Vision-Guided Robot Hearing

Xavier Alameda-Pineda, Radu Horaud

Vision-Guided Robot Hearing

Xavier Alameda-Pineda and Radu Horaud

INRIA Grenoble Rhône-Alpes
655 Av. de l'Europe, 38334, Montbonnot, France
firstname.lastname@inria.fr

International Journal of Robotics Research

Abstract

Natural human-robot interaction (HRI) in complex and unpredictable environments is important with many potential applicatons. While vision-based HRI has been thoroughly investigated, robot hearing and audio-based HRI are emerging research topics in robotics. In typical real-world scenarios, humans are at some distance from the robot and hence the sensory (microphone) data are strongly impaired by background noise, reverberations and competing auditory sources. In this context, the detection and localization of speakers plays a key role that enables several tasks, such as improving the signal-to-noise ratio for speech recognition, speaker recognition, speaker tracking, etc. In this paper we address the problem of how to detect and localize people that are both seen and heard. We introduce a hybrid deterministic/probabilistic model. The deterministic component allows us to map 3D visual data onto an 1D auditory space. The probabilistic component of the model enables the visual features to guide the grouping of the auditory features in order to form audiovisual (AV) objects. The proposed model and the associated algorithms are implemented in real-time (17 FPS) using a stereoscopic camera pair and two microphones embedded into the head of the humanoid robot NAO. We perform experiments with (i) synthetic data, (ii) publicly available data gathered with an audiovisual robotic head, and (iii) data acquired using the NAO robot. The results validate the approach and are an encouragement to investigate how vision and hearing could be further combined for robust HRI.

1 Introduction

For the last decade, robotics researchers have developed the concept of human robot-companions, endowed with such skills as moving in complex and unconstrained environments. While a robot must be able to safely navigate and manipulate objects, it should also be able to interact with people. Obviously, speech communication should play a crucial role in modeling the cognitive aspects of human-robot interaction (HRI). But in typical real-world scenarios, humans that emit speech (as well as other sounds of interest) are at some distance and hence the robot's microphone signals are strongly impaired by noise, reverberations, and interfering sound sources. Compared with other types of hands-free human-machine audio interfaces, e.g., smart phones, in typical HRI scenarios, the human-to-robot distance is much larger. Moreover, the problem is aggravated further as the robot produces significant *ego noise* due to its mechanical drives and electronics. This implies that robot-embodied audition cannot fully exploit state-of-the-art speech recognition techniques and, more generally, HRI based on verbal dialog.

Humans have sophisticated abilities to enhance and disambiguate weak unimodal data based on information fusion from multiple sensory inputs [Anastasio 00, King 09]. In particular, audiovisual fusion is one of the most prominent forms of multimodal data processing and interpretation mechanisms; it plays a crucial role in extracting auditory information from dirty acoustic signals, e.g., the cocktail party

problem [Haykin 05]. In this paper we address the problem of how to detect and localize people that are both seen and heard by a humanoid robot. We are particularly interested in combining vision and hearing in order to identify the activity of people, e.g., emitting speech and non-speech sounds, in informal scenarios.

A typical example of such a scenario is shown in Figure 1 where people sit at some distance from the robot and informally chat which each other and with the robot. The robot's first task (prior to speech recognition, language understanding, and dialog handling) consists in retrieving the time-varying auditory status of the speakers. This allows the robot to turn its attention towards an acoustically active person, precisely determine the position and orientation of his/her face, optimize the emitter-to-receiver acoustic pathway so as to maximize the signal-to-noise ratio (SNR), and eventually retrieve a clean speech signal. We note that this problem cannot be solved within the traditional human-computer interface paradigm which is based on *tethered* interaction, i.e., the user wears a close-range microphone, and which primarily works for a single user and with clean acoustic data. On the contrary, untethered interaction is likely to broaden the range of potential cooperative tasks between robots and people, to allow natural behaviors, and to enable multi-party dialogs.

This paper has the following two main contributions:

- The problem of detection and localization of multiple audiovisual (AV) events is cast into a mixture model. We explore an emitter-to-perceiver acoustic-wave propagation model that allows us to map *both* 3D visual features and 3D sound sources onto the 1D auditory space spanned by interaural time difference (ITD) variable between two microphones. Therefore, visual and auditory cues can potentially be clustered together to form AV events. We derive an expectation-maximization (EM) procedure that exhibits several interesting features: it allows either to put vision and hearing on an equal footing, or to weight their relative importance so that the algorithm can be partially supervised by the more reliable modality, it allows to perform model selection or, more precisely, to estimate the number of AV events, it is robust to outliers, such as visual artifacts and reverberations, it is extremely efficient since it relies on a one-dimensional Gaussian mixture model and since the 3D event locations can be estimated without any additional effort.
- The proposed model and method are implemented in real-time using a stereoscopic camera pair and two microphones embedded into the head of the humanoid companion robot NAO[1]. We describe a modular software architecture based on open-source *Robotics Service Bus* (RSB) middleware[2]. RSB allows to distribute an application over the robot's onboard CPU and an external computer. RSB is event-based: events are equipped with several timestamps, thus handling the synchronization of visual and auditory observations gathered at different sampling rates as well as the synchronization of higher level visual and auditory processing modules. This software architecture allows to implement and test our algorithms without the performance and deployment restrictions imposed by the robot's onboard computing resources. More interestingly, the proposed implementation can be reused with other robots.

The remainder of the paper is organized as follows: Section 2 describes the related work; Section 3 outlines the hybrid deterministic/probabilistic model that we propose; Section 4 describes the methods used for the extraction of auditory and visual features and for audiovisual calibration. Section 5 describes the proposed probabilistic audiovisual fusion model. Sections 6 describes the multimodal inference procedure and Section 7 describes its implementation on the humanoid robot NAO. Section 8 shows the results obtained with the proposed method, and Section 9 draws some conclusions and gives some directions for future work.

[1] NAO is manufactured by Aldebaran Robotics
[2] https://code.cor-lab.org/projects/rsb

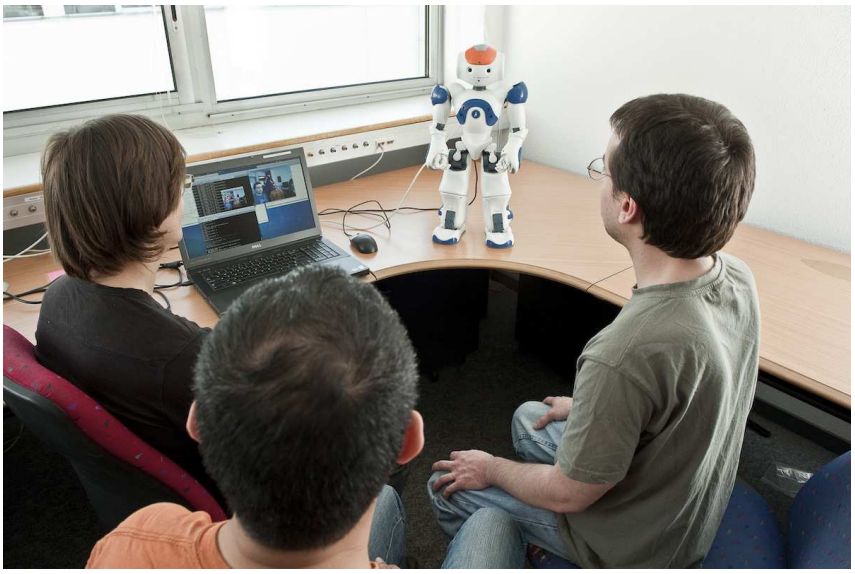


Figure 1: A typical scenario in which a companion humanoid robot (NAO) performs audiovisual fusion in an attempt to assess the auditory status of each one of the speakers in front of the robot and to estimate the 3D locations of their faces. The method uses the robot's onboard cameras and microphones as well as a modular software architecture based on the freely available RSB (robotics service bus) middleware. This allows *untethered interaction* between robots and people. Moreover, RSB allows method implementation using external computing power and without the performance and deployment restrictions imposed by the onboard computing resources.

2 Related Work

While computational models and methods for vision and hearing have mainly been addressed separately, several behavioral, electrophysiological and neuro-imaging studies [Calvert 04], [Ghazanfar 06], [Senkowski 08] postulated that the fusion of different sensorial modalities is an essential component of perception. Nevertheless, computational models for audiovisual fusion and their implementation on robots remain largely unexplored.

The problem of integrating data gathered with physically different sensors, e.g., cameras and microphones, is extremely challenging. Auditory and visual sensor-data correspond to different physical phenomena which must be interpreted in a different way. Relevant visual information must be inferred from the way light is reflected by scene objects and valid auditory information must be inferred from the perceived signals such that they contain the intrinsic properties of one or several emitters. The spatial and temporal distributions of visual and auditory data are also very different. Visual data are spatially dense and continuous in time. Auditory sources are spatially sparse and temporally intermittent. These two modalities are perturbed by different phenomena such as occlusions and self-occlusions (visual data) or ambient noise and echoic environments (auditory data).

In order to address these challenges, several researchers investigated the fusion of auditory and visual cues for a variety of goals, such as event classification [Natarajan 12, Cristani 07], speech recognition [Barker 09], voice-activity detection [Yoshida 12], sound-source separation [Naqvi 10], speaker tracking [Hospedales 08], [Gatica-Perez 07] and speaker diarization [Noulas 12]. However, these approaches are not always suitable for robotic applications, either because of their algorithmic complexity, or because the need of a distributed sensor network. Moreover, some of these methods need a large amount of training data, which considerably limits their practical potential. In particular, in HRI scenarios, user-specific training [Hospedales 08], [Noulas 12], e.g., voice and face recognition, is not practical and is likely to limit the method's range of applications.

Comparatively, much less effort has been devoted to design audiovisual fusion methods in conjunction with HRI. Interesting approaches that propose methods specifically conceived for humanoid robots are available, e.g., speech recognition [Nakadai 04], beat tracking [Itohara 11], [Itohara 12], active audition [Kim 07] or sound recognition [Nakamura 11], to cite a few. All these methods deal with the detection and localization problems by using combinations of off-the-shelf techniques.

Many HRI applications rely on finding speakers in a scene and assessing their status. Providing a robust methodology to simply count how many speakers are in a room, to localize and track them over time, and to identify their auditory status are central audiovisual tasks. There has been a lot of interest on how to temporally aggregate audiovisual information. Audiovisual speaker tracking and diarization were investigated in the recent past, e.g., [Beal 03, Perez 04, Checka 04, Gatica-Perez 07, Hospedales 08, Noulas 12]). This article focuses on speaker detection and localization, two tasks that are complementary to the tasks of tracking and of diarization. Indeed, detection and localization are both needed in order to initialize tracking methods, to periodically re-initialize them, as well as to correct possible drift behavior.

Prior work on speaker detection and localization can be grouped into two categories. On one side, several statistical non-parametric approaches have been developed: [Gurban 06], [Besson 08b] and [Besson 08a] investigate the use of information theoretic methods to associate auditory and visual data in order to detect the active speaker. Similarly, [Barzelay 07] proposes an algorithm matching auditory and visual onsets over time. Even though these approaches show very good performance results, they use user-dedicated cameras, thus limiting the interaction. Moreover, these non-parametric approaches need a lot of training data. One critical outcome of such training is that it is environment-dependent. Consequently, implementing such methods on mobile platforms results in systems with almost no flexibility.

On the other side, several probabilistic approaches were suggested. In [Khalidov 08, Khalidov 11] the conjugate Gaussian mixture model (GMM) for audiovisual fusion is introduced. Two Gaussian mixtures are estimated, one for each modality (vision and auditory) while the two GMM parameter sets are constrained via a common set of *tying parameters*, namely the 3D locations of the AV events being sought. Recently in [Noulas 12], a factorial HMM is proposed to associate auditory, visual and audiovisual features. All these methods simultaneously detect and localize the speakers but they are not suitable for real-time processing, because of their algorithmic complexity. [Kim 07] proposed a Bayesian framework for inferring the position of the active speaker and for combining a sound-source localization technique with a face-tracking algorithm. The reported results are good in the case of one active speaker, but show bad performance for multiple or faraway speakers. This is due to the fact that the proposed probabilistic framework is not able to correctly handle outliers. In [Alameda-Pineda 11], the authors use a GMM to fuse the auditory and visual data by building AV clusters. This probabilistic framework is able to handle outliers thanks to a uniform component in the mixture model.

In this paper we propose a novel multimodal deterministic/probabilistic model for audiovisual detection and localization of speaking people. The proposed method has the following features: (i) it is theoretically sound and robust, (ii) it is designed to process robot-centered data, (iii) it accommodates to different visual and auditory features, (iv) it is robust to noise and outliers, (v) it requires a simple calibration step that must be performed only once, thus guaranteeing the adaptability of the system, (vi) it works in unrestricted indoor environments, (vii) and it is implemented on a commercially available humanoid using an open-source middleware.

