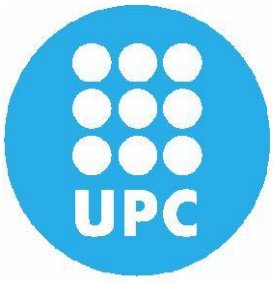

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONA**TECH**

PhD Thesis

Acoustic Event Detection and Localization using Distributed Microphone Arrays

Rupayan Chakraborty

Thesis Advisor:
Dr. Climent Nadeu

Speech Processing Group
TALP Research Center
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya

Barcelona, October 2013

Abstract

Automatic acoustic scene analysis is a complex task that involves several functionalities: detection (time), localization (space), separation, recognition, etc. This thesis focuses on both acoustic event detection (AED) and acoustic source localization (ASL), when several sources may be simultaneously present in a room. In particular, the experimentation work is carried out with a meeting-room scenario. Unlike previous works that either employed models of all possible sound combinations or additionally used video signals, in this thesis, the time overlapping sound problem is tackled by exploiting the signal diversity that results from the usage of multiple microphone array beamformers.

The core of this thesis work is a rather computationally efficient approach that consists of three processing stages. In the first, a set of (null) steering beamformers is used to carry out diverse partial signal separations, by using multiple arbitrarily located linear microphone arrays, each of them composed of a small number of microphones. In the second stage, each of the beamformer output goes through a classification step, which uses models for all the targeted sound classes (HMM-GMM, in the experiments). Then, in a third stage, the classifier scores, either being intra- or inter-array, are combined using a probabilistic criterion (like MAP) or a machine learning fusion technique (fuzzy integral (FI), in the experiments).

The above-mentioned processing scheme is applied in this thesis to a set of complexity-increasing problems, which are defined by the assumptions made regarding identities (plus time endpoints) and/or positions of sounds. In fact, the thesis report starts with the problem of unambiguously mapping the identities to the positions, continues with AED (positions assumed) and ASL (identities assumed), and ends with the integration of AED and ASL in a single system, which does not need any assumption about identities or positions.

The evaluation experiments are carried out in a meeting-room scenario, where two sources are temporally overlapped; one of them is always speech and the other is an acoustic event from a pre-defined set. Two different databases are used, one that is produced by merging signals actually recorded in the UPC's department smart-room, and the other consists of overlapping sound signals directly recorded in the same room and in a rather spontaneous way. From the experimental results with a single array, it can be observed that the proposed

detection system performs better than either the model based system or a blind source separation based system. Moreover, the product rule based combination and the FI based fusion of the scores resulting from the multiple arrays improve the accuracies further. On the other hand, the posterior position assignment is performed with a very small error rate.

Regarding ASL and assuming an accurate AED system output, the 1-source localization performance of the proposed system is slightly better than that of the widely-used SRP-PHAT system, working in an event-based mode, and it even performs significantly better than the latter one in the more complex 2-source scenario. Finally, though the joint system suffers from a slight degradation in terms of classification accuracy with respect to the case where the source positions are known, it shows the advantage of carrying out the two tasks, recognition and localization, with a single system, and it allows the inclusion of information about the prior probabilities of the source positions. It is worth noticing also that, although the acoustic scenario used for experimentation is rather limited, the approach and its formalism were developed for a general case, where the number and identities of sources are not constrained.

Resum

L'anàlisi automatic d'escenes acústiques és una tasca complexa que requereix unes quantes funcionalitats: detecció (temps), localització (espai), separació, reconeixement, etc. Aquesta tesi s'enfoca tant cap a la detecció d'esdeveniments acústics (AED) com a la localització de fonts acústiques (ASL), en el cas en què en una sala hi puguin coexistir diverses fonts acústiques simultàniament. En concret, el treball d'experimentació es du a terme en un escenari de sala de reunions.

El nucli del treball de la tesi rau en un plantejament eficient en termes de càlcul que es basa en tres etapes de processament. En la primera, s'utilitza un conjunt de comformadors de feix per dur a terme diverses separacions parcials de senyals, usant múltiples configuracions lineals de micròfons col·locades arbitràriament, cada una composta d'un nombre petit de micròfons. En la segona etapa, cada una de les sortides dels comformadors passa per un classificador, el qual té models de totes les classes considerades. I llavors, en la tercera etapa, les puntuacions del classificador, ja siguin intra o inter-configuració, es combinen amb un criteri probabilistic (com MAP) o amb una tècnica de fusió amb aprenentatge automàtic (la integral difusa (FI), en els experiments).

L'esquema de processament esmentat s'aplica en aquesta tesi a un conjunt de problemes de complexitat creixent, que queden definits per les suposicions que es fan en relació a les identitats (més els temps d'inici i final) i/o les posicions dels sons. En efecte, l'informe de la tesi comença amb el problema de l'assignació sense ambiguïtat de les identitats a les posicions, continua amb AED (suposant les posicions) i ASL (suposant les identitats), i acaba amb la integració de AED i ASL en un sistema únic que no necesita fer cap suposició respecte les identitats o les posicions.

Els experiments tenen lloc en un escenari de sala de reunions, on hi ha dues fonts superposades en el temps; una és sempre parla i l'altra és un esdeveniment acústic d'entre un conjunt predefinit. S'usen dues bases de dades diferents, una s'ha produït barrejant senyals enregistrats realment en la sala intel·ligent de la UPC, i l'altra consisteix en senyals de sons ensolapats que gravats directament en la sala i d'una manera més aviat espontània. S'observa

dels resultats experimentals amb una sola configuració que el sistema proposat de detecció es comporta millor que el sistema basat en models o el sistema basat en separació cega de fonts. A més a més, tant la combinació basada en la regla producte com la fusió basada en FI de les puntuacions obtingudes dels múltiples arrays milloren encara més la precisió. D'altra banda, l'assignació posterior de posicions té lloc amb una taxa d'error molt petita.

