target both onto a *visual space* and onto an *auditory space*. We exploit the fact that these two distinct observation spaces (visual and auditory) are conditioned by a common set of parameters, namely the (unknown) 3D trajectory of the target. 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 2 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 3. Results obtained with both simulated and real data are shown in section 4 and in section 5 respectively. Section 6 concludes the paper along with a short discussion.

2 Observation Space Alignment Through Multimodal Trajectory Matching

Two cameras and two microphones observe a audiovisual target, e.g., Figure 3. This audiovisual target consists of both 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

$$t_{\min} \leq t_1^{\mathrm{f}} < \ldots < t_m^{\mathrm{f}} < \ldots < t_M^{\mathrm{f}} \leq t_{\max}$$

in the visual observation space, and at times

$$t_{\min} \leq t_1^{\mathrm{g}} < \ldots < t_k^{\mathrm{g}} < \ldots < t_K^{\mathrm{g}} \leq t_{\max}$$

in the auditory observation space. This gives rise two to 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:

$$\begin{cases} \mathcal{F} : \mathbb{S} \to \mathbb{F} \\ \mathcal{G} : \mathbb{S} \to \mathbb{G} \end{cases} \tag{3}$$

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 [31], the mapping $\mathcal{F}$ associates a point $\boldsymbol{s} = (x, y, z) \in \mathbb{S}$ from the 3D scene to a *stereoscopic* visual observation $\boldsymbol{f} = (u, v, d)$:

$$\boldsymbol{f} = \mathcal{F}(\boldsymbol{s}) = \left(\frac{x}{z}, \frac{y}{z}, \frac{1}{z}\right)^{\top} \tag{4}$$

  where (u, v) are the 2D coordinates of a left-image point and d is the *horizontal disparity* between two matched points, i.e., $d = u' - u$ and $v' = v$ with (u', v') being the 2D coordinates of a right-image point that is in stereoscopic correspondence with (u, v). Note that the *projective* mapping defined by (4) is one-to-one and is invertible:

$$\boldsymbol{s} = \mathcal{F}^{-1}(\boldsymbol{f}) = \left(\frac{u}{d}, \frac{v}{d}, \frac{1}{d}\right)^{\top} \tag{5}$$

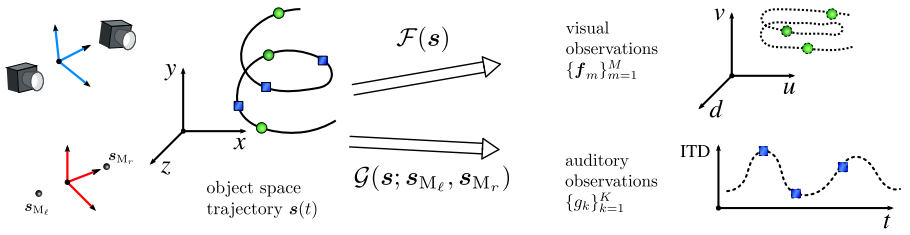


Figure 4: Spatio-temporal approach to the alignment of auditory and visual data. Two cameras and two microphones observe an audio-visual object which moves along an unknown 3D trajectory $\boldsymbol{s}(t)$. Auditory (g_k) and visual ($\boldsymbol{f}_m$) observations are generated by the mappings $\mathcal{G}$ and $\mathcal{F}$. These mappings are defined by the corresponding acoustic and geometric models. Alignment is achieved by finding the relation between the microphone-centred and the camera-centred coordinate frames. Unlike many other audio-visual calibration methods, this alignment is achieved in the 3D space.

- 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, which is the time difference of arrival, also called the *interaural time difference* (ITD) of a sound emitted from $\boldsymbol{s}$ and perceived by the left and right microphones:

$$g = \mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}) = c^{-1}\Big(\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\ell}\| - \|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}\|\Big) \tag{6}$$

  where c 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. It is important to notice that, unlike the binocular visual model, the binaural mapping $\mathcal{G}$ is not injective: There exists a conic surface embedded in the 3D space that is associated to an auditory observation g.

The generative models for auditory and visual observations are depicted in Figure 4.

The problem of aligning auditory and visual observations is conditioned by the mappings $\mathcal{F}$ and $\mathcal{G}$ which must act in the same system of coordinates. If the stereo camera-pair is calibrated and rectified, (4) allows to recover the 3D position of a scene point being viewed by both cameras in a camera-centered coordinate frame. Therefore, *audiovisual calibration* consists in estimating the microphone locations $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$ in this frame. Techniques for stereo calibration are extremely well understood, both from a methodological and practical points of view. Therefore, we assume that the stereo camera-pair has been previously calibrated, hence 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 conditional probabilities which are convex combinations of a normal and an uniform distribution :

$$P(\boldsymbol{f}_m|\boldsymbol{s}_m) = \mu\mathcal{N}(\boldsymbol{f}_m|\mathcal{F}(\boldsymbol{s}_m), \boldsymbol{\Sigma}) + (1-\mu)\mathcal{U}(V) \tag{7}$$

$$P(g_k|\boldsymbol{s}_k) = \lambda\mathcal{N}(g_k|\mathcal{G}(\boldsymbol{s}_k), \sigma) + (1-\lambda)\mathcal{U}(U) \tag{8}$$

where:

$$\mathcal{N}(\boldsymbol{f}_m|\mathcal{F}(\boldsymbol{s}_m), \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{3/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}\|(\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_m)\|^2_{\boldsymbol{\Sigma}}} \tag{9}$$

$$\mathcal{N}(g_k|\mathcal{G}(\boldsymbol{s}_k), \sigma) = \frac{1}{(2\pi\sigma)^{1/2}} e^{-\frac{1}{2\sigma^2}(g_k - \mathcal{G}(\boldsymbol{s}_k))^2} \tag{10}$$

