Egocentric Audio-Visual Scene Analysis. A Machine Learning and Signal Processing Approach

Xavier Alameda-Pineda

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques-Informatique**

Arrêté ministérielle :

Présentée par

Xavier Alameda-Pineda

Thèse dirigée par **Radu Horaud**

préparée au sein **INRIA Grenoble Rhône-Alpes et Université Joseph Fourier**
et de **l'École Doctorale de Mathématiques, Sciences et Technologie de l'Information et Informatique**

Egocentric Audio-Visual Scene Analysis

A Machine Learning and Signal Processing Approach

Thèse soutenue publiquement le ,
devant le jury composé de :

M. Daniel Gatica-Pérez
IDIAP Research Institute, Rapporteur
M. Nicu Sebe
Professeur University of Trento, Rapporteur
M. Josep Ramon Casas
Professeur Universitat Politècnica de Catalunya, Examinateur
M. Laurent Girin
Professeur Grenoble INP, Président
Mme. Florence Forbes
Directeur de recherche à l'INRIA Grenoble Rhône-Alpes, Co-Directeur de thèse
M. Radu Horaud
Directeur de recherche à l'INRIA Grenoble Rhône-Alpes, Directeur de thèse

Analyse Égocentrique de Scènes Audio-Visuelles

Une Approache par Apprentissage Automatique et Traitement du Signal

Xavier Alameda-Pineda

15 Octobre 2013

Résumé

Depuis les vingt dernières années, l'industrie a developpé plusieurs produits commerciaux dotés de capacités auditives et visuelles. La grand majorité de ces produits est composée d'un camescope et d'un microphone embarqué (téléphones portables, tablettes, etc). D'autres, comme la Kinect, sont équipés de capteurs de profondeur et/ou de petits réseaux de microphones. On trouve également des téléphones portables dotés d'un système de vision stéréo. En même temps, plusieurs systèmes orientés recherche sont apparus (par exemple, le robot humanoïde NAO). Du fait que ces systèmes sont compacts, leurs capteurs sont positionés près les uns des autres. En conséquence, ils ne peuvent pas capturer la scène complète, mais qu'un point de vue très particulier de l'interaction sociale en cours. On appelle cela "Analyse Égocentrique de Scènes Audio-Visuelles".

Cette thèse contribue à cette thématique de plusieurs façons. D'abord, en fournissant une base de données publique qui cible des applications comme la reconnaissance d'actions et de gestes, localisation et suivi d'interlocuteurs, analyse du tour de parole, localisation de sources auditives, etc. Cette base a fait l'objet d'une publication [Alameda-Pineda 12b] et a été utilisé en dedans et en dehors de cette thèse. Nous avons aussi travaillé le problème de la detection

d'événements audio-visuels. Nous avons montré comme la confiance en une des modalités (issue de la vision en l'occurrence), peut être modélisée pour biaser la méthode, en donnant lieu à un algorithme d'espérance-maximisation visuellement supervisé [Alameda-Pineda 11]. Cette dernière publication a eu le "Outstanding Paper Award" à la conférence ICMI'11. Ensuite, nous avons modifié l'approche pour cibler la détection audio-visuelle d'interlocuteurs en utilisant le robot humanoïde NAO. Les détails sont publiés dans [Sanchez-Riera 12b]. En parallèle aux travaux en détection audio-visuelle d'interlocuteurs, nous avons developpé une nouvelle approche pour la reconnaissance audio-visuelle de commandes. Nous avons evalué la qualité de plusieurs indices et classeurs [Sanchez-Riera 12a], et confirmé que l'utilisation des données auditives et visuelles favorise la reconnaissance, en comparaison aux méthodes qui n'utilisent que l'audio ou que la vidéo. Plus tard, dans [Alameda-Pineda 13c], nous avons cherché la meilleure méthode pour des ensembles d'entraînement minuscules (5-10 observations par catégorie). Il s'agit d'un problème intéressant, car les systèmes réels ont besoin de s'adapter très rapidement et d'apprendre de nouvelles commandes. Ces systèmes doivent être opérationnels avec très peu d'échantillons pour l'usage publique. Pour finir, nous avons contribué au champ de la localisation de sources sonores, dans le cas particulier des réseaux coplanaires de microphones. C'est une problématique importante, car la géometrie du réseau est arbitraire et inconnue. En conséquence, cela ouvre la voie pour travailler avec des réseaux de microphones dynamiques, qui peuvent adapter leur géométrie pour mieux répondre à certaines tâches. De plus, la conception des produits commerciaux peut être contrainte de façon que les réseaux linéaires ou circulaires ne sont pas bien adaptés. Dans un premier temps, nous avons publié le cadre général et un algorithme dans [Alameda-Pineda 12a]. Nous avons présenté la totalité du modèle géometrique et une méthode plus robuste dans [Alameda-Pineda 13b].

En conclusion, nous avons abordé différents problèmes concernant l'analyse de scènes audio-visuelles, avec des données égocentriques. Les méthodes proposées font partie du domaine de l'apprentissage statistique et discriminative ainsi que du domaine de l'optimisation non-linéaire; toujours appuyées sur une base solide de traitement du signal. Les résultats et les contributions ont été publiés dans de conférences et journaux internationaux de très haut niveau.

Egocentric Audio-Visual Scene Analysis

A Machine Learning and Signal Processing Approach

Xavier Alameda-Pineda

October, 15^{th}, 2013

Abstract

Along the past two decades, the industry has developed several commercial products with audio-visual sensing capabilities. Most of them consists on a video-camera with an embedded microphone (mobile phones, tablets, etc). Other, such as Kinect, include depth sensors and/or small microphone arrays. Also, there are some mobile phones equipped with a stereo camera pair. At the same time, many research-oriented systems became available (e.g., humanoid robots such as NAO). Since all these systems are small in volume, their sensors are close to each other. Therefore, they are not able to capture de global scene, but one point of view of the ongoing social interplay. We refer to this as "Egocentric Audio-Visual Scene Analysis".

This thesis contributes to this field in several aspects. Firstly, by providing a publicly available data set targeting applications such as action/gesture recognition, speaker localization, tracking and diarisation, sound source localization, dialogue modelling, etc. This work has been published in [Alameda-Pineda 12b] and used later on inside and outside the thesis. We also investigated the problem of AV event detection. Published in [Alameda-Pineda 11], we show how the trust on one of the modalities (visual to be precise) can be modelled and used to bias the method, leading to a visually-supervised EM algorithm (ViSEM). This paper got the Outstanding Paper Award at ICMI'11. Afterwards we modified the approach to target audio-visual speaker detection yielding to an on-line method working in the humanoid robot NAO. The details can be found in [Sanchez-Riera 12b].

In parallel to the work on audio-visual speaker detection, we developed a new approach for audio-visual command recognition. In [Sanchez-Riera 12a] we explored different features and classifiers and confirmed that the use of audio-visual data increases the performance when compared to auditory-only and to video-only classifiers. Later, in [Alameda-Pineda 13c] we sought for the best method using tiny training sets (5-10 samples per class). This is interesting because real systems need to adapt and learn new commands from the user. Such systems need to be operational with a few examples for the general public usage. Finally, we contributed to the field of sound source localization, in the particular case of non-coplanar microphone arrays. This is interesting because the geometry of the microphone can be any. Consequently, this opens the door to dynamic microphone arrays that would adapt their geometry to fit some particular tasks. Also, because the design of commercial systems may be subject to certain constraints for which circular or linear arrays are not suited. At a first stage we published in [Alameda-Pineda 12a], where we presented the general framework and one algorithm working up to a certain extent. Later on we submitted the full geometric model together with a much more solid algorithm in [Alameda-Pineda 13b].

In summary, we face different real problems of AV scene analysis using egocentric data. Methods vary from statistical and discriminative learning to non-linear programming, always on top of a solid basis of signal processing. Results and contributions have been peer-reviewed by the international research community and published in international top conferences and journals.

Anàlisi Egocèntrica d'Escenes Audio-Visuals

Estratègies Basades en l'Aprenentatge Automàtic i en el Processat del Senyal

Xavier Alameda i Pineda

15 d'Octubre del 2013

Resum

Durant les darreres dues dècades, la indústria ha desenvolupat diversos productes comercials amb habilitats sensorials auditives i visuals. La gran majoria consisteixen en una càmera i un micròfon encastat (telèfons mòbils, tablets, etc). D'altres, com el Kinect, inclouen sensors de profunditat i/o arrays petits de micròfons. A més a més, hi ha alguns telèfons mòbils amb un equip d'estèreo-visió. Al mateix temps, nombrosos sistemes orientats a la recerca han esdevingut disponibles (e.g., robots humanoides com NAO). Atès que aquests sistemes són petits, els sensors estan a prop els uns dels altres. Per tant, no són capaços de capturar la globalitat de la escena, només un punt de vista interactions socials del moment. Ens referim a això com "Anàisi Egocèntrica d'Escenes Audio-Visuals".

Aquesta tèsi contribueix en diversos aspectes. Primerament, fent pública una base de dades per a applicacions com el recineixement d'accions i gests, la localització i seguit d'interlocutors, l'anàlisi del torn de paraula, la localització de fonts sonores, etc. Això es va publicar à [Alameda-Pineda 12b] i s'ha utilitzat dins i fora de la present tèsi. També hem investigat el problema de la detecció i localització d'events audio-visuals. A [Alameda-Pineda 11], vam mostrar que la confiança en una de les modalitats (la visual en el nostre cas) es pot modelar

i usar per a esbiaixar la metodologia, donant lloc a un algorisme d'esperança-maximització visualment supervisat. Aquest article obtingué el "Outstanding Paper Award" a la conferència ICMI'11. Després, vam modificar l'estratègia per applicar el mètode a la deteció d'interlocutors, produint un algorisme de processat on-line executable en el robot humanoïde NAO. Els detalls es poden trobar a [Sanchez-Riera 12b]. Paral·lelament, als treballs de detecció audio-visual d'interlocutors, hem desenvolupat une nova estratègia per al reconeixement audio-visual de commandes. A [Sanchez-Riera 12a] hem explorat diverses característiques i classificadors i hem confirmat que l'utilització de dades audio-visuals millora la qualitat de reconeixement en comparació amb els mètodes que utilitzent només l'àudio o el vídeo. Més tard, a [Alameda-Pineda 13c], hem cercat el millor metode utilitzant conjunts d'entrenament minúsculs (5-10 observacions per categoria). Això és interessant perquè els sistemes reals s'han d'adaptar i aprendre noves commandes a partir d'exemples de l'usuari. L'utilització massiva d'aquests sistemes, obliga que siguin operacionals amb molt pocs exemples. Per acabar, hem contribuit en l'àrea de la localització de fons sonores, en el cas particular de xarxes no-coplanars de micròfons. El fet que la geometria de la xarxa és arbitraria fa el problema interessant. A més a més, obre la porta a xarxes de micròfons dinàmiques, que podrien adaptar la seva geometria per a satisfer criteris diversos o taskes precises. D'altra banda, el disseny de sistemes comercials pot imposar restriccions, per a les quals les xarxes linears o circulars no són apropiades. D'entrada vam publicar el marc general i un algorisme de localització a [Alameda-Pineda 12a]. Més tard, el model geometric complet i una metodologia de localització molt més robusta es van presentar a [Alameda-Pineda 13b].

Per resumir, ens hem enfrontat a diversos problemes reals de l'anàlis audio-visual de la escena, utilizant dades egocèntriques. Els mètodes usats van des de l'aprenentatge estadístic i discriminatiu fins a l'optimització no-lineal, sempre amb uns bons fonaments de processat del senyal. Els resultats i les contribucions han estat revisades per la comunitat internacional de recerca i publicats en revistes i conferències internacionals d'alt nivell.

Acknowledgments

Radu, parce que sans ton expertise et ton guidage, ça n'aurait pas été possible. Parce qu'on a su se comprendre et produir de beaux résultats. Parce que tu m'as montré ce que c'est la recherche.

Florence, parce que t'as apporté plein de bonnes idées dans nos discussions. Pour ton point de vue frais et judicieux.

Daniel, porque confiaste en mi y porque todavía lo haces. Por nuestras charlas y tu prudente sinceridad.

Nicu, because you were the first to congratulate me, even before myself.

Josep Ramon, per la teva resposta immediata i encoratjadora.

Laurent, parce que tu m'as toujours traité comme l'un de tes pairs.

Philippe, parce que mes premiers pas dans la recherche ont été à côté de toi. Ja vais difficilement oublier ça.

Because, no matter where I will be, I hope we keep in touch and start great scientific collaborations with the seven of you.

Nathalie, Florence et Eric, parce que votre éfficacité, parfois caché, est un des pilliers du bon déroulement du centre, et donc, de ma thèse. Mais surtout, parce qu'on rigole et on se soutient.

Perception Team, and INRIA : Miles, Simone, Ramya, Regis, Gaëtan, Amaël, Kiran, Visesh, Vineeth, George, Quentin, Pierre, Israel, Guru, Vincent, Dionisos, Kaustubh, Lamiae, Laurentiu, Maël, Svetlana, Valentin, Diana, Senan, Darren, Vasil, Ben, and so many other. Because daily life without you would have been extremely boring and uneventful.

Pare, Mare, perquè el vostre suport incondicional ha estat la clau del meu èxit, i per tantes altres coses per a les quals no tinc, i no sé si mai tindré, les paraules per a agraïr-vos-les.

Marc, Eli, Berta, perquè sou la meva joia.

Barcelona, Catalunya, la meva terra, i els meus amics. Perquè és amb vos-altres, no cal que n'escrigui els noms, amb qui navego pel mar de la felicitat. Perquè és igual a on sigui i que faci: esteu sempre amb mi.

Nassos, because they have been three awesome years, talking, bouldering, drinking and laughing.

Kumaï, perquè m'has ensenyat coses de mi mateix que altrament no hagués pogut aprendre.

Jordi, just darrera d'en Kumaï, sé que te l'estimes molt. Perquè poder xerrar amb algú com tu és un plaer.

Soraya, parce que ton énergie est un exemple. Ne la perds pas, jamais.

Chapeau, parce que ton esprit inspire. Parce qu'on rigole, on s'embête et on danse.

Maxime, pour ton sens commun et ton soin des détails.

Antoine, parce que t'es bien et parce qu'on se tient.

Ahmad, pour ta persévérance et ta sagesse, souvent cachée.

Harsimrat, because I learnt a lot from you, about so many things.

Мария-за сериалите, разговорите и шоколада.

Thomas, pour nos randos et nos grimpes, tes commentaires prudents, sincères et judicieux

Raph, parce que je n'ai pas besoin de le dire, car tu le sais. Parce que je n'ai pas assez de place. On en discutera un autre jour...

Лени, заради теб, усмивката ти и твоя суинг.

Ne plaisante jamais avec ces choses-là...

CONTENTS

LIST OF FIGURES

LIST OF TABLES

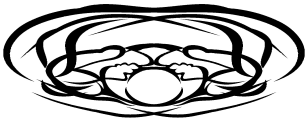
How to Read this Manuscript

This is a quick reference on how to read this manuscript.

- Bibliographic references are denoted as [Alameda-Pineda 11].
- Figures, Tables and other floating objects as well as equations are numbered within the chapter number.
- Equations are referred as (3.7).
- The front matter covers the table of contents, the list of figures and tables, and this guide.
- The main matter of the Thesis's manuscript starts at page 1, until page 103.
- The back matter covers the list of the candidate's publications and the bibliographic references cited along the text.
- Small notes on the margin might be used to easily navigate through the manuscript. They are meant to summarize paragraphs/blocks of text.

Example of margin note

- The end of the chapter is shown by the following sign between horizontal rules.



CHAPTER 1

INTRODUCTION

1.1 Motivation

Multimodal human-robot interaction

In recent years, robots have gradually moved from production and manufacturing environments to populated spaces, such as public spaces, e.g., museums and entertainment parks, offices, hospitals, homes, etc. There is an increasing need to develop robots that are capable of interacting and communicating with people in unstructured, unconstrained and unknown environments in the most natural way. For robots to fulfil interactive tasks, not only they need to recognize humans, human gestures, intentions and speech, they equally need to gather data from different sensing modalities as well as to coordinate their perceptive, communicative and motor skills, i.e., *multimodal human-robot interaction*.

Audio-visual fusion, but why?

Data gathered with different sensory modalities need to be combined in order to extract the semantic content of a complex environment and hence build an internal representation of the real world. Vision and audition are the modalities the most suitable to be used by a robot due to the wide availability of associated sensors, namely cameras and microphones. Combining auditory and visual data is naturally performed by human beings. Indeed, many behavioural, electrophysiological and imaging studies [Calvert 04, Ghazanfar 06, Senkowski 08] postulate that the fusion of different sensorial modalities is an essential component of perception.

Challenges of audio-visual fusion

Immediately, the interesting question on how to fuse information coming from the two modalities (vision and hearing) arises. Because the gathered data correspond to two different physical phenomena, the intrinsic meaning of the information carried in the data is also different. On one side, visual information encodes the reflection of light-rays onto the different surfaces composing the scene. On the other side, auditory information encompasses the variations of the air pressure produced by different emitting devices. Moreover, the spatio-temporal distribution of the data is radically different. While visual information is continuous

in time and space, auditory information is not only sparse in time (existing only when the emitter is active) but also spare is space (coming from the sound sources' position plus the eventual reverberations). In addition, auditory and visual streams are corrupted in different ways. On one hand, visual data suffers from occlusions, self-occlusions, limited field-of-view and lighting conditions. On the other hand, auditory data suffers from microphone noise, reverberations and interferences.

Egocentrinc Audio-Visual Scene Analysis

Humans interact with complex environments in a daily basis. In other words, people solve the audio-visual fusion problem in a natural way thus interpreting auditory and visual input in their everyday life. For instance, they have no difficulties in focusing their attention onto a dialogue between two speakers in an extremely noisy environment, i.e., in the presence of a multitude of other auditory and visual events. More interestingly, human beings perform this task with a set of sensors placed in a small volume compared to the scene space, in other words, the auditory and visual sensors are close to each other. We refer to this as "Egocentric Audio-Visual Scene Analysis". In particular, our interest is to build up solid methodologies whose associated algorithms provide robust AV capabilities to an agent-centered architecture such as a humanoid robot.

Data properties

In all, the problem we address have several attributes in terms of the data used, the methods derived and the results desired. Indeed, the data acquired by the robot is (D1) *egocentric*, that is captured with a sensor network fitting in a small volume, with all sensors placed near to each other, as in the human head, (D2) *multimodal*, or more precisely, audio-visual, thus consisting in image flows and sound tracks, (D3) *corrupted*, as described, by occlusions and bad lighting conditions on the visual side and by noise, reverberations and interferences on the auditory side.

Method properties

Some properties would desirably characterize the methods and their associated algorithms: (M1) *efficient*, so the limited resources of advanced platforms such as humanoid robots are not misused, (M2) *fast*, ensuring that the produced results correspond to the ongoing social interplay and not to expired acts of communication, (M3) *robust*, avoiding that small/medium perturbations of the scene and of the system have a devastating effect on the method's performance, (M4) *adaptable*, making of the robot a widely-usable system able to work in a large variety of environments, (M5) *reliable*, such that other applications can build on them and provide higher-level capabilities to the robot.

Outcome properties

In order to guarantee the quality of the final system, the outcome should be (O1) *temporally coherent*, ensuring that the final system does not lead to a "moody" and overreactive robot, (O2) *spatially consistent*, so that the management of the robot's space is correctly performed, (O3) *semantically meaningful*, providing the opportunity to build a natural answer for the other members of the interaction, e.g., humans.

1.2 Socio-Economic Context

Socio-economic context & impact

Along the past two decades, the industry has developped several comercial products with audio-visual sensing capabilities. Most of them consists on a video-camera with an embedded mircophone (mobile phones, tablets, etc). Other, such as Kinect, include depth sensors and/or small microphone arrays. Also, there are some mobile phones equiped with a stereo camera pair. At the same time, many research-oriented systems became available (e.g., humanoid robots such as NAO). Far from capturing the global scene, these systems are small in volume. Consequently theur sensors are close to each other.

Funding consequences: the EU projects

One of the consequences of this socio-economic context is the Sixth and Seventh Framework Programs (FP6, FP7) of the European Union. Research and research organisations may find gere different ways to get funding for their research. This PhD Thesis is related to two particular projects issued form the FP6 and FP7 calls: the POP project and the HUMAVIPS project, respectively. Thanks to them I have been able to work with high technology devices and collaborate with researchers from different countries.

The POP Project: FP6-IST-027268

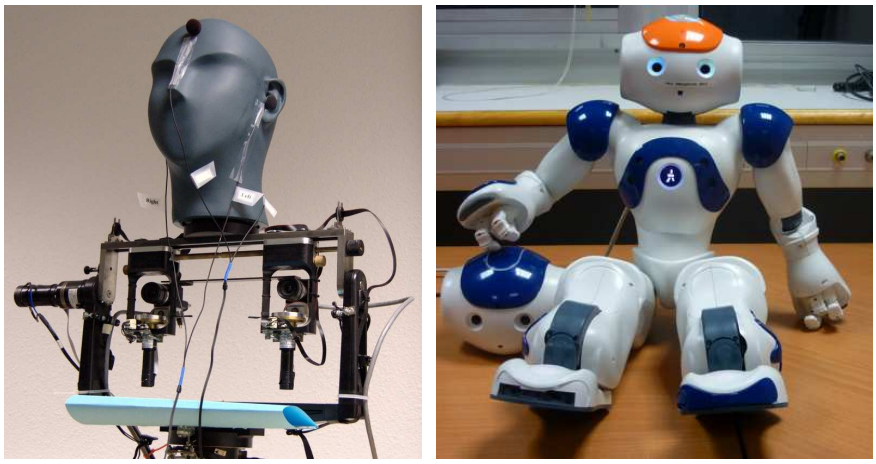
The POP project proposed to develop a new approach, perception on purpose (POP), based on five principles: (i) visual and auditory information should be integrated in both space and time, (ii) active exploration of the environment is required to improve the audiovisual signal-to-noise ratio, (iii) the enormous potential sensory requirements of the entire input array should be rendered manageable by multimodal models of attentional processes, (iv) bottom-up perception should be stabilized by top-down cognitive function and lead to purposeful action and (v) all parts of the system should be underpinned by rigorous mathematical theory, from physical models of low-level binocular and binaural sensory processing to trainable probabilistic models of audiovisual scenes[1].

Thanks to the fact that the Perception Team at INRIA was the coordinator of the POP project, we posses one POPEYE robot (see Figure 1.1a). POPEYE is equipped with four microphones and two cameras providing for auditory and visual sensory faculties. The four microphones are mounted on a dummy-head designed to imitate the filtering properties associated with a real human head. Both cameras and the dummy head are mounted on a four-motor structure that provides for accurate moving capabilities. I used POPEYE several times during my Thesis to acquire data in order to test the algorithms I developped. Hence, POPEYE has been an extremly useful tool for my research

The HUMAVIPS Project: FP7-ICT-247525

The objective of HUMAVIPS has been to endow humanoid robots with audio-visual (AV) abilities: exploration, recognition, and interaction, such that they exhibit adequate behavior when dealing with a group of people. Developed research and technological developments have emphasized the role played by multimodal perception within principled models of human-robot interaction and of humanoid

[1] Extracted from [POP 09].



(a) The POPEYE Robot **(b)** Tha NAO Robot

Figure 1.1: The two robotic platforms available in the Perception Team. (a) The POPEYE robot head: the colour-camera pair as well as two (front and left) out of four microphones are shown in the image. POPEYE was specifically manufactured for the POP project. (b) The humanoid robot NAO with the new audiovisual head that is composed of a synchronized camera pair and two microphones. The new head was designed and manufactured in the framework of the HUMAVIPS project.

behavior. An adequate architecture has implemented auditory and visual skills onto a fully programmable humanoid robot (the consumer robot NAO). A free and open-source software platform has been developed to foster dissemination and to ensure exploitation of the outcomes of HUMAVIPS beyond its lifetime[2].

As in the case of the POP project, INRIA coordinated the HUMAVIPS project. Thanks to it, we we able to work and experiment with the NAO robot (see Figure 1.1b). This has been a rich experience since we understood many issues related to on-line applications: computational resource optimization, information loss, reaction time, ... In addition to the material, the HUMAVIPS project also helped me attending conferences and other research meetings. Amon them, the project scientific meetings and the code camps have been useful to eachange information, impressions and opinions on everyone's work. In all, a very fulfiling experience that enriched my research.

1.3 Research Context and Contributions

Audio-visual signal processing applications

Egocentric Audio-Visual Scene Analysis is related to many research fields such as: speaker detection and localization, source separation, face detection and recognition, voice recognition, action/gesture recognition, identity recognition, dialogue modelling, role identification, feedback recognition, emotion recognition, nod recognition and intention recognition. Numerous researchers investigated the

[2]Extracted from [HUMAVIPS 13].

fusion of auditory and visual cues in a variety of domains such as event classification [Natarajan 12], speech recognition [Barker 09], sound source separation [Naqvi 10], speaker tracking [Hospedales 08], [Gatica-Perez 07] and speaker diarisation [Noulas 12]. However, these approaches are not suitable for robots either because the algorithmic complexity is too high, or because methods use a distributed sensor network or because the amount of training data needed is too high, drastically reducing the robots' adaptableness. Unfortunately, much less effort has been devoted to design audio-visual fusion methods for humanoid robots. Nevertheless, there are some interesting works introducing methods specifically conceived for humanoid robots on speech recognition [Nakadai 04], beat tracking [Itohara 11], [Itohara 12], active audition [Kim 07] and sound recognition [Nakamura 11]. Far from being an exhaustive list, this is representative of what is possible with auditory and visual signal processing. In this PhD I focused on three topics: audio-visual robot command recognition, audio-visual speaker detection and localization and sound source localization. In the following, the main contributions of this PhD are detailed and the structure of the rest of the manuscript is given. Since different applications are addressed, a detailed description of the closest existing work to each of them is given in the corresponding chapter.

An egocentric data set: RAVEL

In order to be able to test the methods and algorithms outcomming from the research process, some data was needed. Together with other members of the HUMAVIPS projects we decided to acquire a dataset: since the existing ones would not fulfil our requirements. We recorded the RAVEL data set and it is fully described in Chapter 2. We required a dataset targeting many different applications, since many members of the project worked in different topics. In addition, recordings should be stereoscopic and (at least) binaural, since, in our group, we are interested in 3D localization from audio and video. This discards all data sets recorded with one camera and/or with one microphone [Hazen 04, Patterson 02]. The third condition is that the data should be egocentric, meaning that all the sensors should be mounted in a device of a small size compared to the scene size. In other words, we could not use the datasets recorded in smart-rooms using spread camera/microphone networks such as [Mostefa 07, Lathoud 05]. Hence, we contributed by providing a publicly available data set targeting applications such as action/gesture recognition, speaker localization, tracking and diarisation, sound source localization, dialogue modelling, etc. This work has been published in [Alameda-Pineda 12b] and used later on inside and outside the thesis.

Audio-visual speaker detection and localisation

The first application we targeted is audio-visual speaker detection and localization. Several mid-level applications such as tracking, beamforming and voice recognition benefit from knowing where the speakers are and when they are speaking. Hence, this is a key point in Egocentric Audio-Visual Scene Analysis. Many work has been done in this field, see Chapter 3 for more details. We contributed by developing a new approach for AV event detection. Published in [Alameda-Pineda 11], we show how the trust on one of the modalities (visual) can be modelled and used to bias the method **AVS1**, leading to a visually-supervised EM algorithm (ViSEM) **AVS2**. This paper got the Outstanding Paper Award at

ICMI'11. Afterwards, we modified the approach to target audio-visual speaker detection yielding to an on-line method working in the humanoid robot NAO **AVS3**. The details can be found in [Sanchez-Riera 12b].

Audio-visual command recognition

The second application we addressed is the audio-visual robot command recognition (see Chapter 4). By means of auditory and visual recordings we recognize a robot command, that is a visual gesture accompanied by a short sentence. This is of vital importance because most of these commands represent a direct and immediate interaction between people or aim to trigger a certain reaction on the recognition system (e.g., robot). Recognition of AV gestures has been of course investigated before inter alia [Mühling 12, Natarajan 12, Ye 12]. We first explored in [Sanchez-Riera 12a] different features and classifiers and confirmed that the use of audio-visual data increases the performance when compared to auditory-only and to video-only classifiers **AVG1**. Later, in [Alameda-Pineda 13c] we sought for the best method using tiny training sets (5-10 samples per class) **AVG2**. This is interesting because real systems need to adapt and learn new commands from the user. Such systems need to be operational with a few examples for the general public usage.

Multichannel sound source localisation

Finally, we found interesting to investigate the field of sound source localization. Because, often, the acquiring devices do not have a regular geometry, we focused our attention to the use of arbitrarily shaped (non-coplanar) microphone arrays. This is interesting because the geometry of the microphone can be any. Consequently, this opens the door to dynamic microphone arrays that would adapt their geometry to fit some particular tasks. Also, because the design of commercial systems may be subject to certain constraints for which circular or linear arrays are not suited. Multichannel time delay estimation for sound source localization has been addressed before in [Chen 03b]. At a first stage we published in [Alameda-Pineda 12a], where we presented the general framework **GTDE1** and a local optimization algorithm **GTDE2**. Later on we published the full geometric model **GTDE3** together with a much more solid algorithm **GTDE4** in [Alameda-Pineda 13b].

In summary, this thesis has several scientific contributions:

1. The acquisition of the **RAVEL** data set, consisting on several scenarios targetting different human-robot-interaction applications. Auditory and visual data was gathered with from an egocentric view of the scene [Alameda-Pineda 12b].
2. A **hybrid probabilistic/deterministic model** for audio-visual data fusion aiming the detection and localisation of speakers in the scene [Alameda-Pineda 13a].
3. The **Motion-Guided** AV fusion algorithm, that maps the motion-related visual observations into the auditory feature space working on egocentric data [Alameda-Pineda 11].
4. The **Face-Guided** AV fusion algorith, inspired from the one above, and de-

signed to work on-line with the humanoid platform NAO [Sanchez-Riera 12b].

5. The **convex-weighting scheme** for AV command recognition, able to cope with classifiers of different nature [Sanchez-Riera 12a].

6. A **benchmarking on tiny training sets** of AV command recognition methods, pushing their limits to get the best trade-off between user-adaptability and recognition performance [Alameda-Pineda 13c].

7. A **sound source localisation** algorithm for non-coplanar microphone arrays from multichannel time delay measurements [Alameda-Pineda 12a].

8. A deep study of the constraints on multichannel time delay estimation associated to the **arbitrary geometry** of a microphone array [Alameda-Pineda 13b].

In conclusion, we faced different real problems of AV scene analysis, using egocentric data. Methods vary from statistical and discriminative learning to non-linear programming, always on top of a solid basis of signal processing. Results and contributions have been peer-reviewed by the international research community and published in international top conferences and journals.

1.4 Related Events: Internships and Workshops

This PhD Thesis originated many events. First, two internships in the Perception Team were related to this Thesis and co-advised between Prof. Radu Horaud and myself. In addition, two workshops were organized, jointly with two other researchers as detailed below.

Workshop organisation

During my PhD I co-organized two Grand Challenges together with Dr. Roman Bednarik[3], Researcher at the School of Computing, University of Eastern Finland and Dr. Kristiina Jokinen[4], Adjunct Professor of Language Technology at University of Helsinki.

Within the framework of the Call for Challenges at ICMI 2012, the D-META Grand Challenge proposed to set up the basis for comparison, analysis, and further improvement of multimodal data annotations and multimodal interactive systems. Such machine learning-based challenges do not exist in the Multimodal Interaction community. The main goal of this Grand Challenge was to foster research and development in multimodal communication and to further elaborate algorithms and techniques for building various multimodal applications. Held by two coupled pillars, method benchmarking and annotation evaluation, the D-META challenge envisioned a starting point for transparent and publicly available application and annotation evaluation on multimodal data sets.

[3] http://cs.joensuu.fi/\tilderbednari/
[4] http://www.ling.helsinki.fi/~kjokinen/

Unfortunately, D-META did not have the expected success. We re-oriented the prupose of the challenge to target a specific applications. This lead to the Multimodal Conversational Analytics (MCA) Grand Challenged, organized in the framework of ICMI 2013. The challenge aims to bring together researchers from across disciplines related to multimodal conversational analytics. The challenge follows the D-META Challenge organized at the ICMI 2012 in Santa Monica. Unfortunately for us, the MCA did not have neither the expected success. Consequently and sadly, we could not extract any conclusions of the research status on the field of MCA.

Master thesis/internships

I co-advised, together with Prof. Radu Horaud, two Masters interships. Firstly we guided Maxime Janvier through the machine hearing literature to investigate several sound recognition techniques and evaluate them in the framework of robot audition. Maxime did a very nice job, which got him the Diploma of Engineer in Signal Processsing and, on top of this, an article published in the International Conference on Humanoid Robots [Janvier 12]. He is, nowadays, since October 2012, a PhD candidate in our Team on the topic of sound recognition for humanoid robots. Secondly, we advised Israel-Dejene Gebru in the field of audio-visual speaker detection. More precisely, we investigated the use of observation-associated relevance information in the framework of probabilistic graphical models. Israel-Dejene Gebru is in the Masters in Telecommunications Engineering at Trento University, and he will join the Perception team as a PhD candidate from October 2013.

1.5 Logistic Context

Perception and Mistis teams

This PhD Thesis was carried out at the Perception Team[5] at INRIA Grenoble Rhône-Alpes[6]. Dr. Radu Horaud, the head of Perception, is Director of Research at INRIA and has been the main supervisor of the present Thesis. Dr. Florence Forbes, the head of the Mistis[7] Team at INRIA and also Director of Research, co-advised the Thesis in conjunction with Dr. Radu Horaud. Their complementary background, in computer vision, signal processing and robotics on one side and in statistical, signal and image processing on the other side, has been of great help and extremely instructive. There is no need to say how much I learnt about the research world from them both, acquiring skill on teaching, lecturing, grant proposal writing, student supervision and paper writing and reviewing.

INRIA has robots

INRIA in general and the Perception Team in particular are extremely well equipped. Indeed, as outlined before, I could perform experiments on data gathered with the POPEYE robot Figure 1.1a as well as with the NAO robot Figure 1.1b. On one side, the POPEYE robot allowed me to test the developed methods on high quality data. Thus, models and algorithms were validated in a

[5] http://perception.inrialpes.fr/
[6] http://www.inria.fr/centre/grenoble
[7] http://mistis.inrialpes.fr/

controlled processing pipeline, in which real-time inherent issues such as computing time and data loss did not have play any role. On the other side, once methods were validated, the NAO robot allowed me to test them under the particularities of real-time processing. Usually, changes on the design and implementation of the algorithms were applied to overcome these challenges.

Computational requirements

A part from these two robotic platforms, the Perception Team possesses several computers, shared between all members, to perform the scientific experiments. In the following, the computational power and implementation issues of each chapter are commented.

The acquisition of the RAVEL data set did no require any particular set up, expect for a PC with two hard-disks (one per image flow). We used the software provided by 4DViews[8] for the visual acquisition and Audacity[9] for the sound recording.

The experiments of Chapter 3 were performed in two different platforms. Regarding the results on synthetic data and on real data, they were obtained by a MATLAB/C++ implementation and run in one core of a HP Z800 Workstations at 2.53 GHz. Approximately, the algorithm needed 30 second per visual frame. The optimized real-time implementation was entirely coded in C++ and running in a PC with an i7 core at 2.5 GHz. The bottleneck of the pipeline was in the image delivering module. Consequently, our algorithm was able to perform at full-rate, that is, 17 frames per second.

The classifiers of Chapter 4 were also trained and tested in the Z800 Workstation. The training and testing on tiny datasets, that is the learning and evaluation of more than 150,000 SVMs were coded using Python to interface the Shogun C++ library. It took no more than one night to train and test all the SVMs. It is worth mentioning that, in these experiments, we needed to use R's implementation of k-means, because the one of MATLAB was not able to handle the amount of features we had, resulting in a memory overflow.

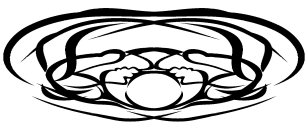
Finally, in Chapter 5, we implemented everything in MATLAB. The slowest methods were the local-optimization grid-based methods, which took one week to estimate the time delays from the 189 positions. Always in the Z800, the global optimization method took less than three days. Notice that this code was not designed for real-time processing. Much faster implementations should be possible with a more adapted programming language such as C++.