3 A Hybrid Deterministic/Probabilistic Model

In this section we introduce a novel deterministic/probabilistic fusion model which is well suited for audiovisual detection and localization of speaking people and that is implementable for real-time applications. The algorithms derived from the model are able to count how many speakers are out there, to locate them and to ascertain whether they speak or not. In other words, we seek the number of speakers, $N \in \mathbb{N}$, their positions $\{\boldsymbol{S}_n\}_{n=1}^N \in \mathbb{S}$ ($\mathbb{S} \subset \mathbb{R}^3$) is the scene and their speaking states $\{e_n\}_{n=1}^N \in \{0,1\}$ (0 – *not speaking* and 1 – *speaking*).

In order to accomplish the detection and localization of speakers, auditory and visual features are extracted from the raw signals (sound signals and images) over a time interval Δt. We assume Δt to be short enough so that the speakers remain approximately in the same 3D location and long enough to capture small displacements and oscillatory movements of their heads and faces. The auditory and visual features extracted during Δt are denoted by $\mathbf{a} = \{a_1, \dots, a_k, \dots, a_K\} \subset \mathbb{A} \subset \mathbb{R}$ and by $\mathbf{v} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_m, \dots, \boldsymbol{v}_M\} \subset \mathbb{V} \subset \mathbb{R}^3$ respectively, where $\mathbb{A}$ and $\mathbb{V}$ are the auditory and visual feature spaces.

We want to estimate N, $\{\boldsymbol{S}_n\}_{n=1}^N$ and $\{e_n\}_{n=1}^N$ that best explain the observed features $\mathbf{a}$ and $\mathbf{v}$. Therefore, we need a framework that encompasses both the hidden and the observed variables and that accounts for the following challenges: (i) the visual and auditory observations lie in mathematically different spaces with different dimensionality, (ii) the object-to-observation assignments are not known in advance, (iii) both visual and auditory observations are contaminated with noise and outliers (irrelevant data), (iv) the relative importance of the two modalities is not known in advance, (v) the position and speaking states of the speakers have to be gauged and (vi) since we want to be able to deal with a time-varying number of speakers, we must estimate the parameter N.

The proposed deterministic/probabilistic framework seeks the desired variables and accounts for the outlined challenges. On one hand, the deterministic components allow to model those characteristics of the scene that are known with precision in advance. They may be the outcome of a very accurate calibration step, or the direct consequence of some geometrical or physical properties of the sensors. On the other hand, the probabilistic components model random effects. For example, noise and outliers, which are a consequence of the complexity of the scene as well as of the feature extraction procedure.

4 The Deterministic Model

In this section we delineate the deterministic components of our model, namely the visual and auditory mappings. Because the scene space, the visual-observation space and the auditory-observation space are different, we need two mappings: the first one, $\mathcal{A} : \mathbb{S} \to \mathbb{A}$, links the scene space to the auditory space and the second one, $\mathcal{V} : \mathbb{S} \to \mathbb{V}$, links the scene space to the visual space. Both mappings are illustrated on Figure 2. An AV object located at $\boldsymbol{S} \in \mathbb{S}$, is mapped on $\mathcal{A}(\boldsymbol{S})$ in the auditory space and on $\mathcal{V}(\boldsymbol{S})$ in the visual space.

The mappings $\mathcal{A}$ and $\mathcal{V}$ provide a link between the two observations spaces as well, namely $\mathcal{A} \circ \mathcal{V}^{-1}$ or $\mathcal{V} \circ \mathcal{A}^{-1}$. While the former maps *images onto sounds*, the latter maps *sounds onto images*. Notice however that the two mappings, $\mathcal{A}$ and $\mathcal{V}$, may or may not be invertible. This essentially depends on the number of sensors being used. With a stereoscopic camera pair (or with any other 3D visual sensor), the inverse visual mapping is well defined because there is a one-to-one correspondence between 3D features and 2D image features. The inverse auditory mapping is more problematic. Recently, it was proved that at least four non-coplanar microphones are needed in order to yield a one-to-one correspondence between the location of an audio-source and the associated interaural time differences [Alameda-Pineda 14]. For example, [Alameda-Pineda 11, Sanchez-Riera 12], $\mathcal{V}$ use a stereo camera pair and two microphones, hence $\mathcal{A} \circ \mathcal{V}^{-1}$ is available. In multi-microphone sound-source localization, e.g., [Nakadai 04], one can map sound locations onto visual features, i.e., $\mathcal{V} \circ \mathcal{A}^{-1}$. In [Khalidov 08, Khalidov 11] the mappings are used within a conjugate generative model to tie the parameters of the probabilistic model, hence $\mathbb{A}$ and $\mathbb{V}$ are used independently of each other. In [Butz 05, Kidron 05, Kidron 07, Liu 08], the scene space is undetermined and the authors learn a common representation space (the scene space) at the same time they learn both mappings.

In this work we use a stereoscopic camera pair, hence we extract 3D visual features and use $\mathcal{A} \circ \mathcal{V}^{-1}$ to map them onto the auditory space (Section 4.2). The auditory features correspond to the interaural time differences (Section 4.1), and a direct path propagation model defines $\mathcal{A}$. The mapping $\mathcal{A} \circ \mathcal{V}^{-1}$ is accurately built from the geometric and physical models estimated through a calibration step (see Section 4.3). Consequently, we are able to map the 3D visual features $\mathbf{v} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_M\}$ onto the 1D

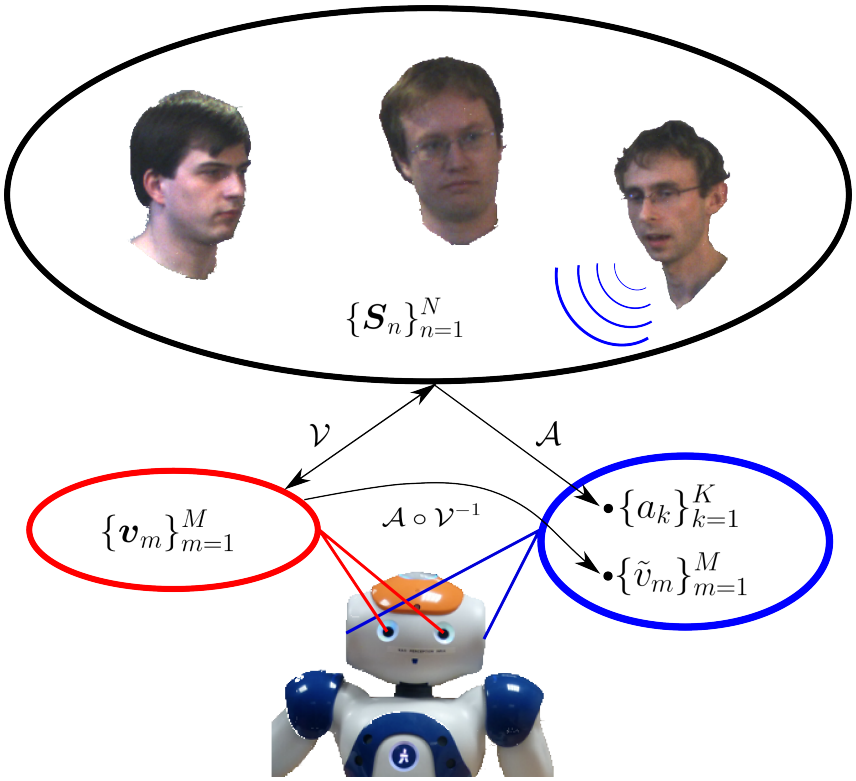


Figure 2: This figure shows the general principle of the proposed method. Audiovisual events ($\boldsymbol{S}$), e.g., speaking persons, are observed with two cameras and two microphones, hence two types of observations are available: 3D binocular features ($\boldsymbol{v}$) and 1D binaural features (a). By combining the inverse visual mapping with the direct auditory mapping, $\mathcal{A} \circ \mathcal{V}^{-1}$, it is possible to project features from the 3D visual space to the 1D auditory space and hence to represent visual and auditory data in the same space.

auditory space $\mathbb{A}$. Hence, a 3D visual feature $\boldsymbol{v}_m$ is mapped onto a 1D feature $\tilde{v}_m \in \mathbb{A}$:

$$\tilde{v}_m = (\mathcal{A} \circ \mathcal{V}^{-1})(\boldsymbol{v}_m). \tag{1}$$

To summarize, the mapping $\mathcal{A} \circ \mathcal{V}^{-1} : \mathbb{V} \to \mathbb{A}$ allows us to project 3D visual features onto the 1D ITD space, and hence to represent both visual and auditory observations in the same space.

4.1 Auditory Features

Auditory observations $\{a_k\}_{k=1}^K$ correspond to interaural time difference (ITD) measurements between the left and right microphones. An ITD corresponds to the time difference of arrival (TDOA), i.e., from the signal emitted by a scene source to the left and right microphones. If we assume a direct acoustic-wave propagation model traveling at a constant velocity, ν, the source-to-ITD mapping $\mathcal{A} : \mathbb{S} \to \mathbb{A}$ writes:

$$\mathcal{A}(\boldsymbol{S}) = \frac{\|\boldsymbol{S} - \boldsymbol{M}_L\| - \|\boldsymbol{S} - \boldsymbol{M}_R\|}{\nu}, \tag{2}$$

where $\boldsymbol{M}_L$ and $\boldsymbol{M}_R$ are the 3D positions of the two microphones and $\boldsymbol{S} \in \mathbb{S}$ is the 3D position of the sound source. It is well known that the inverse of this mapping yields a 2D manifold, namely one sheet of a two-sheet hyperboloid with foci $\boldsymbol{M}_L$ and $\boldsymbol{M}_R$ (see Lemma 1 in [Alameda-Pineda 14]). This means that if an ITD a is observed, the sound source lies on a 2D manifold defined by $\mathcal{A}^{-1}(a)$. In practice,

interaural time differences are estimated using the method of [Christensen 07], which yields very good results that are stable over time. Finally, it is worth noticing that ITDs are real-valued observations lying in the interval $\mathbb{A} = [-A, A] \subset \mathbb{R}$. Indeed, the largest absolute ITD value, A, corresponds to a source located at the interaural axis:

$$A = \nu^{-1} \|\boldsymbol{M}_R - \boldsymbol{M}_L\|. \tag{3}$$

4.2 Visual Features

Visual observations, $\mathbf{v} = \{\boldsymbol{v}_m\}_{m=1}^{M}$ are 3D features extracted from a stereoscopic image pair. In this paper we use two types of visual features: Harris-motion 3D (HM3D) interest points and 3D centres of human faces (F3D). The rationale of using Harris interest points [Harris 88] is that they have high detection repeatability and correspond to highly textured image regions, hence they are good candidates for correlation-based stereo matching. We propose to throw out static interest points and to retain only those points that are likely to correspond to moving scene objects. Indeed, it is well known that moving features correlate well with the presence of auditory sources.

HM3D are general purpose visual features that require the presence of a moving object, e.g., lip and facial motions. Alternatively we also implemented a 3D human face detector (F3D) based on face detection and stereo matching. These facial features are static and hence they do not need a motion detector. They represent a good alternative, in particular for real-time implementations and when the goal is to detect speakers only. These features are described in detail below.

HM3D To obtain Harris-motion 3D points, we start by detecting Harris interest points [Harris 88] in the left and right images. Next, we only consider a subset of these points, namely points where motion is the most likely to occur. For each interest-point (u, v) in the left image, we consider the image intensities at the same location (u, v) over the time interval Δt and we compute the temporal standard deviation $\tau_{(u,v)}$ of the intensity. Assuming stable lighting condition over Δt, we simply classify the interest points into static ($\tau_{(u,v)} \leq \tau_M$) or dynamic ($\tau_{(u,v)} > \tau_M$) where τ_M is a user-defined threshold. Finally, we apply standard stereo matching and stereo reconstruction techniques [Hartley 04] to yield a set of HM3D points $\mathbf{v}$ associated with Δt.

F3D 3D centers of human faces are obtained using the face detector described in [Šochman 05]. More precisely, the left-image and right-image centers of the bounding boxes associated with the face detector are reconstructed in 3D.