En relació amb ASL i suposant una sortida del sistema AED, les prestacions de localització del sistema proposat per a una sola font són lleugerament millors que les del sistema SRP-PHAT treballant en mode esdeveniment, i fins i tot són significativament millors que les d'aquest darrer sistema en el cas més complex de l'escenari de dues fonts. Finalment, tot i que amb el sistema conjunt s'observa una lleugera degradació en termes de precisió de classificació respecte del cas en què es coneixen les posicions de les fonts, aquest té l'avantatge de dur a terme les dues tasques, reconeixement i localització, amb un únic sistema, i permet la inclusió d'informació sobre les probabilitats a priori de les posicions de les fonts. Cal fer notar també que, tot i que l'escenari acústic que s'ha usat en l'experimentació és bastant limitat, el plantejament i el seu formalisme s'han desenvolupat per a un cas general, sense restriccions pel que fa al nombre i les identitats de les fonts.

Resumen

El análisis automático de escenas acústicas es una tarea compleja que requiere unas cuantas funcionalidades: detección (tiempo), localización (espacio), separación, reconocimiento, etc. Esta tesis se enfoca tanto hacia la detección de eventos acústicos (AED) como la localización de fuentes acústicas (ASL), en el caso en que en una sala puedan coexistir diversas fuentes acústicas simultáneamente. En concreto, el trabajo de experimentación se lleva a cabo en un escenario de sala de reuniones.

El núcleo del trabajo de la tesis radica en un planteamiento eficiente en términos de cálculo que se basa en tres etapas de procesamiento. En la primera, se utiliza un conjunto de conformadores de haz para llevar a cabo diversas separaciones parciales de señales, usando múltiples configuraciones lineales de micrófonos colocadas arbitrariamente, cada una compuesta de un número pequeño de micrófonos. En la segunda etapa, cada una de las salidas de los conformadores pasa por un clasificador, el cual tiene modelos de todas las clases consideradas. Y entonces, en la tercera etapa, las puntuaciones del clasificador, ya sean intra o inter-configuración, se combinan con un criterio probabilístico (como MAP) o con una técnica de fusión con aprendizaje automático (la integral difusa (FI), en los experimentos).

El esquema de procesamiento mencionado se aplica en esta tesis a un conjunto de problemas de complejidad creciente, que quedan definidos por las suposiciones que se hacen en relación a las identidades (más los tiempos de inicio y final) y / o las posiciones de los sonidos. En efecto, el informe de la tesis comienza con el problema de la asignación sin ambigüedad de las identidades a las posiciones, continúa con AED (suponiendo las posiciones) y ASL (suponiendo las identidades), y termina con la integración de AED y ASL en un sistema único que no necesita hacer ninguna suposición respecto a las identidades o las posiciones.

Los experimentos tienen lugar en un escenario de sala de reuniones, donde hay dos fuentes superpuestas en el tiempo; una es siempre habla y la otra es un evento acústico de entre un conjunto predefinido. Se usan dos bases de datos diferentes, una se ha producido mezclando señales registradas realmente en la sala inteligente de la UPC, y la otra consiste en

señales de sonidos solapados grabados directamente en la sala y de una manera más bien espontánea. Se observa de los resultados experimentales con una sola configuración que el sistema propuesto de detección se comporta mejor que el sistema basado en modelos o el sistema basado en separación ciega de fuentes. Además, tanto la combinación basada en la regla producto como la fusión basada en FI de las puntuaciones obtenidas de los múltiples arrays mejoran aún más la precisión. Por otra parte, la asignación posterior de posiciones tiene lugar con una tasa de error muy pequeña.

En relación con ASL y suponiendo una salida del sistema AED, las prestaciones de localización del sistema propuesto para una sola fuente son ligeramente mejores que las del sistema SRP-PHAT trabajando en modo evento, e incluso son significativamente mejores que las de este último sistema en el caso más complejo del escenario de dos fuentes. Finalmente, aunque con el sistema conjunto se observa una ligera degradación en términos de precisión de clasificación respecto del caso en que se conocen las posiciones de las fuentes, éste tiene la ventaja de llevar a cabo las dos tareas, reconocimiento y localización, con un único sistema, y permite la inclusión de información sobre las probabilidades a priori de las posiciones de las fuentes. Conviene hacer notar también que, aunque el escenario acústico que se ha usado en la experimentación es bastante limitado, el planteamiento y su formalismo se han desarrollado para un caso general, sin restricciones en cuanto al número y las identidades de las fuentes.

Acknowledgements

I wish to express my sincere gratitude to those who have supported me in one way or the other during this amazing journey.

I am extremely grateful to my research guide, Climent Nadeu, for his valuable guidance and consistent encouragement that I received throughout this research work. This feat is possible because of his unconditional support. A person with an amicable and positive disposition, he has always made himself available to discuss and clarify my doubts despite of his busy schedules. I consider it as a great opportunity to do my PhD thesis under his guidance and to learn from his research expertise. Thanks to him, for all the support, help, and guidance.

I am also grateful to Javier Hernando, the supervisor of the project, in which I have been working during this research. Thanks to him for giving a continuous support. I really have enjoyed a beautiful time with him while attending a conference.

I would like to thank Taras Butko for his help at the beginning of my work. It would have been difficult for me to proceed without his initial help. He encouraged me a lot. However, it was very short, but very nice time that I spend with him in the lab.

I was very fortunate to interact with the person like Alfonso Ortega, Walter Kellermann and Volker Hohmann. In the beginning of my work, all discussions with them helped me a lot. Thanks to them.