1.6 Manuscript Structure

The rest of the manuscript is structured as follows. Chapter 2 describes de RAVEL data set. The scenarios, recordings devices and environments, acquired data, annota-

[8] http://www.4dviews.com/
[9] http://audacity.sourceforge.net/

tion and examples of use are given. Afterwards in Chapter 3, we describe our contribution to audio-visual speaker detection and localisation and present two vision-guided robot hearing algorithms. Chapter 4 is devoted to detail our investigation on audio-visual command recognition. Chapter 5 shows the last experimental contribution of this Thesis: multichannel time delay estimation for sound source localisation using non-coplanar microphone arrays. Finally, Chapter 6 presents the conclusions and the vision of the forthcoming years in the research of Egocentric Audio-Visual Scene Analysis.



CHAPTER 2

THE RAVEL DATA SET

We introduce RAVEL (Robots with Audiovisual Abilities), a publicly available data set which covers examples of Human Robot Interaction (HRI) scenarios. These scenarios are recorded using the audio-visual robot head POPEYE, equipped with two cameras and four microphones, two of which being plugged into the ears of a dummy head. All the recordings were performed in a standard room with no special equipment, thus providing a challenging indoor scenario. This data set provides a basis to test and benchmark methods and algorithms for audio-visual scene analysis with the ultimate goal of enabling robots to interact with people in the most natural way. The data acquisition set up, sensor calibration, data annotation and data content are fully detailed. Moreover, three examples of using the recorded data are provided, illustrating its appropriateness for carrying out a large variety of HRI experiments. The RAVEL data are publicly available at: `http://ravel.humavips.eu/`

2.1 Introduction

The RAVEL data set in a nutshell.

In this chapter we describe the publicly available data set RAVEL (Robots with Audiovisual Abilities). This dataset was recorded in the framework of the HUMAVIPS project. Designed to fulfil the needs of all the partners, the data set consists of three categories: human activities, robot-commands and interaction. A detailed description of the categories and of the scenarios inside the categories is given below. Two of the scenarios are particularly important for this PhD Thesis: the cocktail party problem scenario (which is part of the "interaction" category) used in Chapter 3 and the robot-command category used in Chapter 4. Figure 2.1 presents some snapshots of the recorded scenarios in all three categories. All scenarios were recorded using an audio-visual (AV) robot head, shown in Figure 2.2a, equipped with two cameras and four microphones, which provide multimodal and multichannel synchronized data recordings.

Relevance of RAVEL for the research community.

Researchers working in multimodal human-robot interaction can benefit from RAVEL for several reasons. First of all, four microphones are used in order to be

(a) *Talk on the phone* **(b)** *Stop!* **(c)** *Where is the kitchen?* **(d)** *Cheers!*

(e) *Cocktail party* **(f)** *Hand-shaking* **(g)** *Let me introduce you!* **(h)** *Someone arrives*

Figure 2.1: Scenario examples from the RAVEL data set. (a) Human activity – *talk on the phone*–, (b) Robot command – *stop!*–, (c) Asking the robot for instructions, (d) Human-human interaction, (e) Cocktail party, (f) Human introducing a new person (g) Robot introducing a new person, and (h) New person.

able to study the sound source separation problem; robots will face this problem when interacting with humans and/or other robots. Secondly, the simultaneous recording of stereoscopic image pairs and microphone pairs gives an opportunity to test multimodal fusion methods [Luo 89] in the particular case of visual and auditory data. Moreover, the fact that a human-like robot head is used, makes the data appropriate to test methods intended to be implemented on humanoid robots. Finally, the scenarios are designed to study action and gesture recognition, localization of auditory and visual events, dialogue handling, gender and face detection, and identity recognition. In summary, many different HRI-related applications can be tested and evaluated on this data set.

Contributions and novelty of the RAVEL data set.

The RAVEL data set was published in [Alameda-Pineda 12b] and is novel since it is the first data set devoted to study the human robot interactions consisting of synchronized binocular image sequences and four channel audio tracks. The stability of the acquisition device ensures the repeatability of recordings and, hence, the significance of the experiments using the data set. In addition, the scenarios were designed to benchmark algorithms aiming at different applications as described later on. To the best of our knowledge, there is no equivalent publicly available data set in terms of data quality and scenario design.

The remainder of the chapter is structured as follows. Section 2.2 delineates the related existing data sets. Section 2.3 is devoted to describe the acquisition set up: the recording device, the recording environment and the characteristics of the acquired data. A detailed description of the categories and of the scenarios is given in Section 2.4. Afterwards, the data set annotation procedure is discussed (Section 2.5). Before drawing the conclusions (Section 2.7), some examples of usage of the RAVEL data set are given (Section 2.6).

2.2 Related Work

The RAVEL data set is at the cross-roads of several HRI-related research topics, such as robot vision, audio-visual fusion, sound source separation, dialogue handling, etc. Hence, there are many public data sets related to RAVEL. These data sets are reviewed in this section and the most relevant ones are described.

Related data sets: action segmentation, isolated & continuous action recognition.

Accurate recognition of human actions and gestures is of prime importance in HRI. There are two tasks in performing human actions recognition from visual data: classification of actions and segmentation of actions. There are several available data sets for action recognition. KTH [Schüldt 04], Youtube Action Classification [Liu 09] and Hollywood1 [Laptev 08] are data sets devoted to provide a basis for solving the action classification task. For the segmentation task two data sets are available: Hollywood2 [Marszalek 09] and Coffee and Cigarettes [Willems 09]. All these data sets provide *monocular* image sequences. In contrast, the INRIA XMAS data set [Weinland 06] provides 3D visual hulls and it can be used for the classification and localization tasks. In the INRIA XMAS data set, the actors perform actions in a predefined sequence and are recorded using a complex multiple camera set up that operates in a specially arranged room. The Opportunity data set [OPPORTUNITY 11] serves as a data set for the challenge with the same name. The focus of this challenge is benchmarking the different state-of-the-art action recognition methods. Last, but not least, there are three data sets concerning the daily activities on a "kitchen" scenario namely: the KIT Robo-Kitchen Activity Data Set [Rybok 11], the University of Rochester Activities of Daily Living Data Set [Messing 09] and the TUM Kitchen Data Set [Tenorth 09].

Related data sets: audio-visual speaker detection and tracking, face and voice recognition.

Audio-visual perception [Kim 07, Khalidov 11] is an useful skill for any entity willing to interact with human beings, since it provides for a spatio-temporal representation of an event. There are several existing data sets for the AV research community. In particular, a strong effort has been made to produce a variety of multimodal data sets focusing on faces and speech, like the AV-TIMIT [Hazen 04], GRID [Cooke 07], M2VTS [Pigeon 96], XM2VTSDB [Messer 99], Banca [Bailly-Baillire 03], CUAVE [Patterson 02] or MOBIO [Marcel 10] data sets. These data sets include individual speakers (AV-TIMIT, GRID, M2VTS, MOBIO, XM2VTSDB, Banca) or both individual speakers and speaker pairs (CUAVE). All have been acquired with one close-range fixed camera and one close-range fixed microphone. Two corpora more closely related to RAVEL are the AV16.3 data set [Lathoud 05] and the CAVA data set [Arnaud 08]. Both include a range of situations. From meeting situations where speakers are seated most of the time, to motion situations, where speakers are moving most of the time. The number of speakers may vary over time. Whilst for the AV16.3 data set three fixed cameras and two fixed 8-microphone circular arrays were used, for the CAVA data set two cameras and two microphones were mounted in a person's head. Instead, RAVEL uses an active robot head equipped with far-range cameras and microphones.

(a) The POPEYE robot head. **(b)** Environment set up

Figure 2.2: The acquisition set up at a glance. (a) The POPEYE robot head was used to collect the RAVEL data set. The colour-camera pair as well as two (front and left) out of four microphones are shown in the image. Four motors provide the rotational degrees of freedom and ensure the stability of the device and the repeatability of the recordings. (b) Two views of the recording environment. The POPEYE robot is in one side of the room. As shown, the sequences were shot with and without daylight providing for lighting variations. Whilst two diffuse lights were included in the set up to provide for good illumination, no devices were used to modify neither the illumination changes nor the sound characteristics of the room. Hence, the recordings are affected by all kind of audio and visual interferences and artefacts present in natural indoor scenes.

Related data sets: human-robot-interaction

Concerning human robot interaction data sets, [Zivkovic 08] provides typical robot sensors' data of a "home tour" scenario annotated using human spatial concepts; this allows to evaluate methods trying to semantically describe the geometry of an indoor scene. In [Mohammad 08], the authors present a new audio-visual corpus containing information of two of the modalities used by humans to communicate their emotional states, namely speech and facial expression in the form of dense dynamic 3D face geometries.

Different data sets used different devices to acquire the data, depending on the purpose. In the next section, the acquisition set up used in RAVEL, which includes the recording environment and device, is fully detailed. Furthermore, the type of recorded data is specified as well as its main properties in terms of synchronization and calibration.

2.3 Acquisition Setup

Two main properties of RAVEL's acquisition set up: egocentric data and realistic environment.

Since the purpose of the RAVEL data set is to provide data for benchmarking methods and techniques for solving HRI challenges, two requirements have to be addressed by the set up: an egocentric collection of accurate data and a realistic recording environment. In this section the details of this set up are given, showing that these two requisites are satisfied to a large extent. In a first stage the recording device is described. Afterwards, the acquisition environment is delineated. Finally, the properties of the acquired data in terms of quality, synchrony and calibration are detailed and discussed.

The POPEYE robot head is RAVEL's acquisition device.

The POPEYE robot was designed in the framework of the POP project [POP 09].

This robot is equipped with four microphones and two cameras providing for auditory and visual sensory faculties. The four microphones are mounted on a dummy-head, as shown in Figure 2.2a, designed to imitate the filtering properties associated with a real human head. Both cameras and the dummy head are mounted on a four-motor structure that provides for accurate moving capabilities: pan motion, tilt motion and camera vergence.

Main properties of POPEYE.

The POPEYE robot has several remarkable properties. First of all, since the device is alike the human being, it is possible to carry out psycho-physical studies using the data acquired with this device. Secondly, the use of the dummy head and the four microphones, allows for the comparison between using two microphones and the Head Related Transfer Function (HRTF) against using four microphones without HRTF. Also, the stability and accuracy of the motors ensure the repeatability of the experiments. Finally, the use of cameras and microphones gives to the POPEYE robot head audio-visual sensory capabilities in one device that geometrically links all six sensors.

The recording environment is a regular indoor office

All sequences from the data set were recorded in a regular meeting room, shown in Figure 2.2b. Whilst two diffuse lights were included in the set up to provide for good illumination, no devices were used to modify neither the effects of the sunlight nor the acoustic characteristics of the room. Hence, the recordings are affected by exterior illumination changes, acoustic reverberations, outside noise, and all kind of audio and visual interferences and artefacts present in unconstrained indoor scenes.

Data acquired: technical description of the sensing devices.

For each sequence, we acquired several streams of data distributed in two groups: the *primary* data and the *secondary* data. While the first group is the data acquired using the POPEYE robot's sensors, the second group was acquired by means of devices external to the robot. The *primary* data consists of the audio and video streams captured using POPEYE. Both, left and right, cameras have a resolution of 1024×768 and two operating modes: 8-bit gray-scale images at 30 frames per second (FPS) or 16-bit YUV-color images at 15 FPS. The four Soundman OKM II Classic Solo microphones mounted on the Sennheiser MKE 2002 dummy-head were linked to the computer via the Behringer ADA8000 Ultragain Pro-8 digital external sound card sampling at 48 kHz. The *secondary* data are meant to ease the task of manual annotation for ground-truth. These data consist of one flock of birds (FoB) stream (by Ascension technology) to provide the absolute position of the actor in the scene and up to four wireless close-range microphones PYLE PRO PDWM4400 to capture the audio track of each individual actor.

Synchronization and calibration of the acquisition sensors.

Both cameras were synchronized by an external trigger controlled by software. The audio-visual synchronization was done by means of a clapping device. This device provides an event that is sharp – and hence, easy to detect – in both audio and video signals. The FoB was synchronized to the visual stream in a similar way: with a sharp event in both FoB and video signals. Regarding the visual calibration, the state-of-the-art method described in [Bouguet 08] uses several image-pairs to provide an accurate calibration. The audio-visual calibration is

Table 2.1: Summary of the recorded data size per scenario.

Scenario	Trials	Actors	Video in MB	Audio in MB
AR	12	12	4,899	2,317
RG	11	11	4,825	1,898
AD	6	6	222	173
C	5	4	118	152
CPP	1	1	440	200
MS	7	6	319	361
IP	5	7	327	204
Total	–	–	11,141	5,305

manually done by annotating the position of the microphones with respect to the cyclopean coordinate frame [Hansard 08].

Following the arguments presented in the previous paragraphs it can be concluded that the set up suffices conceptual and technical validation. Hence, the sequences have an intrinsic value when used to benchmark algorithm targeting HRI applications. The next section is devoted to fully detail the recorded scenarios forming the RAVEL data set.

2.4 Data Set Description

The three categories: action recognition, robot gestures and interaction.

The RAVEL data set has three different categories of scenarios. The first one is devoted to study the recognition of actions performed by a human being. With the second category we aim to study the audio-visual recognition of gestures addressed to the robot. Finally, the third category consists of several scenarios; they are examples of human-human interaction and human-robot interaction. Table 2.1 summarizes the amount of trials and actors per scenario as well as the size of the visual and auditory data. Figure 2.1 (a)-(h) shows a snapshot of the different scenarios in the RAVEL data set. The categories of scenarios are described in detail in the following subsections.

2.4.1 Action Recognition [AR]

The task of recognizing human-solo actions is the motivation behind this category; it consists of only one scenario. Twelve actors perform a set of nine actions alone and in front of the robot. There are eight male actors and four female actors. Each actor repeats the set of actions six times in different – random – order, which was prompted in two screens to guide the actor. This provides for various co-articulation effects between subsequent actions. The following is a detailed list of the set of actions: (i) *stand still*, (ii) *walk*, (iii) *turn around*, (iv) *clap*, (v) *talk on the phone*, (vi) *drink*, (vii) *check watch* (analogy in [Weinland 06]), (viii) *scratch head* (analogy in [Weinland 06]) and (ix) *cross arms* (analogy in [Weinland 06]).

Actor	(enters the scene)
Actor	Excuse me, where are the toilets?
Robot	Gentleman/Ladies are to the left/right and straight on 10 meters.
Actor	(leaves the scene)

Script 1: The script encloses the text spoken by the actor as well as by the robot in the "*Asking for directions*" scenario.

2.4.2 Robot Gestures [RG]

Learning to identify different gestures addressed to the robot is another challenge in HRI. Examples of such gestures are: waving, pointing, approaching the robot, etc. This category consists of one scenario in which the actor performs six times the following set of nine gestures: (i) *wave*, (ii) *walk towards the robot*, (iii) *walk away from the robot*, (iv) *gesture for 'stop'*, (v) *gesture to 'turn around'*, (vi) *gesture for 'come here'*, (vii) *point action*, (viii) *head motion for 'yes'* and (ix) *head motion for 'no'*. In all cases, the action is accompanied by some speech corresponding to the gesture. In total, eleven actors (nine male and two female) participated in the recordings. Different English accents are present in the audio tracks which makes the speech processing challenging.

2.4.3 Interaction

This category contains the most interactive part of the data set, i.e. human-human as well as human-robot interaction. Each scenario consists of a natural scene in which several human beings interact with each other and with the robot. In some cases one of the actors and/or the robot act as a passive observer. This category contains six different scenarios detailed in the following. In all cases, a person emulated the robot's behavior.

§ Asking for Directions [AD]

In this scenario an actor asks the robot for directions to the toilets. The robot recognizes the question, performs gender identification and gives the actor the right directions to the appropriate toilets. Six different trials (four male and two female) were performed. The transcript of this scenario is in Script 1.

§ Chatting [C]

We designed this scenario to study the robot as a passive observer in a dialogue. The scenario consists of two people coming into the scene and chatting for some undetermined time, before leaving. There is no fixed script – occasionally two actors speak simultaneously – and the sequences contain several actions, e.g. hand shaking, cheering, etc. Five different trials were recorded.

Figure 2.3: A frame of the CPP sequence representative of the complexity of this scenario.

§ Cocktail Party Problem [CPP]

Reviewed in [Haykin 05], the Cocktail Party Problem has been matter of study for more than fifty years (see [Cherry 53]). In this scenario we simulated the cocktail party effect: five actors freely interact with each other, move around, appear/disappear from the camera field of view, occlude each other and speak. There is also background music and outdoor noise. In summary, this is one of the most challenging scenarios in terms of audio-visual scene analysis, action recognition, speech recognition, dialogue engaging and annotation. In the second half of the sequence the robot performs some movements. Figure 2.3 is a frame of the (left camera of the) CPP scenario. Notice the complexity of the scene in terms of number of people involved, dialogue modelling, etc.

§ Where Is Mr. Smith? [MS]

The scenario was designed to test skills such as face recognition, speech recognition and continuous dialogue. An actor comes into the scene and asks for Mr. Smith. The robot forwards the actor to Mr. Smith's office. However, he is not there and when he arrives, he asks the robot if someone was looking for him. The robot replies according to what happened. The transcript for the scenario is in Script 2. Seven trials (five male and two female) were recorded to provide for gender variability.

§ Introducing People [IP]

This scenario involves a robot interacting with three people in the scene. There are two versions of this scenario: passive and active. In the passive version the camera is static, while in the active version the camera is moving to look directly at speakers' face. Together with the *Cocktail Party Problem* scenario, they are the only exception where the robot is not static in this data set.

In the passive version of the scenario, Actor 1 and Actor 2 interact together with the Robot and each other; Actor 3: only interacts with Actor 1 and Actor 2. The transcript of the passive version is in Script 3. In the active version, Actor 1 and Actor 2 interact with the Robot and each other; Actor 3 enters and leaves room, walking somewhere behind Actor 1 and Actor 2, not looking at the Robot. The transcript of the active version is detailed in Script 4

Actor	(enters and positions him in front of the robot)
Actor	I am looking for Mr. Smith?
Robot	Yes Sir, Mr. Smith is in Room No. 22
Actor	(leaves the scene)
Mr. Smith	(enters the scene)
Mr. Smith	Hello Robot.
Robot	Hello Mr. Smith.
Robot	How can I help you?
Mr. Smith	Haven't you seen somebody looking for me?
Robot	Yes, there was a gentleman looking for you 10 minutes ago.
Mr. Smith	Thank you Bye.
Robot	You are welcome.
Mr. Smith	(leaves the scene)

Script 2: Detail of the text spoken by both actors (Actor and Mr. Smith) as well as the Robot in the "*Where is Mr. Smith?*" scenario.

2.4.4 Background Clutter

Since the RAVEL data set aims to be useful for benchmarking methods working in populated spaces, the first two categories of the data set, action recognition and robot gestures, were collected with two levels of background clutter. The first level corresponds to a controlled scenario in which there are no other actors in the scene and the outdoor and indoor acoustic noise is very limited. During the recording of the scenarios under the second level of background clutter, other actors were allowed to walk around, always behind the main actor. In addition, the extra actors occasionally talked to each other; the amount of outdoor noise was not limited in this case.

2.4.5 Data Download

The RAVEL data set is publicly available at `http://ravel.humavips.eu/` where a general description of the acquisition set up, of the data, and of the scenarios can be found. In addition to the links to the data files, we provide previews for all the recorded sequences for easy browsing previous to data downloading.

2.5 Data Set Annotation

Providing the ground truth is an important task when delivering a new data set; this allows to quantitatively compare the algorithms and techniques using the data. In this section we present two types of annotation data provided together with the data set.

Actor 1	(enters room, positions himself in front of robot and looks at robot)
Actor 1	Hello, I'm Actor 1.
Robot	Hello, I'm Nao. Nice to meet you.
Actor 2	(enters room, positions himself next to Actor 1 and looks at robot)
Robot	Excuse me for a moment.
Robot	Hello, I'm currently talking to Actor 1. Do you know Actor 1?
Actor 2	No, I don't know him.
Robot	Then let me introduce you two. What is your name?
Actor 2	Actor 2
Robot	Actor 2, this is Actor 1. Actor 1 this is Actor 2.
Actor 3	(enters room, positions himself next to Actor 1, looks at Actor 1and Actor 2)
Actor 3	Actor 1 and Actor 2, have you seen Actor 4?
Actor 2	No I'm sorry, we haven't seen her.
Actor 3	Ok, thanks. I'll have to find her myself then. Bye.
Actor 3	(leaves)
Actor 2	Actor 1, (turn heads towards robot)
Actor 1	We have to go too. Bye
Robot	Ok. See you later.

Script 3: Detail of the script of the scenario "*Introducing people - Passive*". The three people interact with the robot. The robot is static in this scenario.

2.5.1 Action/Command Performed

Manual annotation of the performed action. Each frame is labelled to the performed action/gesture, if any.

The first kind of annotation we provided is related to the action recognition and robot gesture scenarios of the data set. This annotation is done using a classical convention, that each frame is assigned a label of the particular action. Since the played action is known only one label is assigned to each frame. Because the annotation we need is not complex a simple annotation tool was designed for this purpose in which a user labels each start and end of each action/gesture in the recordings. The output of that tool is written in the standard ELAN [Brugman 04] annotation format. A screen shot of the annotation tool is shown in Figure 2.4.

2.5.2 Speaker Position and Speaking State

The second kind of annotations concern the interaction part of the data set and consists on the position of the actors (both in the images and in the 3D space) and on the speaking state of the actors. In both cases the annotator uses a semi-automatic tool that outputs an ELAN-readable output file. The semi-automatic procedures used are described in the following.

Speaker's position semi-automatic annotation procedure.

Regarding the annotation of the actors' position, the tracking algorithm de-

Actor 1	(enters room, positions himself in front of robot and looks at robot)
Actor 1	Hello, I'm Actor 1.
Robot	Hello, I'm Nao. Nice to meet you.
Actor 2	(enters room, positions himself next to Actor 1 and looks at robot)
Robot	Excuse me for a moment.
Robot	(turns head towards Actor 2)
Actor 1	(turns head towards Actor 2)
Robot	Hello, I'm currently talking to Actor 1. Do you know Actor 1?
Actor 2	No, I don't know him.
Robot	Then let me introduce you two. What is your name?
Actor 2	Actor 2
Robot	Actor 2 this is Actor 1. (turns head towards Actor 1) Actor 1 this is Actor 2.
Actor 3	(enters room, walks somewhere behind Actor 1 and Actor 2, leaves room)
Actor 1	We have to go now. Bye
Robot	(turns head towards Actor 1)
Robot	Ok. See you later.

Script 4: Detail of the script of the scenario "*Introducing people - Active*". Two out of the three people interact with the robot. The latter is a moving robot.

scribed in [Kalal 12] is used to semi-automatize the process. The annotator is asked for the object's bounding box, which is then tracked along time. At any point, the annotator can reinitialize the tracker to correct its mistakes. Once the object is tracked along the entire left camera image sequence, the correspondent trajectory in the other image is automatically estimated. To do that, the classical approach of maximizing the normalized cross-correlation across the epipolar constraint is used [Hartley 04]. From these correspondence pairs, the 3D location is computed at every frame using the DLT reconstruction procedure [Hartley 04]. The location of the speaker in the images is given in pixels and the position in the 3D space are given in millimeters with respect to the cyclopean coordinate reference frame [Hansard 08].

Speaker's speaking state semi-automatic annotation procedure.

Concerning the speaking state, the state-of-the-art voice activity detector described in [Brookes 11] is used on the per-actor close range microphones. In a second step, the annotator is in charge of discarding all false positives generated by the VAD, leading to a clean speaking state annotation per each actor.

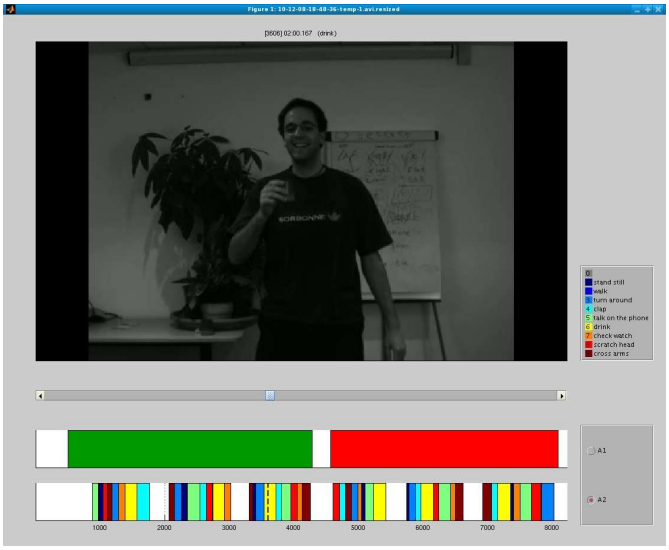


Figure 2.4: The annotation tool screen shot. Two time lines are shown below the image. The first one (top) is used to annotate the level of background clutter. The second one (bottom) details which action is performed at each frame.

2.6 Data Exploitation Examples

In order to prove the relevance and usability of the RAVEL data set, a few data exploitation examples are provided. Three different examples, showing how diverse applications can use the presented data set, are explained in this section: a scene flow extraction method, an event-detection algorithm based on statistical audio-visual fusion techniques and two machine learning-based action recognition methods.

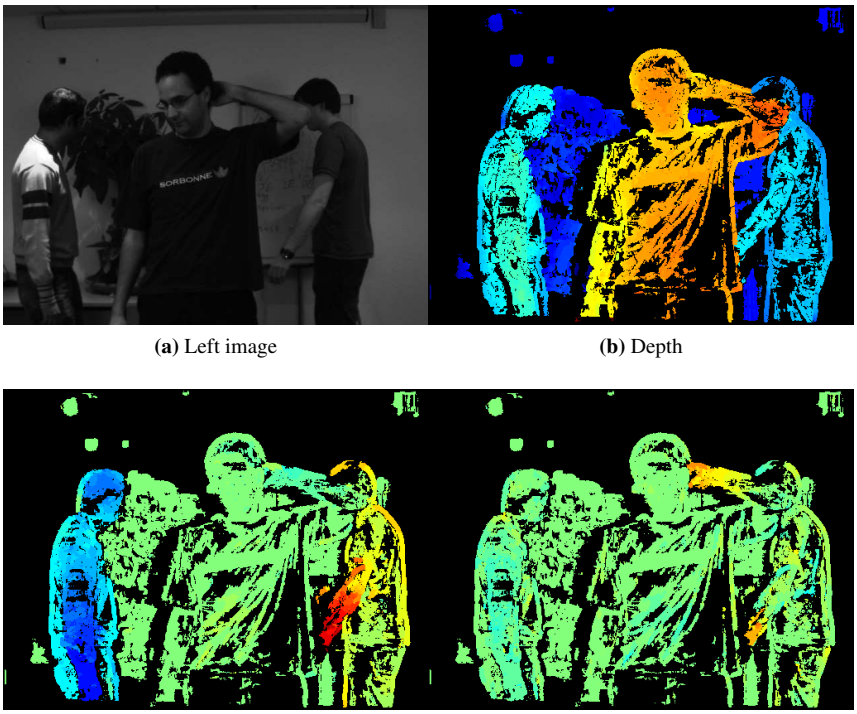
2.6.1 Scene Flow

What is scene flow?

Since the entire data set is captured by synchronized and fully calibrated cameras, it is possible to compute a 3D scene flow [Vedula 05], which is a classical low-level computer vision problem. The 3D scene flow is defined as a motion field such that each reconstructed pixel for a frame has assigned a 3D position and a 3D velocity. It leads to an image correspondence problem, where one has to simultaneously find corresponding pixels between images of a stereo pair and corresponding pixels between subsequent frames.

Scene flow results on RAVEL.

After the 3D reconstruction using the known camera calibration, these correspondences fully determine the 3D scene flow. A projection of a scene flow is shown in Figure 2.5, as a disparity (or depth) map and horizontal and vertical optical flow maps. These results are computed using a recent seed growing algorithm [Čech 11]. The scene flow results can be used for further processing towards the understanding of a dynamic scene.



(a) Left image **(b)** Depth

(c) Horizontal motion **(d)** Vertical motion

Figure 2.5: Results of the scene flow algorithm. The original left image is shown in (a). The actual results are colour coded. For depth map (b), warmer colours are closer to the camera. For horizontal (c) and vertical (d) motion maps, green colour stands for zero motion, while colder colours correspond to right and up motion respectively, warmer colours the opposite direction. Black colour stands for unassigned disparity or optical flow.

2.6.2 Audio-Visual Event Detection

What can we use to detect audio-visual events on RAVEL?

How to detect audio-visual events, i.e. events that are both heard and seen, is a topic of interest for researchers working in multimodal fusion. An entire pipeline – from the raw data to the concept of AV event – is exemplified in this section. This pipeline consists of three modules: visual processing, auditory processing and audio-visual fusion. In the following, the method is roughly described; interested readers can find a more detailed explanation in [Alameda-Pineda 11] and Chapter 3.

Speaker detection on the CPP sequence.

The algorithm was applied onto the *CPP* sequence of the RAVEL data set. Figure 2.6 shows the results of the method in nine frames of the sequence. In this sequence the AV events are people in an informal social gathering. Although the method has some false positives, it correctly detects and localizes 26 objects out of 33 (78.8%).

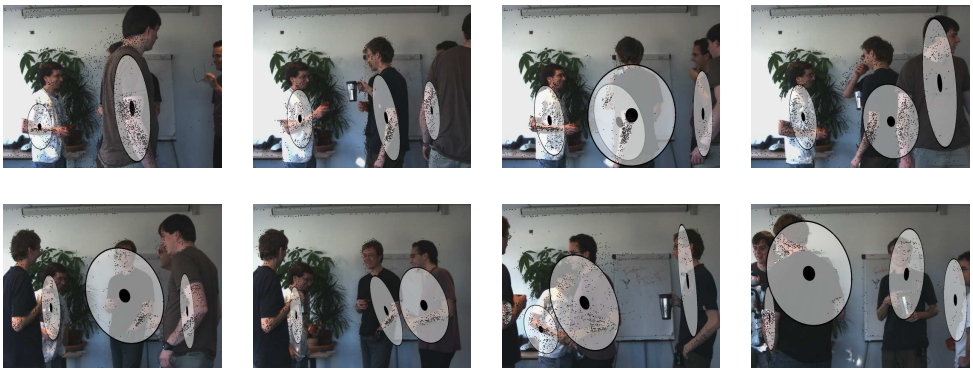


Figure 2.6: A sample of the AV events detected in the *CPP* sequence of the RAVEL data set. The ellipses correspond to the localization of the events in the image plane. The method correctly detects and localizes 26 objects out of 33 (78.8%).

2.6.3 Action Recognition

To demonstrate some of the potentialities of the RAVEL data set we establish a baseline for the Action Recognition subset of RAVEL. In this section we show the performance of the state-of-the-art methods when applied to the RAVEL data set. The results are split depending on the application: either "isolated" recognition or "continuous" recognition.

§ Isolated Recognition

Overview of the Bag-of-words

Among all the different methods to perform isolated action recognition, we decide to use the one described in [Laptev 05]. Its performance is comparable to the state-of-the-art methods and binaries can be easily found and downloaded from here[1]. This method represents an action as a histogram of visual words. Once all the actions are represented, a Support Vector Machine (SVM) is used to learn each class (action) to afterwards determine if an unknown action belongs to one of the classes previously trained. This methodology is known as a Bag-of-Words (BoW) and it can be summarized into five steps: (i) collect set of features for all actions/actors for each video clip, (ii) apply clustering algorithm to these features, for instance, k-means, (iii) apply 1-NN to classify the features of each action into the centroids found by k-means, (iv) obtain a histogram of k bins and (v) Train a classifier with these histograms, for instance, SVM.

Our particular set up: the use of *Laptev* features and SVM

In this experiment the *Laptev* features are used. These features correspond to a set of spatio-temporal Harris detected points described by a concatenation of Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF) descriptors [Dalal 05]. The clustering method to select the k most representative features (visual words) is k-means. We then represent each action as a histogram

[1] http://www.irisa.fr/vista/Equipe/People/Ivan.Laptev.html

Table 2.2: Results of the experiments on isolated action recognition. Recognition rates of the *Laptev* features using different number of clusters for the k-means algorithm for the controlled and normal levels of background clutter.

k	500	1000	2000	3000
Controlled	0.6320	0.6883	0.6926	0.6797
Normal	0.4892	0.4675	0.4762	0.5281

of visual words. Finally a linear multiclass SVM is trained on these histograms. In order to obtain statistically significant results, we evaluate the method's performance using a leave-one-out cross-validation strategy.

Results of isolated action recognition.

In addition to the recognition rates we show several confusion matrices. The ij^{th} position of a confusion matrix represents the amount of instance of the i category classified as the j category. Figures 2.7a, to 2.7f show the confusion matrices when the characters 2, 3 and 11 were tested. The matrices on the top correspond to the scenarios under controlled background clutter and the matrices on the bottom to the scenarios under normal background clutter. We can observe the expected behaviour: the matrices under the controlled conditions report much better results than those under normal conditions. In addition, we observe some variation across different actors on where are the wrong detections. This is caused by two effects: the different ways of performing the actions and the various co-articulations. All together justifies the use of a cross-validation evaluation strategy. Figures 2.7g and 2.7h report the global confusion matrices, from which we can observe that the expected behaviour regarding the performance on controlled vs. normal clutter level, observed before is extensible to the entire data set. Finally, Table 2.2 reports the average recognition rate (that is the average of the diagonal of the confusion matrix) for different values of k.

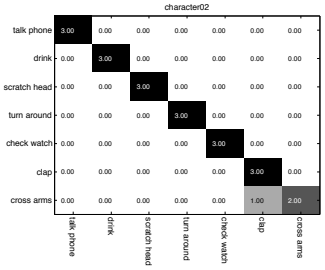
§ Continuous Recognition

Continuous action recognition

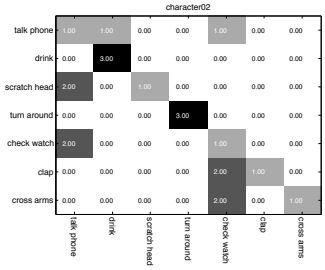
Continuous action recognition, or joint segmentation and classification, refers to the case where a video to be analysed contains a *sequence* of actions played by one actor or by different actors. The order of the actions in the video is not known. Most of the earlier methods assume that the segmentation and classification tasks may be carried out completely independently of each other, i.e., they consider an isolated recognition scenario where the boundaries of action in videos are known a priori. In continuous recognition scenarios the objective is to find the best label sequence for each video frame. The isolated recognition framework of representing an action as a single histogram of visual words can be modified to perform continuous action recognition.

Two approaches for continuous action recognition

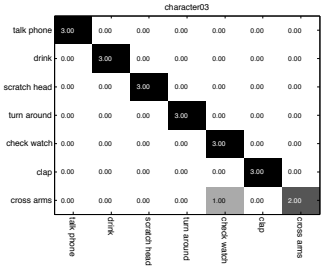
In [Shi 11], the authors propose a method which uses SVM for classification of each action and the temporal segmentation is done efficiently using dynamic programming. Multi-class SVMs are trained on a segmented training set. In the classification stage, actions are searched over several temporal segments at different time scales. Each temporal segment is represented by a single histogram.



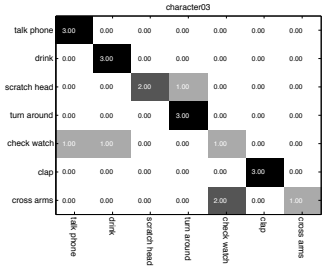
(a) Character 2 - Controlled



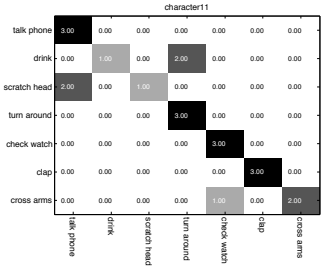
(b) Character 2 - Normal



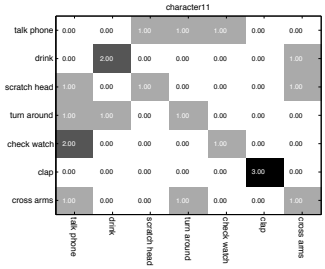
(c) Character 3 - Controlled



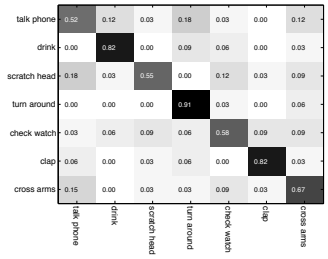
(d) Character 3 - Normal



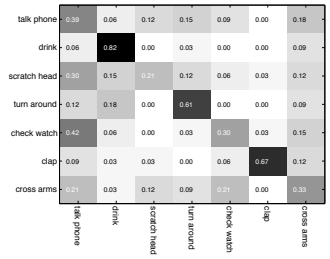
(e) Character 11 - Controlled



(f) Character 11 - Normal



(g) All characters - Controlled



(h) All characters - Normal

Figure 2.7: Confusion matrices with $k = 2000$ clusters using Laptev features. The first column corresponds to the controlled background and the second column to the normal scenario. First row uses the 2^{nd} actor as test, second row uses the 3^{rd} actor, third row uses the 11^{th} actor and last row is the global confusion matrix.