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:

$$P(\mathbf{F}|\boldsymbol{s}) = \prod_{m=1}^{M} P(\boldsymbol{f}_m|\boldsymbol{s}_m) \tag{11}$$

$$P(\boldsymbol{G}|\boldsymbol{s}) = \prod_{k=1}^{K} P(g_k|\boldsymbol{s}_k) \tag{12}$$

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

$$P(\mathbf{s}) \propto \exp\left(-\gamma \sum_{n=1}^{N-1} \frac{\|\boldsymbol{s}_{n+1} - \boldsymbol{s}_n\|^2}{t_{n+1} - t_n}\right) \tag{13}$$
$$\mathbf{s} = \{\boldsymbol{s}_n\}_{n=1}^N, \quad \boldsymbol{s}_n = \boldsymbol{s}(t_n) \in \mathbb{S} \subset \mathbb{R}^3$$

where $\gamma > 0$ is a regularization scalar and the time-stamp set $\{t_n\}_{n=1}^N$ is taken as an ordered union:

$$\{t_n\}_{n=1}^N = \{t_m^{\mathrm{f}}\}_{m=1}^M \cup \{t_k^{\mathrm{g}}\}_{k=1}^K$$

Hence, $N \leq M + K$ since the auditory and visual time-stamps t_m^{f} and t_k^{g} may coincide for some m and k.

The alignment problem is then formulated as the simultaneous inference:

- The estimation of the unknown 3D trajectory $\boldsymbol{s}(t)$, and
- The estimation of 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 maximization of the following log-likelihood function:

$$\{\mathbf{s}^*, \boldsymbol{\theta}^*, \boldsymbol{\psi}^*\} = \underset{\mathbf{s}\in\mathbb{S}^N,\, \boldsymbol{\theta}\in\Theta,\, \boldsymbol{\psi}\in\boldsymbol{\Psi}}{\operatorname{argmax}} \log P(\mathbf{F}, \boldsymbol{G}, \mathbf{s}, \boldsymbol{\theta}\,;\, \boldsymbol{\psi}) \tag{14}$$

with:

$$\begin{aligned}\log P(\mathbf{F}, \boldsymbol{G}, \mathbf{s}, \boldsymbol{\theta}; \boldsymbol{\psi}) = \log P(\mathbf{F}|\mathbf{s}, \boldsymbol{\theta}; \boldsymbol{\psi}) + \log P(\boldsymbol{G}|\mathbf{s}, \boldsymbol{\theta}; \boldsymbol{\psi}) \\ + \log P(\mathbf{s}) + \log P(\boldsymbol{\theta})\end{aligned} \tag{15}$$

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 (7) and (8), and the trajectory likelihood $P(\mathbf{s})$ is given by (13). Microphone locations $\boldsymbol{\theta}$ are assumed to be uniformly distributed over some compact set $\Theta \subset \mathbb{R}^6$:

$$P(\boldsymbol{\theta}) = \mathcal{U}(\Theta) \tag{16}$$

3 Simultaneous Microphone Localization and Trajectory Reconstruction

Formally, (14) 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, (14) is replaced with the *expected complete-data log-likelihood* maximization within the EM algorithm. The maximization of the expected complete-data log-likelihood monotonically increases (14). Therefore, we adopt an alternating optimization approach for the trajectory reconstruction and microphone localization problems: At iteration $q+1$ the microphones' locations $\boldsymbol{\theta}^{(q+1)}$ are estimated using a current estimation of the trajectory $\mathbf{s}^{(q)}$. Next, the trajectory estimation $\mathbf{s}^{(q+1)}$ is updated using the new microphones' locations $\boldsymbol{\theta}^{(q+1)}$.

The expected complete-data log-likelihood to be maximized is:

$$\begin{aligned}Q(\boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{s}^{(q)}, \boldsymbol{\theta}^{(q)}, \boldsymbol{\psi}^{(q)}) = \\ E_{\boldsymbol{A},\boldsymbol{B}}\left[\log P(\mathbf{F}, \boldsymbol{G}, \boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{A}, \boldsymbol{B}; \boldsymbol{\psi})|\mathbf{F}, \boldsymbol{G}, \boldsymbol{s}^{(q)}, \boldsymbol{\theta}^{(q)}; \boldsymbol{\psi}^{(q)}\right]\end{aligned} \tag{17}$$

The expectation is taken over the hidden variables $\boldsymbol{A}$ and $\boldsymbol{B}$. After computing this expectation, we obtain:

$$\begin{aligned} &Q(\boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{s}^{(q)}, \boldsymbol{\theta}^{(q)}, \boldsymbol{\psi}^{(q)}) = \\ &-\sum_{m=1}^{M} \alpha_m^{(q)} \left(\|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_m)\|^2_{\boldsymbol{\Sigma}} + \log|\boldsymbol{\Sigma}| - \log\tfrac{\mu}{1-\mu} \right) \\ &-\sum_{k=1}^{K} \beta_k^{(q)} \left((g_k - \mathcal{G}(\boldsymbol{s}_k;\boldsymbol{\theta}))^2 + 2\log\sigma - \log\tfrac{\lambda}{1-\lambda} \right) \\ &+M\log(1-\mu) + K\log(1-\lambda) - \gamma \sum_{n=1}^{N-1} \tfrac{\|\boldsymbol{s}_{n+1} - \boldsymbol{s}_n\|^2}{t_{n+1}-t_n} \\ &+\log P(\boldsymbol{\theta}), \end{aligned} \tag{18}$$