I would like to mention several persons who were very generous to give their opinions during this research work. First, I would like to give my thanks to Alberto Abad. His comments and suggestions helped me a lot during the exploration of this thesis. I am also thankful to Carlos Segura for making valuable comments during the source localization work. I would like to thank Martin Wolf for sharing his ideas that helped me a lot. I was happy to work in the same group and to attend conferences with Martin Wolf and Henrik Schulz. I would like to thank Diego Lendoiro for providing an excellent computing platform to work. Thanks to Omid, Anna and other research members of the group.

I am also grateful for funding support from the Spanish project SARAI (TEC2010-21040-C02-01), and the continuous help from GPS group.

I am extremely thankful to Utpal Garain, who always guided me like an elder brother. In spite of the distance, we were always in touch, and that has given me immense confidence. His positive attitude in the face of some difficult circumstances amazes and inspires me.

I am thankful to my sister for her support and encouragement. I am also grateful to all the relatives and friends, who have always encouraged me. Thanks everybody for supporting me.

Finally, there is no way I would be where I am today without the unconditional and immeasurable love, support, encouragement and blessings of my parents. I cannot thank them enough for everything they have given me in my life and for always believing in me. They are always an incredible inspiration to me.

CONTENTS

List of Acronyms

AE(s)	Acoustic Event(s)
AED	Acoustic Event Detection
ANN(s)	Artificial Neural Network(s)
ASL	Acoustic Source Localization
ASR	Automatic Speech Recognition
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CHIL	Computer in the Human Interaction Loop
CLEAR	Classification of Event, Activities, and Relationships
CV	Cross Validation
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DOA	Direction of Arrival
EM	Expectation-Maximization
FBE	Filter Bank Energies
FF(BE)	Frequency Filtered (Band Energies)
FFT	Fast Fourier Transform
FI	Fuzzy Integral
FIB	Frequency Invariant Beamforming
FM(s)	Fuzzy Measure(s)
GCC	Generalized Cross Correlation
GMM(s)	Gaussian Mixture Model(s)
GSC	Generalized Sidelobe Cancellers
HMM(s)	Hidden Markov Model(s)
ICA	Independent Component Analysis
MAP	Maximum-a-Posteriori
MSC	Multiple Sidelobe Cancellers
NIST	National Institute of Standards and Technology
NSB	Null Steering Beamformer
PA	Position Assignment
PAR	Position Assignment Rate
PCA	Principal Component Analysis
PHAT	Phase Transformations
RBF	Radial Basis Function
PF	Particle Filter
RT	Rich Transcription
SB	Steered Beamforming
SMB	Sound-Model-Based
SNR	Signal-to-Noise Ratio
SRC	Stochastic Region Contraction
SRP	Steered Response Power

SVM(s)	Support Vector Machine(s)
TD	Time Delay
TDOA	Time Delay of Arrival
WAM	Weighted Arithmetical Mean

List of Figures

Chapter 7:

List of Tables

Chapter 3:

Chapter 4:

Chapter 5:

Chapter 1. Introduction

1.1 Thesis overview and motivations

Acoustic event detection (AED) aims at determining the identity of sounds and their temporal position in audio signals. Actually, in the context of person-machine communication, computers involved in human communication activities have to meet certain requirements and be designed to have minimal possible awareness from the users. Consequently, there is a need of perceptual user interfaces which uses microphone array sensors, and is capable of describing the complete audio scene in the respective environment. One example of such challenging research efforts is the development of smart-rooms. A smart-room is a closed space equipped with multiple microphones and cameras, which are designed to assist and complement human activities. In the case of the audio processing, some of the technologies that may be involved are microphone array signal processing, automatic speech and event recognition, source localization, voice activity detection, speaker identification and verification.

Although speech is usually the most informative AE, other kind of sounds may also carry useful cues for scene understanding. Since in such types of environments the human activity is reflected in a rich variety of acoustic events, either produced by the human body or by objects handled by humans, detection and localization of acoustic events may help to describe complete scene about human activity in the room environment. Additionally, the robustness of automatic speech recognition systems may be increased. For instance, in a meeting/lecture context, we may associate a chair moving or door noise to its start or end, cup clinking to a coffee break, or footsteps to somebody entering or leaving. Furthermore, some of these AEs are tightly coupled with human behaviors or psychological states: paper wrapping may denote tension; laughing, cheerfulness; yawning in the middle of a lecture, boredom; keyboard typing, distraction from the main activity in a meeting; and clapping during a speech, appreciation.

Some acoustic event detection and classification work was carried out at our UPC's lab in the framework of the CHIL EU project ("Computers in the Human Interaction Loop", 2004-2007). The CHIL AEC task was conceived to reliably classify meeting-room sounds like

"door knock", "door open/slam", "steps", "chair moving", "spoon/cup jingle", "paper work", "key jingle", "keyboard typing", "phone ring", "applause", "cough", and "laugh" along with speech and other human noises. The two CLEAR (Classification of Events, Activities and Relationships) evaluation campaigns supported by the CHIL project, which took place in 2006 and 2007, were designed to recognize events, activities, and their relationships in interaction scenarios. The last evaluation campaign showed that, in those seminar conditions, AED is still a challenging problem. In fact, 5 out of 6 submitted systems showed accuracy below 25%, and the best working system got 33.6% accuracy. It was realized that one of the main factors that accounts for those low detection scores is the high degree of overlap between sounds, especially between the targeted acoustic events and speech due to the nature of the recorded testing databases, that were mostly interactive seminars [**1**].