Table 2.3: Accuracy of the continuous recognition methods using artificially merged actions (Weizmann and Hollywood) and actions involving smooth transitions (RAVEL).

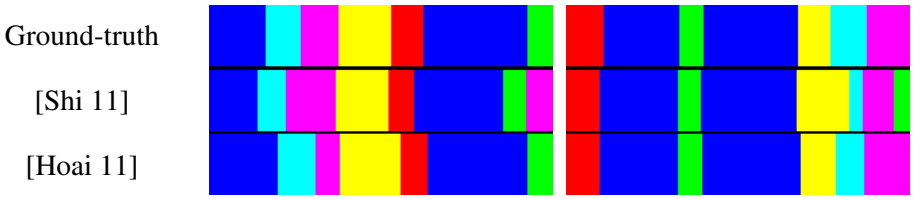
Dataset:	Weizmann	Hollywood	RAVEL
Shi et. al. [Shi 11]	69.7	34.2	55.4
Hoai et. al. [Hoai 11]	87.7	42.2	59.9

The search over the time scale is restricted by the maximum and minimum lengths of actions computed from the training set. Each of these segments are classified by SVM trained on the action classes. This classification yields ordered sets of labels for the segments. To find the best set of labels for the whole video one needs an optimization criteria. In [Shi 11] the optimization criteria is to maximize the sum of the SVM classifier scores computed by concatenating segments over different temporal scales. This optimization is efficiently cast in the dynamic programming framework.

Both [Shi 11] and [Hoai 11] are similar in the way they perform continuous action recognition, i.e., the classification is done at different temporal scales using SVMs, while the segmentation is efficiently done using dynamic programming. The crucial difference between these two methods is the optimization criteria used for dynamic programming. In [Shi 11], the sum of the SVM scores for the concatenated segments is maximized. This ensures the best sequence of labels for the whole video but does not ensure that the best label is assigned to each segment. This problem is overcome in [Hoai 11] where a difference between the SVM score of the winning class label for a segment and the next best label is computed. The sum of these differences computed for each segment is then maximized over concatenated segments at different time scales over the whole video. This optimization is also cast in the dynamic programming framework.

Results of continuous action recognition

Results on the RAVEL dataset using the state-of-art continuous recognition algorithms [Hoai 11, Shi 11] are shown in Table 2.3. The accuracy of the algorithms were measured by percentage of correctly labeled frames. The recognition accuracy is also computed on Weizmann [Gorelick 07] and Hollywood datasets [Laptev 08]. Since these dataset contain isolated actions only, we created a sequence of multiple actions by concatenating single-action clips following the protocol of [Hoai 11]. This concatenation creates abrupt artificial inter-action transitions. In contrast, the RAVEL dataset is recorded continuously in one shot per actor. The actions are played by the actors in random order (given by a hidden prompt) and moreover we did not instruct the actors to come to a rest position after every action. Therefore this dataset is well suitable for the the continuous action recognition. In Figure 2.8 we show the estimated labels of two video sequences by [Hoai 11, Shi 11] in a comparison with the ground-truth segmentation.



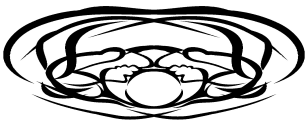
(a) actor 1 **(b)** actor 4

Figure 2.8: Continuous action recognition results on the RAVEL datasets. Colours encodes action labels of frames of the video sequences. Top row shows ground-truth labelling, while two rows below show results of two state-of-the-art algorithms [Shi 11, Hoai 11]. The results are shown for two selected actors.

2.7 Conclusions

This chapter introduces the RAVEL data set, which consists of multimodal (visual and audio) multichannel (two cameras and four microphones) synchronized data sequences. The data set embodies several scenarios designed to study different HRI applications. This new audiovisual corpus is important for two main reasons. On one hand, the stability and characteristics of the acquisition device ensure the quality of the recorded data and the repeatability of the experiments. On the other hand, the amount of data is enough to evaluate the relevance of the contents in order to improve the design of future HRI systems.

The acquisition set up (environment and device) was fully detailed. Technical specifications of the recorded streams (data) were provided. The calibration and synchronization procedures, both visual and audio-visual, were described. The scenarios were detailed; their scripts were provided when applicable. The recorded scenarios fall in three categories representing different groups of applications: action recognition, robot gesture and interaction. Furthermore, the data set annotation method was also described. Finally, three examples of data exploitation were provided: scene flow extraction, audio-visual event detection and action recognition. These prove the usability of the RAVEL data set.



CHAPTER 3

AUDIO-VISUAL

SPEAKER LOCALISATION

Natural human-robot interaction in complex and unpredictable environments is one of the main research lines in robotics, scene understanding and social computing. In typical real-world scenarios, humans are at some distance from the robot and the acquired signals are strongly impaired by noise, reverberations and other interfering sources. In this context, the detection and localisation of speakers plays a key role since it is the pillar on which several tasks (e.g.: speech recognition and speaker tracking) rely. We address the problem of how to detect and localize people that are both seen and heard by a humanoid robot. We introduce a hybrid deterministic/probabilistic model. Indeed, the deterministic component allows us to map the visual information into the auditory space. By means of the probabilistic component, the visual features guide the grouping of the auditory features in order to form AV objects. The proposed model and the associated algorithm 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 performed experiments on (i) synthetic data, (ii) a publicly available data set and (iii) data acquired using the robot. The results we obtained validate the approach and encourage us to further investigate how vision can help robot hearing.

3.1 Introduction

In this chapter, we address the problem of detecting and localizing objects that can be both seen and heard, e.g., people emitting sounds such as speech, sounds produced by foot steps and clothe chafing, etc. This immediately raises the interesting question of how to optimally associate, fuse, and cluster observations that are gathered with physically different sensors, e.g., cameras and microphones, and that live in semantically different spaces, e.g., how to associate *spatiotemporal visual features* with *temporal auditory signals.*

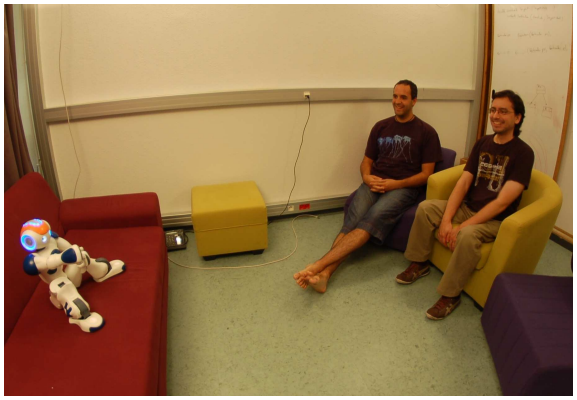


Figure 3.1: A typical scenario in which a companion humanoid robot (NAO) performs audio-visual fusion in an attempt to detect the auditory status of each one of the speakers in the room. The system described processes the raw data gathered with the robot's camera and microphone pairs. The system output is a speaking probability of each one of the actors together with the 3D location of the actors' faces.

Multispeaker detection and localisation

Among all possible applications using audio-visual data, we are interested in detecting multiple speakers in informal scenarios. A typical example of such a scenario is shown in Figure 3.1, in which two people are sitting and chatting in front of the robot. The robot's primary task (prior to speech recognition, language understanding, and dialogue) consists in retrieving the auditory status of several speakers along time. This will allow the robot to concentrate its attention on one of the speakers, *i.e.* , use the speaker's location to optimize the emitter-to-receiver pathway and to attempt to separate the auditory and visual data coming from several speakers. We note that this problem cannot be solved within the traditional human-computer interaction paradigm which is based on *tethered* interaction (the user must wear a close-range microphone) and which primarily works for a single-person-to-robot communication. This considerably limits the the range of potential interactions between robots and people engaged in a cooperative task or simply in a multi-party dialogue. In this chapter we investigate *untethered* interaction thus allowing a robot with its *onboard sensors* to perceive the status of several people at once and to communicate with them in the most natural way. Consequently, we are restricted to the use of egocentric data. The Thesis' contribution to this topic is three-fold.

Contributions

First, a hybrid deterministic/probabilistic model for audio-visual fusion **AVS1**. On one hand, the deterministic components allow us 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. In all, the hybrid model provides a link between the auditory and visual feature spaces and a maximum likelihood framework to estimate the number of speakers, their position and their speaking status. Second, an audio-visual expectation-maximization algorithm that is theoretically sound, efficient, intuitive, and yields very interesting results **AVS2**.

Indeed, it performs clustering in the one-dimensional Interaural Time Difference (ITD) space associated with two microphones and it takes full advantage of a generative model that allows, first to *project* visual observations onto this space and second to *back-project* the detected 1D clusters into the 3D physical space without any additional computational effort. We show experiments performed on the publicly available data set CAVA that were published in [Alameda-Pineda 11]. Third, an original system approach to tackle the problem of on-line audio-visual detection of multiple speakers using the companion humanoid robot NAO[1] **AVS3**. Implemented on top of a platform-independent middleware, the algorithm works on-line with good performance. The 3D positions of the speakers' heads are obtained from the stereo image pair, and Interaural Time Difference (ITD) values are extracted from the auditory data. These cues are then fused in a probabilistic manner in order to compute the speaking status of each person over time (see [Sanchez-Riera 12b]).

The remainder of the chapter is organized as follows: Section 3.2 delineates the related published work, Section 3.3 outlines the hybrid deterministic/probabilistic model, Section 3.4 gives the details of the auditory and visual extracted features, Sections 3.5 and 3.6 describe the multimodal inference procedure as well as its on-line implementation on the humanoid robot NAO, Section 3.7 shows the results we obtained and Section 3.8 draws some conclusions and future work guidelines.

3.2 Related Work

Non-parametric approaches

The existing literature on speaker detection and localisation can be grouped into two main research lines. On one side, many statistical non-parametric approaches have been developed. Indeed, [Gurban 06], [Besson 08b] and [Besson 08a] investigate the use of information theory-based 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. Even though these approaches show very good performance results, they use speaker/object dedicated cameras, thus limiting the interaction. Moreover, the cited non-parametric approaches need a lot of training data. The outcome of such training steps is also environment-dependent. Consequently, implementing such methods on mobile platforms results in systems with almost no practical adaptability.

Parametric approaches

On the other side, several probabilistic approaches have been published. In [Khalidov 08], [Khalidov 11], the authors introduce the notion of conjugate GMM for audio-visual fusion. Two GMMs are estimated, one for each modality (vision and auditory) while the two mixture parameter sets are constrained through 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 audio-visual features. All these methods simultaneously detect

[1] http://www.aldebaran-robotics.com

and localize the speakers but they are not suitable for real-time processing, because of their algorithmic complexity. [Kim 07] proposed a Bayesian framework inferring the position of the active speaker and combining a sound source localisation technique with a face tracking algorithm on a robot. The reported results are good in the case of one active speaker, but show bad performance for multiple/far speakers. This is due to the fact that the proposed probabilistic framework is not able to correctly handle outliers.

Main attributes of our approach

Unlike these recent approaches, we propose a novel hybrid deterministic/probabilistic model for audio-visual detection and localisation of speaking people. Up to the authors' knowledge, we introduce the very first model with the following remarkable attributes all together: (i) theoretically sound and solid, (ii) designed to process egocentric data, (iii) accommodating different visual and auditory features, (iv) robust to noise and outliers, (v) requiring a once-and-forever tiny calibration step guaranteeing the adaptability of the system, (vi) working on unrestricted indoor environments, (vii) handling a variable number of people and (viii) implemented on a humanoid platform.

3.3 A Hybrid Deterministic/Probabilistic Model

Model unknowns

We introduce a multimodal deterministic/probabilistic fusion model for audio-visual detection and localisation of speaking people that is suitable for real-time applications. The algorithms derived from that hybrid model aim to count how many speakers are there, find them in the scene and ascertain when they speak. In other words, we seek for the number of potential speakers, $N \in \mathbb{N}$, their positions $\boldsymbol{S}_n \in \mathbb{S}$ ($\mathbb{S} \subset \mathbb{R}^3$ is the scene space) and their speaking state $e_n \in \{0, 1\}$ (0 – *not speaking* and 1 – *speaking*).

Model observations

In order to accomplish the detection and localization of speakers, auditory and visual features are extracted from the raw signals (sound track and image flow), during a time interval Δt. We assume Δt to be short enough such that the speakers remain approximately in the same 3D location and long enough to capture small displacements and oscillatory movements of the head, hands, torso and legs. The auditory and visual features extracted during Δt are denoted by $\mathbf{a} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k, \ldots, \boldsymbol{a}_K\} \subset \mathbb{A}$ and by $\mathbf{v} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m, \ldots, \boldsymbol{v}_M\} \subset \mathbb{V}$ respectively, where $\mathbb{A}$ ($\mathbb{V}$) is the auditory (visual) feature space.

Formal task

We aim to solve the task from the auditory and visual observations. That is, we want to compute the values of N, $\{\boldsymbol{S}_n\}_{n=1}^N$ and $\{e_n\}_{n=1}^N$, that best explain the extracted features $\mathbf{a}$ and $\mathbf{v}$. Therefore, we need a framework that encompasses all (hidden and observed) variables and that accounts for the following challenges: (i) the visual and auditory observations lie in physically 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, (iv) the relative importance of the two types of data is unassessed, (v) the position and speaking state of the speakers has to be gauged and (vi) since we

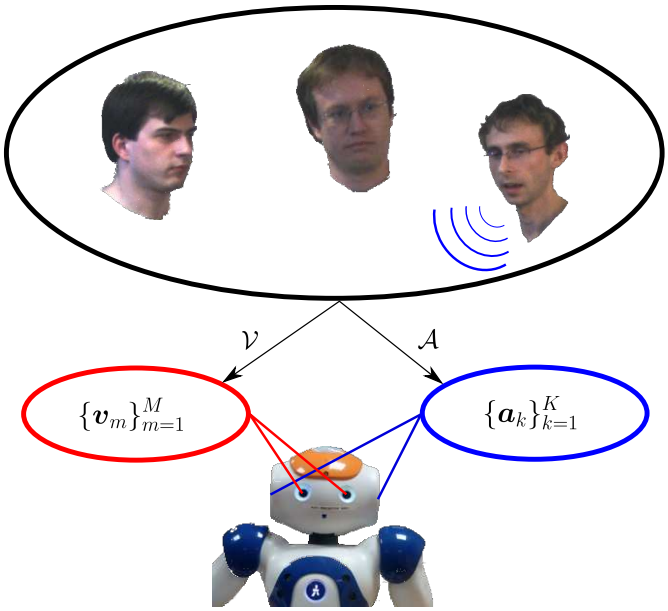


Figure 3.2: Perceptual auditory ($\mathcal{A}$) and visual ($\mathcal{V}$) mappings of NAO. The extracted auditory $\boldsymbol{a}_k$ and lay around $\mathcal{A}(\boldsymbol{S})$ and $\mathcal{V}(\boldsymbol{S})$ respectively. An audio-visual mapping needs to be built to link the two observations spaces.

want to be able to deal with a variable number of AV objects over a long period of time, the number of AV object that are effectively present in the scene must be estimated.

Model decomposition

We propose a hybrid deterministic/probabilistic framework performing audio-visual fusion, seeking for the desired variables and accounting for the outlined challenges. On one hand, the deterministic components allow us 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, the feature noise and outliers, which is a consequence of the contents of the scene as well as the feature extraction procedure.

3.3.1 The Deterministic Model

Visual and auditory mappings

In this section we delineate the deterministic components of our hybrid model: namely the visual and auditory mappings. Because the scene space, the visual space and the auditory 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 represented in Figure 3.2. An AV object placed at $\boldsymbol{S}$ in the scene space, is virtually placed at $\mathcal{A}(\boldsymbol{S})$ in the auditory space and at $\mathcal{V}(\boldsymbol{S})$ in the visual space.

Audio-visual mapping

The definition of $\mathcal{A}$ and $\mathcal{V}$ provide a link between the two observations spaces, which corresponds either to $\mathcal{A} \circ \mathcal{V}^{-1}$ or to $\mathcal{V} \circ \mathcal{A}^{-1}$. Depending on the extracted features and on the sensors, the mappings $\mathcal{A}$ and $\mathcal{V}$ may be invertible. If that is not the case, $\mathcal{A} \circ \mathcal{V}^{-1}$ or $\mathcal{V} \circ \mathcal{A}^{-1}$ should be estimated through a learning procedure. There are several works already published dealing with this problem in different ways. In [Alameda-Pineda 11, Sanchez-Riera 12c], $\mathcal{V}$ is invertible and $\mathcal{A}$ is known, so building $\mathcal{A} \circ \mathcal{V}^{-1}$ is straightforward. In sound source localization approaches (inter alia [Nakadai 04]) $\mathcal{A}$ is invertible and $\mathcal{V}$ is known so $\mathcal{V} \circ \mathcal{A}^{-1}$ is easily constructed. In [Khalidov 08, Khalidov 11], none of the mappings are inverted, but used to tie the parameters of the probabilistic model. So the link between $\mathbb{A}$ and $\mathbb{V}$ is not used explicitly, but implicitly. 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 our case, we chose to extract 3D visual feature points, and represent them in the scene coordinate system (see Section 3.4.2). Thus, the mapping $\mathcal{V}$ is the identity, which is invertible. The auditory features correspond to the Interaural Time Differences (see Section 3.4.1), 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 3.4.3). Consequently, we are able to map the visual features $\mathbf{v} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_M\}$ onto the auditory space $\mathbb{A}$. We will denote the projection of $\boldsymbol{v}_m$ by $\tilde{\boldsymbol{v}}_m$:

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

Summarizing, we use the mapping from $\mathbb{V}$ to $\mathbb{A}$ to map all visual features onto the auditory space. Hence, all extracted features lie, now, in the same space, and we can perform the multimodal fusion in there.

3.3.2 The Probabilistic Model

Hidden variables

Thanks to the link built in the previous section, we obtain a set of projected visual features $\tilde{\mathbf{v}} = \{\tilde{\boldsymbol{v}}_1, \dots, \tilde{\boldsymbol{v}}_M\}$, laying in the same space as the auditory features $\mathbf{a}$. These features need to be grouped to construct audio-visual objects. However, we do not know which observation is generated by which object. Therefore, we introduce two sets of hidden variables $\boldsymbol{Z}$ and $\boldsymbol{W}$:

$$\begin{aligned} \boldsymbol{Z} &= \{Z_1, \dots, Z_m, \dots, Z_M\} \\ \boldsymbol{W} &= \{W_1, \dots, W_k, \dots, W_K\}, \end{aligned}$$

accounting for the observation-to-object assignment. The notation $Z_m = n$ ($m \in \{1, \dots, M\}$, $n \in \{1, \dots, N+1\}$) means that the projected visual observation $\tilde{\boldsymbol{v}}_m$ was either generated by the n^{th} 3D object ($n \in \{1, \dots, N\}$) or it is an outlier ($n = N+1$). Similarly, the variable W_k is associated to the auditory observation $\boldsymbol{a}_k$.

The generative model

We formulate the multimodal probabilistic fusion model under the assumption that all observations $\tilde{\boldsymbol{v}}_m$ and $\boldsymbol{a}_k$ are independent and identically distributed. The n^{th} AV object generates both visual and auditory features normally distributed around $\mathcal{A}(\boldsymbol{S}_n)$ and both the visual and auditory outliers are uniformly distributed in $\mathbb{A}$. Therefore, we write:

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

where Θ contains the Gaussian parameters, that is $\mu_n = \mathcal{A}(\boldsymbol{S}_n)$ and σ_n (the mean and the standard deviation of the n^{th} Gaussian). The exact same rule holds for $\mathrm{P}(\boldsymbol{a}_m|W_m = n, \Theta)$. Thus we can define a generative model for the observations $x \in \mathbb{A}$:

$$\mathrm{p}(x; \Theta) = \sum_{n=1}^{N} \pi_n \mathcal{N}(x; \mu_n, \sigma_n) + \pi_{N+1} \mathcal{U}(x; \mathbb{A}), \tag{3.1}$$

where π_n is the prior probabilities 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:

$$\Theta = \{\pi_1, \ldots, \pi_{N+1}, \mu_1, \ldots, \mu_N, \sigma_1, \ldots, \sigma_N\}. \tag{3.2}$$

Maximum likelihood formulation

Under the probabilistic framework described, the set of parameters is estimated within a maximum likelihood formulation:

$$\mathcal{L}(\tilde{\mathbf{v}}, \mathbf{a}; \Theta) = \sum_{m=1}^{M} \log \mathrm{p}(\tilde{\boldsymbol{v}}_m; \Theta) + \sum_{k=1}^{K} \log \mathrm{p}(\boldsymbol{a}_k; \Theta). \tag{3.3}$$

In other words, the optimal set of parameters is the one maximizing the log-likelihood function (3.3), where p is the generative probabilistic model in (3.1). Unfortunately, direct maximization of (3.3) is an intractable problem. Equivalently, the expected complete-data log-likelihood will be maximized [Dempster 77] (see Section 3.5).

The final goal

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 activity $e_1, \ldots, e_n, \ldots, e_N$. 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 $\tilde{\boldsymbol{v}}_m \leftrightarrow \boldsymbol{v}_m$ 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 3.5.2 and 3.5.3 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 $\mathcal{A} \circ \mathcal{V}^{-1}$.

3.4 Finding Auditory and Visual Features

In this section we describe the auditory (Section 3.4.1) and the visual (Section 3.4.2) features we extract from the raw data. Given this features, the definition of $\mathcal{A}$ and $\mathcal{V}$ is straightforward. However, the computation of the mapping's parameters is done through a calibration procedure detailed in Section 3.4.3.

3.4.1 Auditory Features

An auditory observation $\boldsymbol{a}_k$ corresponds to an Interaural Time Difference (ITD) between the left and right microphones. Because the ITDs are real-valued, the auditory feature space is $\mathbb{A} = \mathbb{R}$. One ITD value corresponds to the different of time of arrival of the sound signal between the left and right microphones. For instance, the sound wave of a speaker located in the left-half of the scene will obviously arrive earlier to the left microphone than to the right microphone. We found that the method proposed in [Christensen 07] yields very good results that are stable over time. The relationship between an auditory source located at $\boldsymbol{S} \in \mathbb{R}^3$ and an ITD observation $\boldsymbol{a}$ depends on the relative position of the acoustic source with respect to the locations of the left and right microphones, $\boldsymbol{M}_L$ and $\boldsymbol{M}_R$. If we assume direct sound propagation and constant sound velocity ν, this relationship is given by the mapping $\mathcal{A} : \mathbb{S} \rightarrow \mathbb{A}$ defined as:

$$\mathcal{A}(\boldsymbol{S}) = \frac{\|\boldsymbol{S} - \boldsymbol{M}_L\| - \|\boldsymbol{S} - \boldsymbol{M}_R\|}{\nu}. \tag{3.4}$$

3.4.2 Visual Features

The visual observations are 3D points extracted using binocular vision. We used two types of features: the Harris-Motion 3D (HM3D) points and the faces 3D (F3D).

HM3D The first kind of features we extracted are called Harris-motion points. We first detect Harris interest points [Harris 88] in the left and right image pairs of the time interval Δt. Second, we only consider a subset of theses points, namely those points where motion occurs. For each interest-point image location (u, v) we consider the image intensities at the same location (u, v) in the subsequent images and we compute a temporal intensity standard deviation $\tau_{(u,v)}$ for each interest point. Assuming stable lighting condition over Δt, we simply classify the interest points into static ($\tau_{(u,v)} \leq \tau_M$) and dynamic ($\tau_{(u,v)} > \tau_M$) where τ_M is a user-defined threshold. Third, we apply a standard stereo matching algorithm and a stereo reconstruction algorithm [Hartley 04] to yield a set of 3D features $\mathbf{v}$ associated with Δt.

F3D The second kind of features are the 3D coordinates of the speakers' faces. They are obtained using the face detector in [Šochman 05]. More precisely,

the center of the bounding box retrieved by the face detector is matched to the right image and the same stereo reconstruction algorithm as in HM3D is used to obtain **v**.

Both 3D features are expressed in cyclopean coordinates [Hansard 08], which are also the scene coordinates. Consequently, the visual mapping $\mathcal{V}$ is the identity mapping. In conclusion, because we are able to accurately model the geometry of the visual sensors, we can assume that $\mathcal{V}$ is invertible and explicitly build the linking mapping $\mathcal{A} \circ \mathcal{V}^{-1}$.

3.4.3 Calibration

In the two previous sections we described the auditory and the visual features respectively. As a consequence, the mappings $\mathcal{A}$ and $\mathcal{V}$ are defined. However, we made implicit use of two, a priory unknown, objects. On one hand the stereo-matching and the 3D reconstruction algorithms need the so-called stereo-calibration. That is, the projection matrices corresponding to the left and right cameras which are estimated using [Bouguet 08]. It is worth to remark that the calibration procedure allows us to accurately represent any point in the field-of-view of both cameras as a 3D point. On the other hand, and in order to use $\mathcal{A}$, we need to know the positions of the microphones $\boldsymbol{M}_L$ and $\boldsymbol{M}_R$ in the scene coordinate frame, which is slightly more complex. Since the scene coordinates are the same as the visual coordinates, we refer to this as "audio-visual calibration". We manually measure the values of $\boldsymbol{M}_L$ and $\boldsymbol{M}_R$ with respect to the stereo rig. However, because these measurements are imprecise, an affine correction model needs to be applied:

Stereo-calibration

Audio-visual calibration

$$\overline{\mathcal{A}}(\boldsymbol{S}) = c_1\,\mathcal{A}(\boldsymbol{S}) + c_0 = c_1\,\frac{\|\boldsymbol{S}-\boldsymbol{M}_L\| - \|\boldsymbol{S}-\boldsymbol{M}_R\|}{\nu} + c_0, \tag{3.5}$$

where c_1 and c_0 are the adjustment coefficients. In order to estimate c_1 and c_0, a person with a speaker held just below the face moves in a zig-zag trajectory in the entire visual field of view of the two cameras. The 3D position of the person's face and the ITD values were extracted. We used white noise because it correlates very well resulting in a single sharp peak in ITD space. In many experiments, we did not observe any effect of reverberations, because the reverberant components are suppressed by the direct component of the long lasting white noise signal. The optimal values for c_1 and c_0, in the least square sense, were computed from these data. Figure 3.3 shows the extracted ITDs (red-circle), the projected faces before (blue) and after (green) the affine correction. We can clearly see how the affine transformation enhanced the audio-visual linking mapping. Hence the projected visual features have the following expression:

Final mapping

$$\tilde{\boldsymbol{v}}_m = (\overline{\mathcal{A}} \circ \mathcal{V}^{-1})(\boldsymbol{v}_m) = c_1\,\frac{\|\boldsymbol{v}_m-\boldsymbol{M}_L\| - \|\boldsymbol{v}_m-\boldsymbol{M}_R\|}{\nu} + c_0. \tag{3.6}$$

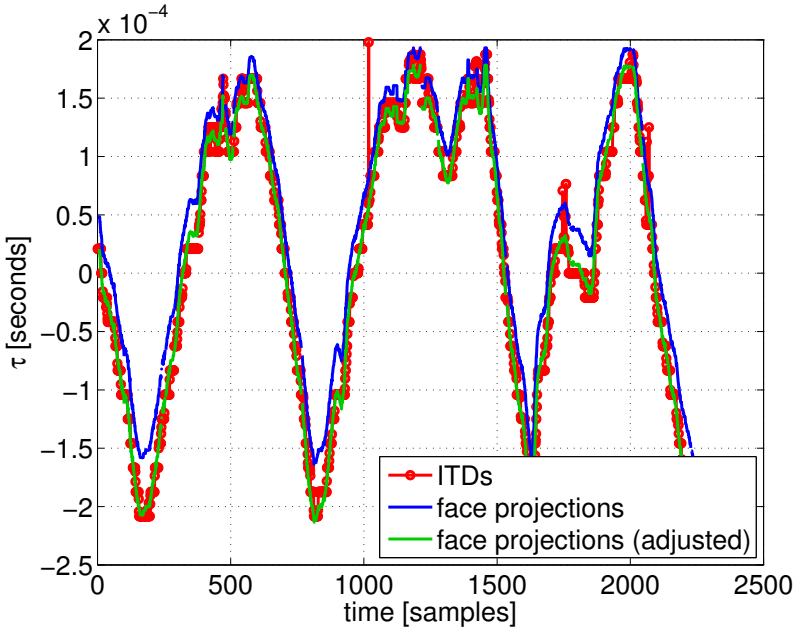


Figure 3.3: Affine correction of the audio-visual calibration. Extracted ITD values are plot in red-circled. F3D features projected into the ITD space using Equation (3.4) are plot in blue. F3D features projected using Equation (3.6), that is after the audio-visual calibration step, are plot in green.

The outlined calibration procedure has three main advantages: (i) it requires very few training data, (ii) it lasts a long period of time and (iii) it is environment-independent, thus guaranteeing the system's adaptability. Indeed, in our case, the calibration ran on a one-minute audio-visual sequence and has been successfully used for the last 18 months in several rooms, including project demonstrations and conference exhibits. Consequently, the robustness of the once-for-all tiny audio-visual calibration step is proved up to a large extent.

3.5 Multimodal Inference

Remaining issues

In Section 3.3 we set up the maximum-likelihood framework to perform AV fusion. The 3D visual features are mapped into the auditory space $\mathbb{A}$ through the audio-visual mapping $(\overline{\mathcal{A}} \circ \mathcal{V}^{-1})$. This mapping takes the form in (3.6) when using the auditory and visual features described in Section 3.4. However, three of the initial issues remain unsolved: (i) the relative importance of each modality, (ii) the estimates for $\boldsymbol{S}_n$ and e_n and (iii) the variable number of AV objects, N. In this Section we described EM-based method solving the ML problem with hidden variables and accounting for these unsolved issues.

3.5.1 Visual Guidance

Relative modality importance

Previous papers do not agree on how to balance the relative importance of each modality. After a deep analysis of the features' statistics, we choose to use the visual information to guide the clustering process of the sparse auditory observations. Indeed, because the HM3D visual features are more dense and have better temporal continuity than the ITD values, we start by fitting a 1D GMM to the projected visual features $\{\tilde{\boldsymbol{v}}_m\}_{m=1}^{M}$. This is done with the standard EM algorithm [Bishop 06]. In the E step of the algorithm the posterior probabilities $\alpha_{mn} = \mathrm{P}(Z_m = n|\tilde{\mathbf{v}}, \Theta)$ are updated via the following formula:

$$\alpha_{mn} = \frac{\pi_n \, \mathrm{P}(\tilde{\boldsymbol{v}}_m|Z_m = n, \Theta)}{\sum_{i=1}^{N+1} \pi_i \, \mathrm{P}(\tilde{\boldsymbol{v}}_m|Z_m = i, \Theta)}. \tag{3.7}$$

The M step is devoted to maximize the expected complete data log-likelihood with respect to the parameters, leading to the standard formulas (with $\bar{\alpha}_n = \sum_{m=1}^{M} \alpha_{mn}$):

$$\begin{aligned}
\pi_n &= \frac{\bar{\alpha}_n}{M}, \\
\mu_n &= \frac{1}{\bar{\alpha}_n} \sum_{m=1}^{M} \alpha_{mn} \tilde{\boldsymbol{v}}_m, \\
\sigma_n^2 &= \frac{1}{\bar{\alpha}_n} \sum_{m=1}^{M} \alpha_{mn} (\tilde{\boldsymbol{v}}_m - \mu_n)^2.
\end{aligned}$$

The visual guidance

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 (3.3) subject to the constraint that the posterior probabilities α_{mn} were previously computed. This leads to *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 applied to enforce the quality and reliability of one of the sensing modalities within a multimodal clustering algorithm. To summarize, the E-step of the algorithm updates only the posterior probabilities of the auditory observations $\beta_{kn} = \mathrm{P}(W_k = n|\mathbf{a}, \Theta)$:

$$\beta_{kn} = \frac{\pi_n \, \mathrm{P}(\boldsymbol{a}_k|W_k = n, \Theta)}{\sum_{i=1}^{N+1} \pi_i \, \mathrm{P}(\boldsymbol{a}_k|W_k = i, \Theta)}, \tag{3.8}$$

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

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

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

$$\mu_n = \frac{1}{\gamma_n} \left(\sum_{m=1}^{M} \alpha_{mn} \, \tilde{\boldsymbol{v}}_m + \sum_{k=1}^{K} \beta_{kn} \, \boldsymbol{a}_k \right), \tag{3.9}$$

$$\sigma_n^2 = \frac{\sum_{m=1}^{M} \alpha_{mn} \, (\tilde{\boldsymbol{v}}_m - \mu_n)^2 + \sum_{k=1}^{K} \beta_{kn} \, (\boldsymbol{a}_k - \mu_n)^2}{\gamma_n}. \tag{3.10}$$

3.5.2 Counting the Number of Speakers

BIC for model selection

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. This is estimating 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:

$$\text{BIC}(\tilde{\boldsymbol{v}}, \mathbf{a}, \Theta_N) = \mathcal{L}\,(\tilde{\boldsymbol{v}}, \mathbf{a}; \Theta_N) - \frac{D_N \, \log(M+K)}{2}, \tag{3.11}$$

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:

$$\hat{N} = \arg\max_{N} \text{BIC}(\tilde{\boldsymbol{v}}, \mathbf{a}, \Theta_N). \tag{3.12}$$

3.5.3 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.

Estimating $\boldsymbol{S}_n$

The locations of the AV objects are estimated thanks to the one-to-one correspondence between 3D visual features and the 1D projected features, $\tilde{\boldsymbol{v}}_m \leftrightarrow \boldsymbol{v}_m$. Indeed, the probabilistic assignments of the projected visual data onto the 1D clusters, α_{mn}, allow us to estimate $\boldsymbol{S}_n$ through:

$$\hat{\boldsymbol{S}}_n = \frac{1}{\bar{\alpha}_n} \sum_{m=1}^{M} \alpha_{mn} \boldsymbol{v}_m. \tag{3.13}$$

Estimating e_n

The auditory activity associated to the n^{th} speaker is estimated as follows (τ_A is a user-defined threshold):

$$\hat{e}_n = \begin{cases} 1 & \text{if } \bar{\beta}_n > \tau_A \\ 0 & \text{otherwise} \end{cases} \tag{3.14}$$

This two formulas account for the last remaining issue: the 3D localization and speaking state estimation of the AV objects. Next section describes some practical considerations to take into account when using this EM-based AV fusion method. Afterward, in Section 3.5.5, we summarize the method by providing an algorithmic scheme of the multimodal inference procedure.

3.5.4 Practical Concerns

Three technical issues

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 law of probability (3.1), 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-dependant knowledge. In all, we must account for three practical concerns: (i) EM initialization, (ii) eventual shortage of observations and (iii) the probabilistic model does not fully correspond to the observations.

EM Initialization

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 $N \in \{0, \ldots, N_{\text{max}}\}$. In the case $N \leq N^{(p)}$, we take the N clusters with the highest weight. 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 $N_{\text{max}} = 10$.

Too few observations

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.