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

$$\alpha_m^{(q)} = \frac{\mu^{(q)}\mathcal{N}(\boldsymbol{f}_m|\mathcal{F}(\boldsymbol{s}_m^{(q)}), \boldsymbol{\Sigma}^{(q)})}{\mu^{(q)}\mathcal{N}(\boldsymbol{f}_m|\mathcal{F}(\boldsymbol{s}_m^{(q)}), \boldsymbol{\Sigma}^{(q)}) + (1-\mu^{(q)})\mathcal{U}(V)} \tag{19}$$

$$\beta_k^{(q)} = \frac{\lambda^{(q)}\mathcal{N}(g_k|\mathcal{G}(\boldsymbol{s}_k^{(q)};\boldsymbol{\theta}^{(q)}), \sigma^{(q)})}{\lambda^{(q)}\mathcal{N}(g_k|\mathcal{G}(\boldsymbol{s}_k^{(q)};\boldsymbol{\theta}^{(q)}), \sigma^{(q)}) + (1-\lambda^{(q)})\mathcal{U}(U)} \tag{20}$$

3.1 The proposed EM algorithm

The optimization of (18) 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 (19) and (20), the M-step is more difficult to achieve because of the presence of the visual and auditory mappings introduced in (3), and explicitly defined by (4) and by (6). Hence, the M-step of the algorithm should be further decomposed into a number of conditional maximization steps, which are described in detail below:

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

$$\boldsymbol{\theta}^{(q+1)} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \left(\sum_{k=1}^{K} \beta_k^{(q)} \left(g_k - \mathcal{G}(\boldsymbol{s}_k^{(q)};\boldsymbol{\theta}) \right)^2 - \log P(\boldsymbol{\theta}) \right) \tag{21}$$

2. Each 3D trajectory point $\boldsymbol{s}_n, 2 \leq n \leq N-1$ is estimated with:

$$\begin{aligned} \boldsymbol{s}_n^{(q+1)} = \underset{\boldsymbol{s}_n\in\mathbb{S}}{\operatorname{argmin}} \Bigg(&\gamma \left(\frac{\|\boldsymbol{s}_{n+1}^{(q)} - \boldsymbol{s}_n\|^2}{t_{n+1}-t_n} + \frac{\|\boldsymbol{s}_n - \boldsymbol{s}_{n-1}^{(q)}\|^2}{t_n - t_{n-1}} \right) \\ &+ \delta_n^{\mathrm{f}} \alpha_m^{(q)} \|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n)\|^2_{\boldsymbol{\Sigma}^{(q)}} \\ &+ \delta_n^{\mathrm{g}} \beta_k^{(q)} \frac{(g_k - \mathcal{G}(\boldsymbol{s}_n;\, \boldsymbol{\theta}^{(q+1)}))^2}{\sigma^{(q)}} \Bigg) \end{aligned} \tag{22}$$

where δ_n^{f} and δ_n^{g} are defined as

$$\delta_n^{\mathrm{f}} = \begin{cases} 1, & \text{if } \exists m : \boldsymbol{f}_m \text{ is observed at } t_n, \\ 0, & \text{otherwise,} \end{cases} \tag{23}$$

$$\delta_n^{\mathrm{g}} = \begin{cases} 1, & \text{if } \exists k : g_k \text{ is observed at } t_n, \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

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

$$\mu^{(q+1)} = \frac{1}{M} \sum_{m=1}^{M} \alpha_m^{(q)} \tag{25}$$

$$\lambda^{(q+1)} = \frac{1}{K} \sum_{k=1}^{K} \beta_k^{(q)} \tag{26}$$

$$\boldsymbol{\Sigma}^{(q+1)} = \frac{\sum\limits_{m=1}^{M} \alpha_m^{(q)} \left(\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_m^{(q+1)})\right) \left(\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_m^{(q+1)})\right)^\top}{\sum_{m=1}^{M} \alpha_m^{(q)}} \tag{27}$$

$$\sigma^{2(q+1)} = \frac{\sum_{k=1}^{K} \beta_k^{(q)} \left(g_k - \mathcal{G}(\boldsymbol{s}_k^{(q+1)}; \boldsymbol{\theta}^{(q+1)})\right)^2}{\sum_{k=1}^{K} \beta_k^{(q)}} \tag{28}$$

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 (27) and (28) correspond to the same calculated 3D trajectory $\mathbf{s}^{(q+1)}$ mapped into the visual and auditory observation spaces $\mathbb{F}$ and $\mathbb{G}$.

Figure 5 provides an outline of the proposed EM algorithm. Below we give details on the initialization and optimization procedures used in practice.