Because both HM3D and F3D are 3D visual features, there is a one-to-one mapping between the scene space and the visual space. Hence, a visual feature $\boldsymbol{v}_m$ can be mapped onto the ITD (auditory) space using (1), or more precisely:

$$\tilde{v} = \frac{\|\mathcal{V}^{-1}(\boldsymbol{v}) - \boldsymbol{M}_L\| - \|\mathcal{V}^{-1}(\boldsymbol{v}) - \boldsymbol{M}_R\|}{\nu}. \tag{4}$$

Whenever the 3D position of a visual feature corresponds to a sound source, (4) allows to predict the ITD value corresponding to that source. In practice of course, noise and outliers do not allow to apply this strategy in a deterministic way and a statistical inference method is necessary.

4.3 Audiovisual Calibration

The mapping defined by (4) requires an explicit representation of $\mathcal{V}$, of its inverse, as well as the 3D coordinates of the two microphones, $\boldsymbol{M}_L$ and $\boldsymbol{M}_R$. In our case, a visual observation $\boldsymbol{v}$ corresponds to a 3D vector (x, y, d), where x, y are the left-image coordinates and d is the horizontal disparity between

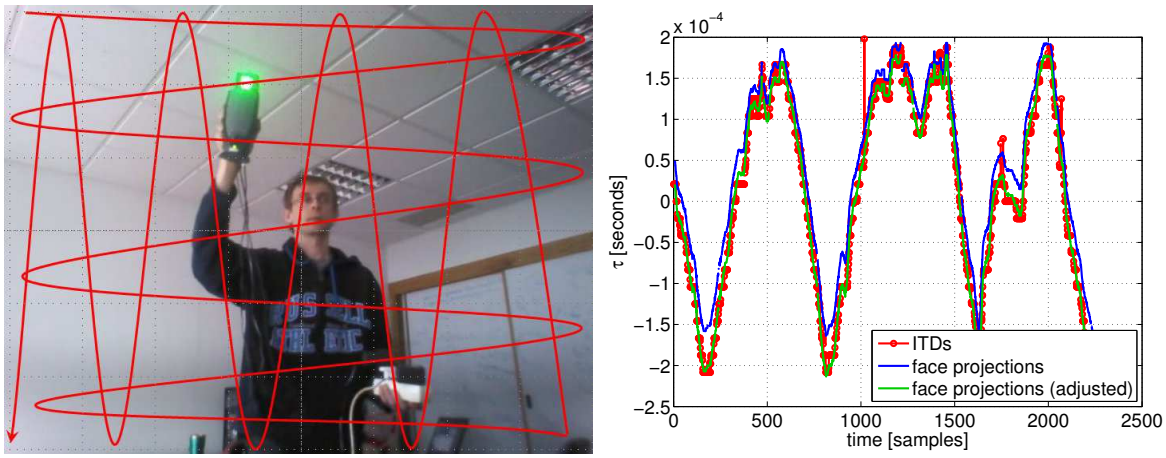


Figure 3: Left: The calibration setup proposed in this paper: an audiovisual target is freely moved in front of the camera-and-microphone setup. Right: the result of augmenting the acoustic propagation model with a linear regression model, in order to compensate for the effects induced by the presence of the robot head. ITD values (red dots), trajectory based on (4) (blue) and on (8) (green).

the left-image and right-image coordinates. Assuming a rectified stereoscopic image pair, the mapping $\mathcal{V} : \mathbb{S} \to \mathbb{V}$ and its inverse write:

$$\mathcal{V}(X, Y, Z) = \left(\frac{X}{Z}, \frac{Y}{Z}, \frac{b}{Z}\right), \tag{5}$$

$$\mathcal{V}^{-1}(x, y, d) = \left(\frac{xb}{d}, \frac{yb}{d}, \frac{b}{d}\right), \tag{6}$$

where (X, Y, Z) are the 3D coordinates of $\boldsymbol{v}$, and b is the stereo baseline (the distance between the projection centres of the two cameras).

To complete the characterization of (4), we need to estimate the microphone locations, $\boldsymbol{M}_L$ and $\boldsymbol{M}_R$, in the camera frame. To do this we use an audiovisual target, e.g., figure 3-left, composed of a loud-speaker equipped with a LED. The precise location of the loud-speaker is estimated using 3D reconstruction with the stereoscopic camera pair. Moreover, the ITD of the white-noise signal emitted by the speaker is estimated using the two microphones. We use white noise because it yields a single sharp ITD value in the cross-correlation function obtained from the two microphone signals. This provides (ITD, $\mathcal{V}^{-1}(\boldsymbol{v})$) pairs that are plugged into (4). By moving the audiovisual target so that it covers the entire field of view of the cameras, we obtain an over-constrained non-linear system of equations in $\boldsymbol{M}_L$ and $\boldsymbol{M}_R$. This is solved using the method proposed in [Khalidov 13].

In the particular case of the NAO robot, we have noticed that the above procedure does not always yield very accurate results. In order to compensate for the acoustic propagation effects due to the presence of the robot head, we introduce a slightly more complex model, namely:

$$\text{ITD} = c_1 \frac{\|\boldsymbol{S} - \boldsymbol{M}_L\| - \|\boldsymbol{S} - \boldsymbol{M}_R\|}{\nu} + c_0, \tag{7}$$

where c_1 and c_0 are two adjustment coefficients. If the microphone positions are known, the adjustment coefficients can be easily estimated via linear regression. In practice we alternate between the estimation of the microphone positions [Khalidov 13] and the estimation of the adjustment coefficients. To account for this audiovisual calibration, (4) is replaced with:

$$\tilde{v} = c_1 \frac{\|\mathcal{V}^{-1}(\boldsymbol{v}) - \boldsymbol{M}_L\| - \|\mathcal{V}^{-1}(\boldsymbol{v}) - \boldsymbol{M}_R\|}{\nu} + c_0. \tag{8}$$

Figure 3 shows a snapshot of the practical calibration setup (left) and the result of calibration (right). The red circles correspond to ITD observations associated with white noise emitted by the loud-speaker. The blue curve shows the trajectory of $\tilde{v}$ predicted with (4) while the green curve shows the trajectory predicted with (8). The audiovisual calibration process that we just described has a number of interesting features: (i) it is very easy to set up, (ii) it uses training data that are quite easy to produce, and (iii) it can be applied to a large number of sensor configurations, including modern 3D visual sensors equipped with a microphone array.

5 The Probabilistic Model

The deterministic model that we just presented allows us to represent visual and auditory observations in the same 1D space. In this section we introduce a probabilistic audiovisual clustering model that allows the inference of 3D audiovisual events. We start by introducing two sets of hidden variables, $\boldsymbol{Z} = \{Z_1, \ldots, Z_m, \ldots, Z_M\}$ and $\boldsymbol{W} = \{W_1, \ldots, W_k, \ldots, W_K\}$ accounting for the observation-to-object assignments. The notation $Z_m = n$, with $m \in \{1, \ldots, M\}$ and $n \in \{1, \ldots, N+1\}$, means that the projected visual observation $\tilde{v}_m$ either is generated by the n^{th} 3D object ($n \in \{1, \ldots, N\}$) or is an outlier ($n = N+1$). Similarly, the variable W_k is associated to the auditory observation a_k.

We formulate the multimodal probabilistic fusion model under the assumption that all observations $\tilde{v}_m$ and a_k are independent and identically distributed. Because the auditory observations and the projected visual observations belong to the auditory space, the statistical model is defined over the same (auditory) space and not in two distinct (visual and the auditory) spaces. Indeed, we assume that the n^{th} AV object generates both visual and auditory features normally distributed around $\mathcal{A}(\boldsymbol{S}_n)$ and that both the visual and auditory outliers are uniformly distributed in $\mathbb{A}$. Therefore, we write:

$$\mathrm{P}(\tilde{v}_m | Z_m = n, \Theta_n) = \begin{cases} \mathcal{N}(\tilde{v}_m; \mu_n, \sigma_n) & n = 1, \ldots, N \\ \mathcal{U}(\tilde{v}_m; \mathbb{A}) & n = N+1, \end{cases} \tag{9}$$

where Θ_n designates the Gaussian parameters, $\mu_n = \mathcal{A}(\boldsymbol{S}_n)$ and σ_n (the mean and the standard deviation of the n^{th} Gaussian). Similarly we have:

$$\mathrm{P}(a_k | W_k = n, \Theta_n) = \begin{cases} \mathcal{N}(a_k; \mu_n, \sigma_n) & n = 1, \ldots, N \\ \mathcal{U}(a_k; \mathbb{A}) & n = N+1. \end{cases} \tag{10}$$

Hence, we can define a generative model for any observation $o \in \mathbb{A}$:

$$\mathrm{p}(o; \Theta) = \sum_{n=1}^{N} \pi_n \mathcal{N}(o; \mu_n, \sigma_n) + \pi_{N+1} \mathcal{U}(o; \mathbb{A}), \tag{11}$$

where π_n is the prior probability of the n^{th} mixture component. That is, $\pi_n = \mathrm{P}(Z_m = n) = \mathrm{P}(W_k = n)$, $\forall n, m, k$. The prior probabilities satisfy $\sum_{n=1}^{N+1} \pi_n = 1$. Summarizing, the model parameters are:

$$\Theta = \{\pi_1, \ldots, \pi_{N+1}, \Theta_1, \ldots, \Theta_N\}. \tag{12}$$

Because the statistical model (11) is a Gaussian mixture, we are left with the problem of how to define a Gaussian density on a bounded domain, $\mathbb{A} \subset \mathbb{R}$, We assume that the variances are small enough such that almost all the probability mass of the Gaussian components is contained in $\mathbb{A}$. Therefore, we ignore the tails of the Gaussians and use the model in (11) with no further modification.

Under this formulation, the set of parameters may be estimated via maximum likelihood:

$$\mathcal{L}(\tilde{\mathbf{v}}, \mathbf{a}; \Theta) = \sum_{m=1}^{M} \log \mathrm{p}(\tilde{v}_m; \Theta) + \sum_{k=1}^{K} \log \mathrm{p}(a_k; \Theta). \tag{13}$$

In other words, the optimal set of parameters is the one maximizing the log-likelihood function (13), where p is the generative probabilistic model in (11). Unfortunately, direct maximization of (13) is an intractable problem. Equivalently, the expected complete-data log-likelihood will be maximized [Dempster 77] (see Section 6).

We recall that the ultimate goal is to determine the number N of AV events, their 3D locations $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_n, \ldots, \boldsymbol{S}_N$ as well as their auditory activities $e_1, \ldots, e_n, \ldots, e_N$. However, the 3D location parameters can be computed only indirectly, once the multimodal mixture's parameters Θ have been estimated. Indeed, once the auditory and visual observations are grouped in $\mathbb{A}$, the $\tilde{v}_m \leftrightarrow \boldsymbol{v}_m$ correspondences are used to infer the locations $\boldsymbol{S}_n$ of the AV objects and the grouping of the auditory observations $\mathbf{a}$ is used to infer the speaking state e_n of the AV objects. The choice of N as well as the formulas for $\boldsymbol{S}_n$ and e_n are given in Sections 6.2 and 6.3 respectively. Before these details are given and in order to fix ideas, we devote next section to describe the auditory and visual features, justify the existence of $\mathcal{V}^{-1}$ and detail the calibration procedure leading to a highly accurate mapping $\mathcal{A} \circ \mathcal{V}^{-1}$.

6 Multimodal Inference

Section 5 described a maximum-likelihood framework to perform audiovisual fusion. The 3D visual features are mapped onto the auditory space $\mathbb{A}$ through the audiovisual mapping ($\mathcal{A} \circ \mathcal{V}^{-1}$). This mapping takes the form (8) when under precise audiovisual calibration. In this section we address the following issues (i) the relative importance of each modality, (ii) robust estimation of the problem's variables $\boldsymbol{S}_n$ and e_n, and (iii) online estimation of the number of audiovisual events present in the scene over a short time interval, N. In this section we describe the proposed EM procedure.