Spurious clusters

Finally, we need to face the fact that the probabilistic model does not fully represent the observations. Indeed, we observed the existence of spurious clusters. Although the 3D visual observations associated with these clusters may be uniformly distributed, their projections onto the auditory space $\tilde{\boldsymbol{v}}_m$ may form a spurious cluster. Hence these clusters are characterized by having their points distributed near some hyperboloid in the 3D space (hyperboloids are the level surfaces of the linking mapping defined in (3.6)). As a consequence, the volume of the back-projected 3D cluster is small. We discard those clusters whose covariance matrix has a small determinant. Similarly as in (3.13), the clusters' covariance matrix is estimated via:

$$\hat{\boldsymbol{\Sigma}}_n = \frac{1}{\bar{\alpha}_n} \sum_{m=1}^{M} \alpha_{mn} \left(\boldsymbol{v}_m - \hat{\boldsymbol{S}}_n\right) \left(\boldsymbol{v}_m - \hat{\boldsymbol{S}}_n\right)^\top. \tag{3.15}$$

3.5.5 Motion-Guided Robot Hearing

Algorithm 3.1 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 (3.6). Then, for $N \in \{1, \ldots, N_{\text{max}}\}$ it iterates through the following steps: (a) Initialize a model with N components using the output of the previous time interval (Section 3.5.4), (b) apply EM using the selected N to model the 1D projections of the visual data (Section 3.5.1), (c) apply the *vision-guided EM fusion* algorithm to both the auditory and projected visual data (Section 3.5.1) in order to perform audio-visual clustering, and (d) compute the BIC score associated with the current model, i.e., (3.11). This allows the algorithm to select the model with the highest BIC score, i.e., (3.12). The post-processing step is then applied to the selected model (Section 3.5.4) prior to computing the final output (Section 3.5.3).

3.6 Implementation on NAO

The previous multimodal inference algorithm has desirable statistical properties and good performance (see Section 3.7). Since our final aim is to have a stable

Algorithm 3.1 Motion-Guided Robot Hearing

1: **Input:** HM3D, $\{\boldsymbol{v}_m\}_{m=1}^{M}$, and ITD, $\{\boldsymbol{a}_k\}_{k=1}^{K}$, features.
2: **Output:** Number of AV events $\hat{N}$, 3D localization $\left\{\hat{\boldsymbol{S}}_n\right\}_{n=1}^{\hat{N}}$ and auditory status $\{\hat{e}_n\}_{n=1}^{\hat{N}}$.
3: Map the visual features onto the auditory space, $\tilde{\boldsymbol{v}}_m = (\overline{\mathcal{A}} \circ \mathcal{V}^{-1})(\boldsymbol{v}_m)$ (3.6).
4: **for** $N = 1 \rightarrow N_{\text{max}}$ **do**
5: **(a)** Initialize the model with N clusters (Section 3.5.4).
6: **(b)** Apply EM clustering to $\{\tilde{\boldsymbol{v}}_m\}_{m=1}^{M}$ (Section 3.5.1).
7: **(c)** Apply the *Vision-guided EM fusion* algorithm to cluster the audio-visual data (Section 3.5.1).
8: **(d)** Compute the BIC score (3.11).
9: **end for**
10: Estimate the number of clusters based on the BIC score (3.12).
11: Post-processing (Section 3.5.4).
12: Compute the final outputs $\{\hat{\boldsymbol{S}}_n\}_{n=1}^{\hat{N}}$ and $\{\hat{e}_n\}_{n=1}^{\hat{N}}$ (Section 3.5.3).

component working on a humanoid robot (i.e., able to interact with other components), we reduced the computational load of the AV fusion algorithm. Indeed, we adapted the method described in Section 3.5 to achieve a light on-line algorithm working on mobile robotic platforms.

Adapting to NAO

In order to reduce the complexity, we substituted the Harris-Motion 3D point detector (HM3D) with the face 3D detector (F3D), described in Section 3.4.2. 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, 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 anymore. Third, since the visual features are not used as observations in the EM, but to initialize it, the complexity of the *vision-guide EM fusion* algorithm is $\mathcal{O}(NK)$ instead of $\mathcal{O}\left(N(K+M)\right)$. 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 not need to estimate them through (3.13).

3.6.1 Face-Guided Robot Hearing

The resulting procedure is called *Face-Guided Robot Hearing* and it is summarized in Algorithm 3.2 below. It takes as input the detected heads $(\boldsymbol{S}_1, \ldots, \boldsymbol{S}_N)$ and the auditory (**a**) observations gathered during a time interval Δt. The algorithm's output is the estimated auditory activity $\{\hat{e}_n\}_{n=1}^{N}$.

Algorithm 3.2 Face-Guided Robot Hearing

1: **Input:** Faces' position $\{\boldsymbol{S}_n\}_{n=1}^{N}$ and auditory $\{\boldsymbol{a}_k\}_{k=1}^{K}$ features.
2: **Output:** AV objects' auditory status $\{\hat{e}_n\}_{n=1}^{\hat{N}}$.
3: Map the detected heads onto the auditory space, $\mu_n = (\overline{\mathcal{A}} \circ \mathcal{V}^{-1})(\boldsymbol{S}_n)$ (3.6).
4: Apply EM clustering to $\{\boldsymbol{a}_k\}_{k=1}^{K}$ (Section 3.5.1).
5: Compute the final outputs $\{\hat{e}_n\}_{n=1}^{\hat{N}}$ (Section 3.5.3).

3.6.2 System Architecture

The RSB middleware

We implemented 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 was chosen instead of NAO's native framework NAOqi and we could implement and test our algorithm remotely without performance and deployment restrictions imposed by the robot platform. Moreover, the resulting implementation can be reused for other robots.

Synchronization tools

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 hence well-suited to synchronize events which originate from the same source (in the world) but suffered from 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

Figure 3.4: Within this work we used a new audio-visual head that is composed of a synchronized camera pair and two microphones. This "orange" head replaces the former "blue" head and is fully interfaced by the RSB middleware previously described in this section.

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.

3.6.3 Modular Structure

The implementation is divided into components shown in the pipeline of Figure 3.6. Components are color-coded: modules provided by the RSB middleware (white), auditory (red) and visual (green) processing, audio-visual fusion (purple) and the visualization tool (blue) described at the end of this Section.

Visual processing

The visual processing is composed by five modules. *Left video* and *Right video* stream the images received at 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 audio-visual head for NAO was used for this implementation. The new head (see Figure 3.4) is equipped with a pair of cameras and four microphones, thus providing a synchronized VGA stereoscopic image flow as well as four audio channels.

Auditory processing

The auditory component consists of three modules. Interleaved audio samples coming from the four microphones of NAO are streamed by the *Interleaved audio* module. The four channels are deinterleaved 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).

Audio-visual fusion

Both visual and auditory features flow until the *Audio-visual synchronization* module; the *TimeFrame* strategy is used here to find the ITD values coming from

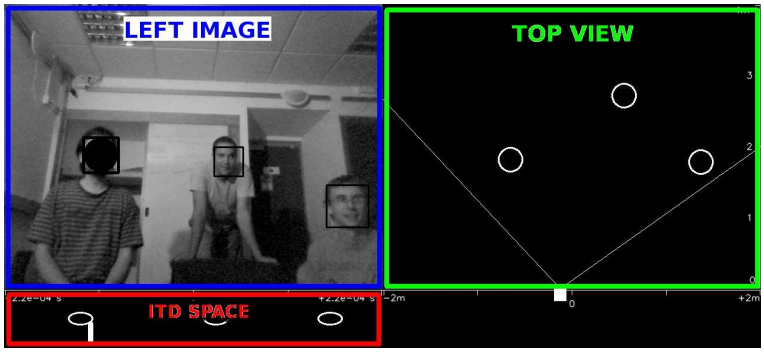


Figure 3.5: Snapshot of the visualization tool. The top-left (blue-framed) image is the original left image plus one bounding box per detected face. In addition, an intensity-coded circle appears when the speaker is active. The darker the colour is, the higher the speaking probability is. The top-right (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.

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.

Visualization

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 3.5. The image consists of three 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 codding the emitting sound probability, the higher it is, the darker the circle. 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.

3.6.4 Implementation Details

Some details need to be specified regarding the implementation of the face-guided robot hearing method. First, the integration window F and the frame shift f of the ITD extraction procedure. 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. We also used an activity threshold: when the energy of the sound signals is lower than $E_A = 0.001$, the window is not processed. Thus saving computational time for other components in the system when there are no emitted sounds. Notice that this parameter could be controlled by a higher level module which would learn

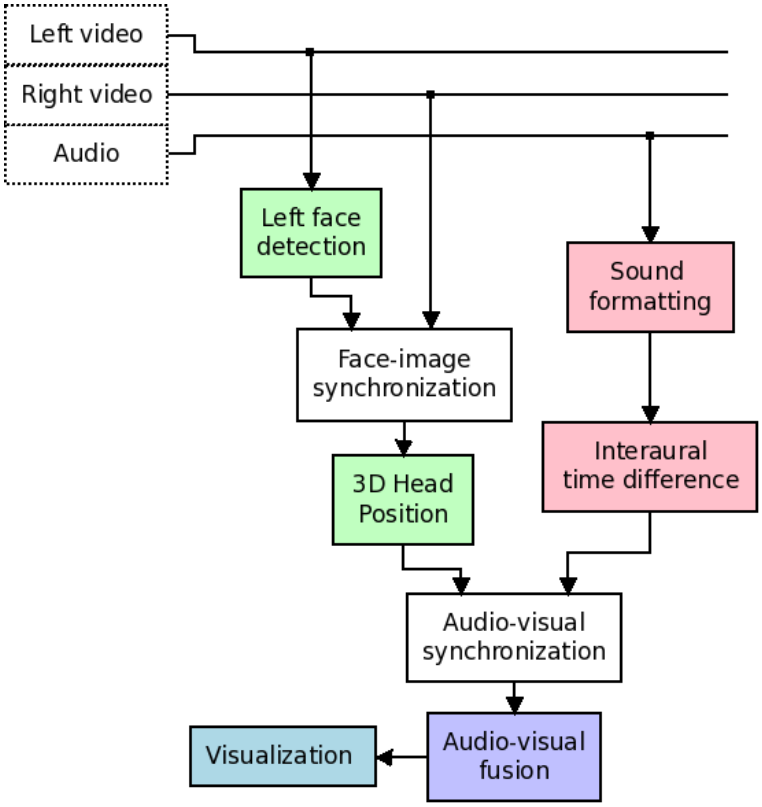


Figure 3.6: Modular structure of the *Face-Guided Robot Hearing* procedure implemented on NAO. There are five types of modules: streaming & synchronization (white), visual processing (green), auditory processing (red), audio-visual fusion (purple) and visualization (blue).

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 $\tau_A = K/(N+2)$. The entire pipeline was running on a laptop with an i7 processor at 2.5 GHz.

3.7 Results

In order to evaluate the proposed approach, we ran three sets of experiments. First, we evaluated the Multimodal Inference method described in Section 3.5 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.

An example of auditory and visual observations gathered within Δt of the publicly available data set

In all our experiments we used a time interval of 6 visual frames, $\Delta t = 0.4s$; time in which approximately 2,000 HM3D observations and 20 auditory observations are extracted. A typical set of visual and auditory observations are shown in Figures 3.7 and 3.8. Indeed, Figure 3.7 focuses on the extraction of the HM3D features: the Harris interest point detection, filtered by motion, matched between images and reconstructed in 3D. Figure 3.8 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 of 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 pics' variances, or from sensor noise which is sparse along the ITD space.

Evaluation metric

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 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.

3.7.1 Results on Synthetic Data

Synthetic sequences

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 3.1 shows the visual evaluation of the method when tested

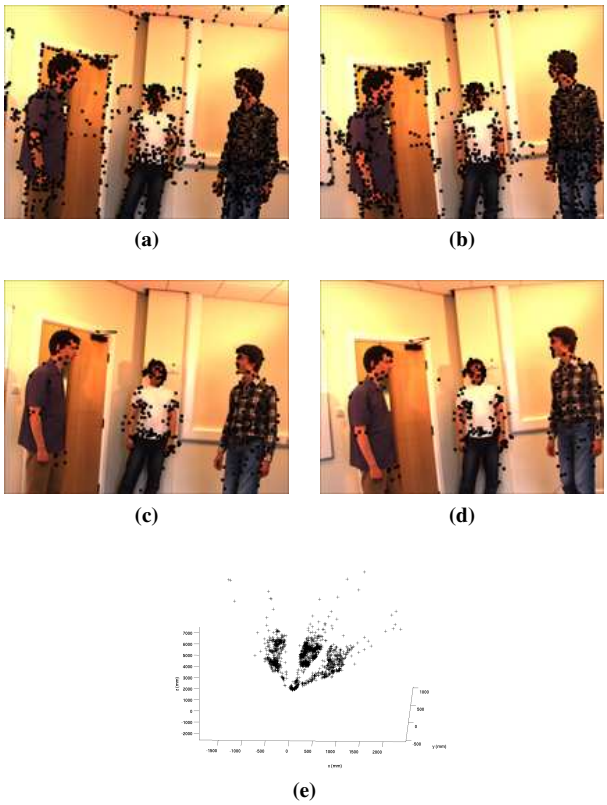
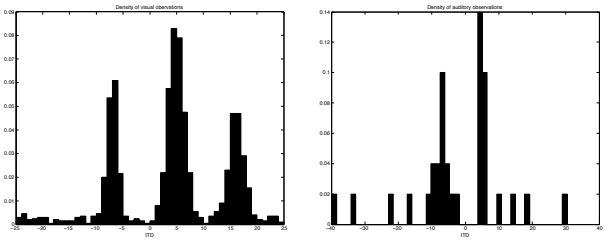


Figure 3.7: Interest points as detected in the left (a) and right (b) images. Dynamic interest points detected in the left (c) and the right (d) images. (e) HM3D visual observations, $\{\boldsymbol{v}_m\}_{m=1}^{M}$. Most of the background (hence static) points are filtered out from (a) to (c) and from (b) to (d). It is worth noticing that the reconstructed HM3D features suffer from reconstruction errors.

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. First, we observe that the right detection rate is always above 65%, increasing to 96% in the case where there are 3 visible static clusters. We also observe that the fact that the number of AV objects in the scene varies does not impact the localization error. The effect on the localization error is due, hence, to the dynamic character of the scene; if the AV objects move or not. The third observation is that both the dynamic character of the scene and the varying number of clusters have a lot of impact on the detection rate.

Results on synthetic data

Table 3.2 shows the auditory evaluation of the method when tested with syn-



(a) **(b)**

Figure 3.8: Observation densities in the auditory space $\mathbb{A}$: (a) of the projected HM3D features, $\{\tilde{\boldsymbol{v}}_m\}_{m=1}^{M}$, and (b) of the ITD features, $\{\boldsymbol{a}_k\}_{k=1}^{K}$. 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 3.1: 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	43	139 (34.1%)	269 (65.9%)	0.10
StaVar	46	69 (30.1%)	160 (69.9%)	0.03
DynVar	40	82 (35.9%)	147 (64.1%)	0.11

thetic 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.

3.7.2 Results on Real Data

The CTMS3 sequence of CAVA

The *Motion-Guided Robot Hearing* method was tested on the CTMS3 sequence of the CAVA data set [Arnaud 08]. The CAVA (*computational audio-visual anal-*

Table 3.2: 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	144	56 (21.2%)	208 (78.8%)
StaVar	53	33 (18.8%)	143 (81.2%)
DynVar	56	34 (19.7%)	139 (80.3%)

ysis) data set was specifically recorded to test various real-world audio-visual scenarios. The CTMS3 sequence[2] 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.

Results on CTMS3

Figure 3.9 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 3.9a 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 3.9b shows a failure of the *ridgeline* method, used to merge Gaussian components, where two different clusters are associated into one. Figure 3.9c 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 3.9d clearly shows that our approach cannot deal with occluded objects, because of the instantaneous processing of egocentric data, the person occluded will never be detected. Figures 3.9e, 3.9f and 3.9g are examples of success. The three speakers are localised and their auditory status correctly guesses. 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 3.9h and 3.9i 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%).

3.7.3 Results on NAO

To validate the *Face-Guided Robot Hearing* method using NAO, we performed a set of experiments with five different scenarios. The scenarios were recorded in a room around 5×5 meters with just a sofa and 3 chairs where NAO and the other persons sat respectively. We designed five scenarios 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.

The NAO scenarios

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 (FoV) 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 standing instead of sitting. These two scenarios are useful

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

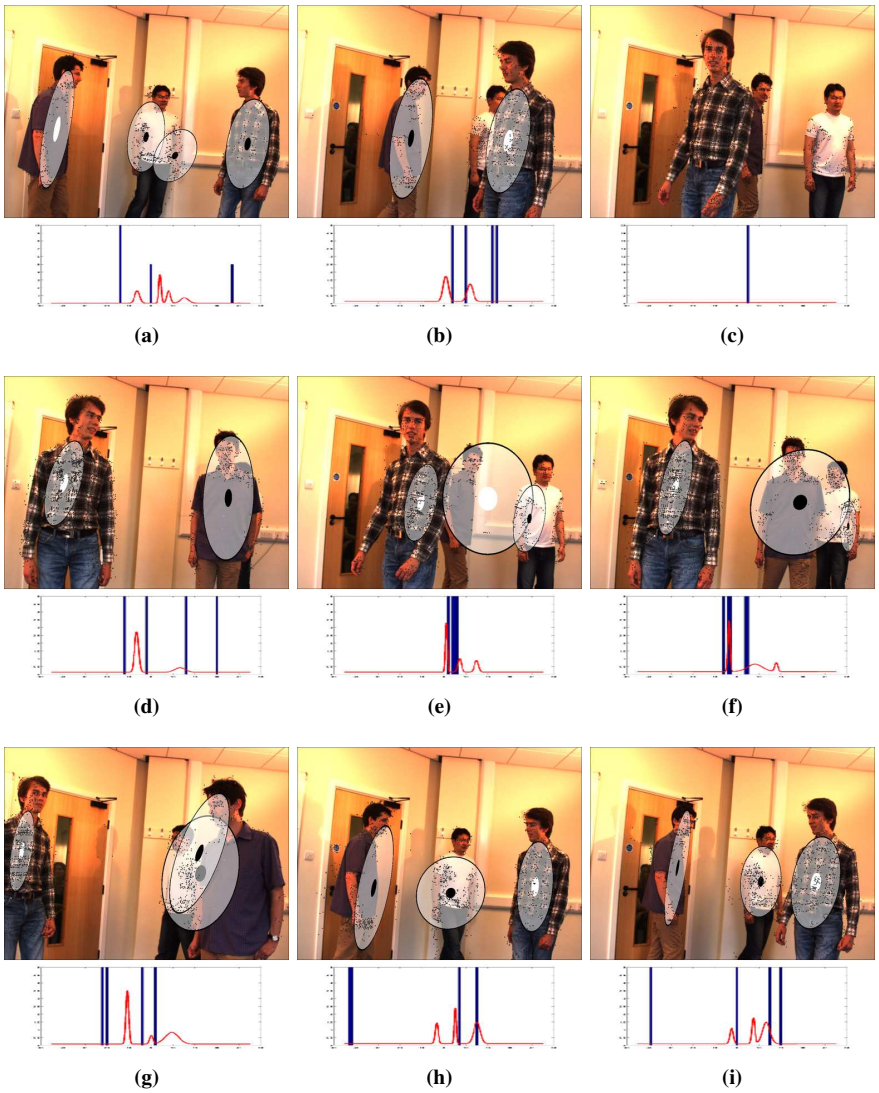


Figure 3.9: Results obtained with the CTMS3 sequence from the CAVA data set. The ellipses correspond to the 3D covariance matrices projected onto the image. The circle at each ellipse center illustrates the auditory activity: speaker emitting a sound (white) or being silent (black) during each time interval. The plot associated with each image shows the auditory observations as well as the fitted 1D mixture model.

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 FoV. This scenario is used to test if people speaking outside the FoV affect the performance of the algorithm. In the last scenario (**S5**) the three people are in the FoV, 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.

Results on **S1**-**S5**.

In Figure 3.10 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 3.10a 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 3.10b, in which among the three people, only one speaks. A failure of the ITD extractor is shown in Figure 3.10c, where the actor in the left is speaking, but no ITDs are extracted. In Figure 3.10d 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 3.10e 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 3.10f, Figure 3.10g and Figure 3.10h show the effect of reverberations. While in Figure 3.10h 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 3.10f and Figure 3.10g), hence demonstrating the robustness of the proposed approach.

Table 3.3 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.

	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 3.3: 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).

3.8 Conclusions and Future Work

Overview

This chapter 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

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

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

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

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

Figure 3.10: 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 3.6.3 for how to interpret the images above.

associated with the visual and auditory sensors: the audio-visual mapping $(\overline{\mathcal{A}} \circ \mathcal{V})$ allows us to transform the visual features from the 3D space to an 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 is validated on a humanoid robot and interfaced through the RSB middleware leading to a platform-independent implementation.

Visual guidance, the spirit

The main novelty of the approach is the visual guidance. Indeed, we derived to EM-based 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 audio-visual calibration procedure that provides an accurate projection mapping $(\overline{\mathcal{A}} \circ \mathcal{V})$. The calibration is specifically designed for robotic usage since it requires very few data, it is long-lasting and environment-independent.

Solved issues

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.

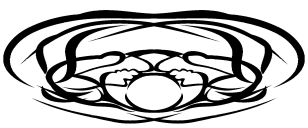
Practical advantages

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 audio-visual 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 audio-visual data flow from two microphones mounted inside the head of a companion robot with noisy fans and two cameras at a rate of 17 Hz.

Future work

There are several possible ways to improve and to extend our 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 weight the relative importance of the two modalities is under investigation. Our algorithm can also accommodate other types of visual cues, such as 2D or 3D optical flow, body detectors, etc., or auditory cues, such as Interaural Level Differences. We used one pair of microphones, but the method can be easily extended to several microphone pairs. 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 is highly interfered by another speaker/sound source. Moreover, this future dynamic model selection should be extended to provide for audio-visual tracking capabilities, since they enhance the temporal coherence of the perceived audio-visual scene.



CHAPTER 4

AUDIO-VISUAL

COMMAND RECOGNITION

In this chapter we address the problem of audio-visual command recognition. Such commands consist on a combination of a gesture and a short sentence. For instance, someone asks the companion robot to get closer. In addition, because such commands play an important role in human-robot communication, it is desirable that robots interacting with people have an excellent command recognition system available. We propose a normalized convex weighting scheme to perform multimodal classification. The method is able to merge multiple monomodal classifiers of different nature. Moreover, because this commands are culture-, language- and user-dependant, methods need to learn from very few examples. We present a benchmark of several command recognition methods using tiny training sets. All experiments are conducted on the publicly available data set Ravel.

4.1 Introduction

What an Audio-Visual command is and why their recognition is relevant.

For the last decade, human-computer interaction methods have rapidly evolved towards flexible multimodal systems; There is a clear need to understand human commands. In this context, we are interested in the recognition of audio-visual commands, that is a combination of a gesture and a short phrase. For instance, a person asks his/her companion robot to perform a task. In addition, because such commands play an important role in human-robot communication, it is desirable that robots interacting with people have an excellent command recognition system available. In this chapter we present the results of our research on AV command recognition with special emphasis on the particularities associated to humanoid robots.

Challenges associated to AV command recognition

The AV command recognition task is challenging for several reasons. First, the data may be corrupted. On one side auditory recordings may be contaminated by reverberations, noise and interferences. On the other side, visual data might suffer from occlusions or bad lighting conditions. Hence, proper cues need to be

extracted from the data flow. Second, despite the empirical advantages of combining auditory and visual cues, there no agreement on a common audio-visual representation. Thus, we need to seek for the most appropriate representation in our case. Third, because these commands are culture-, language- and user-dependant, methods need to constantly adapt. Consequently we are interested on the methods able to learn from very few examples.

Thesis contributions

The contribution of this thesis to AV command recognition is two-fold. First we proposed a multimodal normalized convex weighting scheme. Based on a high-performance and solid learning technique, the method uses binocular and monocular visual features and monaural auditory features. Audio-visual recognition is performed at a later stage in which the classification scores from audio and video are combined to get the final audio-visual score, yielding to important increasings on the Average Recognition Rate (ARR). This work was published in [Sanchez-Riera 12a] **AVG1**. Second we analysed the performance of different AV learning methods when trained on tiny training sets, in order to seek for the method with the highest adaptableness. These experiments and their conclusions were published in [Alameda-Pineda 13c] **AVG2**.

This chapter is structured as follows. Section 4.2 describes the published work related to the topic. Section 4.3 delineates the auditory and visual features we built upon. These features are used to construct per-instance and per-frame representations (see Section 4.4). The general framework for AV categorization is described in Section 4.5, before detailing the experimental conditions, i.e., the dataset and the evaluation metric in Section 4.6. Section 4.7 introduces the normalized Convex Weighting Scheme. Afterward, different AV learning schemes are benchmarked on tiny training sets (Section 4.8). Finally, some conclusions and a few exciting future guidelines are drawn in Section 4.9.

4.2 Related Work

Audio-visual discriminative classification approaches can be grouped depending on the way the audio-visual command is represented. *Early Fusion* applies when the representation is audio-visual, *i.e.*, one observation vector corresponds to joint audio-visual information. *Late Fusion* applies when two different observations represent the modalities (auditory and visual). In the following, we present the existing literature on audio-visual discriminative classification.

Related work on early fusion

Early Fusion: In [Jiang 09] an audio-visual representation named short-term audio-visual atom is proposed. It is a concatenation of color/texture, motion and auditory features. Targeting semantic concept detection, the method is evaluated on a dataset of 3,000 sequences. In [Monaci 06], the authors learn dictionaries of multimodal features extracted from the raw data by means of generative functions. More recently, a different way to combine audio-visual features at an early stage is proposed in [Ye 12], where a bipartite graph quantizes features coming from auditory and visual channels. The authors evaluate the audio-visual

event detection performance on a dataset of about 9000 sequences. The authors in [Liu 08] target a speech detection application, and perform the audio-visual fusion at a feature level as well. Principal component analysis (PCA) features are taken from the face images and Mel Frequency Cepstral Coefficients (MFCC) are the auditory features. Both types of features are then projected in a joint subspace using canonical correlation analysis (CCA). A Gaussian mixture model (GMM) is used for classification. In [Mühling 12] audio-visual video concept detection is targeted and the approach consists of concatenating the visual and auditory descriptors, thus forming an audio-visual representation. Tests are performed on a dataset of around 45,000 videos. The reader is referred to [Luo 08] for a study using support vector machines (SVM) that compares feature-level fusion techniques to classification-level fusion techniques.

Related work on late fusion

Late Fusion: Also in [Mühling 12], the auditory and visual representation are fused through Multiple Kernel Learning (MKL). This technique is popular because the relative relevance of different kernels is learned from the data. A two-stage strategy is proposed in [Natarajan 12]. First, MKL is used to classify auditory and visual features separately. Second, the normalized scores are merged using a Bayesian model. This is tested in a dataset of about 45,000 videos. In [Wu 10] several auditory and visual features are computed. Afterwards, they are classified separately and a convex combination of the unimodal classification scores allows to choose the best audio-visual score. The method is tested on a dataset of 900 videos and 12 classes. In [Lopes 06] two methods based on feature selection are compared. The complete set of audio-visual features is a 3000-dimensional vector, from which 35 to 70 features are selected. Tests are performed on a data set with 15 training instances per class. Among the classification-level fusion methods we remark [Lacheze 09], in which the authors experiment different combination strategies for object detection. Visual features are based on texture description and entropy-based variable-size patches. Auditory features correspond to the energy of the signal's gammatone filter bank decomposition. Monocular video and monaural audio are used and there is a strong need of uniform visual background. Object recognition based SVMs is used in [Saenko 08] where a probabilistic method combining posterior class probability output by each classifier is proposed; mainly, this means that each modality is trained separately and then combined. SIFT descriptors are used as visual features and a commercial speech recognizer is used to classify the incoming audio signal. A different approach from the ones mentioned so far is described [Xiong 05] which finds sport highlights using a coupled hidden Markov model (CHMM). Several video features are used such as quantization average motion vectors and colour. On the auditory side, the authors chose to use MFCC features. Both these features train a CHMM to perform the classification.

From the presented literature review we extract several conclusions:

- Most of the approaches use MFCC features to describe the sound track. This choice is also supported by many decades of fruitful research in speech recognition.

- There is no clear agreement on the visual feature to use though. As a matter of fact, there are some papers extracting many features and performing a dimensionality reduction step before the classification takes place.
- The use of audio-visual data clearly enhances performance. Albeit, there is no agreement neither on when to fuse the two modalities nor on how to classify them. Indeed, *Early fusion* and *Late fusion* approaches have similar ARR depending on the application targeted, the data used for evaluation and the classifier (SVM, GMM, HMM, etc).
- In the *Late fusion* literature, when two different classifiers are used, the fusion scheme is usually specifically built for the particular classifiers. No general scheme is used to unify the output of two completely different (unknown?) classifiers.
- In the majority of the published works, the training set is large. Thousands of videos are used to learn the chosen AV command models.

Consequently, we addressed the problem of AV command recognition in the following manner. First, we designed the normalized convex weighting scheme able to deal with classifiers of different nature that use various types of visual features. The audio track is described using the features designed for speech recognition, MFCC, in agreement to the vast majority of the existing works. A variety of classifiers are tested in order to evaluate this new *Late Fusion* scheme. Second, one particular combination among the previously ones is chosen to be evaluated on tiny training sets together with other existing monomodal and multimodal methods.

4.3 Audio and Visual Features

In this Section we describe the auditory and visual descriptors we used along this chapter. It is not our aim to develop new descriptors/representations, but to investigate how to fuse both modalities and how to overcome the challenges associated to the use of humanoid robots. Section 4.3.1 is devoted to the auditory descriptor and Section 4.3.2 is devoted to the visual descriptor.

4.3.1 The Auditory Features

The auditory features: Mel Frequency Cepstral Coefficients

The auditory stream is represented by the Mel Frequency Cepstral Coefficients (MFCC), widely used for speech/sound analysis and recognition (see [Ramasubramanian 11, Rabiner 11]). It has proven extremely good performance on speech recognition, specially when used together with the Hidden Markov Models. They are computed following the three steps: (i) perform the short-time Fourier transform (STFT), (ii) map the power spectrum onto the Mel scale and (iii) take the discrete cosine transform of these mapped powers. The are three main parameters

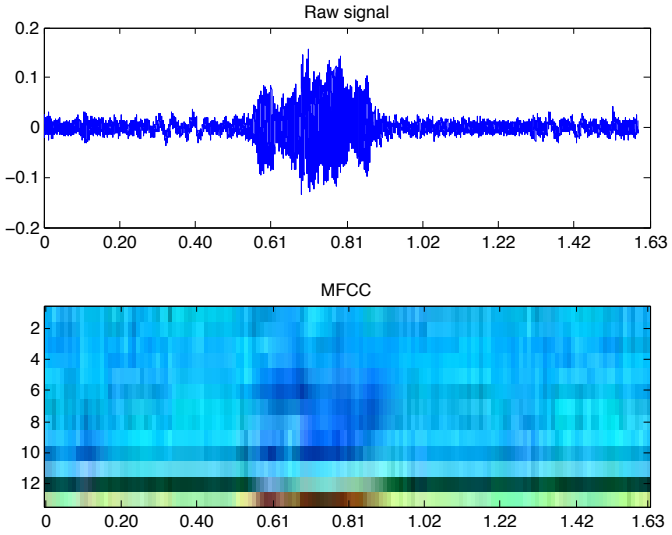


Figure 4.1: Mel Frequency Cepstral Coefficients for one voice-command instance. From the raw signal (top) to the extracted MFCC (bottom).

associated with MFCC features. First, the frame size defines the length of the STFT (denoted by W). Second, the frame shift (F) determines the time between two consecutive STFT windows. Third, the amount of cepstral coefficients (D), that sets the dimension of the output MFCC representation. The parameters to compute the MFCC features were set to the standard ones in speech recognition: $W = 21.3$ ms, $F = W/2$ and $D = 13$. We note by m_n the n^{th} extracted feature for $n = 1, \ldots, N_m$, where $N_m = \left\lceil \frac{T-W}{F} \right\rceil$ is the number of time frames and T the sequence length. Figure 4.1 shows an example of the MFCC features for a particular instance of an AV command.

4.3.2 The Visual Features

As we pointed out from the literature, there is no clear agreement on the optimal visual feature for action description. This, we chose to use two different visual descriptors, because of their very good performance on visual gesture/action recognition.

The first visual descriptor: scene-flow based features.

The first one was proposed in [Sanchez-Riera 12c] and it is based on the scene flow, which is the 3D equivalent of the optical flow [Čech 11]. The scene flow is represented by the optical flow plus the depth at each image position. Together with the camera calibration, this is equivalent to a vector field of 3D position and associated 3D velocities. This intrinsic representation is potentially less sensitive to changes of texture and illumination than the intensity images. Moreover, the notion of depth allows to focus on the actor, while discarding any activity from the background. We assume that the actor of interest is the person closest to the camera. This is a reasonable assumption, since it holds in most of the human-

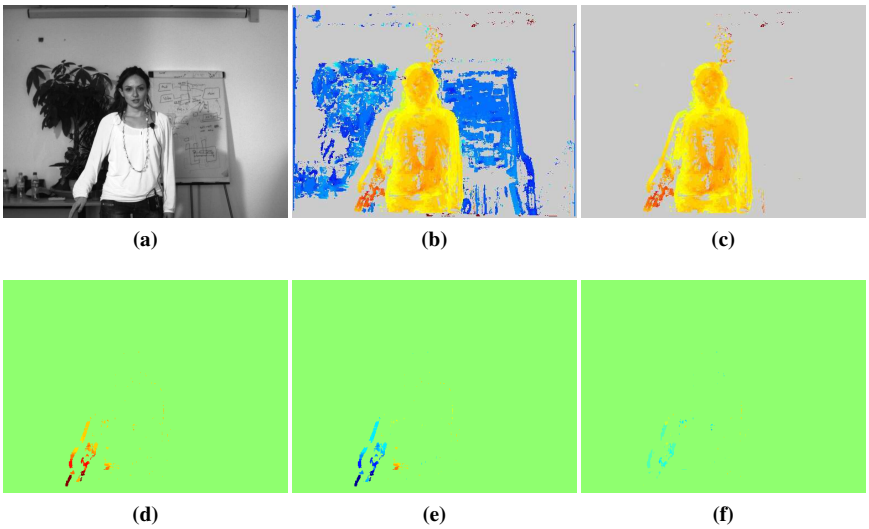


Figure 4.2: Construction of the proposed descriptor. The actor's face is detected from the left input image (a). The raw disparity map (b) is segmented, such that all pixels having the lower disparity than the actor's face are discarded (c). The descriptor is then computed for all remaining pixels undergoing non-zero motion, such that it consists of the pixel's position relative to the face, it's disparity (d), and horizontal (e) and vertical (f) components of optical flow.

robot-interaction applications and on movies. The final descriptor consists on the position and disparity relatives to the actor's face plus the optical flow (see Figure 4.2 for an example and [Sanchez-Riera 12c] for detailed explanation). $s_{n,p}$ denotes the p^{th} scene flow feature extracted from the n^{th} image, for $p = 1, \ldots, P_n$, $n = 1, \ldots, N_v$, where N_v is the number of images and P_n is the number of scene flow features extracted from the n^{th} image.

The second visual descriptor: spatio-temporal interest points

The second descriptor, STIPs, was proposed in [Laptev 05]. This descriptor consists on the histogram of gradients concatenated to the histogram of optical flow (HOG-HOF), applied at Harris 3D interest points. Notice that, while the first descriptor uses stereo-vision and has low dimension (5), the second uses monocular vision and is high dimensional (200). Furthermore, while the former has a semi-dense nature, the latter is sparse in the spatio-temporal domain. $l_{n,q}$ denotes the q^{th} scene flow feature extracted from the n^{th} image, for $q = 1, \ldots, Q_n$, $n = 1, \ldots, N_v$, where N_v is the number of images and Q_n is the number of scene flow features extracted from the n^{th} image.

4.4 Per-instance and Per-frame Representations

We are interested in two types of representations, namely: *per-instance* and *per-frame*. While the first one corresponds to one vector representing the entire command, the second one corresponds to a sequence of "instantaneous" vectors.

4.4.1 Per-instance Representations

The per-instance representation: averaging along the command

The per-instance representations are useful in the framework of the Bag-of-Words (BoW) paradigm, which consists of five different steps: (i) extract local descriptors, (ii) cluster them to get a vocabulary, (iii) map each of the descriptors to the vocabulary, (iv) build a histogram of word occurrence and (v) feed these histogram-based representations to a classifier. During the first three steps, a codebook of size K is built. Subsequent steps are used to represent instances and learn a classifier from these representations. Later for recognition, an unlabelled audio-visual sequence is first represented as a histogram which is fed to the classifier to estimate the sequence's class.