3.2 Initialization

It is well known that the initialization procedure 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 $\boldsymbol{\psi}^{(0)}$ will reduce the number of iterations needed by the algorithm and hence will reduce the overall elapsed time to find the optimal values. Proper initialization will also help 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 $\mathcal{G}$, i.e., (4) and (6) 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 $\{t_k\}_{k=1}^K$.

input : Auditory observations $\{g_k\}_{k=1}^K$ with timestamps $\{t_k^{\mathrm{g}}\}_{k=1}^K$ and visual observations $\{\boldsymbol{f}_m\}_{m=1}^M$ with timestamps $\{t_m^{\mathrm{f}}\}_{m=1}^M$;
output : 3D microphone locations $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$ and audiovisual target trajectory $\boldsymbol{s}(t)$;

1: Initialize $\boldsymbol{\theta}^{(0)}$, $\mathbf{s}^{(0)}$ and $\boldsymbol{\psi}^{(0)}$ using the procedure described in Section 3.2;
2: **E-step**. Update the posterior probabilities using (19) and (20);
3: **CM-step-1**. Update the microphone locations $\boldsymbol{\theta}^{(q+1)} = \boldsymbol{s}_{\mathrm{M}_\ell}^{(q+1)}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}^{(q+1)}$ by minimizing (21);
4: **CM-step-2**. Update the 3D trajectory $\mathbf{s}^{(q+1)}$ through the minimization of (22);
5: **CM-step-3**. Update the mixtures' parameters through (25)-(28).
6: Terminate on convergence, otherwise $q \leftarrow q+1$ and go to Step 2.

Figure 5: An EM algorithm for simultaneous microphone localization and trajectory reconstruction. The standard M-step is replaced by three conditional maximization (CM) steps.

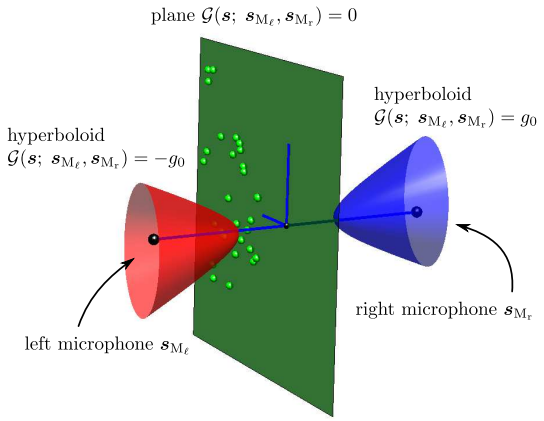


Figure 6: Geometric properties of the auditory mapping $\mathcal{G}$. Isosurfaces $\mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}) = g_0$ are one-sheet hyperboloids for $g_0 \neq 0$. Isosurface $\mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}) = 0$ is a plane perpendicular to the line connecting the left and right microphones. When the microphone locations are not known, this plane can be reconstructed by the observations that have ITD values equal to zero (displayed as bright dots on the surface).

- The initial microphone locations $\boldsymbol{\theta}^{(0)} = \{\boldsymbol{s}_{\mathrm{M}_\ell}^{(0)}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}^{(0)}\}$ are calculated as follows. First, notice that $g = 0$ in (6) corresponds to the plane $\mathcal{M} =$

$\{\boldsymbol{s} \mid \mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}) = 0\}$, i.e, a plane orthogonal to the line-segment joining the two microphones and located at the midpoint of this line-segment (see Figure 6). Suppose that the audio-visual calibration sequence was recorded so that there are observations of the audiovisual target that lie in this plane, they correspond to auditory observations g_k, i.e., interaural time differences, equal to zero. The associated 3D positions $\boldsymbol{s}_k$ can be estimated by stereoscopic triangulation and interpolation as just explained. We gather a number of such 3D positions such that $g_k = 0$ (or at least $|g_k| < \varepsilon$ for some small value ε) to fit a plane $\mathcal{M}$. This plane contains the middle point between the two microphones $\boldsymbol{s}_\mathrm{M} = (\boldsymbol{s}_{\mathrm{M}_\ell} + \boldsymbol{s}_{\mathrm{M}_\mathrm{r}})/2$ and is perpendicular to the line connecting the two microphone locations $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$. Therefore, we initialize the middle point $\boldsymbol{s}_\mathrm{M}$ on the plane $\mathcal{M}$ and choose $\boldsymbol{s}^{(0)}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}^{(0)}_{\mathrm{M}_\mathrm{r}}$ to be symmetric with respect to $\mathcal{M}$ and such that $\boldsymbol{\theta}^{(0)} = \{\boldsymbol{s}^{(0)}_{\mathrm{M}_\ell}, \boldsymbol{s}^{(0)}_{\mathrm{M}_\mathrm{r}}\}$ lies within the compact support Θ. 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, see (6) and Figure 6.

- 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 (7) and (8) are defined based on setup specifications. The size of the auditory domain is defined by the maximum values of ITDs that can be observed, while the visual domain size depends on the parameters associated with the stereoscopic calibration and on the observed scene limits.

3.3 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 (21) and (22). The minimization of (21) does not admit a closed-form solution, so we use the simultaneous perturbation stochastic approximation (SPSA) algorithm proposed in [35]. This algorithm is an iterative zero-order optimization method that works well in case of noisy data that can lead to degeneracies in higher-order methods. SPSA is known to combine both local and global convergence properties, while keeping optimization iterations efficient. In practice, SPSA turned out to be much more efficient for this optimization task, than gradient descent, quasi-Newton and Newton-Raphson methods, especially for data with high noise levels.

Various choices can be made for the prior distribution on microphone locations in (21), depending on what kind of information one possesses about the setup and how precise it is. In our experiments we use a uniform distibution on the compact parameter space Θ.