During that time, the PhD thesis entitled "Acoustic event detection and classification", from A. Temko [**2**], already included a first attempt to tackle the overlapping problem at the model level, using together models for isolated sounds and models for overlapped sounds [**3**]. A subsequent PhD thesis entitled "Feature selection for multimodal acoustic event detection", by T. Butko [**4**], among other contributions, showed a noticeable improvement in the detection rate of overlapped acoustic events that have a visual correlate, thanks to the inclusion of video features, and using different types of fusion, at the feature level and at the decision level [**5**]. Moreover, a real-time system is implemented in UPC's smart-room, where AED task is carried out using the model based approach, and ASL is based on SRP-PHAT algorithm. In summary, these previous works have dealt with the overlapping problem mostly either at the model level or by fusion with the additional video modality that is not affected by acoustic interference.

As the overlapping problem still remains, and there are other alternatives to solve it, source separation prior to detection could be the reasonable one. Motivated by this fact, the main effort in this thesis work is to develop an AED system to detect the acoustic events, which may be overlapped in time, as well as to identify them. For that purpose, all the arrays are used. The chosen methodology consists of performing some kind of source separation at the front end so that the overlapping problem could be tackled at signal level in order to avoid the burden of a huge number of models for overlapping events when the number of events is

not low. Both blind source separation techniques and beamforming techniques are employed to carry out a separation of acoustic sources that, though it is partial, allows to achieve a satisfactory detection accuracy. As real-time processing is a constraint put on the systems in our application, beamforming techniques will get a strong emphasys in our work. They will be developed to take advantage of the room setup. To integrate the scores from all the arrays, either a product rule based combination or fuzzy integral fusion is used in the decision block of the AED system. Moreover, within the same framework, the permutation problem is solved; since each hypothesized event is assigned to a given source position.

Since ASL is a very important task in the meeting-room acoustic scene analysis, a novel localization technique is explored in this thesis that can work with the sound recognition system. Most ASL approaches rely on some kind of measurement of the acoustic energy as a function of space, so the identity of the sound is not used at all. Once the identities and the endpoints of the simultaneous sounds are known, the proposed technique uses the statistical models of those sounds to compute a likelihood score for each model and each beamformer. Those scores are subsequently combined to find the MAP-optimal positions in the room. Although the proposed localization system can work alone, but focusing to a more integrated approach, an attempt is made to jointly recognize and localize several simultaneous acoustic events that take place in a meeting room environment, by developing a computationally efficient technique that employs multiple arbitrarily-located small microphone arrays.

1.2 Thesis objectives

The goal of this thesis work is designing efficient algorithms for acoustic event detection and localization in meeting-room environments. The thesis work contemplates several main research directions: multi-microphone array signal processing for source separation, sound-model-based event localization, combining assistive technologies like AED and ASL in a room environment, source ambiguity resolution, and classifier fusion at the decision stage.

Therefore, in this thesis work, one of the main objectives will be to develop a system that detects and identifies several possible overlapping acoustic events. The experimental work will be carried out in the particular case where there are only two possible acoustic sources, a meeting-room acoustic event and speech, which may be overlapped in time. The chosen methodology consists of performing source separation prior to the detection. Both beamforming and blind source separation techniques will be employed to carry out a separation of acoustic sources. Beamforming techniques will get a strong emphasis in our work, as real-time processing is the constraint that is put on the systems in our application. In addition, an optimal strategy for combining the processed signals acquired from the various arrays will be designed. Under the similar acoustic scenario, the performance of the proposed source separation based AED will be compared with the already implemented and baseline model based AED system.

On the other hand, once we have the output of the AED system, there is the need of a posterior identification of which one of the source positions given by the ASL system corresponds to speech and which one corresponds to the detected AE. That requires an additional position assignment system, which will also be developed in the thesis work.

Information regarding the source positions is necessary to analyze an acoustic scene. This particular task is performed by an additional ASL system. Most ASL approaches rely on some kind of measurement of the acoustic energy as a function of space, so the identity of the sound is not used at all. In this thesis work, a novel source localization technique will be proposed and developed which shows that the information about the content of the signals that are captured by distant and distributed microphones may be effectively used for localization of the sources in the space domain. In fact, instead of relying only on energy-like based

measures, a similarity measure delivered by a classifier is proposed. As the classifier uses models for the different sound classes like in the detection task, this method is called as sound-model-based (SMB) localization. The validity of proposed localization method will be verified in a scenario consisting of meeting-room acoustic events, either isolated or overlapped with speech.

The functionalities like detection, localization and separation are assigned to different sub-systems. However, it can be expected that an integrated approach, where all the functionalities are developed jointly, can offer advantages. Therefore, focusing to a more integrated approach, systems to jointly recognize and localize several simultaneous acoustic events will be developed. The system will also be developed with a computationally efficient technique that employs multiple arbitrarily-located small microphone arrays.

1.3 Thesis outline

The thesis is organized as follows.

Chapter 2 presents a state of the art in the area of acoustic event detection from the application point of view. It also reports one of the main problems, the current event detection systems has, is the temporal overlapping, and its different possible solutions. The source separation based approach, which is based on either beamforming or BSS and applied in different speech technologies, is briefly discussed. In this Chapter, a brief state of the art for acoustic source localization is also presented.

Chapter 3 describes the overlapped acoustic event detection system. A proposal of source separation based AED, which based on beamforming and BSS are presented. The possibility of using two different types of beamforming: frequency dependent, and frequency invariant is also discussed. The overall scheme of the system that could be used either for detection, or for localization, or for source ambiguity resolution, or jointly for both detection and localization is presented. A Fuzzy Integral based fusion approach to combine the several information sources at the decision stage for the improvement of the performance of the overall system is also discussed. In addition, the acoustic scenario and the databases, which will be used in the thesis work, are also presented.