6.1 Visually-guided Inference

An interesting problem that has barely been addressed, is how to balance the relative importance of each one of the two modalities. An analysis of both the physical nature and the statistics of the auditory and visual features, led us to choose a visually-guided audio-clustering process. Indeed, visual features, e.g. HM3D and F3D, are spatially dense and enjoy a better temporal continuity than auditory data, in spite of the fact that visual data are often corrupted by occlusions. We refer to Figure 8 for a typical example of audiovisual data associated with the problem at hand. Therefore, inference from visual data are likely to be statistically more consistent than inference from auditory data. The temporal consistency arises from the fact that visual observations are meaningful along the entire sequence, while auditory observations carry useful information only when at least one of the speakers emits a sound. For all these reasons, we start by fitting a 1D GMM to the audio-space-projected visual features $\{\tilde{v}_m\}_{m=1}^{M}$. This is done with a standard EM algorithm [Bishop 06]. In the E-step the posterior probabilities $\alpha_{mn} = \mathrm{P}(Z_m = n | \tilde{\mathbf{v}}, \Theta)$ are updated via:

$$\alpha_{mn} = \frac{\pi_n \, \mathrm{P}(\tilde{v}_m | Z_m = n, \Theta)}{\sum_{i=1}^{N+1} \pi_i \, \mathrm{P}(\tilde{v}_m | Z_m = i, \Theta)}. \tag{14}$$

The M-step maximizes the expected complete-data log-likelihood with respect to the model parameters, leading to:

$$\pi_n = \frac{\bar{\alpha}_n}{M} \quad n = 1, \ldots, N+1, \tag{15}$$

$$\mu_n = \frac{1}{\bar{\alpha}_n} \sum_{m=1}^{M} \alpha_{mn} \tilde{v}_m \quad n = 1, \ldots, N, \tag{16}$$

$$\sigma_n^2 = \frac{1}{\bar{\alpha}_n} \sum_{m=1}^{M} \alpha_{mn} (\tilde{v}_m - \mu_n)^2 \quad n = 1, \ldots, N, \tag{17}$$

with $\bar{\alpha}_n = \sum_{m=1}^{M} \alpha_{mn}$. Once the model is fitted to the projected visual data, i.e., the visual information has already been probabilistically assigned to the N objects, the clustering process proceeds by including the auditory information. Hence, we are faced with a constrained maximum-likelihood estimation problem: maximize (13) subject to the constraint that the posterior probabilities α_{mn} were previously computed. This leads to the *vision-guided EM fusion algorithm* in which the E-step only updates the posterior probabilities associated with the auditory observations while those associated with the visual observations remain unchanged. This semi-supervision strategy was introduced in the context of text classification [Nigam 00, Miller 03]. Here it is extended to enforce both the quality and the reliability of one of the sensing modalities, within a clustering-based fusion algorithm. Moreover, we enforce the same generative assumption on both modalities, namely both the projected visual features and the auditory features follow the model in (11) with the same parameters, as opposed to having different parameters for auditory and for visual data. Unfortunately, the use of an audio-dedicated parameter set leads to statistically inconsistent ML estimates, due to the scarcity of the auditory observations. To summarize, the E-step of the algorithm updates only the posterior probabilities of the auditory observations $\beta_{kn} = \mathrm{P}(W_k = n|\mathbf{a}, \Theta)$:

$$\beta_{kn} = \frac{\pi_n \, \mathrm{P}(a_k|W_k = n, \Theta)}{\sum_{i=1}^{N+1} \pi_i \, \mathrm{P}(a_k|W_k = i, \Theta)}, \tag{18}$$

while keeping the visual posterior probabilities, α_{mn}, constant. The M-step has a closed-form solution and the prior probabilities are updated with:

$$\pi_n = \frac{\gamma_n}{M + K}, \quad n = 1, \ldots, N + 1, \tag{19}$$

where $\gamma_n = \bar{\alpha}_n + \bar{\beta}_n$ and $\bar{\beta}_n = \sum_{k=1}^{K} \beta_{kn}$. The means and variances of the current model are estimated by combining the two modalities:

$$\mu_n = \frac{1}{\gamma_n} \left(\sum_{m=1}^{M} \alpha_{mn} \, \tilde{v}_m + \sum_{k=1}^{K} \beta_{kn} \, a_k \right) \quad n = 1, \ldots, N, \tag{20}$$

$$\sigma_n^2 = \frac{\sum_{m=1}^{M} \alpha_{mn} \, (\tilde{v}_m - \mu_n)^2 + \sum_{k=1}^{K} \beta_{kn} \, (a_k - \mu_n)^2}{\gamma_n} \quad n = 1, \ldots, N. \tag{21}$$

We stress that the parameters of the generative model are defined in the auditory space since both the projected visual features and the auditory features belong to this space. Therefore, the output of the entire clustering process is twofold. Firstly, the observations are soft-assigned to clusters through α_{mn} and β_{kn}, for visual and auditory observations respectively. Secondly, the model parameters, $\pi_1, \ldots, \pi_{N+1}$, $\mu_1, \ldots, \mu_N$ and $\sigma_1, \ldots, \sigma_N$, are estimated to fit the observations.

6.2 Finding the Number of Events

Since we do not know the value of N, a reasonable way to proceed is to estimate the parameters Θ_N for different values of N using the method delineated in the previous section. Once we estimated the maximum likelihood parameters for models with different number of AV objects, we need a criterion to choose which is the best one. In other words, we need to estimate the number of AV objects (clusters) in the scene. BIC [Schwarz 78] is a well known criterion to choose among several maximum likelihood statistical models. BIC is often chosen for this type of tasks due to its attractive consistency properties [Keribin 00]. It is appropriate to use this criterion in our framework, due to the fact that the statistical models after the *vision-guided EM algorithm*, fit the AV data in an ML sense. In our case, choosing among these models is equivalent to estimate the number of AV events $\hat{N}$. The formula to compute the BIC score is:

$$\mathrm{BIC}(\tilde{v}, \mathbf{a}, \Theta_N) = \mathcal{L}\,(\tilde{v}, \mathbf{a}; \Theta_N) - \frac{D_N \, \log(M + K)}{2}, \tag{22}$$

where $D_N = 3N$ is the number of free parameters of the model.

The number of AV events is estimated by selecting the statistical model corresponding to the maximum score:

$$\hat{N} = \arg\max_{N} \mathrm{BIC}(\tilde{v}, \mathbf{a}, \Theta_N). \tag{23}$$

6.3 Detection and Localisation

The selection on N leads to the best maximum-likelihood model in the BIC sense. That is, the set of parameters that best explain the auditory and visual observations $\mathbf{a}$ and $\tilde{\mathbf{v}}$. In the following, $\mathbf{v}$ are used to estimate the 3D positions in the scene and $\mathbf{a}$ to estimate the speaking state of each AV object.

The locations of the AV objects are estimated thanks to the one-to-one correspondence between 3D visual features and the 1D projected features. Indeed, even if the mapping $\left(\bar{\mathcal{A}} \circ \mathcal{V}^{-1}\right)$ is not invertible, we know the correspondence between $\tilde{v}_m$ and $\boldsymbol{v}_m$. In all, the probabilistic assignments of the projected visual data onto the 1D clusters, α_{mn}, can be used to estimate $\boldsymbol{S}_n$ through:

$$\hat{\boldsymbol{S}}_n = \frac{1}{\bar{\alpha}_n} \sum_{m=1}^{M} \alpha_{mn} \boldsymbol{v}_m. \tag{24}$$

More precisely, the 1D clustering parameters are used to infer 3D clusters. This intuitive formulation can be seen as a very simple case of the *label transfer* [Liu 11]. The auditory activity associated to the n^{th} speaker is estimated as follows (τ_A is a user-defined threshold):

$$\hat{e}_n = \begin{cases} 1 & \text{if } \bar{\beta}_n > \tau_A \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

These two formulae account for the last remaining issue: the 3D localization and speaking state estimation of the AV objects. The next section describes some practical considerations to be taken into account when using this EM-based AV fusion method. Afterwards, in Section 6.5, we summarize the method by providing an algorithmic scheme of the multimodal inference procedure.

6.4 Implementation Details

Even though the EM algorithm has proved to be the proper (and extremely powerful) methodology to solve ML problems with hidden variables, in practice we need to overcome two main hurdles. First, since the log-likelihood function has many local maxima and EM is a local optimization technique, a very good initialization is required. Second, because real data is finite and may not strictly follow the generative model in (11), the consistency properties of the EM algorithm do not guarantee that the model chosen by BIC is meaningful regarding the application. Thus, a post-processing step is needed in order to include the application-dependent knowledge. In all, we must account for three practical concerns: (i) EM initialization, (ii) possible shortage of observations and (iii) the probabilistic model does not fully correspond to the observations.

It is reasonable to assume that the dynamics of the AV objects are somehow constrained. In other words, the positions of the objects at a time interval are close to the positions at the previous time interval. Hence, we use the model computed in the previous time interval to initialize the EM based procedure. More precisely, if we denote by $N^{(p)}$ the number of AV objects found in the previous time interval, we initialize a new 1D GMM with N clusters, for $N \in \{0, \ldots, N_{\max}\}$. In the case $N \leq N^{(p)}$, we take the N clusters with the highest weights. For $N > N^{(p)}$, we incrementally split a cluster at its mean into two clusters. The cluster to be split is selected on the basis of a high Davies-Bouldin index [Davies 79]:

$$DW_i = \max_{j \neq i} \frac{\sigma_i + \sigma_j}{\|\mu_i - \mu_j\|}.$$

We chose to split the cluster into two clusters in order to detect AV objects that have recently appeared in the scene, either because they were outside the field of view, or because they were occluded by another AV object. This provides us with a good initialization. In our case the maximum number of AV objects is $N_{\max} = 10$.

A shortage of observations usually leads to clusters whose interactions may describe an overall pattern, instead of different components. We solve this problem by merging some of the mixture's components. There are several techniques to merge clusters within a mixture model (see [Hennig 10]). Since the components to be merged lie around the same position and have similar spread, the *ridgeline* method [Ray 05] best solves our problem.

Finally, we need to face the fact that the probabilistic model may not fully represent the observations. This is due to the non-injectivity of the audiovisual mapping $\overline{\mathcal{A}} \circ \mathcal{V}^{-1}$. We remind that the manifold defined by $\mathcal{A}^{-1}$ is a two-sheet hyperboloid [Alameda-Pineda 14]. Consequently, any 3D visual feature lying on this hyperboloid are mapped to the same value in the ITD space. In other words, two visual features that are far away from each other in the 3D space may be projected to a similar ITD value. This behavior may create *spurious clusters* in the ITD space. Since spurious clusters consist of points lying on a hyperboloid, the volume occupied by their 3D features is small. Therefore, we can easily identify (and discard) them by computing (and thresholding) the determinant of the covariance matrix estimated via a formula derived from (24):

$$\hat{\boldsymbol{\Sigma}}_n = \frac{1}{\bar{\alpha}_n} \sum_{m=1}^{M} \alpha_{mn} \left(\boldsymbol{v}_m - \hat{\boldsymbol{S}}_n \right) \left(\boldsymbol{v}_m - \hat{\boldsymbol{S}}_n \right)^\top . \tag{26}$$

6.5 Robot Hearing Guided by Visual Motion

Algorithm 1 below summarizes the proposed method. It takes as input the visual (MH3D) and auditory (ITD) observations gathered during a time interval Δt. The algorithm's output is the estimated number of clusters $\hat{N}$, the estimated 3D positions of the AV events $\{\hat{\boldsymbol{S}}_n\}_{n=1}^{\hat{N}}$ as well as their estimated auditory activity $\{\hat{e}_n\}_{n=1}^{\hat{N}}$. Because the grouping process is supervised by the HM3D features, we name the procedure *Motion-Guided Robot Hearing*. The algorithm starts by mapping the visual observations onto the auditory space by means of the linking mapping defined in (8). Then, for $N \in \{1, \ldots, N_{\max}\}$ it iterates through the following steps: (a) Initialize a model with N components using the output of the previous time interval (Section 6.4), (b) apply EM using the selected N to model the 1D projections of the visual data (Section 6.1), (c) apply the *vision-guided EM fusion* algorithm to both the auditory and projected visual data (Section 6.1) in order to perform audiovisual clustering, and (d) compute the BIC score associated with the current model, i.e., (22). This allows the algorithm to select the model with the highest BIC score, i.e., (23). The post-processing step is then applied to the selected model (Section 6.4) prior to computing the final output (Section 6.3).

7 Audiovisual Inference with NAO

The multimodal inference algorithm presented above has desirable statistical properties and good performance (see Section 8). Since our final aim is to have a stable component working on a humanoid robot (i.e., able to interact with other components), we reduce the computational load of the AV fusion algorithm. Indeed, we adapt the method described in Section 6 to achieve a light on-line algorithm working on mobile robotic platforms.