The 3D points $\{\mathbf{s}_n^{(q+1)}\}_{n=1}^N$ are estimated as the minimizers of (22) (one minimization must be carried out 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. As auditory and visual observations can both be available for certain trajectory points $\boldsymbol{s}_n$, we make use of the efficient conjugate M-step proposed in [20] for such cases. In practice, it is sufficient to update only a certain amount of points at iteration (q) that give highest values of (22). This way the algorithm can be significantly speeded up.

4 Experiments with Simulated Data

We evaluated the performance of our algorithm on simulated data. A spiral 3D trajectory of audiovisual object was simulated using:

$$\boldsymbol{s}(t) = (30t\cos(3t), 30t\sin(3t), 100t)^\top \tag{29}$$

where $t \in [5\pi, 9\pi]$. This trajectory was chosen to get the ITD values and associated visual disparities at various depths and angles. We imitated the natural limits to the visual field of view which restricts visual observations to lie within a fixed conic volume. 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 (7) and (8). Detector failure levels $1 - \pi_*$ and $1 - \lambda_*$ are taken to be equal to 0.05 for both modalities. Detector noise is taken normally distributed with covariance matrix $\boldsymbol{\Sigma}$ and variance σ for visual and auditory data respectively. Different settings were considered depending on the amount of noise and its nature. The summary of the corresponding values of σ and $\boldsymbol{\Sigma}$ is given in Table 1. As soon as auditory observations (ITDs) are often 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 different frequencies: 25Hz for video and 75Hz for audio. This results in total of about *non synchronized* $M = 3000$ video observations and $K = 9000$ audio observations. Some examples of the generated data in auditory and visual domains are shown in Figure 7.

We ran 100 optimization iterations of the alignment algorithm with the regularization constant set to $\gamma = 100$. This choice corresponds to smooth trajectories in the 3D scene space and filters out all the abrupt changes that are due to noise. We chose Aitken criterion [36] with tolerance 0.01 to stop the microphone optimization iteration (21). To increase the algorithm speed we

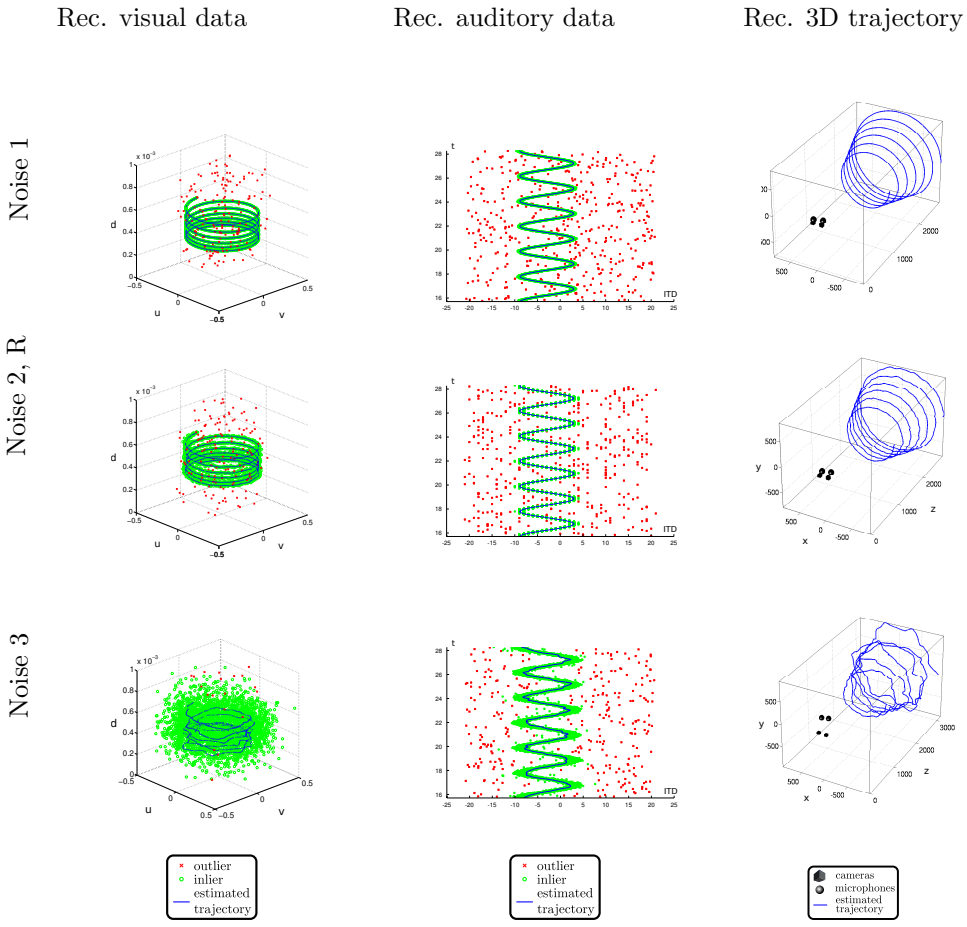


Figure 7: Auditory and visual spaces alignment results for simulated data experiments for different setting of noise levels. First two columns show results of classification of visual and auditory observations respectively into inliers (∘) and outliers (×). Third column shows camera and microphone locations 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.

Name	σ	$\mathbf{\Sigma}$	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 1: Simulated data sets and the corresponding auditory (σ) and visual ($\mathbf{\Sigma}$) (co-)variance values. A version with discretized (rounded) auditory observations is considered in each case.

performed trajectory regularization using visual data only which has a closed form solution, followed by optimization (22) only for the 100 worst nodes. This did not have any impact on the convergence speed, though reduced a lot the computational time.