Chapter 4 presents a system that resolves the source ambiguity when several acoustic sources are simultaneously active in a meeting-room scenario, and both the position of the sources and the identity of the time-overlapped sound classes have already been estimated is presented. The problem of source ambiguity is discussed at the beginning of the Chapter. Then a position assignment system for resolving that problem is presented. Combination of microphones at both the signal and the classifier level is also investigated. It is also verified how the inclusion of speech model in the classifier along with the AE models improved the performance of the system. The experimental results obtained with the available database, are reported. The performance of the system with two different types of beamforming is also investigated in this Chapter.

Chapter 5 presents the acoustic event classification and detection system based on beamforming based source separation assuming the source positions are known. A MAP based

classification and detection is described. Position assignment (PA) system for resolving the source ambiguity is also discussed within the same framework. In the experimental section, metrics for the evaluation of the above-mentioned system are described. Classification, detection and the position assignment results are also reported with the discussion.

Chapter 6 presents a new acoustic event localization system assuming the sources have already been detected, i.e. the identities and the end-points of the sources are known. A SRP-PHAT based source localization system, which is considered as a baseline one is described. The MAP criterion for the newly proposed localization technique is discussed. The metrics used for the evaluation of the localization systems is presented in the experimental section. The results with the proposed system and a comparison with the baseline SRP-PHAT based localization system are also reported.

In Chapter 7, a joint approach for detection and localization is presented. At the beginning, an integrated method for classification plus direction-of-arrival estimation is described. Then the joint method for classification and 2D localization is presented. In the experimental section, the metrics and the experimental results with discussion are presented.

Finally, in Chapter 8, conclusions are presented. The main achievements are summarized in this Chapter. Several promising future directions are also highlighted.

Chapter 2. Overlapped acoustic event detection

2.1 Chapter overview

In this Chapter, an overview of various concepts and techniques that will be the basis of the different works in this thesis is presented. In the initial sections, the state of the art for acoustic event detection is presented from the application point of view. In this context, the problem of signal overlapping in the current state of the art AED system is also discussed with some possible solutions. A brief overview of source separation based techniques that are mostly used in speech technologies including acoustic event detection, is presented. The microphone array signal processing technique like beamforming is discussed from application point of view. In the later part of this Chapter, a brief state of the art regarding acoustic source localization is presented.

2.2 Acoustic event detection

Acoustic Event Detection (AED) is the task related to identify different classes of sounds and their temporal positions. Initially, it was started as a task of segregating few audio sources in computational auditory scene analysis problem [**6**] [**7**] and segmenting an audio stream into a small number of acoustically compact categories [**8**]. Many of the existing contributions in AED are intended either for indexing and retrieval of multimedia documents [**9**] [**10**] [**11**] or to improve the robustness of speech recognition task [**12**] [**13**]. For an instance, in [**14**], ten key audio effects are taken into consideration: applause, car-racing, cheer, car-crash, explosion, gunshot, helicopter, laughter, plane, and siren. In [**15**], author proposed to detect violent events in feature films by analyzing environmental sounds such as gunfire, engines, and explosions. Extraction of semantic information with annotations of basketball multimedia document is presented in [**16**]. In [**17**], authors proposed a method for automatically extracting highlights of TV baseball programs. Audio segmentation (AS) in the broadcast news domain is a very specific application of AED, where the speech data exhibit considerable diversity, ranging from clean to noisy speech interspersed with music, commercials, sports etc. This line of research was started in [**18**], then some works for speech/music discrimination from radio stations in [**19**] [**20**], recognition of a broader set of acoustic classes in [**9**], difficult problem of speech music discrimination for singing segment [**21**], and mixed sound detection in the recent years shows that it is still a challenging problem [**22**] [**23**] [**24**]. Audio segmentation was one of the tracks in the recent Albayzin Evaluation campaign, held in 2010 and 2012, consist of segmenting a broadcast audio document (TV channels and radio) and assign labels for each segment indicating the presence of speech, music and/or noise [**25**] [**26**].

Detection of acoustic events has been carried out in other applications like monitoring and surveillance. For instance, AED is performed in living environments [**27**] [**28**], hospitals [**29**], kitchen-rooms [**30**], bathroom [**31**] [**32**], public places. In [**33**], the authors propose an activity recognition system based on detection of acoustic events caused by residents within the living space. The efforts have been made to assist interdependent people in aging society today, particularly to the elderly living alone at home [**33**] [**34**]. An example of event detection for an audio based surveillance system is presented in [**35**] , and detection and localization of

selected acoustic events in acoustic field for smart surveillance applications is presented in [**36**].

A specific case that have received some attention in recent years, mainly due to the work done in the context of the Computers in the Human Interaction Loop (CHIL) EU project [**37**] consists of the description of sounds that take place in a meeting-room environment. Within the context of ambient intelligence, AED applied to give a contextual description of a meeting scenario was pioneered by [**2**]. Moreover, AED has been adopted as a semantically relevant technology in CHIL European project [**37**], and in several evaluation campaigns CLEAR 2006 and CLEAR 2007 [**38**] [**39**] were international efforts to evaluate systems designed to recognize events, activities, and their relationships in interactive scenarios like lectures or meetings.