In order to reduce the complexity, we substitute the Harris-Motion 3D point detector (HM3D) with the face 3D detector (F3D), described in Section 4.2. F3D replaces hundreds of HM3D points with a few face locations in 3D, $\{\boldsymbol{v}_m\}_{m=1}^{M}$. We then consider that the potential speakers correspond to the detected faces. Hence we set $N = M$ and $\boldsymbol{S}_n = \boldsymbol{v}_n$, $n = 1, \ldots, N$. This has several crucial consequences. First,

Algorithm 1 Motion-Guided Robot Hearing

Input: HM3D, $\{\boldsymbol{v}_m\}_{m=1}^{M}$, and ITD, $\{a_k\}_{k=1}^{K}$, features.
Output: Number of AV events $\hat{N}$, 3D localization $\left\{\hat{\boldsymbol{S}}_n\right\}_{n=1}^{\hat{N}}$ and auditory status $\{\hat{e}_n\}_{n=1}^{\hat{N}}$.
Map the visual features onto the auditory space, $\tilde{v}_m = (\overline{\mathcal{A}} \circ \mathcal{V}^{-1})(\boldsymbol{v}_m)$ (8).
for $N = 1 \rightarrow N_{\max}$ **do**
  (a) Initialize the model with N clusters (Section 6.4).
  (b) Apply EM clustering to $\{\tilde{v}_m\}_{m=1}^{M}$ (Section 6.1).
  (c) Apply the *Vision-guided EM fusion* algorithm to cluster the audiovisual data (Section 6.1).
  (d) Compute the BIC score (22).
end for
Estimate the number of clusters based on the BIC score (23).
Post-processing (Section 6.4).
Compute the final outputs $\{\hat{\boldsymbol{S}}_n\}_{n=1}^{\hat{N}}$ and $\{\hat{e}_n\}_{n=1}^{\hat{N}}$ (Section 6.3).

the number of AV objects corresponds to the number of detected faces; the model selection step is not needed and the EM algorithm does not have to run $N_{\max}$ times, but just once. Second, because the visual features provide a good initialization for the EM (by setting $\mu_n = (\overline{\mathcal{A}} \circ \mathcal{V}^{-1})(\boldsymbol{S}_n)$), the visual EM is not required and the hidden variables $\boldsymbol{Z}$ do not make sense any more. Third, since the visual features are not used as observations in the EM, but to initialize it, the complexity of the *vision-guided EM fusion* algorithm is $\mathcal{O}(NK)$ instead of $\mathcal{O}(N(K+M))$. This important because the number of HM3D points is much bigger than the number of ITD values, i.e., $M \gg K$. Last, because the visual features provide the $\boldsymbol{S}_n$'s, there is no need to estimate them through (24).

7.1 Face-guided Robot Hearing

The resulting procedure is called *Face-Guided Robot Hearing* and it is summarized in Algorithm 2 below. The algorithm's input are the detected heads $(\boldsymbol{S}_1, \ldots, \boldsymbol{S}_N)$ and the auditory observations $(\mathbf{a})$ gathered during the time interval Δt. Its output is the estimated auditory activity $\{\hat{e}_n\}_{n=1}^{N}$.

Algorithm 2 Face-guided robot hearing

Input: Faces' position $\{\boldsymbol{S}_n\}_{n=1}^{N}$ and auditory $\{a_k\}_{k=1}^{K}$ features.
Output: AV objects' auditory status $\{\hat{e}_n\}_{n=1}^{\hat{N}}$.
Map the detected heads onto the auditory space, $\mu_n = (\overline{\mathcal{A}} \circ \mathcal{V}^{-1})(\boldsymbol{S}_n)$ (8).
Apply EM clustering to $\{a_k\}_{k=1}^{K}$ (Section 6.1).
Compute the final outputs $\{\hat{e}_n\}_{n=1}^{\hat{N}}$ (Section 6.3).

7.2 System Description and Architecture

We implement our method using several components which are connected by a middleware called Robotics Services Bus (RSB) [Wienke 11]. RSB is a platform-independent event-driven middleware specifically designed for the needs of distributed robotic applications. It is based on a logically unified bus which can span over several transport mechanisms like network or in-process communication. The bus is hierarchically structured using scopes on which events can be published with a common root scope. Through the unified bus, full introspection of the event flow between all components is easily possible. Consequently, several tools exist which can record the event flow and replay it later, so that application development can largely be done without a running robot. RSB events are automatically equipped with several timestamps, which provide for introspection and synchronization abilities. Because of these reasons RSB is

chosen instead of NAO's native framework NAOqi and we can implement and test our algorithm on an external processing unit without performance and deployment restrictions imposed by the robot platform. Moreover, the resulting implementation can be reused for other robots.

One tool available in the RSB ecosystem is an event synchronizer, which synchronizes events based on the attached timestamps with the aim to free application developers from such a generic task. However, several possibilities of how to synchronize events exist and need to be chosen based on the intended application scenario. For this reason, the synchronizer implements several strategies, each of them synchronizing events from several scopes into a resulting compound event containing a set of events from the original scopes. We used two strategies for the implementation. The *ApproximateTime* strategy is based on the algorithm available in [ROS 12] and outputs sets of events containing exactly one event from each scope. The algorithm tries to minimize the time between the earliest and the latest event in each set and is hence well-suited to synchronize events which originate from the same source (in the world) but undergo perception or processing delays in a way that they have non-equal timestamps. The second algorithm, *TimeFrame*, declares one scope as the primary event source and for each event received here, all events received on other scopes are attached that lie in a specific time frame around the timestamp of the source event.

ApproximateTime is used in our case to synchronize the results from the left and right camera as frames in general form matching entities but due to independent grabbing of both cameras have slightly different timestamps. Results from the stereo matching process are synchronized with ITD values using the *TimeFrame* strategy because the integration time for generating ITD values is much smaller than for a vision frame and hence multiple ITD values belong to a single vision result.

7.3 Modular Structure

The implementation is modular and divided into several components, as shwon on Figure 6. The dashed boxes correspond to data delivered by the robot's sensors and processed by the embedded computing unit. All the other modules are executed on an external computer. For clarity, the components are color-coded: modules provided by the RSB middleware (white), auditory (red) and visual (green) processing, audiovisual fusion (purple) and the visualization tool (blue) described at the end of this section.

The visual processing is composed by five modules. *Left video* and *Right video* stream the images received from the left and right cameras. The *Left face detection* module extracts the faces from the left image. These are then synchronized with the right image in *Face-image synchronization*, using the *ApproximateTime* strategy. The *F3D Extraction* module computes the F3D features. A new audiovisual head for NAO was used for this implementation. The new head (see Figure 4) is equipped with a pair of cameras and four microphones, thus providing a synchronized VGA stereoscopic image flow as well as four audio channels. Nevertheless, only two out of the four microphone signals can be exploited, because two microphones (the front and back ones) are recording fan and electronic noise from inside the robot head, and hence their signal to noise ratios are too low to be used in practice.

The auditory component consists of three modules. Interleaved audio samples coming from the microphones of NAO are streamed by the *Interleaved audio* module. The channels are separated by the *Sound deinterleaving* module, which outputs the auditory flows corresponding to the left and right microphones. These flows are stored into two circular buffers in order to extract the ITD values (*ITD extraction* module).

Both visual and auditory features flow until the *audiovisual synchronization* module; the *TimeFrame* strategy is used here to find the ITD values coming from the audio pipeline associated to the 3D positions of the faces coming from the visual processing. These synchronized events feed the *Face-guided robot hearing* module, which is in charge of estimating the speaking state of each face, e_n.

Finally, we developed the module *Visualization*, in order to get a better insight of the proposed algorithm. A snapshot of this visualization tool can be seen in Figure 5. The image consists of three

Figure 4: Within this work we use a prototype audiovisual head that embeds a binocular and synchronized camera pair, as well as two microphones. The "orange" head, used instead of the standard "blue" head, is fully interfaced by the RSB middleware described in this section.

parts. The top-left part with a blue frame is the original left image plus one rectangle per detected face. In addition to the face's bounding box, a solid circle is plot on the face of the actor encoding the emitting sound probability. The higher the probability is, the darker the circle (this feature is exploited in Figure 10). The top-right part, framed in green, is a bird-view of the scene, in which the detected heads appear as circles. The bottom-left part, with a red frame, represents the ITD space. There, both the mapped heads (ellipses) and the histogram of ITD values are plot.

7.4 Implementation

Some details need to be specified regarding the implementation of the face-guided robot hearing method. The ITDs are extracted using a sliding window of length W, with a window shift of length f. The bigger the integration window is the more reliable the ITD values are and the more expensive its computation becomes. Similarly, the smaller f is the more ITD observations are extracted and the more computational load we have. A good compromise between low computational load, high rate, and reliability of ITD values was found for $W = 150$ ms and $f = 20$ ms. In order to save computational power, we discard those windows in which the average energy of the sound signal is below $E_A = 0.001$. Notice that this parameter could be controlled by a higher level module which would learn the characteristics of the scene and infer the level of background noise. We initialize $\sigma_n^2 = 10^{-9}$, since we found this value big enough to take into account the noise in the ITD values and small enough to discriminate speakers that are close to each other. The threshold τ_A has to take into account how many audio observations (K) are gathered during the current time interval Δt as well as the number of potential audible AV objects (N). For instance, if there is just one potential AV object, most of the audio observations should be assigned to it, whereas if there are three of them the audio observations may be distributed among them (in case all of them emit sounds). The threshold τ_A was experimentally set to $\tau_A = K/(N + 2)$. When there are more than 4 people speaking at the same time, the system is using both, the computing power of the NAO robot and the computation power of a i7 processor at 2.5 GHz.

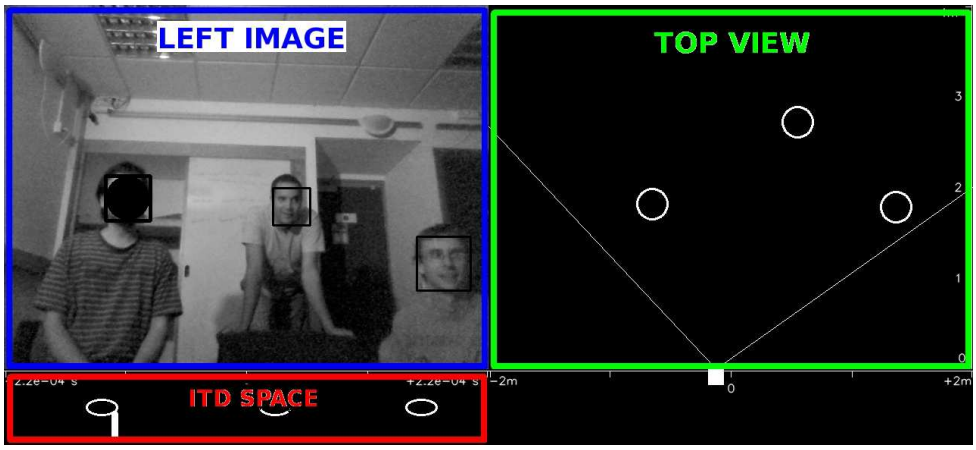


Figure 5: Snapshot of the visualization tool. The top-left (blue-framed) image is the original left image plus one bounding box per detected face. In addition, an intensity-coded circle appears when the speaker is active. The darker the color is, the higher the speaking probability is (this feature is exploited in Figure 10). The top-right (green-framed) image corresponds to the bird-view of the scene, in which each circle corresponds to a detected head. The bottom-left (red-framed) image represents the ITD space. The projected faces are represented by an ellipse and the histogram of extracted ITD values is plot.

8 Results

In order to evaluate the proposed approach, we ran three sets of experiments. First, we evaluated the Multimodal Inference method described in Section 6 on synthetic data. This allowed us to assess the quality of the model on a controlled scenario, where the feature extraction did not play any role. Second, we evaluated the *Motion-Guided Robot Hearing* method on a publicly available dataset, thus assessing the quality of the entire approach. Finally, we evaluated the *Face-Guided Robot Hearing* implemented on NAO, which proves that the proposed hybrid deterministic/probabilistic framework is suitable for robot applications.

In all our experiments we used a time interval of 6 visual frames, that is $\Delta t = 0.4$ s. During this time interval, approximately 2,000 HM3D observations and 20 auditory observations are extracted. A typical set of visual and auditory observations is shown in Figures 7 and 8. Indeed, Figure 7 focuses on the extraction of the HM3D features: the Harris interest point detection, filtered by motion, matched between images and reconstructed in 3D. Figure 8 shows the very same 3D features projected in to the ITD space. Also, the ITD values extracted during the same time interval are shown. These are the input features to the *Motion-Guided Robot Hearing* procedure. Notice that both auditory and visual data are corrupted by noise and by outliers. Visual data suffer from reconstruction errors either from wrong matches or from noisy detection. Auditory data suffer from reverberations, which enlarge the peaks' variances, or from sensor noise which is sparse along the ITD space.