The choice of BoW is justified by the vast literature proving efficiency and robustness. The power of BoW rises from the quality and quantity of the descriptors as well as the discriminability of the classifier. We denote by h^m the Bag-of-Words representation corresponding to the MFCC features $\{m_n\}_{n=1}^{N_m}$ and write, $h^m = \text{BoW}\left(\{m_n\}_{n=1}^{N_m}\right)$. Similarly, we write h^s (respectively h^l) to denote the Bag-of-Words representation corresponding to the scene-flow features (respectively the Laptev features), and write $h^s = \text{BoW}\left(\{s_{n,p}\}_{p=1,n=1}^{P_n,N_v}\right)$ (respectively $h^l = \text{BoW}\left(\{l_{n,q}\}_{q=1,n=1}^{Q_n,N_v}\right)$).

4.4.2 Per-frame Representations

The per-frame representation: a sequence of instantaneous vectors

The auditory per-frame representations correspond to the MFCC features, since they already provide a sequence of instantaneous vectors. In the visual case, we use per-frame Bag-of-Words representations, that is to say, we generate one histogram per image: $f_n^s = \text{BoW}\left(\{s_{n,p}\}_{p=1}^{P_n}\right)$ and $f_n^l = \text{BoW}\left(\{l_{n,q}\}_{q=1}^{Q_n}\right)$. However because, the HOG-HOF descriptor is sparse (there are many empty frames), it is not well suited for a per-frame representation.

4.5 Audio-Visual Categorization

A multiclass classifier consists of a discriminant function $f : \mathcal{X} \times \mathcal{C} \rightarrow \mathbb{R}$, where $\mathcal{X}$ is the feature space, $\mathcal{C} = \{1, \ldots, C\}$ is the set of labels and C is the number of classes. Here $\mathcal{X}$ represents a generic feature space. Given a feature vector (or sequence of feature vectors) $\boldsymbol{x} \in \mathcal{X}$, $f(\boldsymbol{x}; c)$ is the score of classifying $\boldsymbol{x}$ as c. The higher the score is, the more likely c is the class of $\boldsymbol{x}$. Hence, a new unlabeled observation $\boldsymbol{x} \in \mathcal{X}$ is classified as:

$$c^*(\boldsymbol{x}) = \arg\max_{c \in \mathcal{C}} f(\boldsymbol{x}; c).$$

In the following, $\mathbf{X}$ will denote the training set, i.e. a set of feature vectors $\mathbf{X} = \{\boldsymbol{x}^n\}_{n=1}^N$ which class is known, and that is used to train the classifiers. The set of training festures of class c will be denoted by $\mathbf{X}_c$.

4.5.1 Per-instance Learning: Support Vector Machines

In order to classify the commands from their per-instance representations, we used Support Vector Machines. Widely studied, SVMs have proven excellent discriminability power when used in combination with the BoW representation. SVM's are a discriminative binary classification method, based on a function $h_c(\boldsymbol{x})$ learnt from a set of positive and negative examples. The points satisfying $h_c(\boldsymbol{x}) = 0$ form a hyperplane in the space induced by a kernel function $k(\cdot, \cdot)$. $h(\boldsymbol{x}) > 0$ means that $\boldsymbol{x}$ should be classified as positive and $h(\boldsymbol{x}) < 0$ as negative. We refer the reader to [Bishop 06] for details on the formulation. Importantly, a parameter $\mathcal{Q}$ regulates the amount of allowed misclassification in the training set, such that SVMs deal with overlapping classes. One way to extend SVMs to the multiclass classification problem is to train the function $h_c(\boldsymbol{x})$ with $\mathbf{X}_c$ as positive examples and $\mathbf{X} \setminus \mathbf{X}_c$ as negative examples. In all, the function f is defined as:

$$f(\boldsymbol{x}; c) = \sum_{n=1}^{N} \beta_{n,c} k(\boldsymbol{x}, \boldsymbol{x}_n),$$

where $\{\boldsymbol{x}_n\}_{n=1}^{N}$ is the training set, $k(\cdot, \cdot)$ is the kernel function and $\beta_{n,c} \in \mathbb{R}$ are computed during the training phase.

4.5.2 Per-frame Learning: Hidden Markov Models

The Hidden Markov Models (HMM) belong to the family of graphical models. In a HMM the observations depend on a hidden discrete random variable usually called state, taking values from 1 to S. The probability of the observations given the state value is called emission probability. The state is assumed to be Markovian, that is, the state at time t only depends on the state at time $t-1$. In addition, we could constrain the dynamics of the HMM, forcing the states to happen in order, i.e. state s before the state $s+1$; this is usually known as *left-to-right* HMM. The emission probability is usually GMM. One model ξ_c per class is learnt (through an EM algorithm). The model consists of the parameters of the emission probability and the parameters modelling the Markovian dynamics. The function f is the log-likelihood of the model:

$$f(\boldsymbol{x}; c) = \ln \mathrm{p}(\boldsymbol{x}|\xi_c).$$

We refer the reader to [Rabiner 11, Bishop 06] for more details about HMM.

4.6 Experimental conditions

4.6.1 The Data Set

The experimental validation is performed on the "Robot Gestures" scenario of the Ravel data set (see Chapter 2 and [Alameda-Pineda 12b]). We use eight sequences, each one containing three instances of the nine command categories.

The set of voice-and-gesture commands are the followings: (i) *wave* ("Hello!"), (ii) *walk towards the robot* ("I am coming."), (iii) *walk away from the robot* ("Bye!"), (iv) *stop* hand-wave ("Stop!"), (v) *turn around* gesture ("Turn around."), (vi) *come here* gesture ("Come here."), (vii) *point* action ("Look!"), (viii) head motion for *yes* ("Yes") and (ix) head motion for *no* ("No"). In all cases, the human gesture is accompanied by the speech corresponding to the command, shown above in brackets. Notice that all the actors in the data set are non-native English speakers of five different nationalities, hence there is a large variability in the pronunciation. This data set is used in Section 4.7 and in Section 4.8 to evaluate and compare the proposed methods.

4.6.2 Evaluation Metric

Evaluating multicategory classifiers means providing the confusion matrices. The ij^{th} entry of such matrix contains how many instances of the i^{th} class have been classified as class j. By averaging the elements of the diagonal, one obtains the average recognition rate (ARR) of the classifier. Moreover, in order to obtain statistically significant results and properly evaluate the different classifiers, a cross-validation strategy is applied. The dataset is split into a training subset and a testing subset, and this is repeated several times. Once the framework is set, one can focus on the contributions of this Thesis to the field of AV scene understanding.

4.7 The Normalized Convex Weighting Scheme

This Section contributes to the field in a double manner. On one side we need to understand how the different representations and classifiers we have chosen perform when used alone (Section 4.7.1). On the other side, we present the normalized Convex Weighting Scheme for AV command recognition (Section 4.7.2).

4.7.1 Monomodal Categorization

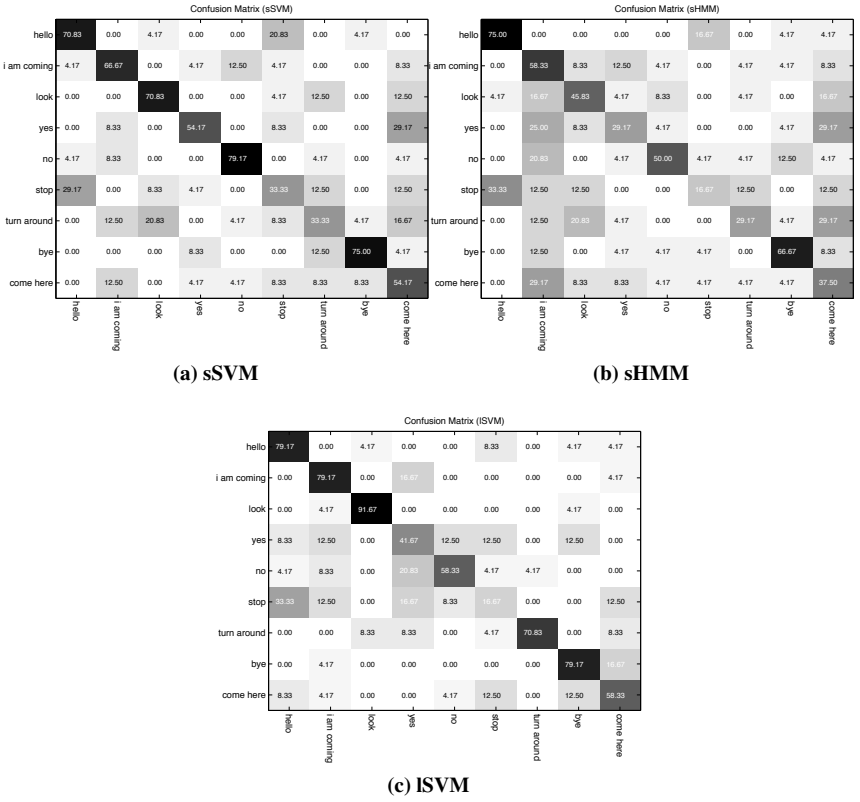
In a first stage we would like to evaluate the different combinations of features and classifiers. We evaluated five different monomodal representation/classifier pairs, as shown in Table 4.1. We remark that the feature space is different for each of the cases in the previous table. Indeed, in the case of **lSVM**, **sSVM** and **aSVM** the dimension of $\mathcal{X}$ is the number of histogram bins of teh BoW representation, which was set to 500. $\mathcal{X}$ has dimension $K \times N_v$ in the case of **sHMM**, where K is the dimension of the per-frame BoW, set to 20. Finally, $\mathcal{X}$ is $13 \times N_m$-dimensional in the case of **aHMM**, since we use the classical configuration of 13 MFCC coefficients.

Table 4.1: The different combinations of features and representations/classifiers evaluated.

		Representation/Classifier	
		Per-instance	Per-frame
Features	[Laptev 05]	h^l+SVM (**lSVM**)	Not used[1]
	[Sanchez-Riera 12c]	h^s + SVM (**sSVM**)	h^s_n+HMM (**sHMM**)
	MFCC	h^m + SVM (**aSVM**)	m_n+HMM (**aHMM**)

§ Results on Visual Categorization

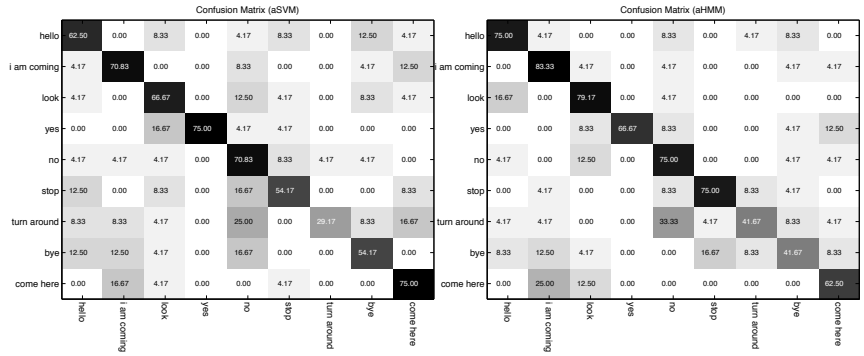
Figures 4.3a, 4.3b and 4.3c show the confusion matrices for the **sSVM**, the **sHMM** and the **lSVM** classifiers. Notice that **sSVM** performs very well in five out of nine gestures (*Hello*, *I am coming*, *Look*, *No* and *Bye*), well in two gestures (*Yes* and *Come here*) and poorly in two gestures (*Stop* and *Turn around*). However, the **sHMM** classifier performs poorly compared to **sSVM**. Indeed, it gets good results for most of the actions, very good results for just two of them (*Hello* and *Bye*) and poor results in three gestures (*Yes*, *Stop* and *Turn around*). Notice also that there is no much difference between the **sSVM** and the **lSVM** classifiers. Actually, they mainly differ in two of the gestures *No* and *Turn around*.



(a) sSVM **(b) sHMM**

(c) lSVM

Figure 4.3: Confusion matrix of the three visual classifiers: (a) **sSVM**, (b) **sHMM** and (c) **lSVM**.



(a) aSVM **(b) aHMM**

Figure 4.4: Confusion matrix of the two auditory classifiers: (a) **aSVM** and (b) **aHMM**.

§ Results on Auditory Categorization

The confusion matrix of **aSVM** and **aHMM** can be seen in figures 4.4a and 4.4b respectively. The SVM-based classifier performs very good in six out of nine actions, good in two of them (*Bye* and *Stop*) and poorly just in one command (*Turn around*). However the **aHMM** classifier has very good performance everywhere except for the *Bye* and *Turn around* commands.

4.7.2 Multimodal Categorization: the normalized Convex Weighting Scheme

§ The Method

An overview

Until here we evaluated the auditory and visual representations separately. This gives us a lower bound on what we should expect from multimodal classifiers. The classifier we propose consists on a convex combination of monomodal classifiers. More precisely, it is built in three steps: (i) train separate monomodal classifiers, (ii) normalize the classification scores and (iii) train the convex weighting parameter.

Detailed explanation

The outcome of the first step are two monomodal classifiers whose discriminant functions will be noted by: $f^{\mathbf{A}}$ and $f^{\mathbf{V}}$ for audio and video respectively. Afterwards, the normalization step takes place: for each modality separately, the mean (denoted $\mu_c^{\mathbf{A}}$ and $\mu_c^{\mathbf{V}}$, respectively) and the spread (denoted $\sigma_c^{\mathbf{A}}$ and $\sigma_c^{\mathbf{V}}$) of the classification scores are computed. Last, a multimodal discrimination function is defined by taking a convex combination of the normalized monomodal scores:

$$f^{\mathbf{WS}}(\boldsymbol{x}^{AV}; c) = \lambda \frac{f^{\mathbf{A}}(\boldsymbol{x}^{A}; c) - \mu_c^{\mathbf{A}}}{\sigma_c^{\mathbf{A}}} + (1-\lambda)\frac{f^{\mathbf{V}}(\boldsymbol{x}^{V}; c) - \mu_c^{\mathbf{V}}}{\sigma_c^{\mathbf{V}}},$$

where $\boldsymbol{x}^{AV}$ contains both the auditory representation $\boldsymbol{x}^{A}$ and the visual representation $\boldsymbol{x}^{V}$. The normalization step is necessary to bring all discriminant functions

Meaning of the weighting factor λ

to a similar value range. The weighting parameter λ can be set either by hand or learnt from the data. The value of λ determines the trust we put on each modality. Actually, some cases deserve a special mention:

$\lambda = 0$ is equivalent to audio-based classification.

$\lambda = 0.5$ the auditory and visual scores stand on equal foot,

$\lambda = 1$ is equivalent to vision-based classification, and

In general, $\lambda > 0.5$ means that we put more trust on the visual classification score, whereas $\lambda < 0.5$ means that we do it with the auditory score. This way of combining the two classifiers allows us to evaluate the relative trust we put on the modalities

§ Results on Audio-Visual Categorization

Analysis of the performance as a function of λ

It is worth to notice that, for instance, some actions that are difficult to recognize by **sSVM** as *Stop* or *Come here* are easily classified by **aSVM**, and viceversa with the actions *Bye* or *Hello*. This supports the idea that the combined classifier should outperform both unimodal classifiers. Figure 4.5 show the ARR of the combined classifier as a function of the weighting parameter l. Please remark that when using the same underlying model (curves **aSVM-sSVM**, **aSVM-lSVM** and **aHMM-sHMM**) the maximum performance of the combined classifier is achieved for values of l around 0.5. However, when the temporal classifier is combined with a non-temporal classifier, the maximum performance of the combined tends to shift towards the modality with temporal modeling. Furthermore, the temporally modeled classifier performs much better when some global information (coming from the non-temporal classifier) is taken into account. In that sense, it is worth to notice that, except for one case (left side of **aSVM-sHMM**), the combined classifiers outperform the unimodal ones when a small weight is given to the other modality's classifier.

Detailed performance at the optimum working point

Finally, Figures 4.6a-4.6f show the confusion matrix of all the multimodal classifiers for the optimal weighting parameter l. Generally speaking, we notice that the performance of these combined classifiers improved with respect to the unimodal ones. Most of them obtain outstanding results for some of the commands and very good results for most of the commands. We need to remark that the **aHMM-sSVM** and the **aHMM-lSVM** classifiers achieve an ARR of 77%, that represents a considerable increment respect the ARR computed on each modality independently.

4.8 Audio-Visual Command Recognition on Tiny Training Sets

Why tiny training sets?

In the previous experiments we noticed that (i) multimodal classifiers work systematically better than monomodal ones and (ii) many of the representations seem

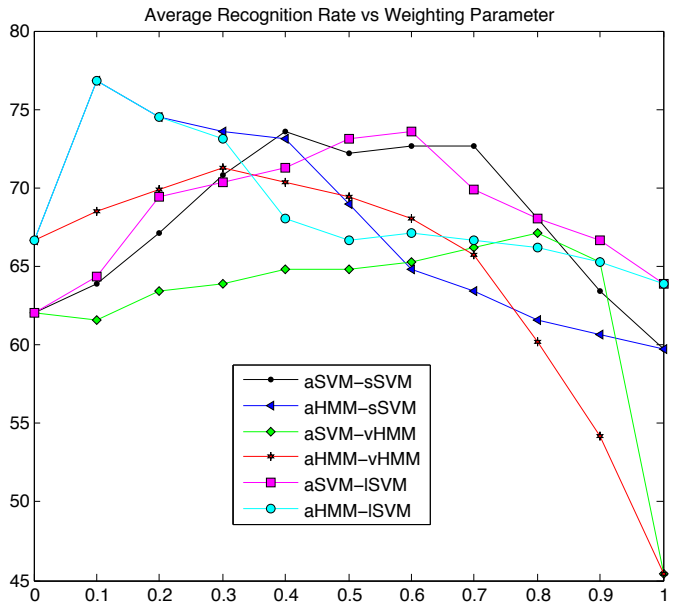


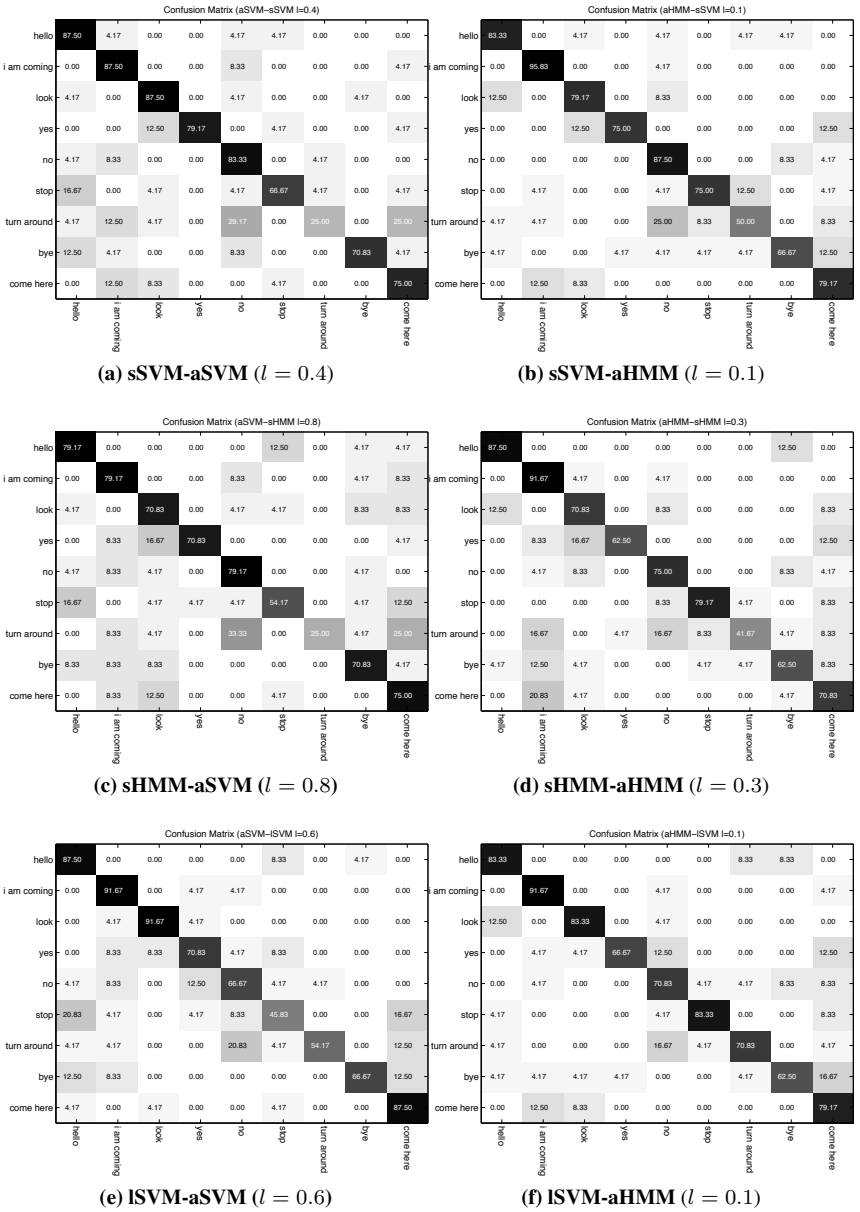
Figure 4.5: Average Recognition Rate as a function of the multimodal weighting parameter l.

to have similar optimal ARR. From a user point of view, it is convenient that the system efficiently learns from a very reduced number of examples. The focus of this contribution is on the performance of different audio-visual discriminative classifiers using tiny training sets. This is important because these AV commands are culture-, language- and user-dependant. Therefore, methods need to constantly adapt and learn from very few examples.

More precisely, we would like to answer three research questions: (1) which is the best classification method? (2) how the methods' accuracy vary when reducing the size of the training set? (3) does the benchmark correspond to the ones obtained using larger training sets? To answer them, we conducted an extensive set of experiments on the RAVEL data set, thus assessing the quality of different approaches and setting a basis for method comparison. For the sake of generality, we ran the experiments with signals acquired using one colour camera and one microphone, the minimal sensor configuration needed to perform audio-visual classification. This disables the use of the scene-flow visual descriptor. In any event, there is no need to keep on comparing both features since we obtained similar performances. Moreover, since we are interested in tiny training sets, we discard the use of the HMM classifier (and hence the per-frame representation) to avoid overfitting problems[2].

We compared five SVM-based methods, with the following discriminant functions:

[2]One may think that regularization techniques may help in this case. However, in our particular framework, the dimension of the feature space is too big compared to the size of the training set, and the overfitting problems remain.



(a) sSVM-aSVM ($l = 0.4$) **(b) sSVM-aHMM** ($l = 0.1$)

(c) sHMM-aSVM ($l = 0.8$) **(d) sHMM-aHMM** ($l = 0.3$)

(e) lSVM-aSVM ($l = 0.6$) **(f) lSVM-aHMM** ($l = 0.1$)

Figure 4.6: Confusion matrix of the optimal combined classifiers: (a) **sSVM-aSVM** ($l = 0.4$), (b) **sSVM-aHMM** ($l = 0.1$), (c) **sHMM-aSVM** ($l = 0.8$), (d) **sHMM-aHMM** ($l = 0.3$), (e) **lSVM-aSVM** ($l = 0.6$) and (f) **aHMM,lSVM** ($l = 0.1$).

aSVM: Audio-only

$$f^{\mathbf{A}}(\boldsymbol{x}^A, c) = \sum_{n=1}^{N} \beta_{n,c}^{\mathbf{AS}} \; k(\boldsymbol{x}^A, \boldsymbol{x}_n^A).$$

lSVM: Video-only

$$f^{\mathbf{V}}(\boldsymbol{x}^V, c) = \sum_{n=1}^{N} \beta_{n,c}^{\mathbf{V}} \; k(\boldsymbol{x}^V, \boldsymbol{x}_n^V).$$

cSVM: Audio-visual concatenation

$$f^{\mathbf{CAT}}(\boldsymbol{x}^{AV}, c) = \sum_{n=1}^{N} \beta_{n,c}^{\mathbf{CAT}} \; k(\boldsymbol{x}^{AV}, \boldsymbol{x}_n^{AV}).$$

wsSVM: The convex Weighting Scheme described in the previous Section[3]

$$f^{\mathbf{WS}}(\boldsymbol{x}^{AV}, c) = \lambda f^{\mathbf{A}}(\boldsymbol{x}^A, c) + (1-\lambda) f^{\mathbf{V}}(\boldsymbol{x}^V, c)$$

mkSVM: The Multiple Kernel framework

$$f^{\mathbf{MK}}(\boldsymbol{x}^{AV}, c) = \sum_{n=1}^{N} \beta_{n,c}^{\mathbf{MK}} (\mu k_A(\boldsymbol{x}^A, \boldsymbol{x}_n^A) + (1-\mu) k_V(\boldsymbol{x}^V, \boldsymbol{x}_n^V)).$$

Notice that the **aSVM** and **lSVM** use only auditory and visual data respectively. Thus, these two methods do not perform any fusion. On the contrary, **cSVM** performs *early fusion*, and **mkSVM** and **wsSVM** perform *late fusion*. The difference between **wsSVM** and **mkSVM** is that, while the first one estimates the SVM coefficients and λ in two different stages, the second performs a joint optimization. In addition, **wsSVM** trains two SVMs instead of one as **mkSVM**, thus twice the number of parameters. A priory, **wsSVM** is faster but less accurate than **mkSVM**.

4.8.1 Audio-Visual Categorization

The (tiny) training sets

We evaluated the methods splitting actor-wise the dataset into a training subset and a testing subset several times, following a standard cross-validation strategy. We named the experiments **E**n, where n is the number of actors in the training set. Hence, **E**n is the average of $\binom{8}{n}$ different training sets, in which there are $3n$ observations per class. We conducted experiments for values of $n = 3, 4, 5, 6, 7$, so a total of 218 different training sets.

The different tested kernels

Since this is the first work (up to the authors' knowledge) that compares audio-visual command classification methods on tiny datasets, we believe necessary to test different possibilities regarding the kernels used and their parameters. The

[3] Because we are using only SVM, there is no need for the normlization step.

Table 4.2: Accuracy results (%) of the methods **aSVM**, **lSVM** and **cSVM** on training sets of different sizes. Bold indicates the best kernel choice.

E	M\ k	L	P	G	C	S
E7	**aSVM**	65.3	65.3	64.8	**71.3**	64.8
	lSVM	59.3	64.4	64.8	**69.0**	65.3
	cSVM	74.1	78.2	78.2	**84.3**	77.3
E6	**aSVM**	62.2	63.4	64.2	**68.2**	62.4
	lSVM	58.9	63.3	64.3	**68.5**	64.4
	cSVM	73.4	75.9	75.9	**81.5**	75.9
E5	**aSVM**	60.2	60.9	61.7	**66.0**	60.3
	lSVM	58.2	61.6	62.5	**65.9**	62.7
	cSVM	72.0	73.5	73.5	**78.8**	73.8
E4	**aSVM**	56.0	57.6	58.6	**63.2**	57.6
	lSVM	56.6	59.6	60.6	**63.7**	60.7
	cSVM	69.9	71.5	71.5	**76.0**	71.7
E3	**aSVM**	49.0	52.6	54.4	**58.8**	54.2
	lSVM	54.9	57.0	57.8	**61.0**	57.8
	cSVM	66.7	67.9	67.9	**72.4**	69.0

tested kernels are: **[L]** linear $k_L(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^{\mathrm{t}}\boldsymbol{x}'$, **[P]** polynomial $k_P(\boldsymbol{x}, \boldsymbol{x}'; d) = (\boldsymbol{x}^{\mathrm{t}}\boldsymbol{x}' + 1)^d$, **[G]** Gaussian $k_G(\boldsymbol{x}, \boldsymbol{x}'; \sigma^2) = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/2\sigma^2)$, **[C]** χ^2 $k_{\chi^2}(\boldsymbol{x}, \boldsymbol{x}'; \nu) = \exp\left(-\frac{1}{\nu}\sum_{k=1}^{K}\frac{(x_k - x'_k)^2}{x_k + x'_k}\right)$ (where $\boldsymbol{x} = (x_1, \ldots, x_K)$) and **[S]** sigmoid $k_S(\boldsymbol{x}, \boldsymbol{x}'; a, c) = \tanh(a\boldsymbol{x}^{\mathrm{t}}\boldsymbol{x}' + c)$. The kernel parameters are: $d \in \{2, 3, 4, 5, 6\}$, $\sigma^2, \nu \in \{10^{-2}, 10^{-1.5}, 10^{-1}, 10^{-0.5}, 10^{0}, 10^{0.5}, 10^{1}\}$ and $a = 20$, $c \in \{-0.5, -0.25, 0, 0.25, 0.5\}$. The codebook size was set to $K = 500$.

For each sub-experiment (training set) and for all choices of kernel(s) and kernel parameter(s), the five methods were evaluated. Notice that for each sub-experiment there are 25 kernel choices for the methods **aSVM**, **lSVM** and **cSVM** and 625 for the methods **wsSVM** and **mkSVM**. In summary, we trained more than 150,000 SVMs[4] to present this study.

4.8.2 Benchmark Results

How to read the tables

In order to compare different methods and kernels we compute the global accuracy of the classifiers, *i.e.*, the percentage of correct classifications. Tables 4.2 and 4.3 show the global accuracies for all the experiments we conducted. Before going into the numeric details we explain how these results are presented in there. First, **M** denotes the method (**aSVM**, **lSVM**, **cSVM**, **wsSVM** or **mkSVM**), k indicates the kernel used (**L**, **P**, **G**, **C** or **S**) and **E** refers to the experiment (**E3**, **E4**, **E5**, **E6** or **E7**). Second, each entry of the table corresponds to the best kernel parameter. Last, the numbers in bold denote the best kernel(s) choice given an experiment **E** and a method **M**.

[4] At this point we would like to mention the MKL C++ library SHOGUN [Sonnenburg 10] and thank the reactivity of its developers, specially Sergey Lisitsyn.

Table 4.3: Accuracy results (%) of the methods **wsSVM** and **mkSVM** on training sets of different sizes. Bold indicates the best kernel choice.

	M	wsSVM					mkSVM				
E	k	**L**	**P**	**G**	**C**	**S**	**L**	**P**	**G**	**C**	**S**
	L	71.8	77.3	79.2	78.2	78.7	76.4	71.8	71.8	69.0	65.3
	P	65.3	66.7	68.1	68.1	68.5	65.3	75.9	75.9	75.9	73.1
E7	**G**	78.2	79.2	80.6	78.2	77.8	71.8	75.9	75.9	75.9	74.5
	C	71.3	71.3	71.3	71.3	71.3	71.3	75.9	80.6	**81.0**	77.3
	S	71.8	76.9	79.6	80.6	**81.5**	64.8	74.5	79.6	77.8	77.8
	L	66.0	68.8	69.8	70.4	70.3	74.2	71.0	70.8	68.5	64.4
	P	63.4	63.7	64.1	64.0	64.2	63.4	74.7	74.8	74.9	70.6
E6	**G**	74.9	**77.2**	76.2	77.0	72.2	70.6	74.5	74.7	75.5	74.3
	C	68.2	68.2	68.2	68.2	68.2	68.2	76.1	79.8	**79.8**	77.0
	S	69.8	72.4	72.4	72.4	72.7	62.4	72.8	76.4	76.9	75.6
	L	62.5	64.8	65.1	65.8	65.2	72.1	68.8	68.7	65.9	62.7
	P	60.9	60.9	60.9	60.9	60.9	60.9	73.2	73.3	72.8	69.5
E5	**G**	72.1	**73.1**	72.6	72.7	68.7	68.6	73.1	73.3	73.4	73.2
	C	66.0	66.0	66.0	66.0	66.0	66.0	74.7	78.4	**78.4**	75.7
	S	65.1	65.5	65.8	65.7	66.0	60.3	64.7	74.2	75.3	74.2
	L	57.6	59.5	59.7	60.4	59.9	69.6	66.5	66.6	63.7	60.7
	P	57.6	57.6	57.6	57.6	57.6	57.6	70.7	70.7	70.5	67.6
E4	**G**	65.5	**67.1**	66.3	66.7	64.2	66.4	70.6	71.1	71.4	71.2
	C	63.2	63.2	63.2	63.2	63.2	63.2	69.1	75.4	**76.3**	73.7
	S	59.9	60.2	60.4	60.8	60.6	57.6	57.8	71.7	72.7	71.7
	L	49.1	49.4	49.6	49.6	49.7	61.1	63.9	63.9	61.0	57.8
	P	52.6	52.6	52.6	52.6	52.6	52.6	66.9	67.0	67.2	65.1
E3	**G**	57.3	**59.9**	59.2	59.8	57.9	63.8	66.7	67.3	68.5	67.2
	C	58.8	58.8	58.8	58.8	58.8	58.8	59.4	72.2	**72.6**	70.7
	S	55.0	55.2	55.6	55.4	55.3	54.2	54.2	68.4	69.7	68.5

Table 4.4: Average time per multiclass classifier.

Method	**aSVM/lSVM**	**cSVM**	**wsSVM**	**mkSVM**
Time Spent [s]	0.79	0.86	1.59	3.27

Results of **aSVM**, **lSVM** and **cSVM**

Table 4.2 shows the accuracy results of three of the methods, namely: **aSVM**, **lSVM** (*no fusion*) and **cSVM** (*early fusion*). We first notice that the audio-visual method performs systematically better than both unimodal approaches. Second, there is no significant difference between methods **aSVM** and **lSVM**. It is also worth noticing how the accuracy of the classifiers decreases when the size of the training set decreases. Indeed, when there is not enough training data, the classifier does not capture the underlying structure of the data, thus causing an accuracy drop.

Results of **wsSVM** and **mkSVM**

Table 4.3 shows the performance of the methods **wsSVM** and **mkSVM**. The columns and rows correspond to the kernel used on visual and auditory data respectively. We remark in the first place that **mkSVM** works better than any unimodal classifier. However, **wsSVM** does not: its accuracy is roughly the same as the unimodal classifiers on the smallest training sets. It is also worth to notice that the **mkSVM** and the **cSVM** methods are comparable and both perform better than the **wsSVM** approach. This last statement is in disagreement with

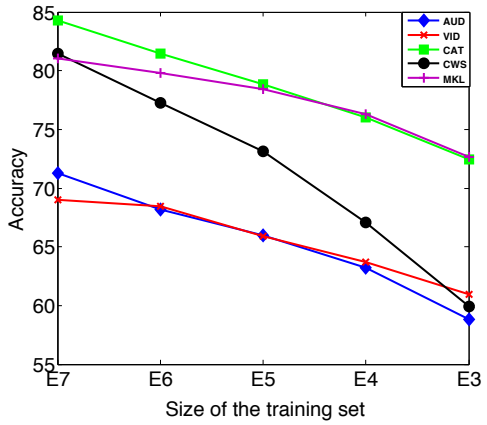


Figure 4.7: χ^2's accuracy for different methods as a function of the training set's size.

[Mühling 12], where **mkSVM** outperforms **cSVM**. Albeit, the experimental conditions are not the same. Indeed, both the size of the training set and the number of classes are smaller here. **wsSVM** shows bad accuracy for smaller datasets compared to **mkSVM** or **cSVM** because **wsSVM** has to train twice the number of parameters than **mkSVM** and **cSVM**. Moreover, the training of those parameters is performed in each modality independently, not allowing, for instance, the auditory information compensate for visual misrepresentations. Hence, when the size of the training set is reduced, the accuracy drop of **wsSVM** is stressed.

χ^2 is the best kernel, let us take a deeper look

We would also like to remark that the best kernel to use is the χ^2 with most of the tested methods. This statement goes accordingly with the existing literature, and there is a simple explanation for that. When using histograms, differences on full bins are less important that differences in almost empty bins. This kind of touch is exactly what the χ^2 kernel accounts for.

In order to present a final comparison, Figure 4.7 shows the accuracy results of the five methods using the χ^2 kernel on the different experiments. From this plot it is clear that (i) audio-visual fusion increases the accuracy, (ii) **mkSVM** and **cSVM** perform equivalently and better than **wsSVM** and (iii) when the size of the training set decreases, the accuracy drops, specially in the case of **wsSVM**

Computation time

Table 4.4 shows the average time spent on the training and testing of one multiclass classifier for the benchmarked methods. As expected, unimodal classifiers are the fastest, closely followed by **cSVM**. **mkSVM** is the slowest method, spending more than twice the time used by the **wsSVM** method.

4.9 Conclusions and Future Work

Summary

In this chapter we presented the contributions to the field of AV command recognition. An AV command is a gesture accompanied by a short sentence or by a

word. For example waving while saying “hello”. We set up a learning framework to understand (i) weather the combination of the two modalities enhances the recognition and (ii) which method performs the best when trained with very few examples.