The comparative update efficiency is illustrated with the microphone distance evolution curve in Figure 8. Curve lengths correspond to the total number of microphone optimization iterations performed during all optimization iterations. The lengths are different because of the Aitken stopping criterion employed to detect the algorithm convergence. Each curve represents the evolution of distances $\|\hat{\boldsymbol{s}}_{\mathrm{M}_\ell} - \boldsymbol{s}^*_{\mathrm{M}_\ell}\| + \|\hat{\boldsymbol{s}}_{\mathrm{M}_\mathrm{r}} - \boldsymbol{s}^*_{\mathrm{M}_\mathrm{r}}\|$ from the estimated microphone locations to the ground truth values. Solid line without markers shows algorithm performance in the case of "ideal" observations, i.e. observations without noise. Even provided the initial microphone locations were very approximate (about 25 cm away from the ground truth location for each microphone), the estimates in this case gradually converge to their optimal values and the final estimate for each microphone is less than 1.5 mm away from the ground truth location. The more the amount of noise increases, the more iterations are required for the algorithm to converge and the less accurate the estimates become. Finally, in the case of the Noise 3 setting we notice that the algorithm fails to identify microphone positions and the approximate estimation error is about 20cm for each of the microphones. However, experiments with real data show that observation (co-)variances typically correspond to the Noise 1 case, where the algorithm shows good convergence.

A summary on estimated microphone positions in simulated data experiments is given in Table 2. Very high precision (less than 1.5 mm in each coordinate) is observed in case of noiseless data. As the amount of noise increases (Noise 1 data sets), we notice error increase in y axis direction. This can be explained from the geometry of the problem (see Figure 6). All the observations lie in the field of view which is defined by the camera pair. Taking into account the fact that for each auditory space value there exists a whole surface of 3D points that are mapped to this value, the error function (21) becomes less sensitive to microphone position changes in certain directions. This also explains why

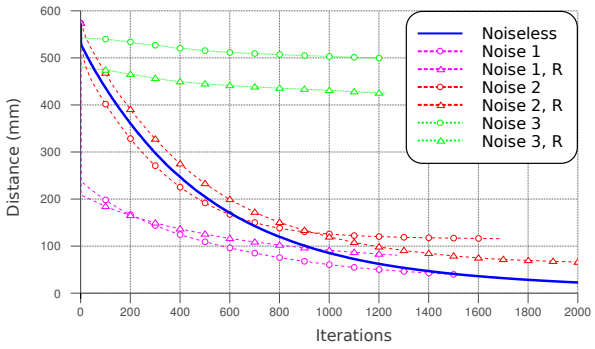


Figure 8: Microphone distance evolution curves for the simulated data sets. Iterations are stopped based on Aitken criterion.

Name	Left microphone	Right microphone
Ground truth	−85, 120, 10	75, 110, −15
Noiseless	−84.9, 121.3, 10.2	75.1, 111.3, −14.8
Noise 1	−82.2, 138.3, 15	77.7, 128.7, −9.9
Noise 1, R	−85.7, 160, 9	75.7, 150.4, −16
Noise 2	−83.5, 134, 66	72.5, 124.4, 40.9
Noise 2, R	−83.1, 138.5, 36.8	76, 128.9, 11.7
Noise 3	−57.6, 344.8, 112.6	77.3, 338.7, 87.4
Noise 3, R	−23.7, 296.7, 111.5	112.7, 290.6, 88.5

Table 2: Ground truth ($\boldsymbol{s}^*_{\mathrm{M}_\ell}$, $\boldsymbol{s}^*_{\mathrm{M}_\mathrm{r}}$) and estimated ($\hat{\boldsymbol{s}}_{\mathrm{M}_\ell}$, $\hat{\boldsymbol{s}}_{\mathrm{M}_\mathrm{r}}$) microphone positions (in millimetres) for simulated data experiments.

Name	Average distance (mm)	Max distance (mm)
Noise 1	2.28	27.91
Noise 1, R	2.73	31.04
Noise 2	12.77	35.2
Noise 2, R	12.65	32.27
Noise 3	215.22	406.73
Noise 3, R	118.11	346.98

Table 3: Average and maximum distance between points of the ground truth and estimated trajectories (in millimetres) for simulated data experiments.

microphone position optimization is so hard to achieve: there exists a whole set of positions that form a ridge, for which the likelihood function values are very close to the optimal one. Further increase of the amount of noise leads to error increase in z axis direction (Noise 2 data sets) and finally, in x axis direction (Noise 3 data sets).

We note that in our real data experiments the observed noise levels correspond to the Noise 1 data set. One possibility to improve the results for the cases where noise levels are higher would be to increase the field of view and thus obtain more correspondences between auditory and visual trajectories. This could be performed, for example, using motors to turn cameras to explore different parts of the space.

To qualitatively assess the algorithm, we included the alignment results for some of the simulated data sets in Figure 7. We note that the observations are significantly contaminated by noise, which prevents normal regularization-based methods to perform well. Our method removes the outliers from the the two data sets, succeeds in aligning the unsynchronized auditory and visual data and reconstructs the trajectory that nicely matches data in both observation spaces.

Average and maximum distances between points in the ground truth and estimated trajectories are given in Table 3. As the amount of noise increases, the deviations become more significant. At the same time we notice that discretization has certain effect of regularization in cases of high noise levels.