To quantitatively evaluate the localization results, we compute a distance matrix between the detected clusters and the ground-truth clusters. The cluster-to-cluster distance corresponds to the Euclidean distance between cluster means. Let $\mathbf{D}$ be the distance matrix, then entry $D_{ij} = \|\mu_i - \hat{\mu}_j\|$ is the distance from the i^{th} ground-truth cluster to the j^{th} detected cluster. Next, we associate at most one ground-truth cluster to each detected cluster. The assignment procedure is as follows. For each detected cluster we compute its ground-truth nearest cluster. If it is not closer than a threshold τ_{loc} we mark it as a *false positive*, otherwise we assign the detected cluster to the ground-truth cluster. Then, for each ground-truth cluster we determine how many detected clusters are assigned to it. If there is none, we mark the ground-truth cluster as *false negative*. Finally, for each of the remaining ground-truth clusters, we select the closest (*true positive*) detected cluster among the ones assigned to the ground-truth cluster and we

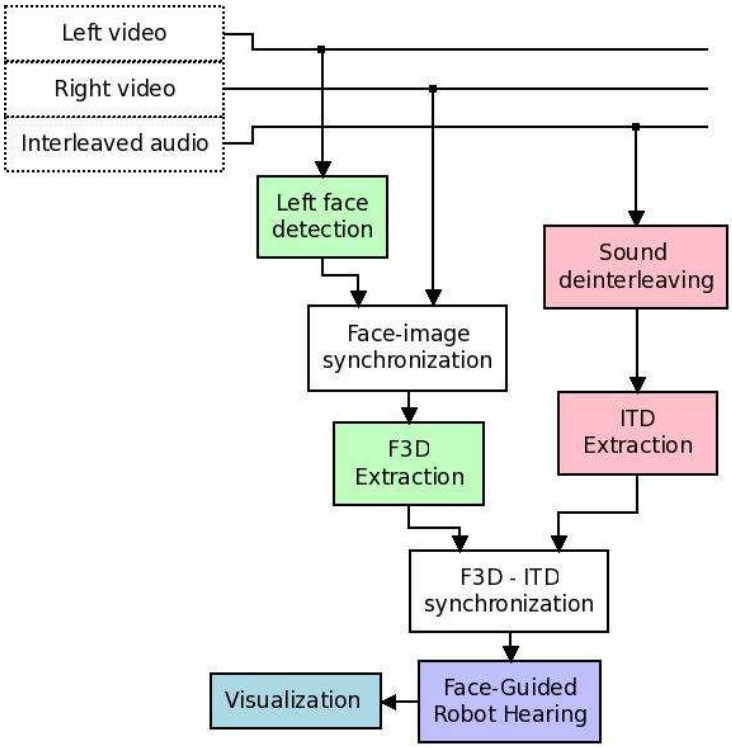


Figure 6: Modular structure of the *Face-Guided Robot Hearing* method using the RSB middleware. There are five types of modules: streaming & synchronization (white), visual processing (green), auditory processing (red), audiovisual fusion (purple) and visualization (blue).

mark the remaining ones as *false positives*. We can evaluate the localization error and the auditory state for those clusters that have been correctly detected . The localization error corresponds to the Euclidean distance between the means. Notice that by choosing τ_{loc}, we fix the maximum localization error allowed. The auditory state is counted as *false positive* if detected audible when silent, *false positive* if detected silent when audible and *true positive* otherwise. τ_{loc} was set to 0.35 m in all the experiments.

8.1 Results on Synthetic Data

Four synthetic sequences containing one to three AV objects were generated. These objects can move and they are not necessarily visible/audible along the entire sequence. Table 1 shows the visual evaluation of the method when tested with synthetic sequences. The sequence code name describes the dynamic character of the sequence (*Sta* means static and *Dyn* means dynamic) and the varying number of AV objects in the scene (*Con* means constant number of AV objects and *Var* means varying number of AV objects). The columns show different evaluation quantities: FP (*false positives*), i.e., AV objects found that do not really exist, FN (*false negatives*), i.e., present AV objects that were not found, TP (*true positives*) and ALE (average localization error). Recall that we can compute the localization error just for the true positives.

We observe that the dynamic nature of the scene and the variable number of speakers have different impact on the performance of the method. On one side, we remark that ALE increases for highly dynamic scenes. This is a natural effect given that we consider all observations gathered during Δt. Clearly, this integration window does not have any visible effect when the scene is static. On the other side, when the number of speakers varies, the number of TP decreases compared to the case with constant number of objects. This is due to the disagreement between the evaluation metric and the nature of the data. Indeed,

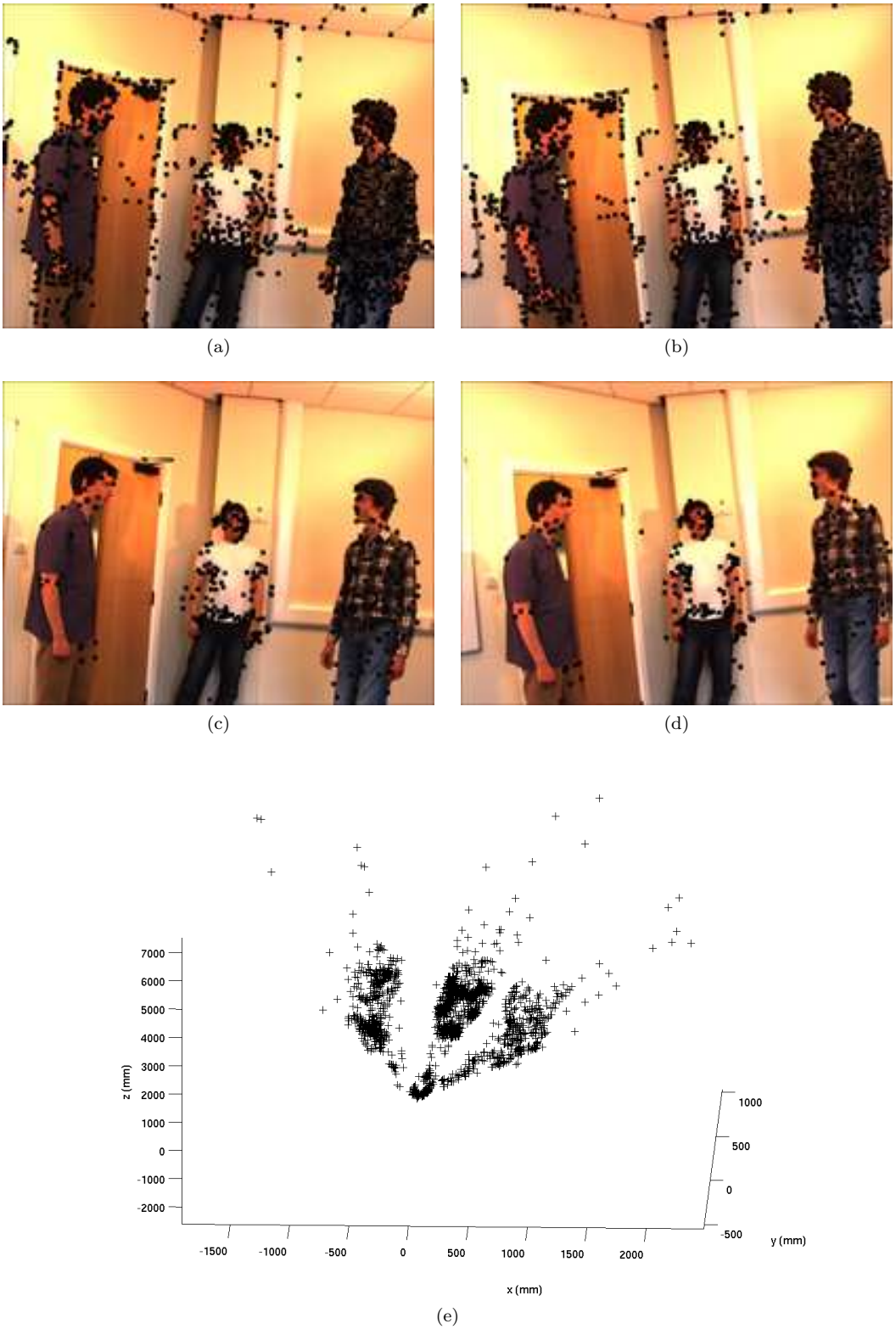
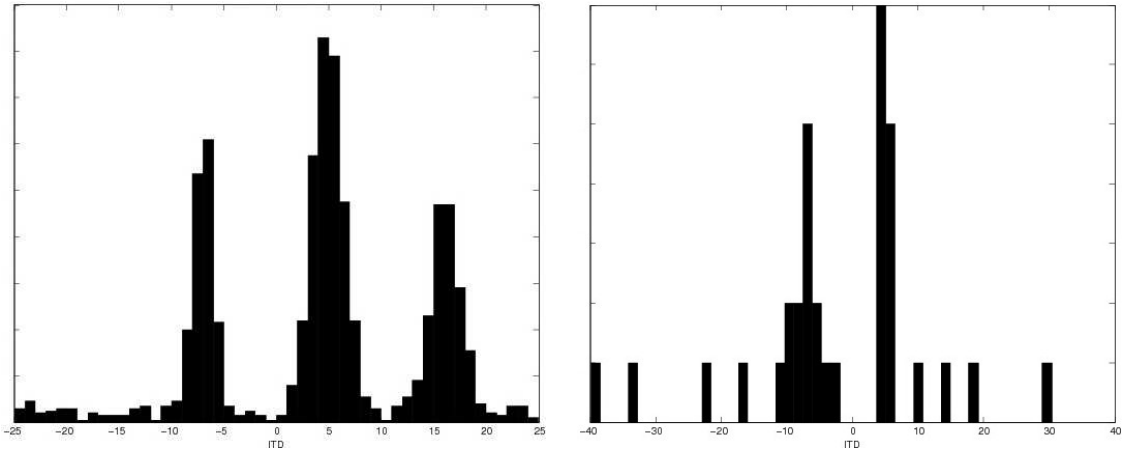


Figure 7: Interest points as detected in the left (a) and right (b) images. Dynamic interest points detected in the left (c) and the right (d) images. (e) HM3D visual observations, $\{\boldsymbol{v}_m\}_{m=1}^{M}$. Most of the background (hence static) points are filtered out from (a) to (c) and from (b) to (d). It is worth noticing that the reconstructed HM3D features suffer from reconstruction errors.



(a) (b)

Figure 8: Observation densities in the auditory space $\mathbb{A}$: (a) of the projected HM3D features, $\{\tilde{v}_m\}_{m=1}^{M}$, and (b) of the ITD features, $\{a_k\}_{k=1}^{K}$. In this particular example, we observe three moving objects (corresponding to the three people in the images). In addition, two of them are emitting sound (left and middle) and one is silent (right). We remark that auditory as well as visual observations are contaminated by noise (enlarging the Gaussian variances) and by outliers (uniformly distributed in the auditory feature space).

Table 1: Visual evaluation of results obtained with synthetic sequences. *Sta*/*Dyn* states for static or dynamic scene; the AV objects move or do not move. *Var*/*Con* states for varying or constant number of AV objects. FP stands for false positives, FN for false negatives, TP for true positives and ALE for average localization error (expressed in meters).

Seq.	FP	FN	TP	ALE [m]
StaCon	12	16 (3.9%)	392 (96.1%)	0.03
DynCon	37	36 (8.8%)	372 (91.2%)	0.10
StaVar	46	69 (30.1%)	160 (69.9%)	0.03
DynVar	40	82 (35.9%)	147 (64.1%)	0.11

we assume that the speakers are visible or not visible in a binary fashion. This assumption is extremely natural from a modeling point of view. However, the number of clusters is estimated statistically. As mentioned in Section 6.2, BIC is a consistent criterion, meaning that the BIC score converges to the right choice when the amount of data tends to infinity. In practical situations, not only the amount of data is finite, but also their effect can be progressive. In our particular case, the appearance of one speaker in the scene is done gradually. For the first frames, the speaker generates a small amount of visual observations since it is *entering* the scene. Therefore, BIC will choose the wrong number of clusters for several frames, thus producing several FN. Similarly, when the object *leaves* the scene, the behavior of BIC will produce some FP.

Table 2 shows the auditory evaluation of the method when tested with synthetic sequences. The remarkable achievement is the high number of right detections, around 80%, in all cases. This means that neither the dynamic character of the scene nor the fact that the number of AV objects varies have an impact on sound detection. It is also true that the number of false positives is large in all the cases.

Table 2: Audio evaluation of the results obtained with synthetic sequences. *Sta*/*Dyn* states for static or dynamic scene; the AV objects move or do not move. *Var*/*Con* states for varying or constant number of AV objects.

Seq.	FP	FN	TP
StaCon	161	33 (13.4%)	214 (86.6%)
DynCon	127	43 (16.7%)	215 (83.3%)
StaVar	53	33 (18.8%)	143 (81.2%)
DynVar	56	34 (19.7%)	139 (80.3%)

8.2 Results with Datasets

The *Motion-Guided Robot Hearing* method was tested on the CTMS3 sequence of the CAVA data set [Arnaud 08]. The CAVA (*computational audiovisual analysis*) data set was specifically recorded to test various real-world audiovisual scenarios. The CTMS3 sequence[3] consists on three people freely moving in a room and taking speaking turns. Two of them count in English (one, two, three, ...) while the third one counts in Chinese. The recorded signals, both auditory and visual, enclose the difficulties found in natural situations. Hence, this is a very challenging sequence: People come in and out the visual field of the two cameras, hide each other, etc. Aside from the speech sounds, there are acoustic reverberations and non-speech sounds such as those emitted by foot steps and clothe chafing. Occasionally, two people speak simultaneously.