First contribution: the **wsSVM** method

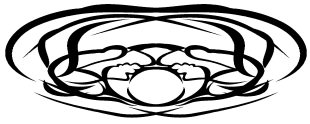
More precisely, in our first contribution we introduced the **wsSVM** method: a late fusion technique able to combine classifiers of different nature. Based on a high-performance and solid learning technique, the method uses binocular/monocular visual features and monaural auditory features. Audio-visual recognition is performed at a later stage in which the classification scores from audio and video are combined to get the final audio-visual score, yielding to important increasings on the ARR. The method is tested in leaving-one-out fashion on a publicly available data set. Results show the importance of using both modalities for recognition. Moreover, from Figure 4.5 it is clear that the choice of the feature/classifier does not have a crucial effect on the performance of the **wsSVM** method. Thus, the classifier will not be chosen regarding its performance, but following other criteria such as training/testing time, overfitting effects, etc.

Second contribution: learning on tiny training sets

Our second contribution deals exactly with this kind of questions, looking towards a real-world scenario. Among all the properties desirable for such applications, we chose to evaluate their performance, speed and user-adaptivity. Since the first two are provided by the BoW+SVM paradigm, we focused on reducing the size of the training set, thus looking for the method yielding the highest user-adaptability. We presented an extensive set of experiments providing for a solid benchmark framework, between three state-of-the-art methods. At the light of these results we answer now the original research questions. In our particular set up, the best trade off between speed, robustness and user-adaptivity is given by **cSVM**. When the size of the training set is reduced, all methods experiment an accuracy drop, as expected. We remark that this drop is much more stressed in the case of **wsSVM**. Finally, the results show that **cSVM** and **mkSVM** react similarly when reducing the size of the training set.

Future work

This work can be extended in several ways. First, by conducting experiments on datasets with higher number of classes. In addition, we would like to perform tests on other audio-visual command datasets recorded in different languages and countries, providing for a large variety of gesture and speech utterances, thus evaluating the cultural influence on the proposed approaches. This will throw the basis for future work towards a continuous audio-visual command recognition method. Last, the methods should be tested on a robot platform with limited resources to produce a robust and highly-adaptive recognition method yielding to a real-world scenario application.



CHAPTER 5

MULTICHANNEL SOUND SOURCE LOCALISATION

This chapter addresses the sound source localisation (SSL) problem from multichannel time delay estimates (TDE) using non-coplanar microphone arrays. The problem is cast into a non-linear constrained optimisation task. On one side, the cost function is derived from the signal model. On the other side, the constraint ensure that the resulting time delay estimate is feasible, i.e., corresponds to a location in the sound source space. The geometry associated to the propagation model guarantees the uniqueness of the sound source position for any feasible set of time delays. Moreover, a localisation mapping is built, thus providing a closed-form to recover the sound source position. Two optimization techniques are proposed to solve the multichannel TDE-SSL problem. We report and extensive set of experiments and comparisons with state-of-the-art methods on simulated and real data in the presence of noise and reverberations, that validate the introduced model and the proposed algorithms

5.1 Introduction

The importance of Time Delay Estimation

For the last decades, source localisation from time delay estimates (TDEs) has proven to be an extremely useful methodology with a variety of applications in such diverse fields as aeronautics, telecommunications and robotics. This problem is highly related to the one of estimating time delays. We are particularly interested in the development of a general-purpose TDE-based method for sound-source localisation in indoor environments, e.g, human-robot interaction, ad-hoc teleconferencing using microphone arrays, etc. This type of consumer-oriented applications are extremely challenging for several reasons: (i) there may be several sound sources and their number varies over time, (ii) regular rooms are echoic, thus leading to reverberations, and (iii) the microphones are often embedded in devices (robot heads, smart phones, etc.) generating high-level noise.

Contributions

During this thesis we introduced several contributions to the field of sound source localisation from time delay estimates. First, we cast the time delay estimation problem for sound source localisation using non-coplanar microphone arrays into a non-linear constrained optimization method **GTDE1**. Second, a local minimization algorithm was proposed to solve the task **GTDE2**. Both contributions were published in [Alameda-Pineda 12a]. Third, the full geometric model – the responsible of the optimization constraints – was derived **GTDE3**. Finally, a global optimization algorithm was introduced in **GTDE4** and published in [Alameda-Pineda 13b] together with the geometric model.

Chapter structure

The remaining of the chapter is structured as follows. Section 5.2 describes the state-of-the-art on the topic. Section 5.3 contains the basics of the approach, namely, the signal model and the propagation model. Section 5.4 presents the first contribution, that is, the full geometric model, together with the formal proofs. Section 5.5 castes the estimation task into a non-linear constrained multivariate optimization problem: our second contribution to this field. Sections 5.6 and 5.7 respectively introduce the grid-based local optimization technique and the global optimization technique to solve the TDE task. Finally, the conclusions and future work guidelines are drawn in Section 5.9.

5.2 Related Work

The time delay estimation (TDE) problem applied to sound source localisation (SSL) has been very well investigated. We grouped the existing works depending on how the received auditory signals are used for localisation. We name *Bichannel SSL* to the group of approaches using two microphones. *Multilateration* denotes the methods using more than two microphones, but performing SSL from binaural cues extracted from all microphone pairs. Finally, *Multichannel SSL* states for those algorithms that localize the sound source from multichannel measurements.

Bichannel TDE-SSL

Most of the algorithms performing sound source localisation with one pair of microphones lead to a method estimating the azimuth of the sound source, see [Liu 08, Mandel 07, Viste 03, Woodruff 12]. There are some approaches, however, able to localize the direction (azimuth-elevation) of the sound source from binaural cues [Kullaib 09, Deleforge 12a]. Other approaches track the sound sources using binaural cues [Keyrouz 06, Keyrouz 07]. In all of these cases, models and methods are specifically designed for the binaural case, i.e., using two microphones. Unfortunately, it is not obvious how to extend most of the cited methods to use more than two microphones. However, one can always estimate the time delays pair-wise and then perform sound source localisation from these pair-wise TDE. This procedure is called *multilateration*.

Multilateration

As outlined before, methods performing multilateration separate the estimation of the time delays from the localisation of the sound source in two different steps. Moreover, most of the approaches emphasize the localisation module, borrowing an off-the-shelf time delay estimation algorithm. The localisation task is cast into a least squares problem in several previous works [Smith 87, Brandstein 97a, Brandstein 97b, Friedlander 87, Huang 01, Canclini 13], with improved algorithms handling the hard-least-square problem [Beck 08] and dealing with time delay estimate outliers [Galati 06]. Two recent methods include the reverberations in the model in order to learn the effect they have on the input signals. On one side, in [Brutti 08] the authors use the acoustic maps together with the GCC-PHAT technique to localize sound sources from TDE's. On the other side, the model in [Ribeiro 10] includes the reverberations in order to enhance the localisation performance while using a uniform circular array of microphones. Another set of methods perform the SSL task in a maximum likelihood framework [Chen 03a, Sheng 05, So 08, Urruela 04, Strobel 99, Zhang 07, Zhang 08]. A geometry-based localisation method is described in [Chan 94]. We refer the reader to [Seco 09, Pertilä 09] for two nice surveys on multilateration. The main advantage of splitting the task into time delay estimation and sound source localisation is that one can try many combinations of TDE algorithms and SSL algorithms. The main disadvantage of such two-step framework is that the obtained time delay estimates may be inconsistent. Indeed, independently estimated time delays may be in disagreement with the geometry of the microphone array. For example, in a linear array with three microphones, the sound wave cannot reach the second microphone after reaching the first and the third microphones. The consistent orders are $1 \rightarrow 2 \rightarrow 3$, $3 \rightarrow 2 \rightarrow 1$ and $2 \rightarrow 1, 3$, but $1 \rightarrow 3 \rightarrow 2$ is not physically possible since 2 is in the propagation path from 1 to 3. This may be avoided by considering the estimation of all the time delays at once, that is, *multichannel TDE*.

Multichannel TDE-SSL (I)

The problem of localizing a sound source from multichannel measurements has been very well investigated and a recent review can be found in [Chen 06]. Methods addressing *multichannel* TDE can be roughly divided into two categories: methods estimating the acoustic impulse responses and methods exploiting the redundancy among several microphones. The estimation of the impulse responses from the raw data is an extremely challenging task, since the effects of the environment as well as those of the microphone array are coupled. [Doclo 03, Salvati 13, Huang 03] estimate these responses by means of the generalized eigenvalue decomposition. In [Lim 13] the authors use canonical correlation analysis, claiming robustness towards low SNR. Because all these methods do not make use of any information related neither to the microphone array nor to the acoustic environment, the impulse responses are learnt directly from the data. Consequently, all these methods need a lot of training data, and complex training phases need to be run before using the algorithm. Furthermore, since the methods do not decouple the microphone array from the acoustic environment, but learn the effects of both all together, a new training is required for every new environment. This is an undesirable feature for mobile platforms that need high adaptableness to new

unconstrained environments.

Multichannel TDE-SSL (II)

The second category is represented by [Chen 03b] where a multichannel criterion based on cross-correlation is proposed to estimate time delays using a *linear* microphone array. This approach was extended in [He 13b] by using temporal prediction. Also [Chen 03b] is proven to be equivalent to two information theory-based criterion for TDE [He 13a, Benesty 07], under some statistical assumptions. However all these methods are designed for the particular case of linear microphone arrays. There exists approaches performing SSL designed for other array geometries such as circular [Pavlidi 13] and spherical [Sasaki 12]. In all these cases, the geometry array is directly encoded into the cost function/probability model. As a consequence, most of these frameworks are not easily extendible to arbitrarily-shaped microphone arrays.

Unlike existing approaches for multichannel TDE, we did not include the geometry of the array into the minimization criterion, because it is not known in advance. Hence, we developed a new framework introducing several contributions. First, we analysed the geometry of the problem derived from the direct path propagation model. This lead to a full characterization of the *feasible time delays* (those corresponding to a position in the ambient space), to the *uniqueness of the sound source position* and to a *closed-form* solution for source *localisation*. Second, we cast the TDE problem into a multivariate optimization problem under these feasibility constraints. This formulation allowed us to look for the best minimization technique to solve the TDE task. Third, we used a local optimization procedure using grid-based initialization to evaluate the geometric model in the framework of sound source localisation. Last, we proposed a global optimization technique to reduce the computing time associated to the grid-based method.

5.3 Signal and Propagation Models

Signal model

In this Section we describe the signal model and the propagation model allowing to relate time delays with the relative position between source and microphones. We introduce the following notations: the position of the sound source $\boldsymbol{S} \in \mathbb{R}^N$, the number of microphones M, as well as their positions, $\{\boldsymbol{M}_m\}_{m=1}^{m=M} \in \mathbb{R}^N$. Let $x(t)$ be the signal emitted by the source. The signal received at the m^{th} microphone writes:

$$x_m(t) = x(t - t_m) + n_m(t), \tag{5.1}$$

Propagation model

where n_m is the noise associated with the m^{th} microphone and t_m is the time-of-arrival from the source to that microphone. The microphones' noise signals are assumed to be zero-mean independent Gaussian random processes. Throughout the chapter, constant sound propagation speed is assumed, denoted by ν. Hence we write $t_m = \|\boldsymbol{S} - \boldsymbol{M}_m\|/\nu$. Using this model, the expression for the time delay between the m^{th} and the n^{th} microphones, denoted by $t_{m,n}$, writes:

$$t_{m,n} = t_n - t_m = \frac{\|\boldsymbol{S} - \boldsymbol{M}_n\| - \|\boldsymbol{S} - \boldsymbol{M}_m\|}{\nu}. \tag{5.2}$$

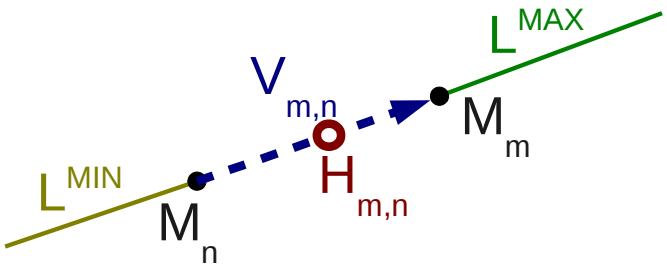


Figure 5.1: Geometry associated with the two microphone case, located at $\boldsymbol{M}_m$ and $\boldsymbol{M}_n$ (see Lemma 1). $H_{m,n}$ is the mid-point of the the microphones (in red) and $V_{m,n}$ the vector $\boldsymbol{M}_m - \boldsymbol{M}_n$ (in dashed-blue). $L^{\text{MAX}}_{m,n}$ and $L^{\text{MIN}}_{m,n}$ are the two half lines represented in green and yellow respectively.

5.4 Time Delay Feasibility

Overview of the main result

We recall that the task is to estimate the time delays to further localize the sound source. In this Section, we provide the three main theoretical results: (i) the conditions under which a set of *time delays correspond to a sound source* (such sets will be called *feasible* sets) and (ii) *the uniqueness of the sound source position* for any feasible set and (iii) a *closed-formula for localisation*, i.e., to retrieve the position of the sound source from a feasible set. Even if, in practice the problem is set in the ambient space, $\mathbb{R}^3$, the theory presented here is valid in $\mathbb{R}^N, N \geq 2$. In the following, Section 5.4.1 studies the case of two microphones and Section 5.4.2 exploits the geometry of the M microphone case.

5.4.1 The Case of Two Microphones

Characterization of $\boldsymbol{S}$ satisfying $\hat{t}_{m,n} = t_{m,n}(\boldsymbol{S})$

We start by formally characterizing the set of possible sound-source locations in the case of two microphones located at $\boldsymbol{M}_m$ and $\boldsymbol{M}_n$. For a given time delay $\hat{t}_{m,n}$, we characterize $\boldsymbol{S}$ satisfying $\hat{t}_{m,n} = t_{m,n}(\boldsymbol{S})$. Because (5.2) is a hyperboloid in $\mathbb{R}^N$, this equation embeds the *hyperbolic geometry* of the problem. For completeness, we state the following lemma (Figure 5.1):

Lemma 1 *The set of sound-source locations $\boldsymbol{S} \in \mathbb{R}^N$ satisfying $t_{m,n}(\boldsymbol{S}) = \hat{t}_{m,n}$ is:*

(i). empty if $|\hat{t}_{m,n}| > t^*_{m,n}$, where $t^*_{m,n} = \|\boldsymbol{M}_m - \boldsymbol{M}_n\|/\nu$,

(ii). the half line $\boldsymbol{L}^{MAX}_{m,n}$ (or $\boldsymbol{L}^{MIN}_{m,n}$), if $\hat{t}_{m,n} = t^*_{m,n}$ (or if $\hat{t}_{m,n} = -t^*_{m,n}$), where $\boldsymbol{L}^{MAX}_{m,n} = \{\boldsymbol{H}_{m,n} + \mu \boldsymbol{V}_{m,n}\}$, $\boldsymbol{L}^{MIN}_{m,n} = \{\boldsymbol{H}_{m,n} - \mu \boldsymbol{V}_{m,n}\}$, $\mu \geq 1/2$, $\boldsymbol{H}_{m,n} = (\boldsymbol{M}_m + \boldsymbol{M}_n)/2$ and $\boldsymbol{V}_{m,n} = \boldsymbol{M}_m - \boldsymbol{M}_n$,

(iii). the hyperplane passing by $\boldsymbol{H}_{m,n}$ perpendicular to $\boldsymbol{V}_{m,n}$, if $\hat{t}_{m,n} = 0$ or

(iv). one sheet of a two-sheet hyperboloid with foci $\boldsymbol{M}_m$ and $\boldsymbol{M}_n$ for other values of $\hat{t}_{m,n}$.

Proof: Using the triangular inequality, it is easy to see $-t^*_{m,n} \leq t_{m,n}(\boldsymbol{S}) \leq t^*_{m,n}$, $\forall \boldsymbol{S} \in \mathbb{R}^N$, which proves (i). (ii) is proven by rewriting $\boldsymbol{S} = \boldsymbol{H}_{m,n} + \mu_1 \boldsymbol{V}_{m,n} + \sum_{k=2}^{N} \mu_k \boldsymbol{W}_k$, where $(\boldsymbol{V}_{m,n}, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_N)$ is an orthogonal basis of $\mathbb{R}^N$, and deriving with respect to the μ_i's. In order to prove (iii) and (iv) and without loss of generality, we can assume $\boldsymbol{M}_m = \boldsymbol{e}_1$, $\boldsymbol{M}_n = -\boldsymbol{e}_1$ and $\nu = 1$, where $\boldsymbol{e}_1$ is the first element of the canonical basis of $\mathbb{R}^N$. Equation (5.2) rewrites:

$$(\hat{t}_{m,n})^2 + 4x_1 = -2\hat{t}_{m,n} \left((x_1 + 1)^2 + \sum_{k=2}^{N} x_k^2 \right)^{\frac{1}{2}}, \tag{5.3}$$

where $(x_1, \ldots, x_N)^{\mathrm{t}}$ are the coordinates of $\boldsymbol{S}$. By squaring the previous equation we obtain:

$$a(4-a) + 4a \sum_{k=2}^{N} x_k^2 - 4(4-a)x_1^2 = 0, \tag{5.4}$$

where $a = \left(\hat{t}_{m,n}\right)^2$. Notice that if $\hat{t}_{m,n} = 0$ we get $x_1 = 0$, which corresponds to the statement in (iii). For the rest of values of a, that is $0 < a < (t^*_{m,n})^2 = 4$, equation (5.4) represents a two-sheet hyperboloid, since all coefficients are strictly positive except the one of x_1^2, that is strictly negative. We can rewrite (5.4) as:

$$x_1^2 = \frac{a(4-a) + 4a \sum_{k=2}^{N} x_k^2}{4(4-a)}. \tag{5.5}$$

We observe that the set of solutions of (5.4) can be split into two subsets $\mathcal{S}^+_{m,n}$ and $\mathcal{S}^-_{m,n}$ parametrized by $(x_2, \ldots, x_N)$, corresponding to the two solutions for x_1 of equation (5.5). These two sets are the two sheets of the hyperboloid defined in (5.4). Moreover, one can easily verify that $t_{m,n}\left(\mathcal{S}^+_{m,n}\right) = -t_{m,n}\left(\mathcal{S}^-_{m,n}\right)$, so either $t_{m,n}\left(\mathcal{S}^+_{m,n}\right) = \hat{t}_{m,n}$ or $t_{m,n}\left(\mathcal{S}^-_{m,n}\right) = \hat{t}_{m,n}$, but not both. Hence the set of points $\boldsymbol{S}$ satisfying $t_{m,n}(\boldsymbol{S}) = \hat{t}_{m,n}$ is either $\mathcal{S}^+_{m,n}$ or $\mathcal{S}^-_{m,n}$, so one sheet of a two-sheet hyperboloid.

Understanding the spurious solutions

We remark that some solutions of equation (5.4) are not solutions of (5.3). Because (5.4) only depends on $a = (\hat{t}_{m,n})^2$ and not on $\hat{t}_{m,n}$, the sign of $\hat{t}_{m,n}$ is irrelevant for the solutions of (5.4). Consequently, the solutions of (5.4) contain, in addition to the *genuine* solutions (those of (5.3)), a set of *spurious* solutions satisfying (5.3) with $-\hat{t}_{m,n}$ instead of $\hat{t}_{m,n}$. However, since $t_{m,n}(\mathcal{S}^+_{m,n}) = -t_{m,n}(\mathcal{S}^-_{m,n})$, we are able to disambiguate the *genuine* solutions from the *spurious* ones. The previous Lemma performs a deep geometrical analysis of the consequences of equation (5.2) on $\boldsymbol{S}$. The next natural step is to consider the consequences of all the equations (5.2) at once, analysing the geometry of the most general microphone set up.

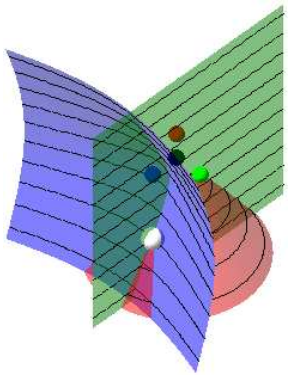


Figure 5.2: Localization of the source using four microphones. Their position is shown in black ($\boldsymbol{M}_1$), blue ($\boldsymbol{M}_2$), red ($\boldsymbol{M}_3$) and green ($\boldsymbol{M}_4$). The sound source is placed in the white marker. The blue hyperboloid corresponds to $\hat{t}_{1,2}$, the red to $\hat{t}_{1,3}$ and the green to $\hat{t}_{1,4}$. The intersection of the hyperboloids corresponds to the sound source position.

5.4.2 The Case of M Microphones in General Position

Linear coupling between time delays

In this Section we characterize the set of possible sound-source locations in the case of M microphones. We first notice that if a set of time delays $\hat{\boldsymbol{t}} = \{\hat{t}_{m,n}\}_{m=1,n=1}^{m=M,n=M} \in \mathbb{R}^{M^2}$ satisfies (5.2) $\forall m, n$, then the time delays are coupled by $\hat{t}_{m,n} = -\hat{t}_{1,m} + \hat{t}_{1,n}$. Hence, we only need to consider the time delays $\boldsymbol{t} = (t_{1,2}, \ldots, t_{1,M})$ which lie in a $(M-1)$-dimensional vector subspace $\mathcal{W} \subset \mathbb{R}^{M^2}$.

The original system of equations

Consequently, there are $M-1$ equations of the form (5.2). Geometrically, this is equivalent to seek the intersection of $M-1$ hyperboloids in $\mathbb{R}^N$ (see Figure 5.2). Algebraically, this is equivalent to solve a system on $M-1$ non-linear equations in N unknowns. In general, this leads to search for the roots of a high-degree polynomial. However, in our case the hyperboloids share one focus, namely $\boldsymbol{M}_1$. As it will be shown below, the problem in this case reduces to solving a second-degree polynomial plus a linear system of equations. The $M-1$ equations write:

$$\begin{cases} \nu\hat{t}_{1,2} &= \|\boldsymbol{S}-\boldsymbol{M}_2\| - \|\boldsymbol{S}-\boldsymbol{M}_1\| \\ &\vdots \\ \nu\hat{t}_{1,M} &= \|\boldsymbol{S}-\boldsymbol{M}_M\| - \|\boldsymbol{S}-\boldsymbol{M}_1\| \end{cases} . \tag{5.6}$$

Summary of the reasoning

Because the M microphones are in general position (they do not lie in the same hyperplane), we have $M \geq N+1$, hence the number of equations is greater or equal than the number of unknowns. We now provide the conditions on $\hat{\boldsymbol{t}}$ under which (5.6) yields a real and unique solution for $\boldsymbol{S}$. More precisely, firstly we provide a necessary condition on $\hat{\boldsymbol{t}}$ for (5.6) to have real solutions, secondly we prove the uniqueness of the solution and build a mapping to recover the solution $\boldsymbol{S}$, and thirdly we provide a necessary and sufficient condition on $\hat{\boldsymbol{t}}$ for (5.6) to have a real and unique solution.

Expressing the original system in matrix form

Notice that each equation in (5.6) is equivalent to $(\nu\hat{t}_{1,m} + \|\boldsymbol{S} - \boldsymbol{M}_1\|)^2 = \|\boldsymbol{S}-\boldsymbol{M}_m\|^2$, from which we obtain $-2(\boldsymbol{M}_1-\boldsymbol{M}_m)^{\mathrm{t}}\boldsymbol{S}+p_{1,m}\|\boldsymbol{S}-\boldsymbol{M}_1\|+q_{1,m} = 0$, where $p_{1,m} = 2\nu\hat{t}_{1,m}$ and $q_{1,m} = \nu^2(\hat{t}_{1,m})^2+\|\boldsymbol{M}_1\|^2-\|\boldsymbol{M}_m\|^2$. Hence, (5.6) can now be written in matrix form:

$$\mathbf{M}\boldsymbol{S} + \boldsymbol{P}\|\boldsymbol{S} - \boldsymbol{M}_1\| + \boldsymbol{Q} = 0, \tag{5.7}$$

where $\mathbf{M} \in \mathbb{R}^{(M-1)\times N}$ is a matrix with its m^{th} row, $1 \leq m \leq M-1$, equal to $(\boldsymbol{M}_{m+1} - \boldsymbol{M}_1)^{\mathrm{t}}$, $\boldsymbol{P} = (p_{1,2}, \ldots, p_{1,M})^{\mathrm{t}}$ and $\boldsymbol{Q} = (q_{1,2}, \ldots, q_{1,M})^{\mathrm{t}}$. Notice that $\boldsymbol{P}$ and $\boldsymbol{Q}$ depend on $\hat{\boldsymbol{t}}$.

Splitting (5.7) into two subsystems

Without loss of generality and because the points $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_M$ do not lie in the same hyperplane, we assume that $\mathbf{M}$ can be written as a concatenation of an invertible matrix $\mathbf{M}_L \in \mathbb{R}^{N\times N}$ and a matrix $\mathbf{M}_E \in \mathbb{R}^{(M-N-1)\times N}$ such that $\mathbf{M} = \begin{pmatrix} \mathbf{M}_L \\ \mathbf{M}_E \end{pmatrix}$. Similarly $\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P}_L \\ \boldsymbol{P}_E \end{pmatrix}$ and $\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_L \\ \boldsymbol{Q}_E \end{pmatrix}$. Thus, (5.7) rewrites:

$$\begin{aligned} \mathbf{M}_L\boldsymbol{S} + \boldsymbol{P}_L\|\boldsymbol{S} - \boldsymbol{M}_1\| + \boldsymbol{Q}_L &= 0, & (5.8)\\ \mathbf{M}_E\boldsymbol{S} + \boldsymbol{P}_E\|\boldsymbol{S} - \boldsymbol{M}_1\| + \boldsymbol{Q}_E &= 0, & (5.9) \end{aligned}$$

where $\boldsymbol{P}_L$, $\boldsymbol{Q}_L$ are vectors in $\mathbb{R}^N$ and $\boldsymbol{P}_E$, $\boldsymbol{Q}_E$ are vectors in $\mathbb{R}^{M-N-1}$. If we decompose $\hat{\boldsymbol{t}}$ into $\hat{\boldsymbol{t}}_L$ and $\hat{\boldsymbol{t}}_E$, we observe that $\boldsymbol{P}_L$ and $\boldsymbol{W}_L$ depend only on $\hat{\boldsymbol{t}}_L$ and that $\boldsymbol{P}_E$ and $\boldsymbol{Q}_E$ depend only on $\hat{\boldsymbol{t}}_E$. Notice that (5.6) is strictly equivalent to (5.8)-(5.9). In the following, (5.8) will be used for defining the necessary conditions on $\hat{\boldsymbol{t}}$ as well as localizing the sound source. The study of (5.9) is reported further on. By introducing a scalar variable w, (5.8) can be written as:

The first subsystem is the intersection between a straight line and a quadric

$$\begin{aligned} \mathbf{M}_L\boldsymbol{S} + w\boldsymbol{P}_L + \boldsymbol{Q}_L &= 0, & (5.10)\\ \|\boldsymbol{S} - \boldsymbol{M}_1\|^2 - w^2 &= 0. & (5.11) \end{aligned}$$

We remark that the system (5.10)-(5.11) is defined in the $(\boldsymbol{S}, w)$ space. Notice that (5.10) represents a straight line and (5.11) represents quadric. Hence the solution to (5.10)-(5.11) is the intersection of a straight line and a quadric. In such systems there are two possible configurations: (i) the quadric contains the straight line, and there are an infinite number of solutions, or (ii) the straight line crosses the quadric, and there are two (maybe complex) solutions. In fact, the first case, (i), does not occur. Notice that the quadric is a two-sheet hyperboloid. Because two-sheet hyperboloids are not ruled surfaces, (5.11) does not contain any straight line. Consequently the system has two (maybe complex) solutions.

Solving (5.10)-(5.11)

In order to solve (5.10)-(5.11), we first rewrite (5.10) as

$$\boldsymbol{S} = \boldsymbol{A}w + \boldsymbol{B}, \tag{5.12}$$

where $\boldsymbol{A} = -\mathbf{M}_L^{-1}\boldsymbol{P}_L$ and $\boldsymbol{B} = -\mathbf{M}_L^{-1}\boldsymbol{Q}_L$, and then substitute $\boldsymbol{S}$ from (5.12) into (5.11) obtaining:

$$(\|\boldsymbol{A}\|^2 - 1)w^2 + 2\langle \boldsymbol{A}, \boldsymbol{B} - \boldsymbol{M}_1\rangle w + \|\boldsymbol{B} - \boldsymbol{M}_1\|^2 = 0. \tag{5.13}$$

We are interested in the real solutions, that is, $\boldsymbol{S} \in \mathbb{R}^N$. Because $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^N$, the solutions of (5.10)-(5.11) are real, if and only if, the solutions to (5.13) are real too. Equivalently, the discriminant of (5.13) has to be non-negative. Hence the solutions to (5.10)-(5.11) are real if and only if $\hat{\boldsymbol{t}}$ satisfies:

The solutions need to be real: deriving the necessary condition for $\hat{\boldsymbol{t}}$

$$\Delta(\hat{\boldsymbol{t}}) := \langle \boldsymbol{A}, \boldsymbol{B} - \boldsymbol{M}_1 \rangle^2 - \|\boldsymbol{B} - \boldsymbol{M}_1\|^2 (\|\boldsymbol{A}\|^2 - 1) \geq 0. \tag{5.14}$$

The previous equation is a *necessary condition* for (5.10)-(5.11) to have real solutions. Albeit, we are interested in the solutions of (5.8). Obviously, if $\boldsymbol{S}$ is a solution of (5.8), then $(\boldsymbol{S}, \|\boldsymbol{S} - \boldsymbol{M}_1\|)$ is a solution of (5.10)-(5.11). However, the reciprocal is not true; these two systems are not equivalent. Indeed, since $\Delta(\hat{\boldsymbol{t}}) = \Delta(-\hat{\boldsymbol{t}})$, one of the solutions of (5.10)-(5.11) is the solution of (5.8) and the other is the solution of (5.8) replacing $\hat{\boldsymbol{t}}$ by $-\hat{\boldsymbol{t}}$. In other words, the two solutions of (5.10)-(5.11), namely $(\boldsymbol{S}^+, w^+)$ and $(\boldsymbol{S}^-, w^-)$, satisfy either:

$$\begin{cases} \boldsymbol{t}(\boldsymbol{S}^+) = \hat{\boldsymbol{t}} \\ \boldsymbol{t}(\boldsymbol{S}^-) = -\hat{\boldsymbol{t}} \end{cases} \text{ or } \begin{cases} \boldsymbol{t}(\boldsymbol{S}^+) = -\hat{\boldsymbol{t}} \\ \boldsymbol{t}(\boldsymbol{S}^-) = \hat{\boldsymbol{t}} \end{cases}$$

Consequently, the solution to (5.8) is *unique*. Moreover, we can use (5.12) to define the following *localisation mapping*, which retrieves the sound-source position from a feasible $\hat{\boldsymbol{t}}$:

Symmetry in the $\boldsymbol{t}$ space: the solution of (5.10)-(5.11) is unique and has a closed-form

$$L(\hat{\boldsymbol{t}}) := \begin{cases} \boldsymbol{S}^+ = \boldsymbol{A}w^+ + \boldsymbol{B} & \text{if } \boldsymbol{t}(\boldsymbol{S}^+) = \hat{\boldsymbol{t}} \\ \boldsymbol{S}^- = \boldsymbol{A}w^- + \boldsymbol{B} & \text{otherwise.} \end{cases} \tag{5.15}$$

Until now we provided the condition for equation (5.8) to have real solutions, the uniqueness of the solution and a localisation mapping. However, the original system includes also equation (5.9). In fact, (5.9) adds $M - N - 1$ constraints onto $\hat{\boldsymbol{t}}$. Indeed, if $(L(\hat{\boldsymbol{t}}), \|L(\hat{\boldsymbol{t}}) - \boldsymbol{M}_1\|)$ is the solution to (5.8), then in order to be a solution of (5.8)-(5.9), it has to satisfy:

The sufficient condition for $\hat{\boldsymbol{t}}$

$$E(\hat{\boldsymbol{t}}) := \mathbf{M}_E L(\hat{\boldsymbol{t}}) + \boldsymbol{P}_E \|L(\hat{\boldsymbol{t}}) - \boldsymbol{M}_1\| + \boldsymbol{Q}_E = 0. \tag{5.16}$$

Moreover, the reciprocal is true. Summarizing, the system (5.8)-(5.9) has a unique solution $L(\hat{\boldsymbol{t}})$ if and only if $\Delta(\hat{\boldsymbol{t}}) \geq 0$ and $E(\hat{\boldsymbol{t}}) = 0$.

The mappings Δ, E and L are explicitly constructed solely from the microphone locations $\mathfrak{M}$. Hence, these mappings are not only an interesting mathematical finding in its own right, but also useful from a computational perspective. In addition, the mappings Δ and E can be understood from two points of view. Geometrically, they characterize the *time delays corresponding to a sound source*. Algebraically, Δ and E represent the *feasibility* constraint to the time delay estimation problem, i.e., the time delay estimate should satisfy the necessary and sufficient conditions for the existence of $\boldsymbol{S}$. L has to be understood as the *closed-form solution for localisation*, allowing to recover $\boldsymbol{S}$ from any feasible $\hat{\boldsymbol{t}}$.

Summary

In all, this can be formalized as:

The formal result

Theorem 1 *Let $\mathfrak{M} = \{\boldsymbol{M}_m\}_{m=1}^M \subset \mathbb{R}^N$ be a set of known points in general position (i.e., not lying in the same hyperplane). Let be $\nu \in \mathbb{R}$, $\nu > 0$. Consider also an unknown point $\boldsymbol{S} \in \mathbb{R}^N$. The following statements hold:*

Geometric Characterization *The set of feasible values for $\hat{\boldsymbol{t}}$ (i.e., the values satisfying (5.2) $\forall m, n$) is a bounded N-dimensional manifold with boundary, denoted by $\mathcal{T}$, contained in a $M-1$ vector subspace $\mathcal{W} \subset \mathbb{R}^{M^2}$.*

Algebraic Characterization *Moreover, there exist two mapping $\Delta : \mathcal{W} \rightarrow \mathbb{R}$ and $E : \mathcal{W} \rightarrow \mathbb{R}^{M-N-1}$, built solely from $\mathfrak{M}$, such that $\forall \hat{\boldsymbol{t}} \in \mathcal{W}$:*

$$\hat{\boldsymbol{t}} \in \mathcal{T} \Leftrightarrow (\Delta(\hat{\boldsymbol{t}}) \geq 0 \text{ and } E(\hat{\boldsymbol{t}}) = 0).$$

Localization *There exists a mapping $L : \mathcal{T} \rightarrow \mathbb{R}^N$, such that $\boldsymbol{S} = L(\hat{\boldsymbol{t}})$ satisfies (5.6), $\forall \hat{\boldsymbol{t}} \in \mathcal{T}$.*

Proof: During this section we have already proven the stated result. The *Algebraic Characterization* is proven in equations (5.14) and (5.16). Notice that, because E is defined from $\mathcal{W}$ to $\mathbb{R}^{M-N-1}$, the dimension of the feasible values is reduced to $M-1-(M-N-1) = N$. Thus $\hat{\boldsymbol{t}}$ lay in a N-dimensional manifold $\mathcal{T}$. The boundary of the manifold is considered because the points satisfying $\Delta(\hat{\boldsymbol{t}}) = 0$ are feasible points. More precisely, the boundary of the manifold is defined as:

$$\partial\mathcal{T} = \{\hat{\boldsymbol{t}} \in \mathcal{W} | \Delta(\hat{\boldsymbol{t}}) \text{ and } E(\hat{\boldsymbol{t}}) = 0\}.$$

Last, $\mathcal{T}$ is bounded because $\hat{\boldsymbol{t}}$ are bounded too (see Lemma 1). Consequently we also proved the *Geometric Characterization*. The *Localization* mapping L is naturally built (5.15). The reader may verify the remaining details.

5.5 TDE-SSL as Non-linear Constrained Optimization

TDE as an optimization problem

In this section we show how the TDE estimation problem can be cast into a multivariated non-linear optimization problem. We first derive a criterion to choose the best value for $\boldsymbol{t}$ and the formally set up the optimization task.