2.3 The problem of temporal overlapping

The detection of single acoustic event sources is performed with relatively high accuracy in the CHIL works [**1**] [**2**] [**37**]. However, in the CLEAR 2007 evaluation campaign [**1**], where acoustic event detection (AED) was carried out with meeting-room seminars, it became clear that time overlapping of acoustic events is a major source of detection errors (70%). The problem of acoustic overlaps is closely related to the ‘‘cocktail party” problem [**40**]. In that problem, one usually tries to separate one speech source from the others; however, in AED it could be to separate acoustic events from speech or one event from the others. The former case is very relevant in the context of meeting room AED, where one source could always be speech. Temporal overlaps are matter of research in other speech processing areas, like in speaker recognition or diarization. The problem of several simultaneous speakers has been considered in the NIST RT-09 [**41**] evaluation campaign, where the involved tasks (e.g. speaker diarization) have been evaluated on overlapped speaker segments as well. In fact, the overlap problem has recently gained a strong interest in speech processing [**42**] [**43**] . For instance, in [**44**], the authors propose several different features and investigate their effectiveness for detection of overlaps of two and more speakers. Also, some improvement in detection of speech overlaps for speaker diarization is shown in [**45**] [**46**] [**47**].

2.3.1 Different solutions in previous UPC work

The detection of overlapping events may be dealt with different approaches, either at the signal level, at the model level, at the feature level or at the decision level. In the UPC lab, although some attempts have been made to tackle the overlapping problem at the level of model, including additional video modality for the improvement of the detection of the overlapped events that have visual correlate, but not tackled at the signal level.

2.3.1.1 Model based approach

In [**3**], a model based approach was adopted for detection of events in a realistic meeting-room scenario with two sources, one of which is always speech, and the other one is a different acoustic event from a list of 11 pre-defined events. Thus, apart from the mono-event acoustic models, additional acoustic models were considered for each AE overlapped with speech, so the number of models was doubled (i.e. 22) as shown in Figure 1. That approach is used in the current real-time system implemented in the Universitat Politècnica de Catalunya (UPC)'s smart-room, which includes both AED and ASL [**48**]. In this approach, the AED system requires a model for each event class whether it is an isolated or an overlapped event. So there must be a different model for each possible combination of classes as depicted in Figure 1.

There is a limitation of this approach to be considered as a general solution of the overlapping problem. For the general case, the number of models required may be high since it depends on total number of event classes and the number of events which can exist simultaneously. Though this approach is feasible in some limited scenario, it may not be the case for other scenarios where the numbers of events, or the number of simultaneous sources, or both are large.

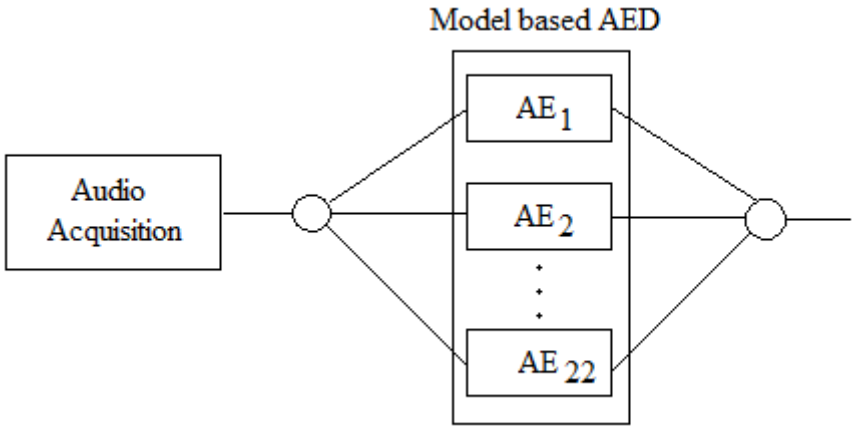

Figure 1: Model based acoustic event detection system

Although the estimated position of the sources is provided by an ASL system, there is still an ambiguity regarding the correspondence between identified events and located sources. As an example, in the most general case, AED have two detected events, i.e. *E* (one of the possible AE) and "sp", and two source positions are provided by ASL: P_1 and P_2. To have a complete description of the acoustic scene in the room, there is still the need of assigning each one of the two positions to each one of the two AEs. In this context, the system requires a posterior identification of which one of the source positions given by the ASL system corresponds to speech and which one corresponds to the detected AE.

2.3.1.2 Using extra modalities

In [**5**], multimodal techniques are used to deal with the problem of signal overlaps in the meeting-room signals. The video modality is used along with the audio modality. Two different strategies for fusion of audio and video modalities have been employed: feature-level fusion is based on concatenating feature vectors from different modalities into one super vector; and decision-level fusion, where each modality acts as an independent "expert", giving its opinion about the unknown acoustic event (AE). Decision-level fusion is carried out using weighted arithmetical mean (WAM) and fuzzy integral (FI) approaches. In [**4**] [**5**], a number of features are extracted from video recordings by means of object detection, motion analysis, and multi-camera person tracking to represent the visual counterpart of AEs. From the audio

perspective, the video modality has an attractive property: the disturbing acoustic noise usually does not have a correlate in the video signal. There are several video technologies which provide useful features for AED task. Person tracking features are very useful in this context. It is necessary that multiple cameras are employed to perform tracking of multiple interacting people in the scene. Some AEs are associated with motion of objects around the person like paper wrapping (under the assumption that a paper sheet is distinguishable from the background colour). Color specific motion descriptors, namely the motion history energy (MHE) and image (MHI), have been found useful to describe and recognize actions.

2.3.1.3 Using acoustic localization features

In order to enhance the meeting-room AED recognition rate, acoustic localization features are used in combination with audio spectro-temporal features [**4**]. In the case where the characteristics of the room are known beforehand, the position (*x*, *y*, *z*) of the acoustic source may carry useful information. Indeed, some AEs can only occur at particular locations, like door slam and door knock can only appear near the door, or footsteps and chair moving events take place near the floor. Based on this fact, a set of meta-classes (e.g. “near door” and “far door”, related to the distance of the acoustic source to the door, and “below table”, “on table” and “above table” meta-classes depending on the *z*-coordinate of the detected AE) are defined that depend on the position where each AE can be detected. The height-related meta-classes and their likelihood function modelled via Gaussian Mixture Models can provide useful information. These types of features could be included by using a standard localization system (e.g. SRP-PHAT).