Figure 9 shows the results obtained with nine time intervals chosen to show both successes and failures of our method and to allow to qualitatively evaluate it. Figure 9a shows one extreme case, in which the distribution of the HM3D observations associated to the person with the white T-shirt is clearly not Gaussian. Figure 9b shows a failure of the *ridgeline* method, used to merge Gaussian components, where two different clusters are associated into one. Figure 9c is an example with too few observations. Indeed, the BIC points as optimal the model with no AV objects, thus considering all the observations to be outliers. Figure 9d clearly shows that our approach cannot deal with occluded objects, because of the instantaneous processing of robocentric data, the person occluded will never be detected. Figures 9e, 9f and 9g are examples of success. The three speakers are localised and their auditory status correctly guessed. However, the localisation accuracy is not good in these cases, because one or more covariance matrices are not correctly estimated. The grouping of AV observations is, then, not well conducted. Finally, Figures 9h and 9i show two case in which the *Motion-Guided Robot Hearing* algorithms works perfectly, three people are detected and their speaking activity is correctly assessed from the ITD observations. In average, the method correctly detected 187 out of 213 objects (87.8%) and correctly detected the speaking state in 88 cases out of 147 (59.9%).

8.3 Results with NAO

To validate the *Face-Guided Robot Hearing* method using NAO, we perform a set of experiments with five different scenarios. The scenarios, recorded in a room of size 5×5 meters with one sofa and three chairs. The five scenarios are designed to test the algorithm in different conditions in order to identify its limitations. Each scenario is repeated several times and consists on people counting from one up to sixteen.

In scenario **S1**, only one person is in the room sitting in front of the robot and counting. In the rest of the scenarios (**S2**-**S5**) three persons are in the room. People are not always in the field of view of the cameras and sometimes they move. In scenario **S2** three persons are sitting and counting alternatively one after the other. The configuration of scenario **S3** is similar to the one of **S2**, but one person is

[3] http://perception.inrialpes.fr/CAVA_Dataset/Site/data.html#CTMS3

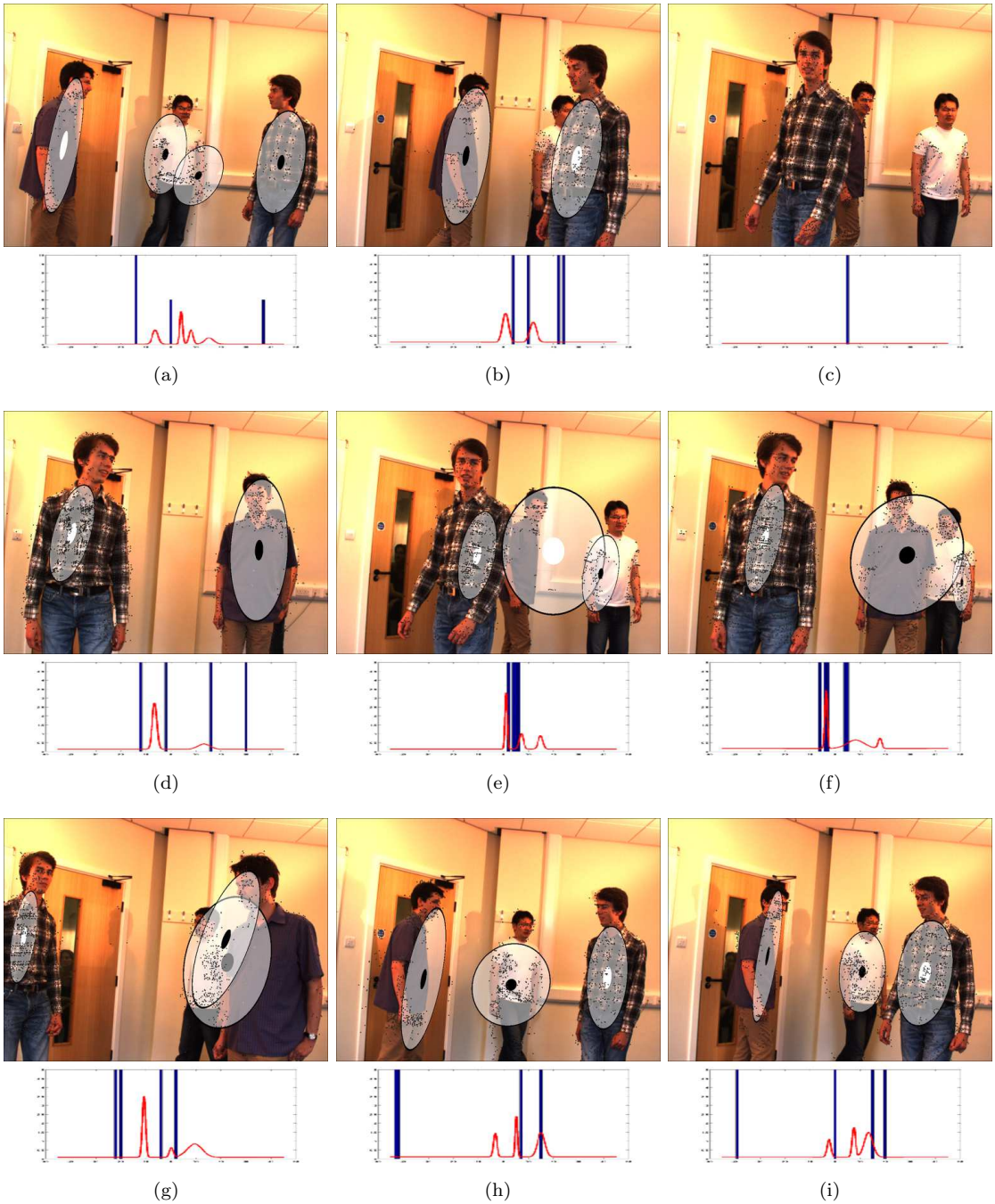


Figure 9: Results obtained with the CTMS3 sequence from the CAVA data set. The ellipses correspond to the 3D covariance matrices projected onto the image. A circle at each ellipse center illustrates the auditory activity: speaker emitting a sound (white) or being silent (black) during each time interval. The plots below the images show the interaural time difference observations as well as the estimated 1D GMM.

	FP	FN	TP
S1	13	23 (13.4%)	149 (86.6%)
S2	22	31 (14.9%)	176 (85.1%)
S3	19	20 (11.3%)	157 (88.7%)
S4	37	12 (6.7%)	166 (93.3%)
S5	53	32 (19.0%)	136 (81.0%)

Table 3: Quantitative evaluation of the proposed approach for the five scenarios. The columns represent, in order: the amount of correct detections (CD), the amount of false positives (FP), the amount of false negatives (FN) and the total number of counts (Total).

standing instead of sitting. These two scenarios are useful to determine the precision of the ITDs and experimentally see if the difference of height (elevation) affects the quality of the extracted ITDs. The scenario **S4** is different from **S2** and **S3** because one of the actors is outside the field of view. This scenario is used to test if people speaking outside the field of view affect the performance of the algorithm. In the last scenario (**S5**) the three people are in the field of view, but they count and speak independently of the other actors. Furthermore, one of them is moving while speaking. With **S5**, we aim to test the robustness of the method to dynamic scenes.

In Figure 10 we show several snapshots of our visualization tool. These frames are selected from the different scenarios aiming to show both the successes and the failures of the implemented system. Figure 10a shows an example of perfect alignment between the ITDs and the mapped face, leading to a high speaking probability. A similar situation is presented in Figure 10b, in which among the three people, only one speaks. A failure of the ITD extractor is shown in Figure 10c, where the actor in the left is speaking, but no ITDs are extracted. In Figure 10d we can see how the face detector does not work correctly: two faces are missing, one because of the great distance between the robot and the speaker, and the other because it is partially out of the field of view. Figure 10e shows a snapshot of an AV-fusion failure, in which the extracted ITDs are not significant enough to set a high speaking probability. The Figure 10f, Figure 10g and Figure 10h show the effect of reverberations. While in Figure 10h we see that the reverberations lead to the wrong conclusion that the actor on the right is speaking, we also see that the statistical framework is able to handle reverberations (Figure 10f and Figure 10g), hence demonstrating the robustness of the proposed approach.

Table 3 shows the results obtained on scenarios (that were manually annotated). First of all we notice the small amount of false negatives: the system misses very few speakers. A part from the first scenario (easy conditions), we observe some false positives. These false positives are due to reverberations. Indeed, we notice how the percentage of FP is severe in **S5**. This is due to the fact that high reverberant sounds (like hand claps) are also present in the audio stream of this scenario. We believe that an ITD extraction method more robust to reverberations will lead to more reliable ITD values, which in turn will lead to a better active speaker detector. It is also worth to notice that actors in different elevations and non-visible actors do not affect the performance of the proposed system, since the results obtained in scenarios **S2** to **S4** are comparable.

9 Conclusions and Future Work

This paper introduces a multimodal hybrid probabilistic/deterministic framework for simultaneous detection and localization of speakers. On one hand, the deterministic component takes advantage of the geometric and physical properties associated with the visual and auditory sensors: the audiovisual mapping $(\overline{\mathcal{A}} \circ \mathcal{V}^{-1})$ allows us to transform the visual features from the 3D space to a 1D auditory space. On the other hand, the probabilistic model deals with the observation-to-speaker assignments, the noise and the outliers. We propose a new multimodal clustering algorithm based on a 1D Gaussian mixture model, an initialization procedure, and a model selection procedure based on the BIC score. The method

(a) **S1** (b) **S2**

(c) **S4** (d) **S5**

(e) **S5** (f) **S2**

(g) **S3** (h) **S3**

Figure 10: Snapshots of the visualization tool. Frames selected among the five scenarios to show the method's strengths and weaknesses. The faces' bounding box are shown superposed to the original image (top-left). The bird-view of the scene is shown in the top-right part of each subimage. The histogram of ITD values as well as the projected faces are shown in the bottom-left. See Section 7.3 for how to interpret the images above.

is validated on a humanoid robot and interfaced through the RSB middleware leading to a platform-independent implementation.

The main novelty of the approach is the visual guidance. Indeed, we derived two EM procedures for *Motion-Guided* and *Face-Guided* robot hearing. Both algorithms provide the number of speakers, localize them and ascertain their speaking status. In other words, we show how one of the two modalities can be used to supervise the clustering process. This is possible thanks to the audiovisual calibration procedure that provides an accurate projection mapping ($\overline{\mathcal{A}} \circ \mathcal{V}^{-1}$). The calibration is specifically designed for robotic usage since it requires very few data, it is long-lasting and environment-independent.

The presented method solves several open methodological issues: (i) it fuses and clusters visual and auditory observations that lie in physically different spaces with different dimensionality, (ii) it models and estimates the object-to-observation assignments that are not known, (iii) it handles noise and outliers mixed with both visual and auditory observations whose statistical properties change across modalities, (iv) it weights the relative importance of the two types of data, (v) it estimates the number of AV objects that are effectively present in the scene during a short time interval and (vi) it gauges the position and speaking state of the potential speakers.

One prominent feature of our algorithm is its robustness. It can deal with various kinds of perturbations, such as the noise and outlier encountered in unrestricted physical spaces. We illustrated the effectiveness and robustness of our algorithm using challenging audiovisual sequences from a publicly available data set as well as using the humanoid robot NAO in regular indoor environments. We demonstrated good performance on different scenarios involving several actors, moving actors and non-visible actors. Interfaced by means of the RSB middleware, the *Face-Guided Robot Hearing* method processes the audiovisual data flow from two microphones mounted inside the head of a companion robot with noisy fans, and two cameras which deliver synchronized image sequences at 17FPS.

Since all the sensors are onboard of the robot, the proposed method delivers agent-centered speaker localization. Clearly, ego-motion is a problem with this kind of approaches, because if the robot moves while it gathers data, the localization performance is naturally affected. While not done in this paper, feedback from the robot's proprioceptive sensors can be used to compensate for ego-motion.