5.5.1 A Criterion for Multichannel TDE

The criterion: linear prediction error

The criterion used in [Chen 03b] was built from the theory of linear predictors. The authors presented this criterion in the framework of linear microphone arrays. We here generalise it to arbitrarily-shaped microphone arrays following a similar line of though. Given the M received signals $\{x_m(t)\}_{m=1}^{m=M}$, we would like to estimate the time delays between them. As explained before, only $M-1$ of the delays are independent. Without loss of generality we choose the delays $t_{1,2}, \ldots, t_{1,m}, \ldots, t_{1,M}$. We select $x_1(t)$ as the reference signal and set the following prediction error:

$$e_{\boldsymbol{c},\boldsymbol{t}}(t) = x_1(t) - \sum_{m=2}^{M} c_{1,m}\, x_m(t + t_{1,m}), \tag{5.17}$$

where $\boldsymbol{c} = (c_{1,2}, \ldots, c_{1,m}, \ldots, c_{1,M})^{\mathrm{t}}$ is the vector of the prediction coefficients and $\boldsymbol{t} = (t_{1,2}, \ldots, t_{1,m}, \ldots, t_{1,M})^{\mathrm{t}}$ is the vector of the prediction time delays. Notice also that, when $\boldsymbol{t}$ takes the true value, the signals $x_m(t+t_{1,m})$ and $x_n(t+t_{1,n})$ are on phase. The criterion to minimize is the expected energy of the prediction error (5.17), leading to a (right now) unconstrained optimization problem:

$$(\boldsymbol{c}^*, \boldsymbol{t}^*) = \arg\min_{\boldsymbol{c},\boldsymbol{t}} \tilde{J}(\boldsymbol{c}, \boldsymbol{t}) = \arg\min_{\boldsymbol{c},\boldsymbol{t}} \mathrm{E}\left\{ e^2_{\boldsymbol{c},\boldsymbol{t}}(t) \right\},$$

where $\mathrm{E}\{\cdot\}$ denotes the expectation. The optimal value for $\boldsymbol{c}$ is $\boldsymbol{c}^*(\boldsymbol{t}) = \tilde{\mathbf{R}}^{-1}(\boldsymbol{t})\boldsymbol{r}(\boldsymbol{t})$, with:

$$\tilde{\mathbf{R}}(\boldsymbol{t}) = \begin{pmatrix} R_{2,2}(0) & R_{2,3}(t_{1,3} - t_{1,2}) & \cdots \\ R_{2,3}(t_{1,3} - t_{1,2}) & R_{3,3}(0) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

and

$$\boldsymbol{r}(\boldsymbol{t}) = \begin{pmatrix} R_{1,2}(-t_{1,2}) & R_{1,3}(-t_{1,3}) & \ldots \end{pmatrix}^{\mathrm{t}},$$

where we denoted by $R_{i,j}(\tau) = \mathrm{E}\{x_i(t)x_j(t-\tau)\}$ the cross-correlation functions. We assume that the direct propagation path is dominant with respect to the reverberant paths. Hence, $R_{i,j}(\tau)$ has its maxima at the true value of $t_{i,j}$. By setting $\boldsymbol{c} = \boldsymbol{c}^*$, the optimization problem becomes:

$$\boldsymbol{t}^* = \arg\min_{\boldsymbol{t}} \tilde{J}(\boldsymbol{c}^*(\boldsymbol{t}), \boldsymbol{t}) = \arg\min_{\boldsymbol{t}} \left\{ R_{1,1}(0) - \boldsymbol{r}^{\mathrm{t}}(\boldsymbol{t})\tilde{\mathbf{R}}^{-1}(\boldsymbol{t})\boldsymbol{r}(\boldsymbol{t}) \right\}. \quad (5.18)$$

An equivalent criterion

In addition, it can be shown (see Section 5.A) that this optimization problem is equivalent to the following one:

$$\boldsymbol{t}^* = \arg\min_{\boldsymbol{t}} J(\boldsymbol{t}), \quad (5.19)$$

where $J(\boldsymbol{t}) = \det(\mathbf{R}(\boldsymbol{t}))$ with $\mathbf{R}(\boldsymbol{t}) \in \mathbb{R}^{M\times M}$ being the real matrix of normalized cross-correlation functions evaluated at $\boldsymbol{t}$. That is $\mathbf{R}(\boldsymbol{t}) = [\rho_{i,j}(\boldsymbol{t})]_{ij}$ with:

$$\rho_{i,j}(\boldsymbol{t}) = \frac{\mathrm{E}\{x_i(t+t_{1,i})x_j(t+t_{1,j})\}}{\sqrt{E_i E_j}},$$

where $E_i = R_{i,i}(0) = \mathrm{E}\left\{x_i^2(t)\right\}$ is the energy of the i^{th} signal. This is how the time delay estimation problem is cast into a non-linear optimization problem. The problem is multivariate due to the fact that the signal model does not encode the geometry of the array, as in [Chen 03b]. Moreover, we could use the feasibility constraints derived in the previous section to constrain the minimization.

5.5.2 The Non-linear Constrained Optimization Problem

The optimization task

In Section 5.4 we characterized the *feasible* values of $\boldsymbol{t}$ (i.e., those corresponding to a sound source position) and in the previous section we introduced a criterion to choose the best value for $\hat{\boldsymbol{t}}$. Of course, the next step is to look for the best value among the feasible ones. We call this operation geometrically-constrained time delay estimation and it is naturally cast into the following *non-linear constrained optimization* problem:

$$\begin{cases} \min\limits_{\boldsymbol{t}} J(\boldsymbol{t}), \\ \text{s.t.} \quad \boldsymbol{t} \in \mathcal{W}, \quad -\boldsymbol{t}^* \leq \boldsymbol{t} \leq \boldsymbol{t}^*, \\ \qquad \Delta\left(\boldsymbol{t}\right) \geq 0, \quad E\left(\boldsymbol{t}\right) = 0, \end{cases} \tag{5.20}$$

where $\mathcal{W}$, $\boldsymbol{t}^*$, Δ and E are defined in Section 5.4. This new formulation was first published in [Alameda-Pineda 12a] and then used again in [Alameda-Pineda 13b].

5.6 Local Optimization

Local optimization using a log-barrier dual interior point method

The third contribution is a local optimization method with grid-based initialisation to estimate the time delays. Here, the minimization of (5.20) is carried out using a publicly available MATLAB implementation [Carbonetto 08] of the *log-barrier dual interior point* (DIP) method [Boyd 04]. This method is designed for continuous convex optimization problems with non-linear constraints. Indeed, the inequality constraint of the problem, $\Delta(\boldsymbol{t}) \geq 0$, is added to the cost by means of a log-barrier function. The optimization problem in (5.20) is converted into a sequence of problems indexed by a real parameter $\mu \geq 0$:

$$\begin{cases} \min\limits_{\boldsymbol{t}} J(\boldsymbol{t}) - \mu \log(\Delta(\boldsymbol{t})), \\ \text{s.t.} \quad \boldsymbol{t} \in \mathcal{W}, \quad -\boldsymbol{t}^* \leq \boldsymbol{t} \leq \boldsymbol{t}^*, \quad E\left(\boldsymbol{t}\right) = 0. \end{cases} \tag{5.21}$$

The solutions of these problems form a sequence of optimal solutions $\hat{\boldsymbol{t}}_\mu$ also indexed by μ. Moreover, $\hat{\boldsymbol{t}}_\mu \rightarrow \hat{\boldsymbol{t}}$ when $\mu \rightarrow 0$. In other words, the solution of (5.21) is close enough to the solution of (5.20) for a value of μ small enough. Once the inequality is included in the cost function, a gradient-based optimization method is applied and the parameter μ of the algorithms decreases with the iterations. In order to increase the convergence speed and the accuracy, we computed the analytical gradient and Hessian of the original cost function J and constraint function Δ. This can be found in Sections 5.B and 5.C respectively. As any gradient-based technique, the DIP method is likely to fail in finding the global optimum of non-convex problems such as (5.21). To overcome this issue, our algorithm starts from several initial points, i.e., the set $\mathcal{S}^I = \{\boldsymbol{t}_i^I\}_{i=1}^P$. For each one of these initializations a local minimum is found, then the minimum over these local minima is selected.

Grid-based global optimization

5.7 Global Optimization

Branch & bound for global optimization of Lipschitz functions

In this section we present our last contribution to the field: a global optimization technique to solve (5.20). If the functions $\rho_{i,j}$ are continuously differentiable, the cost function J is Lipschitz continuous in the compact set $-\boldsymbol{t}^* \leq \boldsymbol{t} \leq \boldsymbol{t}^*$, and hence a branch and bound (B&B) global optimization algorithm is appropriate. The skeleton of the B&B method is shown in Algorithm 5.1. The input is the Lipschitz constant L, and a list $\mathcal{I} = \{(\boldsymbol{t}^{(i)}, s^{(i)})\}_{i=1}^{I}$ of initial sets, where $(\boldsymbol{t}^{(i)}, s^{(i)})$ represents the cube of center $\boldsymbol{t}^{(i)}$ and side $2s^{(i)}$. The Branch and Bound routines are alternated until convergence. While the Branch method splits the sets in $\mathcal{I}$ into cubes of side $s^{(i)}$ (half of the original size), the Bound method bound the cost function on the recently created cubes. The Bound routine is shown in Algorithm 5.2. The upper and lower bound of all the sets in $\mathcal{I}$ are computed. Those sets whose lower bound is bigger than the minimum of the upper bounds are discarded, since the optimum cannot lie inside those sets. In order to stop the iterative procedure we could use at least two possible convergence criteria: the size of the sets in $\mathcal{I}$ (as a precision for the solution) and the variation of the minimum cost function during the algorithm. After convergence, we choose the set in $\mathcal{I}$ with minimum cost among those satisfying the constraint. If no such set exists in $\mathcal{I}$, we rerun the B&B algorithm on $\mathcal{O}$.

Algorithm 5.1 Branch and Bound

Input: The Lipschitz constant L and a list of sets $\mathcal{I}$
Output: A list of potential solutions $\mathcal{I}$ and a list of discarded solutions $\mathcal{O}$.
repeat
  (a) $\mathcal{I}$ = Branch($\mathcal{I}$)
  (b) $[\mathcal{I}, \mathcal{U}]$ = Bound($\mathcal{I}$,L)
  (c) $\mathcal{O} = \mathcal{O} \cup \mathcal{U}$
until Convergence

Algorithm 5.2 Bound subroutine of Algorithm 5.1

Input: The Lipschitz constant L and a list of sets $\mathcal{I}$
Output: A list of currently inlier sets $\mathcal{I}$ and a list of outlier sets $\mathcal{U}$.
for $(\boldsymbol{t}^{(i)}, s^{(i)}) \in \mathcal{I}$ **do**
  (a) $l^{(i)} = J(\boldsymbol{t}^{(i)}) - s^{(i)}L$.
  (b) $u^{(i)} = J(\boldsymbol{t}^{(i)}) + s^{(i)}L$.
  (c) $\tau = \min_{i=1,\ldots,|\mathcal{I}|} u^{(i)}$.
end for
for $i = 1, \ldots, |\mathcal{I}|$ **do**
  if $l^{(i)} > \tau$ **then**
    Move $(\boldsymbol{t}^{(i)}, s^{(i)})$ from $\mathcal{I}$ to $\mathcal{U}$.
  end if
end for

Table 5.1: Results obtained on simulated data with $SNR = 0$ dB. The second column corresponds to the values of T_{60} in seconds. The six remaining columns correspond to each of the evaluated methods. For each combination SNR, T_{60} and method there are three values: the proportion of inliers (angular error < 30 degrees), the inlier angular error mean and standard deviation.

SNR	T_{60}	*pi-tde*	*tde*	*i-gtde*	*fg-gtde*	*sg-gtde*	*bb-gtde*
0	0.0	53.7%	38.3%	80.3%	75.3%	46.9%	82.1%
		11.31	15.89	15.75	10.54	11.63	9.59
		5.55	7.47	7.11	4.57	5.54	3.66
	0.1	53.3%	36.2%	77.4%	75.4%	46.7%	82.8%
		12.47	16.15	15.94	11.55	12.58	10.49
		6.17	7.30	7.18	5.26	6.17	4.47
	0.2	44.3%	33.3%	62.4%	67.5%	40.9%	73.8%
		14.60	17.01	16.49	13.54	14.79	12.65
		6.92	7.46	7.38	6.51	6.98	6.14
	0.4	30.3%	23.3%	41.2%	44.6%	27.9%	48.3%
		16.81	17.94	17.04	15.53	17.05	14.99
		7.33	7.30	7.50	7.21	7.33	7.09
	0.6	23.6%	19.2%	30.2%	33.4%	22.2%	35.7%
		17.67	18.60	17.76	16.25	17.67	16.10
		7.60	7.69	7.48	7.22	7.13	7.30

5.8 Results

Evaluation protocol

In order to accurately validate the two optimization algorithms, we developed a formal evaluation protocol using simulated and real data. The set up is the same in both cases: a $4 \times 4 \times 4$ meter room with an array of four microphones at (in meters) $\boldsymbol{M}_1 = (2.0, 2.1, 1.83)^{\mathrm{t}}$, $\boldsymbol{M}_2 = (1.8, 2.1, 1.83)^{\mathrm{t}}$, $\boldsymbol{M}_3 = (1.9, 2.2, 1.97)^{\mathrm{t}}$ and $\boldsymbol{M}_4 = (1.9, 2.0, 1.97)^{\mathrm{t}}$ and the sound source at 189 different positions on a 1.7 m radius sphere around the microphones. The source emitted speech fragments randomly chosen from [Garofolo 93]. One hundred millisecond cuts of these sounds are the input of the evaluated methods. In the simulated case, we control two parameters. First, the SNR, regulating the amount of noise added to the received signals, and taking the following values (in dB): -10, -5 and 0. Secondly the T_{60}, used in the Image-Source Model [Lehmann 08] (available at [Lehmann 12]) to control the amount of reverberations, taking the following values (in s): 0, 0.1, 0.2, 0.4 and 0.6. In the real case, we used the acquisition protocol defined in [Deleforge 12b], replacing the dummy head by the tetrahedron microphone array.

Several algorithms are compared. *pi-tde* is a pair-wise independent time delay estimation by maximizing the cross-correlation function. *tde* corresponds to the time delay estimation based on unconstrained optimization of (5.19), that is the generalisation of the method in [Chen 03b] to arbitrarly-shaped microphone arrays. The method is initialized on a uniform grid, $\mathcal{S}^I$ of $P = 4096$ points. The bounds of the grid are defined by the geometry of the proble (see Lemma 1). *i-gtde* consists on taking the minimum of the cost function at the subset of feasible

Table 5.2: Results obtained on simulated data with $SNR = -5$ dB. The columns rows have the same meaning as in Table 5.1

SNR	T_{60}	*pi-tde*	*tde*	*i-gtde*	*fg-gtde*	*sg-gtde*	*bb-gtde*
-5	0.0	41.4%	37.5%	78.3%	80.4%	39.3%	84.1%
		13.24	16.76	15.80	11.74	13.41	10.46
		6.11	7.38	7.12	5.52	6.41	4.64
	0.1	41.5%	37.0%	74.0%	77.9%	40.8%	82.7%
		14.23	16.92	16.09	12.99	14.58	11.58
		6.71	7.36	7.15	6.35	6.74	5.45
	0.2	32.7%	31.5%	57.7%	61.9%	34.4%	68.6%
		16.50	17.71	16.48	14.74	16.60	13.91
		7.09	7.41	7.35	6.91	7.21	6.75
	0.4	21.9%	21.4%	34.8%	36.8%	23.6%	41.1%
		18.12	18.48	17.53	16.54	17.76	16.07
		7.15	7.30	7.55	7.63	7.33	7.44
	0.6	16.9%	16.7%	26.9%	28.0%	18.7%	29.8%
		18.08	18.74	17.82	17.11	18.19	16.97
		7.43	7.26	7.45	7.38	7.28	7.35

poitns of $\mathcal{S}^I$, that we will denote $\mathcal{S}^F$. *fg-gtde* consists on the log-barrier method described in Section 5.6, initialized on $\mathcal{S}^F$. *sg-gtde* is the very same log-barrier method, intialized on a sparse grid. We conjecture that the global minimum of J corresponds to local maxima of the functions $\rho_{1,m}$. Thus, for each microphone pair $(1, m)$, we extract K local maxima of $\rho_{1,m}$ to construct a grid with all possible combinations of these values, ending up with K^{M-1} points. *bb-gtde* is the B&B algorithm described in Section 5.7, initialized with a big cube covering the feasible domain. All these algorithms provide a time delay estimate, $\hat{\boldsymbol{t}}$, used to retrieve the sound-source position using the localisation mapping (5.15).

Results on simulated data

Tables 5.1, 5.2 and 5.3 show the localisation results on simulated data for different values of SNR. The first two columns correspond to the value of SNR and of T_{60} respectively. The six other columns correspond to the evaluation methods described in the previous paragraph. For each combination SNR-T_{60}-Method, three quantities are given, namely: the percentage of localisation inliers (angular error less than 30^0), the angular error mean of inliers, and their standard deviation (both in degrees). We first observe that all methods behave as expected when increasing the level of noise and reverberations. Indeed, their performance strictly decreases with the amount of noise and reverberations. Secondly, we notive that the *tde* method has very bad performance. In other words, using the generalisation of the criterion derived by [Chen 03b] without any additional information about the microphone array gives very bad results, even in easy scenarios. Thirdly, we notice that mtehods *pi-tde* and *sg-gtde* have comparable results. The differences are found in the percentage of localisation inliers. While in easy conditions of noise and reverberations the independent estimations works better, in harder conditions the performance of *pi-tde* decreases much faster than *sg-gtde*. Fourthly, the relative situation of *i-gtde* and *fg-gtde* is quite similar as the

Table 5.3: Results obtained on simulated data with $SNR = -10$ dB. The columns and rows have the same meaning as in Table 5.1.

SNR	T_{60}	*pi-tde*	*tde*	*i-gtde*	*fg-gtde*	*sg-gtde*	*bb-gtde*
-10	0.0	29.6%	33.4%	60.4%	66.6%	31.0%	77.5%
		17.04	17.28	16.65	14.69	17.13	13.45
		7.19	7.51	7.30	6.90	7.36	6.56
	0.1	28.9%	29.2%	51.3%	56.5%	29.7%	66.6%
		17.76	17.75	16.90	16.01	17.85	14.35
		7.17	7.62	7.31	7.19	7.31	6.85
	0.2	20.8%	21.6%	35.6%	36.3%	22.1%	44.5%
		18.92	18.67	17.86	17.27	18.46	16.53
		7.49	7.34	7.33	7.37	7.00	7.36
	0.4	14.7%	14.3%	21.4%	20.6%	14.5%	24.8%
		18.98	18.95	18.76	18.28	19.17	18.29
		7.58	7.21	7.29	7.30	7.41	7.29
	0.6	12.5%	11.6%	16.3%	15.8%	13.2%	19.0%
		19.25	19.54	19.03	18.61	19.65	18.63
		7.22	7.03	7.34	7.20	7.26	7.26

Table 5.4: Results obtained on real data. The rows have the same meaning as in Table 5.1.

pi-tde	*tde*	*i-gtde*	*fg-gtde*	*sg-gtde*	*bb-gtde*
13.24%	13.37%	16.69%	27.48%	12.86%	22.28%
18.80	18.50	19.34	16.04	19.03	17.51
7.09	7.15	7.00	7.55	6.88	7.53

previous case. Counterintuitively, *i-gtde* performs better than *fg-gtde* in easy scenarios. However, we observe that the performance of *i-gtde* decreases faster thant the performance of *fg-gtde*. Thus the robustness of *fg-gtde* to noise and reverberations is higher than the robustness of *i-gtde*. Finally we remark that, both, the performance and the robustness of *bb-gtde* are higher than the performance and robustness of any other of the tested methods. Generally speaking, we could say that methods involving the constraints derived from the propagation model are more robust than the others.

Results on real data

Table 5.4 presents the results on real data. The rows have the same meaning than in Table 5.1. In this case we observe that only good global optimization techniques, i.e., *fg-gtde* and *bb-gtde*, show decent results. Indeed, they both perform much better than the rest. We also remark that, contrarily to the simulated case, in the real scenario *fg-gtde* outperforms *bb-gtde*. Last we notice that the results on real data roughly correspond to the simulated case with $T_{60} = 0.6$ s and $SNR = -5$ dB, which is a very challenging scenario.

Effects of the noise and the reverberations

In general, the methods perform as expected with respect to the environmental conditions. That is, the higher the SNR value the better the methods estimate the time delays, the higher the percentage of inliers and the lower the localisation error. We can also observe a clear trend with respect to the reverberation level: the methods' performance decreases with T_{60}. However the SNR and the

T_{60} have different effects on the function to minimize. On one side, the sensor noise decorrelates the microphones' signals leading to much more (and randomly spread) local minima and increasing the value of the true minimum. If this effect is extreme, the hope for a good estimate decreases fast. On the other side, the reverberations produce only a few strong local minima. This perturbation is systematic given the source position in the room. Hence, there is hope to learn the effect of such reverberations in order to improve the quality of the estimates. These types of perturbations (noise and reverberations) of the function to minimize have clearly different effects on the results.

5.9 Conclusions and Future Work

In this chapter, we addressed the problem of time delay estimation for sound source localisation using non-coplanar microphone arrays. The starting point is the signal and propagation models, from which we are able to analyze the geometry of the problem. This analysis describes the feasible values of the time delays, i.e., those that correspond to a position in the source space. Cast into an optimization problem, the time delay estimation is subject to the feasibility conditions. On one hand, they represent a geometric characterization and on the other hand they provide for two explicit mappings constraining the optimization problems. Furthermore, two different approaches are proposed to solve the TDE and sound source localisation tasks. These approaches are evaluated using simulated data and data recorded in a natural indoor environment. From the extensive experiments on both simulated and real data, we conclude that both methods outperform the state-of-the-art, thus validating the geometric model as well as the optimization procedures.

The are several ways to extend this work. As outlined before, it would be very useful to learn the effect the reverberations have on the objective function as in [Ribeiro 10]. Also, it is worth to consider the multiple source case, following approaches like [Chen 03a]. Besides that, a frequency decomposition stage may be useful to avoid the analysis in non-informative frequency bands [Valin 06]. Thirdly, by evaluating the model in the framework of dynamic sound sources. Fourth, adapting the methodology into a calibration task, where the position of the sound source may be known, but not the microphones' position. Finally, performing experiments using a large number of microphones and evaluating the influence of their positions.

5.A Criteria Equivalence

In Section 5.5 we stated that the optimization criteria from Equations (5.18) and (5.19) are equivalent. We recall their expressions:

$$\tilde{J}(\boldsymbol{t}) = R_{1,1}(0) - \boldsymbol{r}^{\mathrm{t}}(\boldsymbol{t})\tilde{\mathbf{R}}^{-1}(\boldsymbol{t})\boldsymbol{r}(\boldsymbol{t}) \qquad J(\boldsymbol{t}) = \det\left(\mathbf{R}(\boldsymbol{t})\right).$$

In order to prove this statement, we will start from the expression of $\tilde{J}$ and recover the expression of J. Notice that we can rewrite $\tilde{J}$ as:

$$J = \det\left(\mathbf{R}\right) = \det\left(\tilde{\mathbf{R}}\right) - \sum_{i=2}^{M}(-1)^{i+1}\rho_{1,i}\det\left(\tilde{\mathbf{R}}_i\right),$$

where

$$\tilde{\mathbf{R}}_i = \begin{pmatrix} \boldsymbol{r} & \tilde{\boldsymbol{R}}_1 & \cdots & \tilde{\boldsymbol{R}}_{i-1} & \tilde{\boldsymbol{R}}_{i+1} & \cdots & \tilde{\boldsymbol{R}}_M \end{pmatrix}.$$

and $\tilde{\boldsymbol{R}}_i$ is the i^{th} column of the matrix $\tilde{\mathbf{R}}$. The reader should make the difference between the matrix $\tilde{\mathbf{R}}_i$ and the vector $\tilde{\boldsymbol{R}}_i$. We can further develop $\det\left(\tilde{\mathbf{R}}_i\right)$:

$$\det\left(\tilde{\mathbf{R}}_i\right) = \sum_{j=2}^{M}(-1)^{j+1}\rho_{1,j}\det(\tilde{\mathbf{R}}_{i,j}),$$

where $\det(\tilde{\mathbf{R}}_{i,j})$ is the ij^{th} minor of $\tilde{\mathbf{R}}$ (since $\tilde{\mathbf{R}}$ is symmetric). Finally,

$$\begin{aligned}
\det\left(\mathbf{R}\right) &= \det\left(\tilde{\mathbf{R}}\right) - \sum_{i,j=2}^{M}(-1)^{i+j}\rho_{1,i}\rho_{1,j}\det\left(\tilde{\mathbf{R}}_{i,j}\right) \\
&= \det\left(\tilde{\mathbf{R}}\right)\left(1 - \sum_{i,j=2}^{M}(-1)^{i+j}\rho_{1,i}\rho_{1,j}\frac{\det\left(\tilde{\mathbf{R}}_{i,j}\right)}{\det\left(\tilde{\mathbf{R}}\right)}\right) \\
&= \det\left(\tilde{\mathbf{R}}\right)\left(1 - \sum_{i,j=2}^{M}(-1)^{i+j}R_{1,i}R_{1,j}\frac{\det\left(\tilde{\mathbf{R}}_{i,j}\right)}{\det\left(\tilde{\mathbf{R}}\right)E_1\sqrt{E_i}\sqrt{E_j}}\right) \\
&= \frac{\det\left(\tilde{\mathbf{R}}\right)}{E_1}\left(E_1 - \sum_{i,j=2}^{M}(-1)^{i+j}R_{1,i}\left(\tilde{\mathbf{R}}^{-1}\right)_{i,j}R_{1,j}\right) \\
&= \frac{\det\left(\tilde{\mathbf{R}}\right)}{E_1}\left(E_1 - \boldsymbol{r}^{\mathrm{t}}\tilde{\mathbf{R}}^{-1}\boldsymbol{r}\right) \\
\Rightarrow J &= \frac{\det\left(\tilde{\mathbf{R}}\right)}{E_1}\tilde{J}
\end{aligned}$$

It is worth to notice that both criteria are equivalent if and only if the quantity $\det(\mathbf{R})/E_1$ is well defined and strictly positive. Since $E_1 > 0$ (because of the

sensor's noise), this is equivalent to say that both criteria are equivalent if and only if $\det(\mathbf{R}) > 0$. We refer the reader to [Chen 03b], since the positiveness of that determinant is proven there.

5.B The Derivatives of the Cost Function

The interior point algorithm relies on the use of the gradient and the Hessian of both, the objective function and the constraint(s). Providing the analytic expression for them would lead to a much more efficient and precise algorithm than estimating them using finite differences. Hence, this section is devoted to the derivation of both, the gradient and the Hessian of the criterion.

In order to do that we need three laws of matrix calculus. Let $\mathbf{Y} : \mathbb{R} \rightarrow \mathbb{R}^{M\times M}$, be a matrix function depending on y, the following formulas hold:

- $\dfrac{\partial \det(\mathbf{Y}(y))}{\partial y} = \det(\mathbf{Y}(y))\text{trace}\left(\mathbf{Y}(y)^{-1}\dfrac{\partial \mathbf{Y}(y)}{\partial y}\right)$
- $\dfrac{\partial \text{trace}\ (\mathbf{Y}(y))}{\partial y} = \text{trace}\left(\dfrac{\partial \mathbf{Y}(y)}{\partial y}\right)$
- $\dfrac{\partial \mathbf{Y}(y)^{-1}}{\partial y} = -\mathbf{Y}(y)^{-1}\dfrac{\partial \mathbf{Y}(y)}{\partial y}\mathbf{Y}(y)^{-1}$

Recall that the function we want to derivative is $J = \det(\mathbf{R})$. From the rules of matrix calculus we have:

$$\frac{\partial J}{\partial t_{1,k}} = \frac{\partial \det(\mathbf{R})}{\partial t_{1,k}} = \det(\mathbf{R})\,\text{trace}\left(\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial t_{1,k}}\right). \tag{5.22}$$

In addition we can compute:

$$\begin{aligned}\frac{\partial^2 \tilde{J}}{\partial t_{1,j}\partial t_{1,k}} &= \frac{\partial}{\partial t_{1,j}}\left(\det(\mathbf{R})\,\text{trace}\left(\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial t_{1,k}}\right)\right)\\ &= \det(\mathbf{R})\,\text{trace}\left(\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial t_{1,j}}\right)\text{trace}\left(\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial t_{1,k}}\right) +\\ &\quad \det(\mathbf{R})\,\text{trace}\left(-\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial t_{1,j}}\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial t_{1,k}} + \mathbf{R}^{-1}\frac{\partial^2 \mathbf{R}}{\partial t_{1,j}\partial t_{1,k}}\right).\end{aligned}$$

Hence, to be able to evaluate the gradient and the Hessian of $\tilde{J}$ we need to compute the first and second derivatives of the matrix $\mathbf{R}$. For clarity purposes, we first rewrite the derivatives of $\mathbf{R}$ in terms of the derivatives of $\tilde{\mathbf{R}}$ and $\boldsymbol{r}$ to finally compute these last ones. Notice that:

$$\mathbf{R} = \mathbf{D}\left(\begin{array}{c|c} E_1 & \boldsymbol{r}^{\mathrm{t}} \\ \hline \boldsymbol{r} & \tilde{\mathbf{R}} \end{array}\right)\mathbf{D}, \tag{5.23}$$

where $\mathbf{D} = \operatorname{diag}\left(E_1^{-1/2}, \ldots, E_M^{-1/2}\right)$ is a diagonal matrix containing the square roots of the signals' energy $E_i = R_{i,i}(0)$. Since the matrix $\mathbf{D}$ does not depend on $\boldsymbol{t}$, the derivatives of $\mathbf{R}$ look like:

$$\frac{\partial \mathbf{R}}{\partial t_{1,k}} = \mathbf{D} \left(\begin{array}{c|c} 0 & \frac{\partial \boldsymbol{r}}{\partial t_{1,k}}^{\mathrm{t}} \\ \hline \frac{\partial \boldsymbol{r}}{\partial t_{1,k}} & \frac{\partial \tilde{\mathbf{R}}}{\partial t_{1,k}} \end{array} \right) \mathbf{D}, \tag{5.24}$$

and

$$\frac{\partial^2 \mathbf{R}}{\partial t_{1,j} \partial t_{1,k}} = \mathbf{D} \left(\begin{array}{c|c} 0 & \frac{\partial^2 \boldsymbol{r}}{\partial t_{1,j} \partial t_{1,k}}^{\mathrm{t}} \\ \hline \frac{\partial^2 \boldsymbol{r}}{\partial t_{1,j} \partial t_{1,k}} & \frac{\partial^2 \tilde{\mathbf{R}}}{\partial t_{1,j} \partial t_{1,k}} \end{array} \right) \mathbf{D}. \tag{5.25}$$

The partial derivative of $\tilde{\mathbf{R}}$ is matrix filled with zeros expect for its k^{th} row and its k^{th} column that are equal to the following vector:

$$\left(\ldots, R'_{k-1,k}(t_{1,k} - t_{1,k-1}), 0, -R'_{k+1,k}(t_{1,k+1} - t_{1,k}), \ldots\right)^{\mathrm{t}}.$$

The partial derivative of $\boldsymbol{r}$ is:

$$\frac{\partial \boldsymbol{r}}{\partial t_{1,k}} = \left(0, \ldots, R'_{1,k}(t_{1,k}), \ldots, 0\right)^{\mathrm{t}}. \tag{5.26}$$

We will differentiate two cases when computing the second derivative:

$j = k$ This will fill the diagonal of the Hessian matrix. Notice that:

$$\frac{\partial^2 \boldsymbol{r}}{\partial t_{1,k}^2} = \left(0, \ldots, R''_{1,k}(t_{1,k}), 0, \ldots\right)^{\mathrm{t}} \tag{5.27}$$

and that the partial second derivative of $\tilde{\mathbf{R}}$ is matrix filled with zeros expect for its k^{th} row and its k^{th} column that are equal to the following vector:

$$\left(\ldots, R''_{k-1,k}(t_{1,k} - t_{1,k-1}), 0, R''_{k+1,k}(t_{1,k+1} - t_{1,k}), \ldots\right)^{\mathrm{t}}.$$

$j > k$ This fills the lower triangular matrix of the Hessian (and the upper triangular since we assume that the hessian is symmetric, i.e., that J is twice continuously differentiable). The second derivative of $\boldsymbol{r}$ is null in this case, however the second derivative of $\tilde{\mathbf{R}}$ is not. Actually just two positions in the second derivative are not necessarily null: the jk^{th} and the kj^{th} being $-R''_{k,j}(t_{1,k} - t_{1,j})$.

5.C The Derivatives of the Constraint

As we have done in the previous Section for the criterion, in this section we compute the formulae for the first and the second derivatives of the non-linear constraint Δ. Recall the expression from (5.14):

$$\Delta = \langle \boldsymbol{A}, \boldsymbol{B} - \boldsymbol{M}_1 \rangle^2 - \|\boldsymbol{B} - \boldsymbol{M}_1\|^2 \left(\|\boldsymbol{A}\|^2 - 1 \right),$$

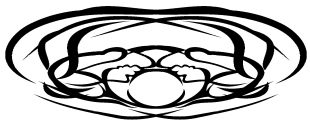
where $\boldsymbol{A} = -2\nu \mathbf{M}^{\ddagger} \boldsymbol{t}$ and $\boldsymbol{B} = \mathbf{M}^{\ddagger} \left(\boldsymbol{K} - \nu^2 \boldsymbol{t}^2 \right)$. It is easy to show that:

$$\begin{aligned} \nabla\Delta &= 2 \Big(\langle \boldsymbol{A}, \boldsymbol{B} - \boldsymbol{M}_1 \rangle \left(\mathbf{J}^{\mathrm{t}}_{\boldsymbol{A}} (\boldsymbol{B} - \boldsymbol{M}_1) + \mathbf{J}^{\mathrm{t}}_{\boldsymbol{B}} \boldsymbol{A} \right) - \\ &- (\|\boldsymbol{A}\|^2 - 1) \mathbf{J}^{\mathrm{t}}_{\boldsymbol{B}} (\boldsymbol{B} - \boldsymbol{M}_1) - \|\boldsymbol{B} - \boldsymbol{M}_1\|^2 \mathbf{J}^{\mathrm{t}}_{\boldsymbol{A}} \boldsymbol{A} \Big) \end{aligned} \tag{5.28}$$

where $\mathbf{J}_{\boldsymbol{A}} = -2\nu \mathbf{M}^{\ddagger}$ and $\mathbf{J}_{\boldsymbol{B}} = -2\nu^2 \mathbf{M}^{\ddagger} \operatorname{diag}(\boldsymbol{t})$. We can also compute the Hessian of Δ:

$$\begin{aligned} \mathbf{H}\Delta &= 2 \Bigg(\left(\mathbf{J}^{\mathrm{t}}_{\boldsymbol{A}} (\boldsymbol{B} - \boldsymbol{M}_1) + \mathbf{J}^{\mathrm{t}}_{\boldsymbol{B}} \boldsymbol{A} \right) \left(\mathbf{J}^{\mathrm{t}}_{\boldsymbol{A}} (\boldsymbol{B} - \boldsymbol{M}_1) + \mathbf{J}^{\mathrm{t}}_{\boldsymbol{B}} \boldsymbol{A} \right)^{\mathrm{t}} + \\ &+ \langle \boldsymbol{A}, \boldsymbol{B} - \boldsymbol{M}_1 \rangle \left(\mathbf{J}^{\mathrm{t}}_{\boldsymbol{A}} \mathbf{J}_{\boldsymbol{B}} + \mathbf{D} + \mathbf{J}^{\mathrm{t}}_{\boldsymbol{B}} \mathbf{J}_{\boldsymbol{A}} \right) - \\ &- \Big[2(\mathbf{J}^{\mathrm{t}}_{\boldsymbol{B}} (\boldsymbol{B} - \boldsymbol{M}_1))(\mathbf{J}^{\mathrm{t}}_{\boldsymbol{A}} \boldsymbol{A})^{\mathrm{t}} + (\|\boldsymbol{A}\|^2 - 1)(\mathbf{E} + \mathbf{J}^{\mathrm{t}}_{\boldsymbol{B}} \mathbf{J}_{\boldsymbol{B}}) + \\ &+ 2(\mathbf{J}^{\mathrm{t}}_{\boldsymbol{A}} \boldsymbol{A})(\mathbf{J}^{\mathrm{t}}_{\boldsymbol{B}} (\boldsymbol{B} - \boldsymbol{M}_1))^{\mathrm{t}} + \|\boldsymbol{B} - \boldsymbol{M}_1\|^2 \mathbf{J}^{\mathrm{t}}_{\boldsymbol{A}} \mathbf{J}_{\boldsymbol{A}} \Big] \Bigg) \end{aligned} \tag{5.29}$$

where $\mathbf{D} = -2\nu^2 \operatorname{Diag}(\mathbf{M}^{\dagger} \boldsymbol{A})$ and $\mathbf{E} = -2\nu^2 \operatorname{Diag}(\mathbf{M}^{\dagger} (\boldsymbol{B} - \boldsymbol{M}_1))$.