5 Experiments with 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 3. It should be emphasized that the acquisitions were made in a normal office room with *no* special arrangements to remove fan noise or reverberations.

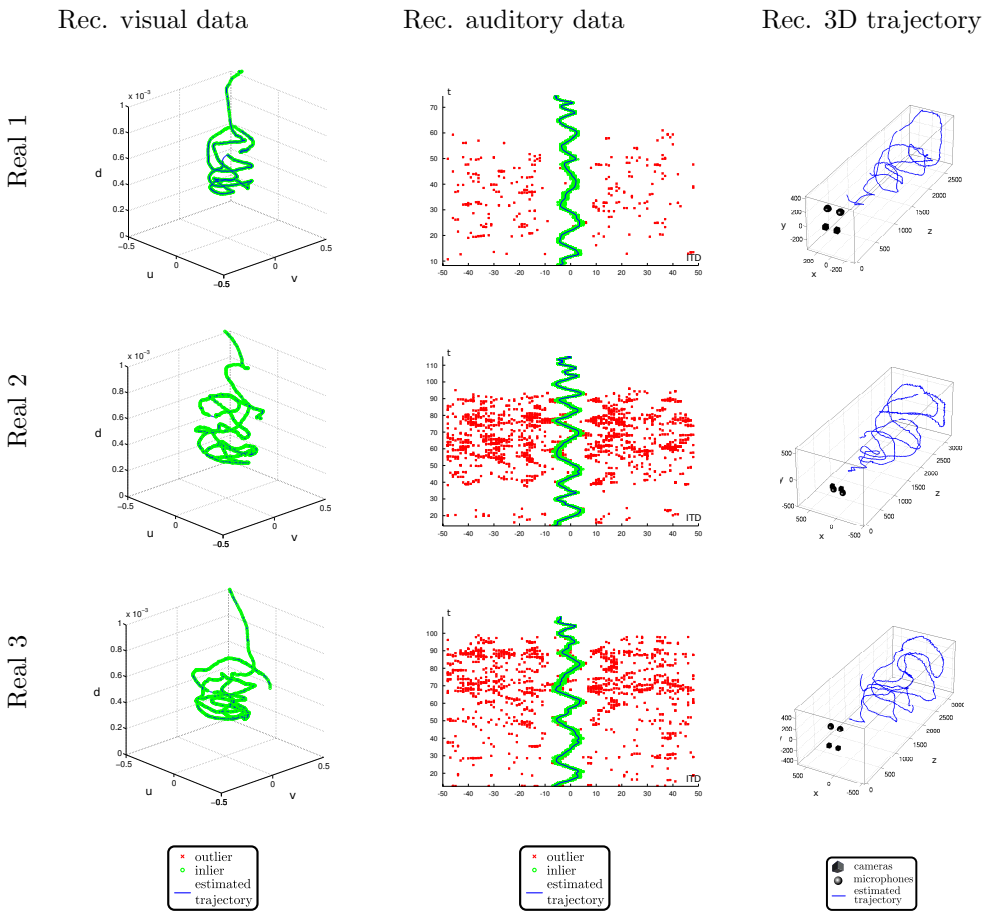


Figure 9: Auditory and visual spaces alignment results for real data experiments for different setting of the field of view (FOV). First two columns show results of classification of visual and auditory observations respectively into inliers (○) and outliers (×). Third column shows camera and microphone 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.

The auditory and visual observations were gathered using the following techniques. The ITD values were calculated using the method described in [37]. First, left and right microphone signals are processed by a filter bank that separates them into different frequency bands. Second, a cross-correlogram is computed for every frequency band, the results are integrated and analyzed to obtain an ITD value. The 3D visual observations were obtained using standard 3D reconstruction techniques [31] based on matched features in the left and right images. Examples of auditory and visual data are shown in Figure 9.

Name	Left microphone	Right microphone
Ground truth	−96.5, 215.2, 0.86	91.5, 204.7, −0.15
Real 1	−109.9, 313.6, 15.2	65.7, 313, 15.7
Real 2	−97.1, −5.8, −5.2	70.9, 5.8, −2.9
Real 3	−99.8, 399, 16.6	66.8, 401, 19.3

Table 4: Ground truth ($\boldsymbol{s}^*_{\mathrm{M}_\ell}$, $\boldsymbol{s}^*_{\mathrm{M}_\mathrm{r}}$) and estimated ($\hat{\boldsymbol{s}}_{\mathrm{M}_\ell}$, $\hat{\boldsymbol{s}}_{\mathrm{M}_\mathrm{r}}$) microphone positions (in millimetres) for real data experiments.

We note that the observations obtained in physically unconstrained environments are significantly contaminated by noise, which prevents normal regularization-based methods to perform well. Nevertheless, our framework allows to extract smooth trajectories based on observations classified as *inliers* by the proposed EM algorithm: The inliers are shown with *green circles* ($\circ$) in Figure 9 while the observations classified as *outliers* are rejected. Outliers are shown with *red crosses* ($\times$).

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. The real-data ground truth on microphone positions was evaluated using an additional stereo camera pair and special markers attached to the sensors. The real-data results resemble qualitatively to those obtained with the simulated data experiments in low-noise cases (see Figure 9). The image of the estimated 3D scene trajectory fits nicely the data in both visual and auditory domains, which means that our method provides a very good observation space alignment. In particular, we note that high noise levels in the auditory modality did not influence our method and the alignment in Real 2 and Real 3 scenarios is as good as in the Real 1 scenario. However, the estimated 3D positions of the microphones are slightly different than the ground-truth ones. A summary on the estimated positions is given in Table 4. One may notice that significant deviations occur in the$y$-coordinate.