There are several possible ways to improve and to extend the proposed method method. Our current implementation relies more on the visual data than on the auditory data, although there are many situations where the auditory data are more reliable. The problem of how to better weight the relative importance of the two modalities is an interesting topic. For instance, estimating two sets of parameters, i.e., fitting two mixture models, one per modality, could lead to a richer statistical description of the observed data. Similarly, speakers who are not in the visual field of view could be located and detected by giving more weight to the auditory modality. However, both cases require that the auditory observations are more reliable, possibly by using more than two microphones. Our algorithm can also accommodate other types of visual cues, such as 2D or 3D optical flow [Cech 11], body detectors, etc., or auditory cues, such as interaural level difference (ILD) and voice activity detection (VAD). In this paper we used one pair of microphones, but the method can be easily extended to several microphone pairs, e.g., [Alameda-Pineda 14]. Each microphone pair yields one ITD space and combining these 1D spaces would provide a much more robust algorithm. Finally, another interesting direction of research is to design a dynamic model that would allow to initialize the parameters in one time interval based on the information extracted in several previous time intervals. Such a model would necessarily involve dynamic model selection, and would certainly help to guess the right number of AV objects, particularly in situations where a cluster is occluded but still in the visual scene, or a speaker voice is highly interfered by another speaker/sound source. Moreover, dynamic model selection may be extended to provide for audiovisual tracking capabilities, such as to enhance the temporal coherence of the perceived audiovisual scene.

Acknowledgments

This work was funded by the HUMAVIPS FP7 European Project FP7-ICT-247525.

References

[Alameda-Pineda 11] X. Alameda-Pineda, V. Khalidov, R. Horaud & F. Forbes. *Finding audio-visual events in informal social gatherings*. In Proceedings of the ACM/IEEE International Conference on Multimodal Interaction, 2011.

[Alameda-Pineda 14] X. Alameda-Pineda & R. Horaud. *A Geometric Approach to Sound Source Localization from Time-Delay Estimates*. IEEE Transactions on Audio, Speech and Language Processing, vol. 22, no. 6, pages 1082–1095, June 2014.

[Anastasio 00] T. J. Anastasio, P. E. Patton & K. E. Belkacem-Boussaid. *Using Bayes' Rule to Model Multisensory Enhancement in the Superior Colliculus*. Neural Computation, vol. 12, no. 5, pages 1165–1187, 2000.

[Arnaud 08] E. Arnaud, H. Christensen, Y.-C. Lu, J. Barker, V. Khalidov, M. Hansard, B. Holveck, H. Mathieu, R. Narasimha, E. Taillant, F. Forbes & R. P. Horaud. *The CAVA corpus: synchronised stereoscopic and binaural datasets with head movements*. In Proceedings of the ACM/IEEE International Conference on Multimodal Interfaces, 2008. http://perception.inrialpes.fr/CAVA_Dataset/.

[Barker 09] J. Barker & X. Shao. *Energetic and Informational Masking Effects in an Audiovisual Speech Recognition System*. Audio, Speech, and Language Processing, IEEE Transactions on, vol. 17, no. 3, pages 446–458, 2009.

[Barzelay 07] Z. Barzelay & Y. Schechner. *Harmony in Motion*. In CVPR, 2007.

[Beal 03] M. Beal, N. Jojic & H. Attias. *A graphical model for audiovisual object tracking*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 7, pages 828–836, 2003.

[Besson 08a] P. Besson & M. Kunt. *Hypothesis testing for evaluating a multimodal pattern recognition framework applied to speaker detection*. Journal of NeuroEngineering and Rehabilitation, vol. 5, no. 1, page 11, 2008.

[Besson 08b] P. Besson, V. Popovici, J. Vesin, J. Thiran & M. Kunt. *Extraction of Audio Features Specific to Speech Production for Multimodal Speaker Detection*. Multimedia, IEEE Transactions on, vol. 10, no. 1, pages 63 –73, jan. 2008.

[Bishop 06] C. M. Bishop. Pattern recognition and machine learning (information science and statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[Butz 05] T. Butz & J.-P. Thiran. *From error probability to information theoretic (multimodal) signal processing*. Signal Process., vol. 85, no. 5, pages 875–902, May 2005.

[Calvert 04] G. Calvert, C. Spence & B. E. Stein. The handbook of multisensory processes. MIT Press, 2004.

[Cech 11] J. Cech, J. Sanchez-Riera & R. Horaud. *Scene Flow Estimation by Growing Correspondence Seeds*. In CVPR 2011 - IEEE Conference on Computer Vision and Pattern Recognition, pages 3129–3136, Colorado Springs, United States, June 2011. IEEE Computer Society.

[Checka 04] N. Checka, K. Wilson, M. Siracusa & T. Darrell. *Multiple person and speaker activity tracking with a particle filter*. In Proc. of IEEE Conference on Acoustics, Speech, and Signal Processing, pages 881–884. IEEE, 2004.

[Christensen 07] H. Christensen, N. Ma, S. Wrigley & J. Barker. *Integrating Pitch and Localisation Cues at a Speech Fragment Level*. In Proc. of Interspeech, pages 2769–2772, 2007.

[Cristani 07] M. Cristani, M. Bicego & V. Murino. *Audio-visual event recognition in surveillance video sequences*. IEEE Transactions on Multimedia, vol. 9, no. 2, pages 257–267, 2007.

[Davies 79] D. Davies & D. Bouldin. *A Cluster Separation Measure*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. PAMI-1, no. 2, pages 224–227, January 1979.

[Dempster 77] A. Dempster, N. Laird & D. Rubin. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, no. 1, pages 1–38, 1977.

[Gatica-Perez 07] D. Gatica-Perez, G. Lathoud, J.-M. Odobez & I. McCowan. *Audiovisual probabilistic tracking of multiple speakers in meetings*. IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 2, pages 601–616, 2007.

[Ghazanfar 06] A. A. Ghazanfar & C. E. Schroeder. *Is neocortex essentially multisensory?* Transactions on Cognitive Neuroscience, vol. 10, page 278285, 2006.

[Gurban 06] M. Gurban. *Multimodal speaker localization in a probabilistic framework*. In In Proc. of EUSIPCO, 2006.

[Harris 88] C. Harris & M. Stephens. *A Combined Corner and Edge Detector*. In Proc. of Fourth Alvey Vision Conference, pages 147–151, 1988.

[Hartley 04] R. I. Hartley & A. Zisserman. Multiple view geometry in computer vision. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[Haykin 05] S. Haykin & Z. Chen. *The Cocktail Party Problem*. Journal on Neural Compututation, vol. 17, pages 1875–1902, September 2005.

[Hennig 10] C. Hennig. *Methods for merging Gaussian mixture components*. Advances in Data Analysis and Classification, vol. 4, pages 3–34, 2010. 10.1007/s11634-010-0058-3.

[Hospedales 08] T. Hospedales & S. Vijayakumar. *Structure Inference for Bayesian Multisensory Scene Understanding*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 12, pages 2140–2157, 2008.

[Itohara 11] T. Itohara, T. Otsuka, T. Mizumoto, T. Ogata & H. G. Okuno. *Particle-Filter Based Audio-Visual Beat-Tracking for Music Robot Ensemble with Human Guitarist*. In IROS, 2011.

[Itohara 12] T. Itohara, K. Nakadai, T. Ogata & H. G. Okuno. *Improvement of Audio-Visual Score Following in Robot Ensemble with Human Guitarist*. In IEEE-RAS International Conference on Humanoid Robots, 2012.

[Keribin 00] C. Keribin. *Consistent Estimation of the Order of Mixture Models*. Sankhya Series A, vol. 62, no. 1, pages 49–66, 2000.

[Khalidov 08] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud & R. Horaud. *Detection and Localization of 3D Audio-Visual Objects Using Unsupervised Clustering*. In ICMI '08, pages 217–224, New York, NY, USA, 2008. ACM.

[Khalidov 11] V. Khalidov, F. Forbes & R. Horaud. *Conjugate Mixture Models for Clustering Multimodal Data*. Neural Computation, vol. 23, no. 2, pages 517–557, February 2011.

[Khalidov 13] V. Khalidov, F. Forbes & R. Horaud. *Alignment of Binocular-Binaural Data Using a Moving Audio-Visual Target*. In IEEE Workshop on Multimedia Signal Processing, 2013.

[Kidron 05] E. Kidron, Y. Y. Schechner & M. Elad. *Pixels that Sound*. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01, CVPR '05, pages 88–95, Washington, DC, USA, 2005. IEEE Computer Society.

[Kidron 07] E. Kidron, Y. Schechner & M. Elad. *Cross-Modal Localization via Sparsity*. Trans. Sig. Proc., vol. 55, no. 4, pages 1390–1404, April 2007.

[Kim 07] H. Kim, J. suk Choi & M. Kim. *Human-Robot Interaction in Real Environments by Audio-Visual Integration*. International Journal of Control, Automation and Systems, vol. 5, no. 1, pages 61–69, 2007.

[King 09] A. J. King. *Visual influences on auditory spatial learning*. Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 364, no. 1515, pages 331–339, 2009.

[Liu 08] M. Liu, Y. Fu, & T. S. Huang. *An Audio-Visual Fusion Framework with Joint Dimensionality Reduction*. In Proceedings of the IEEE International Conference on Audio Speech and Signal Processing, 2008.

[Liu 11] C. Liu, J. Yuen & A. Torralba. *Nonparametric Scene Parsing via Label Transfer*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 33, no. 12, pages 2368–2382, Dec 2011.

[Miller 03] D. Miller & J. Browning. *A mixture model and EM-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabeled data sets*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 11, pages 1468 – 1483, nov. 2003.

[Nakadai 04] K. Nakadai, D. Matsuura, H. G. Okuno & H. Tsujino. *Improvement of recognition of simultaneous speech signals using AV integration and scattering theory for humanoid robots*. Speech Communication, pages 97–112, 2004.

[Nakamura 11] T. Nakamura, T. Nagai & N. Iwahashi. *Bag of multimodal LDA models for concept formation*. In Robotics and Automation (ICRA), 2011 IEEE International Conference on, pages 6233–6238, 2011.

[Naqvi 10] S. Naqvi, M. Yu & J. Chambers. *A Multimodal Approach to Blind Source Separation of Moving Sources*. Selected Topics in Signal Processing, IEEE Journal of, vol. 4, no. 5, pages 895–910, 2010.

[Natarajan 12] P. Natarajan, S. Wu, S. N. P. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad & P. Natarajan. *Multimodal feature fusion for robust event detection in web videos*. In Proceedings of the IEEE International Conference on Computer Vision an Pattern Recognition, 2012.

[Nigam 00] K. Nigam, A. McCallum, S. Thrun & T. Mitchell. *Text Classification from Labeled and Unlabeled Documents using EM*. Machine Learning, vol. 39, no. 2-3, pages 103–134, 2000.

[Noulas 12] A. Noulas, G. Englebienne & B. Krose. *Multimodal Speaker Diarization*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 34, no. 1, pages 79–93, 2012.

[Perez 04] P. Perez, J. Vermaak & A. Blake. *Data Fusion for Visual Tracking with Particles.* Proceedings of IEEE, vol. 92, no. 3, pages 495–513, 2004.

[Ray 05] S. Ray & B. G. Lindsay. *The topography of multivariate normal mixtures.* The Annals of Statistics, vol. 33, no. 5, pages 2042–2065, 2005.

[ROS 12] ROS. *message_filters/ApproximateTime.* http://www.ros.org/wiki/message_filters/ApproximateTime, 2012. accessed: 06/21/02012.

[Sanchez-Riera 12] J. Sanchez-Riera, X. Alameda-Pineda, J. Wienke, A. Deleforge, S. Arias, J. Čech, S. Wrede & R. P. Horaud. *Online Multimodal Speaker Detection for Humanoid Robots.* In IEEE International Conference on Humanoid Robotics, Osaka, Japan, November 2012.

[Schwarz 78] G. Schwarz. *Estimating the dimension of a model.* The Annals of Statistics, vol. 6, pages 461–464, 1978.

[Senkowski 08] D. Senkowski, T. R. Schneider, J. J. Foxe & A. K. Engel. *Crossmodal binding through neural coherence: Implications for multisensory processing.* Trends in Neuroscience, vol. 31, no. 8, page 401409, 2008.

[Šochman 05] J. Šochman & J. Matas. *WaldBoost – Learning for Time Constrained Sequential Detection.* In Proceedings of the IEEE Computer Vision and Pattern Recognition, 2005.

[Wienke 11] J. Wienke & S. Wrede. *A Middleware for Collaborative Research in Experimental Robotics.* In 2011 IEEE/SICE Internatinal Symposium on System Integration, Kyoto, Japan, 2011. IEEE, IEEE.

[Yoshida 12] T. Yoshida & K. Nakadai. *Audio-visual voice activity detection based on an utterance state transition model.* Advanced Robotics, vol. 26, no. 10, pages 1183–1201, 2012.