CHAPTER 6

CONCLUSIONS

This PhD was devoted to collect better insights in the processing of auditory and visual signals acquired by means of an egocentric set of sensors for the purpose of scene understanding. Among the different addressable tasks in that broad field, we chose to work in three, namely: audio-visual speaker detection, audio-visual gesture recognition and multichannel sound source localisation. This is challenging because the data is (D1) egocentric, that is, acquired with a sensor network fitting in a small volume, thus providing very similar points of view of the scene, (D2) audio-visual, that is coming from two different modalities, hence requiring fusion strategies able to extract meaningful objects from the data while exploiting the complementarity inherent to the use of two modalities and (D3) corrupted, because unrestricted environments are characterized by visual occlusions, auditory interferences and reverberations and sensor noise. Moreover, when developing robotic applications, the methods should be (M1) efficient, such that the limited resources of the platform are not misused, (M2) fast, ensuring that the output corresponds to the ongoing social interaction, (M3) robust, that is not easily perturbed by noise and interfering artefacts, (M4) adaptable, thus guaranteeing the wide usability of the system, and (M5) reliable, such that higher-level applications can build on them. Similarly, we desire the outcome of such algorithms to be (O1) temporally coherent, providing results that are stable over time, (O2) spatially consistent, to ensure the correct management of the robot's space and (O3) semantically meaningful, such that the resulting interaction is natural and smooth.

6.1 The RAVEL Data Set

In order to properly evaluate the developed methods, we recorded the RAVEL data set, consisting on three categories: action recognition, gesture recognition and interaction. The scenarios on these categories are designed to study action and gesture recognition, localization of auditory and visual events, dialogue handling,

gender and face detection, and identity recognition. RAVEL was recorded with the POPEYE robot, providing a stereo-image flow and a four-channel sound track. The environment set up and recording device are fully detailed ensuring the repeatability of the recordings. Likewise, the scenarios are outlined and their their script is provided. Finally, application examples are shown proving that RAVEL is useful for testing methods on action recognition, scene flow extraction and audio-visual speaker detection.

6.2 Vision-Guided Speaker Detection

The first task addressed is the detection and localisation of speakers using audio-visual data. Fusing audition and vision is challenging regarding this application because data may be corrupted. Occlusions, noise and reverberations present in regular indoor environments make the detection task very difficult. We developed a hybrid probabilistic/deterministic framework **AVS1** to perform multimodal fusion. On one hand, the deterministic components allow us to model those characteristics of the scene that are known with precision in advance. On the other hand, the probabilistic components model random effects, such as noise and outliers.

Through this framework we showed how vision can guide audition leading to vision-guided speaker detection algorithms. Indeed, an EM-based procedure named Motion-Guided Robot Hearing **AVS2** is used to associate sound-emitting directional features to motion-related directional features. Consequently, active sound regions are associated to moving regions creating AV objects. In a later stage, we implemented this on the humanoid robot NAO. To do that, we simplified the EM procedure by using a face detector, thus resulting in a Face-Guided Robot Hearing algorithm **AVS3**. The proposed model was validated using synthetic data and the derived algorithms were evaluated using real data.

The real data used for evaluation satisfies the three properties (D1), (D2) and (D3), as explained in Chapter 3, **AVS2** is (M1) efficient, (M3) robust, (M4) highly adaptable and (M5) reliable. Its real-time implementation is much (M2) faster and equally (M1) efficient, (M3) robust and (M4) adaptable. However, **AVS3** is not as reliable as **AVS2**, because its performance depends on the one of the face detector. None of the presented algorithms ensure the (O1) temporal coherence of their results, since no tracking is involved. Albeit, both offer (O2) spatially consistent, because of the localisation capabilities of the algorithms directly provided by the 3D stereo reconstruction and (O3) semantically meaningful, outputting motion-emitting regions and speaking faces respectively.

6.3 Audio-Visual Command Recognition

We also addressed the task of audio-visual command recognition. These commands are gestures accompanied by a short sentence or by a word. This problem

is challenging because (i) there is no standard audio-visual representation and (ii) these commands are culture-, language- and user-dependant.

First, we evaluate different features and representations based on the state-of-the-art. Immediately, the problem of how to mix two different classifiers raised. We proposed a novel manner to combine two different classifiers **AVG1**. The normalized convex weighting scheme consists on a whitening of the training classification scores before learning the optimal convex combination of the two classifiers. We evaluated the performance of several features, representations and classifiers.

Secondly, we analysed the performance of SVM-based classifiers when trained with tiny data sets **AVG2**. This is useful because it will lead to systems that are adaptable and robust at the same time, i.e., able to build accurate models of the audio-visual commands from very few examples. We evaluated five different methods on training sets from 9 to 21 observations. Preliminary conclusions were expected, showing that the Multiple Kernel Learning framework is the one that better suits that task. However, further experiments with data sets consisting of higher number of classes need to be done to confirm this preliminary conclusions.

In this application the data was as complex as in the previous case: (D1) egocentric, (D2) multimodal and (D3) corrupted. The normalized convex weighting scheme **AVG1** is inherently (M4) adaptable. Its (M1) efficiency, (M2) speed, (M3) robustness and (M5) reliability depend on the classifiers used. HMM are slower and less efficient thatn SVMs, but SVMs are not able to learn the temporal structure of the commands. **AVG2** shows that the best method in our set up is the Multiple Kernel Learning (MKL). The training of such classifier is slightly slower that training a regular SVM, however this is compensated by far with the (M3) robustness given by the SVMs, its (M2) efficiency, since we train one classifier for the two modalities and its (M5) reliability. MKL reported excellent performance results on audio-visual command recognition with tiny training sets. Because we used SVMs there is no point on considering the (O1) temporal coherence or (O2) spatial consistency. However, the semantic meaning of the outcome is indisputable, because the methods provides a label carrying a well-defined content.

6.4 Multichannel Sound Source Localisation

The last application we worked on is the localisation of sound sources by means of multichannel time delay estimates using non-coplanar microphone arrays. Reverberations and microphone noise are the main issues when addressing the estimation of the time delays. Because, in our case, the geometry of the array is not known beforehand, we cannot include it in the cost function directly. Hence, we introduced a framework for localising the sound sources from the estimation of the time delays. A multichannel TDE criterion is proposed for arbitrary microphone arrays, and the TDE problem is cast into a non-linear optimisation

task **GTDE1**. The minimization is carried out by a log-barrier interior-point local method initialized in a grid **GTDE2**. The consequence of the direct path propagation model are deeply analysed and lead to a set of non-linear equations constraining the optimisation task **GTDE3**. Finally, a global optimisation procedure - a branch and bound algorithm - is also proposed **GTD4**.

While in the other applications the data was (D2) multimodal, inhere we used only auditory data. However, the data is still (D1) robocentric and (D3) corrupted by noise and reverberations. Both methods, **GTDE2** and **GTDE4**, are (M3) robust, (M4) highly adaptable and (M5) reliable. However, we did not spend time on developing a smart implementation, Therefore, they are rather (M1) inefficient and (M2) slow. The results produced by the proposed method are not (O1) temporally coherent since they do not track the source. However, they are (O2) spatially consistent, because they provide for the sound source position and (O3) semantically meaningful since they guess which regions in the scene carry auditory content.

6.5 Forthcoming Years

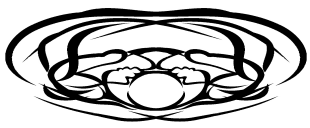
Audio-visual signal processing is a broad research field which has not reached yet its maturity. This topic feeds and is fed from many other research fields such as speech recognition, sound source separation, visual recognition, tracking, dialogue modelling, social computing and human-robot interaction. Consequently, it is a constantly evolving field. Indeed, the learning tools and signal representations used highly depend on the application targeted and even on the data. Moreover, the small number of existing reviews and their recentness are another evidence of the field's youth (see, for instance, [Shivappa 10, Gatica-Perez 09, Vinciarelli 09]). In my humble opinion, in the forthcoming years researchers will draw AV signal processing together and outcome conventions and standards from the deep understanding of the challenges and issues associated to the field. In addition, the interconnections with other research field will become stronger. Summarizing, the topic os AV signal processing will start acquiring the maturity which is lacking of nowadays.

6.6 Final Conclusion

During this PhD we faced real problems by creating adapted models and deriving their computational answers to the research questions. All the applications we addressed were in the framework of egocentric audio-visual scene analysis. Methods were systematically evaluated on synthetic and on real data. It is worth noticing that all the contents of this thesis have been already published/are under review in top international conferences and journals.

Moreover, the experience collected by the candidate exceeds the limits of the contents of that manuscript, as outlined in the introduction. Indeed, the partic-

ipation in international research projects, teaching, the supervision of Masters students, paper reviewing and the organisation of workshops in conferences have also taken place during this PhD, and therefore are part of the acquired experience. In all, this PhD has been a complete research experience, highly enriching, that encourages the candidate to pursue his research career starting a new life as a senior researcher.



PUBLICATIONS

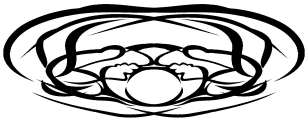
International Journals

- Xavier Alameda-Pineda, and Radu P. Horaud. *Vision-Guided Robot Hearing*, International Journal of Robotics Research, Special Issue on Robot Vision, under review. 2013.
- Xavier Alameda-Pineda, Jordi Sanchez-Riera, Johannes Wienke, Vojtech Franc, Jan Cech, Kaustubh Kulkarni, Antoine Deleforge, and Radu P. Horaud. *Ravel: An annotated corpus for training robots with audiovisual abilities*. Journal on Multimodal User Interfaces, 7(1-2), March 2013.

International Conferences and Workshops

- Jan Cech, Ravi K Mittal, Antoine Deleforge, Jordi Sanchez-Riera, Xavier Alameda-Pineda, Radu P. Horaud. *Active-Speaker Detection and Localization with Microphones and Cameras Embedded into a Robotic Head*. In Proceedings of the International Conference on Humanoid Robotics, Atlanta, GA, 2013.
- Xavier Alameda-Pineda, Radu P. Horaud, and Bernard. Mourrain. *The Geometry of Sounds-Source Localization using Non-Coplanar Microphone Arrays*. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2013.
- Xavier Alameda-Pineda, Jordi Sanchez-Riera, and Radu P. Horaud. *Benchmarking methods for audio-visual recognition using tiny training sets*. In IEEE International Conference on Acoustics, Speech, and Signal Processing, Vancouver, Canada, May 2013. IEEE Signal Processing Society.
- Jordi Sanchez-Riera, Xavier Alameda-Pineda, Johannes Wienke, Antoine Deleforge, Soraya Arias, Jan Cech, Sebastian Wrede, and Radu P. Horaud. *Online multimodal speaker detection for humanoid robots*. In IEEE International Conference on Humanoid Robotics, Osaka, Japan, November 2012.

- Jordi Sanchez-Riera, Xavier Alameda-Pineda, and Radu P. Horaud. *Audio-visual robot command recognition*. In ACM/IEEE International Conference on Multimodal Interaction. ACM, November 2012.
- Maxime Janvier, Xavier Alameda-Pineda, Laurent Girin, and Radu P. Horaud. *Sound-event recognition with a companion humanoid*. In IEEE International Conference on Humanoid Robotics, Osaka, Japan, November 2012.
- Xavier Alameda-Pineda and Radu P. Horaud. *Geometrically-constrained robust time delay estimation using non-coplanar microphone arrays*. In Proceeding of the 20th European Signal Processing Conference (EUSIPCO), Bucharest, Romania, August 2012.
- Xavier Alameda-Pineda, Vasil Khalidov, Radu P. Horaud, and Florence Forbes. *Finding audio-visual events in informal social gatherings*. In Proceedings of the 13th International Conference on Multimodal Interaction, Alicante, Spain, November 2011. ACM. **Oustanding paper award**.
- Julio C. Rolon, Philippe Salembier, and Xavier Alameda-Pineda. *Image compression with generalized lifting and partial knowledge of the signal pdf*. In IEEE International Conference on Image Processing, 2008.



BIBLIOGRAPHY

[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 12a] X. Alameda-Pineda & R. P. Horaud. *Geometrically-constrained Robust Time Delay Estimation Using Non-coplanar Microphone Arrays*. In Proceedings of EUSIPCO, pages 1309–1313, Bucharest, Romania, August 2012.

[Alameda-Pineda 12b] X. Alameda-Pineda, J. Sanchez-Riera, V. Franc, J. Wienke, J. Čech, K. Kulkarni, A. Deleforge & R. P. Horaud. *RAVEL: An Annotated Corpus for Training Robots with Audio Visual Abilities*. Journal of Multimodal User Interfaces, 2012.

[Alameda-Pineda 13a] X. Alameda-Pineda & R. Horaud. *Vision-Guided Robot Hearing*. Submitted to International Journal of Robotics Research, Special Issue on Robot Vision, 2013.

[Alameda-Pineda 13b] X. Alameda-Pineda, R. Horaud & B. Mourrain. *The Geometry of Sounds-Source Localization using Non-Coplanar Microphone ARRAYS*. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2013.

[Alameda-Pineda 13c] X. Alameda-Pineda, J. Sanchez-Riera & R. P. Horaud. *Benchmarking Methods for Audio-Visual Recognitions Using Tiny Training Sets*. In Proceedings of the IEEE International Conference on Audio Speech and Signal Processing, 2013.

[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/`.

[Bailly-Baillire 03] E. Bailly-Baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, F. Porée & B. Ruiz. *The BANCA database and evaluation protocol*. In Proceedings of the International Conference on Audio and Video-Based Biometric Person Authentication, pages 625–638. Springer-Verlag, 2003.

[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.

[Beck 08] A. Beck, P. Stoica & J. Li. *Exact and approximate solutions of source localization problems*. Signal Processing, IEEE Transactions on, vol. 56, no. 5, pages 1770–1778, 2008.

[Benesty 07] J. Benesty, Y. Huang & J. Chen. *Time Delay Estimation via Minimum Entropy*. Signal Processing Letters, IEEE, vol. 14, no. 3, pages 157–160, 2007.

[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.

[Bouguet 08] J.-Y. Bouguet. *Camera Calibration Toolbox for Matlab*, 2008. `http://www.vision.caltech.edu/bouguetj/calib\_doc/`.

[Boyd 04] S. Boyd & L. Vandenberghe. Convex optimization. Cambridge University Press, 2004.

[Brandstein 97a] M. Brandstein, J. Adcock & H. Silverman. *A closed-form location estimator for use with room environment microphone arrays*. Speech and Audio Processing, IEEE Transactions on, vol. 5, no. 1, pages 45–50, 1997.

[Brandstein 97b] M. Brandstein & H. Silverman. *A practical methodology for speech source localization with microphone arrays*. Computer Speech & Language, vol. 11, no. 2, pages 91–126, 1997.

[Brookes 11] M. Brookes. *VOICEBOX: Speech Processing Toolbox for MATLAB*. `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`, 2011.

[Brugman 04] H. Brugman, A. Russel & X. Nijmegen. *Annotating multimedia / multimodal resources with ELAN*. In Proceedings of the International Conference on Language Resources and Evaluation, pages 2065–2068, 2004.

[Brutti 08] A. Brutti, M. Omologo, P. Svaizer & F. Bruno. *Comparison between different sound source localization techniques*. In Hands-Free Speech Communication and Microphone Arrays, pages 69–72, 2008.

[Butz 05] T. Butz & J.-P. Thiran. *From error probability to information theoretic (multi-modal) 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.

[Canclini 13] A. Canclini, E. Antonacci, A. Sarti & S. Tubaro. *Acoustic Source Localization With Distributed Asynchronous Microphone Networks*. Audio, Speech, and Language Processing, IEEE Transactions on, vol. 21, no. 2, pages 439–443, 2013.

[Carbonetto 08] P. Carbonetto. *MATLAB primal-dual interior-point solver for convex programs with constraints*, 2008. http://www.cs.ubc.ca/ pcarbo/convexprog.html.

[Chan 94] Y. Chan & K. Ho. *A simple and efficient estimator for hyperbolic location*. Signal Processing, IEEE Transactions on, vol. 42, no. 8, pages 1905–1915, 1994.

[Chen 03a] J. Chen, K. Yao & R. Hudson. *Acoustic source localization and beamforming: theory and practice*. EURASIP Journal on App. Sig. Proc., pages 359–370, 2003.

[Chen 03b] J. Chen, J. Benesty & Y. Huang. *Robust time delay estimation exploiting redundancy among multiple microphones*. IEEE Transactions on SAP, vol. 11, no. 6, pages 549–557, 2003.

[Chen 06] J. Chen, J. Benesty & Y. A. Huang. *Time Delay Estimation in Room Acoustic Environments: An Overview*. EURASIP Journal on Adv. in Sig. Proc., vol. 2006, no. i, pages 1–20, 2006.

[Cherry 53] E. C. Cherry. *Some experiments on the recognition of speech, with one and with two ears*. Journal of the Acoustical Society of America, vol. 25, no. 5, pages 975–979, 1953.

[Christensen 07] H. Christensen, N. Ma, S. Wrigley & J. Barker. *Integrating pitch and localisation cues at a speech fragment level*. In INTERSPEECH, pages 2769–2772, 2007.

[Cooke 07] M. Cooke, J. Barker, S. Cunningham & X. Shao. *An audio-visual corpus for speech perception and automatic speech recognition (L)*. Speech Communication, vol. 49, no. 5, pages 384–401, 2007.

[Dalal 05] N. Dalal & B. Triggs. *Histograms of Oriented Gradients for Human Detection*. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2005.

[Davies 79] D. Davies & D. Bouldin. *A Cluster Separation Measure*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. PAMI-1, no. 2, page 2247, January 1979.

[Deleforge 12a] A. Deleforge & R. Horaud. *2D sound-source localization on the binaural manifold*. In Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on, pages 1–6, 2012.

[Deleforge 12b] A. Deleforge & R. P. Horaud. *The Cocktail Party Robot: Sound Source Separation and Localisation with an Active Binaural Head*. In IEEE/ACM International Conference on Human Robot Interaction, Boston, Mass, March 2012.

[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, page 1, 1977.

[Doclo 03] S. Doclo & M. Moonen. *Robust Adaptive Time Delay Estimation for Speaker Localization in Noisy and Reverberant Acoustic Environments*. EURASIP Journal on Adv. in Sig. Proc., vol. 2003, no. 11, pages 1110–1124, 2003.

[Friedlander 87] B. Friedlander. *A passive localization algorithm and its accuracy analysis*. Oceanic Engineering, IEEE Journal of, vol. 12, no. 1, pages 234–245, 1987.

[Galati 06] G. Galati, M. Gasbarra, P. Magaro, P. Marco, L. Mene & M. Pici. *New Approaches to Multilateration processing: analysis and field evaluation*. In 2006 European Radar Conference, volume 9, pages 116–119. Ieee, September 2006.

[Garofolo 93] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren & V. Zue. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, 1993. Linguistic Data Consortium, Philadelphia.

[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, page 6016, 2007.

[Gatica-Perez 09] D. Gatica-Perez. *Automatic nonverbal analysis of social interaction in small groups: A review*. Image and Vision Computing, vol. 27, no. 12, pages 1775 – 1787, 2009. Visual and multimodal analysis of human spontaneous behaviour.

[Ghazanfar 06] A. A. Ghazanfar & C. E. Schroeder. *Is neocortex essentially multisensory?* Transactions on Cognitive Neuroscience, vol. 10, page 278–285, 2006.

[Gorelick 07] L. Gorelick, M. Blank, E. Shechtman, M. Irani & R. Basri. *Actions as Space-Time Shapes*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pages 2247–2253, December 2007. `http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`.

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

[Hansard 08] M. Hansard & R. P. Horaud. *Cyclopean Geometry of Binocular Vision*. Journal of the Optical Society of America, vol. 25, no. 9, page 2357–2369, September 2008.

[Harris 88] C. Harris & M. Stephens. *A Combined Corner and Edge Detector*. In Proc. of Fourth Alvey Vision Conference, page 1471, 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.

[Hazen 04] T. J. Hazen, K. Saenko, C.-H. La & J. R. Glass. *A segment-based audio-visual speech recognizer: data collection, development, and initial experiments*. In Proceedings of the ACM International Conference on Multimodal Interfaces, ICMI '04, pages 235–242. ACM, 2004.

[He 13a] H. He, J. Lu, L. Wu & X. Qiu. *Time delay estimation via non-mutual information among multiple microphones*. Applied Acoustics, vol. 74, no. 8, pages 1033 – 1036, 2013.

[He 13b] H. He, L. Wu, J. Lu, X. Qiu & J. Chen. *Time Difference of Arrival Estimation Exploiting Multichannel Spatio-Temporal Prediction*. Audio, Speech, and Language Processing, IEEE Transactions on, vol. 21, no. 3, pages 463–475, 2013.

[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.

[Hoai 11] M. Hoai, Z. Zhong Lan & F. De la Torre. *Joint Segmentation and Classification of Human Actions in Video*. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2011.

[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.

[Huang 01] Y. Huang, J. Benesty, G. Elko & R. Mersereati. *Real-time passive source localization: A practical linear-correction least-squares approach*. Speech and Audio Processing, IEEE Transactions on, vol. 9, no. 8, pages 943–956, 2001.

[Huang 03] Y. Huang & J. Benesty. *A class of frequency-domain adaptive approaches to blind multichannel identification*. Signal Processing, IEEE Transactions on, vol. 51, no. 1, pages 11–24, 2003.

[HUMAVIPS 13] HUMAVIPS. *Humanoids with Auditory and Visual Abilities In Populated Spaces*. http://humavips.eu/, 2010-2013. last visited: July 15th, 2013.

[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.

[Janvier 12] M. Janvier, X. Alameda-Pineda, L. Girin & R. P. Horaud. *Sound-Event Recognition with a Companion Humanoid*. In IEEE International Conference on Humanoid Robotics, Osaka, Japan, November 2012.

[Jiang 09] W. Jiang, C. Cotton, S.-F. Chang, D. Ellis & A. Loui. *Short-term audio-visual atoms for generic video concept classification*. In Proceedings of the ACM International Conference on Multimedia, 2009.

[Kalal 12] Z. Kalal, K. Mikolajczyk & J. Matas. *Tracking-Learning-Detection*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 7, pages 1409 –1422, july 2012.

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

[Keyrouz 06] F. Keyrouz & K. Diepold. *An Enhanced Binaural 3D Sound Localization Algorithm*. 2006 IEEE International Symposium on Signal Processing and Information Technology, pages 662–665, August 2006.

[Keyrouz 07] F. Keyrouz, K. Diepold & S. Keyrouz. *Humanoid Binaural Sound Tracking Using Kalman Filtering and HRTFs*. Robot Motion and Control 2007, pages 329–340, 2007.

[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, page 2174, New York, NY, USA, 2008. ACM.

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

[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.

[Kullaib 09] A. R. Kullaib, M. Al-Mualla & D. Vernon. *2D Binaural Sound Localization: for Urban Search and Rescue Robotics*. In proc. Mobile Robotics, pages 423–435, Istanbul, Turkey, September 2009.

[Lacheze 09] L. Lacheze, Y. Guo, R. Benosman, B. Gas & C. Couverture. *Audio/video fusion for objects recognition*. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009.

[Laptev 05] I. Laptev. *On Space-Time Interest Points*. International Journal of Computer Vision, vol. 64, no. 2-3, pages 107–123, 2005.

[Laptev 08] I. Laptev, M. Marszalek, C. Schmid & B. Rozenfeld. *Learning realistic human actions from movies*. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2008.

[Lathoud 05] G. Lathoud, J. Odobez & D. Gatica-Pérez. *AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking*. In Proceedings of the Workshop on Machine Learning and Multimodal Interaction. Springer Verlag, 2005.

[Lehmann 08] E. A. Lehmann & A. M. Johansson. *Prediction of energy decay in room impulse responses simulated with an image-source model*. JASA, vol. 124, no. 1, pages 269–277, 2008.

[Lehmann 12] E. A. Lehmann. *Matlab code for image-source model in room acoustics*. `http://www.eric-lehmann.com/ism\_code.html`, 2012. accessed November 2011.

[Lim 13] J.-S. Lim & H.-S. Pang. *Time Delay Estimation Method Based on Canonical Correlation Analysis*. Circuits, Systems and Signal Processing, March 2013.

[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 09] J. Liu, J. Luo & M. Shah. *Recognizing Realistic Actions from Videos "in the Wild"*. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2009.

[Lopes 06] J. Lopes & S. Singh. *Audio and Video Feature Fusion for Activity Recognition in Unconstrained Videos*. In Intelligent Data Engineering and Automated Learning, 2006.

[Luo 89] R. C. Luo & M. G. Kay. *Multisensor integration and fusion in intelligent systems*. In Systems, Man and Cybernetics, IEEE Transactions on, volume 19, page 9011, 1989.

[Luo 08] J. Luo, B. Caputo, A. Zweig, J.-H. Bach & J. Anemüller. *Object category detection using audio-visual cues*. In Proceedings of the 6th International Conference on Computer Vision Systems, 2008.

[Mandel 07] M. I. Mandel, D. P. W. Ellis & T. Jebara. *An EM Algorithm for Localizing Multiple Sound Sources in Reverberant Environments*. In Proc. NIPS, pages 953–960, Cambridge, MA, 2007.

[Marcel 10] S. Marcel, C. McCool, P. Matejka, T. Ahonen & J. Cernocky. *Mobile Biometry (MOBIO) Face and Speaker Verification Evaluation*. Idiap-RR Idiap-RR-09-2010, Idiap, 5 2010.

[Marszalek 09] M. Marszalek, I. Laptev & C. Schmid. *Actions in context*. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recogntion, 2009.

[Messer 99] K. Messer, J. Matas, J. Kittler & K. Jonsson. *XM2VTSDB: The Extended M2VTS Database*. In Proceedings of the International Conference on Audio and Video-based Biometric Person Authentication, pages 72–77, 1999.

[Messing 09] R. Messing, C. Pal & H. Kautz. *Activity recognition using the velocity histories of tracked keypoints*. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 2009. IEEE Computer Society.

[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.

[Mohammad 08] Y. Mohammad, Y. Xu, K. Matsumura & T. Nishida. *The H3R Explanation Corpus human-human and base human-robot interaction dataset*. In International Conference on Intelligent Sensors, Sensor Networks and Information Processing, pages 201 –206, dec. 2008.

[Monaci 06] G. Monaci, P. V, B. Mailhé, S. Lesage & R. Gribonval. *Learning Multi-Modal Dictionaries*. IEEE Transactions on Image Processing, 2006.

[Mostefa 07] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Cristoforetti, F. Tobia*et al.* *The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms*. Language Resources and Evaluation, vol. 41, no. 3-4, pages 389–407, 2007.

[Mühling 12] M. Mühling, R. Ewerth, J. Zhou & B. Freisleben. *Multimodal video concept detection via bag of auditory words and multiple kernel learning*. In Proceedings of the International Conference on Advances in Multimedia Modeling, 2012.

[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, page 1034, 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.

[OPPORTUNITY 11] OPPORTUNITY. *Activity and Context Recognition with Opportunistic Sensor Configurations*. `http://www.opportunity-project.eu/`, 2011. last visited: September, 2012.

[Patterson 02] E. K. Patterson, S. Gurbuz, Z. Tufekci & J. N. Gowdy. *CUAVE: A new audio-visual database for multimodal human-computer interface research*. In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, pages 2017–2020, 2002.

[Pavlidi 13] D. Pavlidi, A. Griffin, M. Puigt & A. Mouchtaris. *Real-Time Multiple Sound Source Localization and Counting Using a Circular Microphone Array*. Audio, Speech, and Language Processing, IEEE Transactions on, vol. 21, no. 10, pages 2193–2206, 2013.

[Pertilä 09] P. Pertilä. *Acoustic Source Localization in a Room Environment and at Moderate Distances*. Tampereen teknillinen yliopisto. Julkaisu-Tampere University of Technology. Publication; 794, 2009.

[Pigeon 96] S. Pigeon. *M2vts database*, 1996. `http://www.tele.ucl.ac.be/PROJECTS/M2VTS/`.

[POP 09] POP. *Perception On Purpose*. `http://perception.inrialpes.fr/POP/`, 2006-2009. last visited: July 15th, 2013.

[Rabiner 11] L. R. Rabiner & R. W. Schafer. Theory and applications of digital speech processing. Pearson, 2011.

[Ramasubramanian 11] V. Ramasubramanian, R. Karthik, S. Thiyagarajan & S. Cherla. *Continuous audio analytics by HMM and Viterbi decoding.* In Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing, pages 2396–2399. IEEE, 2011.

[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.

[Ribeiro 10] F. Ribeiro, C. Zhang, D. Florêncio & D. Ba. *Using reverberation to improve range and elevation discrimination for small array sound source localization*. IEEE Transactions on ASLP, vol. 18, no. 7, pages 1781–1792, 2010.

[ROS 12] ROS. *message_filters/ApproximateTime*. `http://www.ros.org/wiki/message_filters/ApproximateTime`, 2012. last visited: June 21st, 2012.

[Rybok 11] L. Rybok, S. Friedberger, U. D. Hanebeck & R. Stiefelhagen. *The KIT Robo-Kitchen Data set for the Evaluation of View-based Activity Recognition Systems*. In Proceedings of the IEEE-RAS International Conference on Humanoid Robots, 2011.

[Saenko 08] K. Saenko & T. Darrell. *Object category recognition using probabilistic fusion of speech and image classifiers*. In Proceedings of the 4th International Conference on Machine Learning for Multimodal Interaction, 2008.

[Salvati 13] D. Salvati & S. Canazza. *Adaptive Time Delay Estimation Using Filter Length Constraints for Source Localization in Reverberant Acoustic Environments*. Signal Processing Letters, IEEE, vol. 20, no. 5, pages 507–510, 2013.

[Sanchez-Riera 12a] J. Sanchez-Riera, X. Alameda-Pineda & R. Horaud. *Audio-Visual Robot Command Recognition: D-META'12 Grand Challenge*. In Proceedings of the International Conference on Multimodal Interaction, 2012.

[Sanchez-Riera 12b] 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.

[Sanchez-Riera 12c] J. Sanchez-Riera, J. Čech & R. Horaud. *Action Recognition robust to Background Clutter by using Stereo Vision*. In In 4th International Workshop on Video Event

Categorization, Tagging and Retrieval (VECTaR), in conjunction with IEEE European Conference on Computer Vision, 2012.

[Sasaki 12] Y. Sasaki, M. Kabasawa, S. Thompson, S. Kagami & K. Oro. *Spherical microphone array for spatial sound localization for a mobile robot*. In Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on, pages 713–718, 2012.

[Schüldt 04] C. Schüldt, I. Laptev & B. Caputo. *Recognizing human actions: A local SVM approach*. In Proceedings of the International Conference on Pattern Recognition, pages 32–36, 2004.

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

[Seco 09] F. Seco, A. Jiménez, C. Prieto, J. Roa & K. Koutsou. *A survey of mathematical methods for indoor localization*. In Intelligent Signal Processing, 2009. WISP 2009. IEEE International Symposium on, numéro x, pages 9–14. IEEE, 2009.

[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 401–409, 2008.

[Sheng 05] X. Sheng & Y.-h. Hu. *Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks*. IEEE Transactions on Signal Processing, vol. 53, no. 1, pages 44–53, January 2005.

[Shi 11] Q. Shi, L. Wang, L. Cheng & A. Smola. *Discriminative Human Action Segmentation and Recognition using SMMs*. International Journal on Computer Vision, vol. 93, no. 1, pages 22–32, Jan 2011.

[Shivappa 10] S. T. Shivappa, B. D. Rao & M. M. Trivedi. *Auvio-Visual Fusion and Tracking With Multilevel Iterative Decoding: Framework and Experimental Evaluation*. In Journal of Selected Topics in Signal Processing, 2010.

[Smith 87] J. Smith & J. Abel. *Closed-form least-squares source location estimation from range-difference measurements*. Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 35, no. 12, pages 1661–1669, 1987.

[So 08] H. So, Y. Chan & F. Chan. *Closed-form formulae for time-difference-of-arrival estimation*. Signal Processing, IEEE Transactions on, vol. 56, no. 6, pages 2614–2620, 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.

[Sonnenburg 10] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. d. Bona, A. Binder, C. Gehl & V. Franc. *The SHOGUN Machine Learning Toolbox*. Journal of Machine Learning Research, vol. 99, pages 1799–1802, August 2010.

[Strobel 99] N. Strobel & R. Rabenstein. *Classification of time delay estimates for robust speaker localization*. In Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings., 1999 IEEE International Conference on, volume 6, pages 3081–3084. IEEE, 1999.

[Tenorth 09] M. Tenorth, J. Bandouch & M. Beetz. *The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition*. In Proceedings of the IEEE International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences in conjunction with the International Conference on Computer Vision, 2009.

[Urruela 04] A. Urruela & J. Riba. *Novel closed-form ML position estimator for hyperbolic location*. In Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on, volume 2, pages ii–149. IEEE, 2004.

[Valin 06] J. Valin & F. Michaud. *Robust 3D localization and tracking of sound sources using beamforming and particle filtering*. IEEE Transactions on ASSP, vol. 2, no. 1, 2006.

[Čech 11] J. Čech, J. Sanchez-Riera & R. P. Horaud. *Scene Flow Estimation by Growing Correspondence Seeds*. In Computer Vision and Pattern Recognition, In the Proceedings of the IEEE International Conference on, 2011.

[Vedula 05] S. Vedula, S. Baker, P. Rander, R. Collins & T. Kanade. *Three-Dimensional Scene Flow*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, 2005.

[Vinciarelli 09] A. Vinciarelli, M. Pantic & H. Bourlard. *Social signal processing: Survey of an emerging domain*. Image and Vision Computing, vol. 27, no. 12, pages 1743 – 1759, 2009. Visual and multimodal analysis of human spontaneous behaviour.

[Viste 03] H. Viste & G. Evangelista. *On the Use of Spatial Cues to Improve Binaural Source Separation*. In proc. DAFx, pages 209–213, 2003.

[Weinland 06] D. Weinland, R. Ronfard & E. Boyer. *Free Viewpoint Action Recognition using Motion History Volumes*. Journal on Computer Vision and Image Understanding, vol. 104, no. 2, pages 249–257, November 2006. `http://4drepository.inrialpes.fr/public/viewgroup/6`.

[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.

[Willems 09] G. Willems, J. H. Becker & T. Tuytelaars. *Exemplar-based Action Recognition in Video*. In Proceedings of the British Machine Vision Conference, 2009.

[Woodruff 12] J. Woodruff & D. Wang. *Binaural Localization of Multiple Sources in Reverberant and Noisy Environments*. IEEE Trans. Acoust., Speech, Signal Process., vol. 20, no. 5, pages 1503–1512, 2012.

[Wu 10] Q. Wu, Z. Wang, F. Deng & D. D. Feng. *Realistic Human Action Recognition with Audio Context*. In Proceedings of the International Conference on Digital Image Computing: Techniques and Applications, 2010.

[Xiong 05] Z. Xiong. *Audio-visual sports highlights extraction using Coupled Hidden Markov Models*. Pattern Anal. Appl., vol. 8, no. 1-2, pages 62–71, 2005.

[Ye 12] G. Ye, I.-H. Jhuo, D. Liu, Y.-G. Jiang, D. Lee & S.-F. Chang. *Joint Audio-Visual Bi-Modal Codewords for Video Event Detection*. In Proceedings of the ACM International Conference on Multimedia Retrieval, 2012.

[Zhang 07] C. Zhang, Z. Zhang & D. Florencio. *Maximum Likelihood Sound Source Localization for Multiple Directional Microphones*. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, numéro 6, pages 125–128. Ieee, 2007.

[Zhang 08] C. Zhang, D. Florêncio, D. E. Ba & Z. Zhang. *Maximum Likelihood Sound Source Localization and Beamforming for Directional Microphone Arrays in Distributed Meetings*. IEEE Transactions on Multimedia, vol. 10, no. 3, pages 538–548, April 2008.

[Zivkovic 08] Z. Zivkovic, O. Booij, B. Krose, E. Topp & H. Christensen. *From Sensors to Human Spatial Concepts: An Annotated Data Set*. IEEE Transactions on Robotics, vol. 24, no. 2, pages 501 –505, april 2008.