The likelihood maximization implies high precision for spatio-temporal observation alignment, but failed to ensure microphone localization error decrease; the microphone distance evolution curves for the three scenarios are depicted in Figure 10. These results differ from the simulated scenario results shown in Figure 8. One reason for such difference is that the actual field of view in real experiments happened to be narrower than in simulated experiments (see ITD domains in Figures 7 and 9). Thus, a good strategy to overcome this issue would be to use camera motors to follow the calibration device without moving the head, or alternatively, rotate the head and keep the calibration device still. In both cases larger areas of space would be covered leading to better microphone localization, while the calibration procedure could be made completely automated.

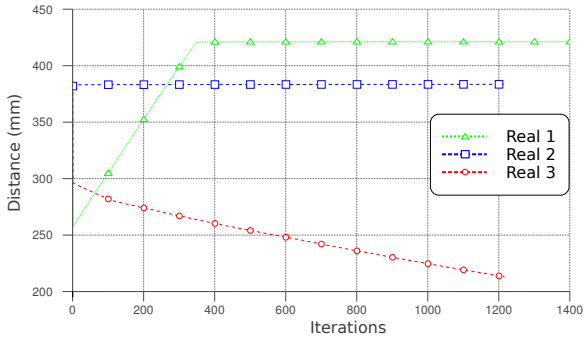


Figure 10: Microphone distance evolution curves for the real data sets. Iterations are stopped based on Aitken criterion.

6 Discussion

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

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 a recently developed EM algorithm [20]. An efficient initialization procedure is proposed as well, 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. Simulated data results showed how important was the choice of the 3D trajectory and thus the size of the field of view. These findings were supported by real data experiments, for which we constructed an audio-visual device aimed to increase the precision and to reduce the noise in the data.

Since the size of the field of view has a great impact on the observation space alignment, one way to improve the results would be to develop an autonomous alignment algorithm that uses controlled motions to create more sophisticated trajectories that cover more of the 3D scene space. This can potentially decrease the estimation error in the y coordinate and is a firm ground to build audio-visual self-calibration algorithms.

References

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

[2] A. Brooks, "Coordinating human-robot communication," Ph.D. dissertation, MIT, 2007.

[3] Y. Fu, R. Li, T. Huang, and M. Danielsen, "Real-time multimodal human-avatar interaction," *Trans. on Cir.Sys.Video*, vol. 18, no. 4, pp. 467–477, 2008.

[4] W. Feng, L. Xie, J. Zeng, and L. Zhi-Qiang, "Audio-visual human recognition using semi-supervised spectral learning and hidden markov models," *J. of Vis. Lang. and Comp.*, no. 20, pp. 188–195, 2009.

[5] S. Petridis and M. Pantic, "Audiovisual discrimination between speech and laughter: Why and when visual information might help," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 216 –234, april 2011.

[6] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 257–267, 2007.

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

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

[9] A. Llagostera Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *IEEE Transactions on Multimedia*, vol. 12, no. 5, pp. 358 –371, aug. 2010.

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

[11] J. Vermaak, M. Ganget, A. Blake, and P. Pérez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," in *Proceedings of the Eighth International Conference on Computer Vision*. IEEE, 2001, pp. 741–746.

[12] 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.

[13] 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.

[14] S. Shivappa, M. Trivedi, and B. Rao, "Audiovisual information fusion in human computer interfaces and intelligent environments: A survey," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1692 –1715, oct. 2010.

[15] B. Stein and T. Stanford, "Multisensory integration: current issues from the perspective of the single neuron," *Nature Reviews Neuroscience*, vol. 9, pp. 255–266, 2008.

[16] A. J. King, "Visual influences on auditory spatial learning," *Phil. Trans. R. Soc. B*, vol. 364, pp. 331–339, 2009.

[17] M. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 828–836, 2003. [Online]. Available: citeseer.ist.psu.edu/beal03graphical.html

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

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

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

[21] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proc. of ICMI*. New York, NY, USA: ACM, 2005, pp. 61–68.

[22] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1154–1164, 2002.

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

[24] V. Raykar and R. Duraiswami, "Automatic position calibration of multiple microphones," in *Proc. of ICASSP*, 2004, pp. 69–72.

[25] S. Birchfield and A. Subramanya, "Microphone array position calibration by basis-point classical multidimensional scaling," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1025 – 1034, 2005.

[26] S. Thrun, "Affine structure from sound," in *Proceedings of Conference on Neural Information Processing Systems (NIPS)*. Cambridge, MA: MIT Press, 2005.

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

[28] P. Aarabi, "Self-localizing dynamic microphone arrays," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 32, no. 4, pp. 474–484, 2002.

[29] J. Chen, R. Hudson, and K. Yao, "Maximum likelihood source localization and unknown sensor location estimation for wide-band signals in the near-field," *IEEE Transactions on Signal Processing*, vol. 50, no. 8, pp. 1843 – 1854, 2002.

[30] A. ODonovan, R. Duraiswami, and J. Neumann, "Microphone arrays as generalized cameras for integrated audio visual processing," in *Proceedings of Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

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

[32] 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.

[33] Z. Barzelay and Y. Schechner, "Harmony in motion," in *CVPR*, 2007.

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

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

[36] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, 2nd ed. New York: Wiley, 2007.

[37] 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.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399