Alternatively, the problem could be tackled at the signal level by separating the signals first before entering the AED system and describes in the following Section.

2.4 Source separation

Another possible solution to the overlapping problem is to have in the front end, a source separation technique that allows the system to separate the desired source from the other interfering sources or noise, followed by detection and recognition of the identity of each overlapped sounds. According to the type of source separation techniques, we can broadly

classify it into two categories. 1. Array processing based separation technique that exploits the information about the positions and orientations of the sources and sensors. 2. Blind source separation (BSS) technique, which separates the signal with or without the aid of the information about the source signals or the mixing process.

2.4.1 Array processing

2.4.1.1 Microphone arrays

A microphone array consists of multiple microphones located at different spatial positions. Using the spatial diversity and the basic principles of sound propagation, the multi-microphone signals can be combined to enhance or reject signals originated from a specific spatial direction. Therefore, it is possible to design the spatial filters using the microphone arrays, which are capable of enhancing a desired signal in the presence of multiple noise sources. Thus, microphone arrays have great potential as a pre-processing stage in many audio processing applications, due to their ability to provide far field audio acquisition and robustness against a wide variety of noises. However, in many cases, the properties of such kind of spatial filters are based on the array geometry and source locations.

The performance of audio processing technologies is affected by many acoustic disturbances. The audio signal recorded by microphones is degraded by the presence of other undesired acoustic sources, the effect of interferences, attenuation, channel distortion or multipath propagation. The impact of those disturbing factors can be augmented in the case of using far field microphones, where the desired signal can be severely attenuated, and significantly degrading the performance of audio technologies.

2.4.1.2 Acoustic disturbances: noise and reverberation

Depending on the application environment, different types of acoustic disturbances may be present. Small enclosures, like a car or an elevator, are characterized by low reverberation and additive noise with the microphones placed relatively close to the source. On the other hand, medium-size space has more adverse condition, since a high level of reverberation is present and the microphones are usually placed at several meters away from the source. This could be the case of an office or a meeting-room, where the main noises are produced by the

participants and the equipments, like computer fans and air-condition machine. In free space, the microphones are usually very close to the speaker, i.e. a cell phone or a door-phone, and reverberation is practically absent. This thesis is developed under the context of a smart-room environment. Therefore, reverberation and noise not related with the acoustic source of interest are expected to be present in such scenarios. A distributed microphones placed at arbitrary locations is employed to capture the signal. The microphones position is assumed fixed and known a-priori. A particular microphone array constellation is not necessary, although some classical approaches need a determined geometry. The microphone placement can have a great impact in the performance of some multi-microphone processing techniques.

Acoustic noise is generally referred to the undesired acoustic signals or disturbances, which is not produced by the acoustic source of interest that we want to detect. Depending on the scenario, the noises can be characterized differently [**49**]. Generally, they are broadly classified in three categories. 1. Non-coherent noise: the noise captured at different locations has little or no correlation. Also known as non-directional noise, it is normally related to thermal noise present at every microphone, which is non-correlated. 2. Coherent noise: A coherent noise (directional noise) consists in a noise coming from a single point that is not reflected by any surface, because multipath would increase the signal scattering, and the coherence function decreases [**50**]. 3. Diffuse noise: also known as homogeneous or isotropic, diffuse noise is common in car or room scenarios. The noise propagates in all directions and many reflections are produced. The coherence between microphones close to the noise source is high and it decreases with distance between microphones [**51**]. The coherence also decreases with increasing frequency.

The propagation of sound throughout multiple paths is a phenomenon commonly known as reverberation. The propagation of sound in closed space is influenced by reflections on the surfaces and scattering by the different objects. This creates indirect propagation paths and enables the multiple delayed versions of the signal reach to the sensors. The number of indirect paths is very high, tending to infinite, thus creating a sound field very similar to a diffuse sound source. However, audio technologies are particularly sensitive to reverberation and it constitutes a challenge to overcome.

2.4.1.3 Multi-microphone array signal processing

In such real and noisy environments, a single microphone captures noise and reverberation. This is one of the very common phenomena that influence almost all the technologies in the field related to speech processing. There are several de-noising algorithms. In most of the cases this is applying a time varying real gain to the short term frequency transformation of an audio frame extracted from the input signal. This class of algorithms is denoted as noise suppression. Various criteria for estimation of this suppression gain (rule) were derived historically: magnitude minimum mean square error (Wiener, 1947) [**52**] [**53**] [**54**] , spectral subtraction (Boll, 1979) [**55**], maximum likelihood (McAllay and Malpass, 1980), short-term MMSE (Ephraim and Malah, 1984), log MMSE (Ephraim and Malah, 1985), etc.

One of the solutions to capture better sound is to use more microphones. The signals from these microphones, combined in certain way, increase the directivity of the device and reduce the captured noise and reverberation. Using multiple microphones allows localization of the sound source and pointing the directivity pattern maximum towards the desired sound source. The concept of algorithmically steering the main lobe or beam of a directivity pattern in a desired direction is called beamforming. The direction the array is steered is called the look direction.

2.4.1.4 Beamforming

The array signal processing operates in a multidimensional space-time domain with the sensor arrays. And, the processor that combines temporal and spatial filtering using sensor arrays is the beamformer [**56**]. Obviously, spatio-temporal filtering is preferable over temporal filtering alone, because desired signal and interference often overlap in time and or frequency, but originates from different spatial position. One of the main purposes of using a microphone array signal processing is the design of spatial filters that able to enhance the desired source from the received signal, while attenuating all other acoustic disturbances. The spatial filtering method is known as beamforming when the signal enhancement is obtained as a linear combination of the signals captured by all microphones. The spatial filtering can be written as a weighted sum of the snapshot $x_q(n)$ and resulting in a beamformed signal $y(n)$:

$$y(n) = \sum_{q=1}^{Q} w_q^* x_q(n) = \mathbf{W}^H \mathbf{x} \tag{2.1}$$

where w_q is the weight for q-th microphone signal. The beamformer response pointing to a given position p at a concrete frequency can be computed as the product of the beamforming weights and the steering vector determined by p for that frequency as follows:

$$\left|H(f,p)\right|^2 = \left|\mathbf{W}^H \mathbf{s}_p\right|^2 \tag{2.2}$$

The task of beamforming then simplifies to find the optimal set of weights **W** that permits the enhancement of the desired sources and attenuation of the non-desired components. There exist many approximations and criteria for the beamformer design, but they can be classified broadly into two: 1) data independent methods, and 2) statistically optimum methods.

Data independent methods design the fixed beamformer patterns independently on the arriving signals according to spatial restrictions or other kinds of restrictions. This can be the case when the position of the source of interest is known a-priori, like in a car or room environment, or when we want to attenuate the noise sources whose positions are already known. There are various schemes to select the weights of the beamformer, each with its own characteristics and limitations. A conventional beamformer is a simple one, sometimes referred as the delay-sum beamformer, with all its weights of equal magnitudes. The phases are selected to steer the array in a particular direction, known as the look direction. The mean output power of the conventional beamformer steered in the look direction is equal to the power of the source in the look direction. In the multiple source scenarios, a technique called null steering beamforming is used to cancel the plane waves arriving from the known directions and thus produce nulls in the response pattern by steering only towards the direction of arrival (DOA) of the desired target source. In general, spatial restrictions can be introduced in the design of the beamformer weights by simply imposing a desired response to concrete directions. The maximum number of restrictions is determined by the number of sensors of the array. In that case, we would like to enhance the response to the signal coming from position

p_0 while also attenuate the sound arriving from *Ns* noise sources at $p_1,\ldots,p_{Ns.}$ The expression with this kind of implementation can be written:

$$\mathbf{W}^H[s_{p_0}s_{p_1}.....s_{p_{N_s}}]=[10.....0] \tag{2.3}$$

or in terms of matrix notation:

$$\mathbf{W}^H\mathbf{S}=\mathbf{e}^T \tag{2.4}$$

Assuming that the noise sources are uncorrelated, that is the noise cross-covariance matrix $\mathrm{R}_n=\sigma\mathrm{I}$, the solution of Eq. 2.4 can be obtained by conditional minimization:

$$\mathbf{W}=\mathbf{S}(\mathbf{S}^H\mathbf{S})^{-1}\mathbf{e} \tag{2.5}$$

This beamforming technique is also known as multiple sidelobe cancellers (MSC) [**56**]. The implementation of this type of beamformer depends on several factors like the array geometry, number of microphone in the arrays, the narrowband or broadband considerations of the signal to be processed. With sufficient number of microphones in a given array geometry, narrowband signals are well suited for the processing by this type of beamformer in time domain. In other words, this type of narrowband assumptions puts frequency dependent limitations on the characteristics of beamformers. Basically, the delay-sum beamformer belongs to a more general class called filter-sum beamformers, in which both the amplitude and phase weights are frequency dependent. The directivity pattern of such kind beamformer with uniformly spaced sensor array depends on the factors like frequency of interest, the inter-element spacing and the number of elements in the array. The dependency in the operating frequency means that the response characteristics (beam width, sidelobe level) will only remain constant for narrow-band signals. For wideband applications, it means that a single linear array is inadequate if frequency invariant beam-pattern is desired. To deal with a wideband signal like speech, one possibility is to process the signal in the frequency domain [**57**]. Hence, the problem of frequency dependency can be overcome by converting the time domain signal into frequency domain and using the narrowband beamforming for each frequency bin [**58**] [**59**] [**60**]. In some applications like in [**61**] [**62**], processing of different frequency bands in different harmonically nested sub-arrays makes the whole system to be

used for broadband applications. Each sub-array is designed as a narrow band beamformer applied to a concrete frequency band typically of one octave. The lowest frequency band is processed by largest sub-arrays and highest frequency band is processed by smallest sub-arrays, in this way reducing the beampattern variation.

In the case of distributed microphone arrays in the room environment like UPC's smart room, the spatial beam-patterns of the array beamformer are expected to be highly dependent on the relative positions of the active sources and the sensors in the room [**63**]. For a total acoustic scene analysis, the combination of the contributions from all microphone arrays is necessary. This type of combination at the signal level is described in [**64**].

The performance of ASR systems in a room environment with distant microphones is strongly affected by reverberation. As the degree of signal distortion varies among acoustic channels (i.e. microphones), the recognition accuracy can benefit from a proper channel selection. Such an attempt is made in [**65**] [**66**], which proves beneficial to the recognition task. Basically channel selection for automatic speech recognition aims to rank the signals according to their quality. To create such ranking, in [**66**] [**67**], authors propose to use posterior probabilities estimated from the N-best hypothesis of each channel.

Practically, microphone array beamforming have to face the problem of multipath propagation due to early reflections and reverberations, wideband signals, non-stationary environments and with the restrictions in array geometry, size and number of sensors due to design constraints [**68**]. In such kind of situations, a frequency invariant beamforming is attempted in [**69**], which uses a numerical approach to construct an optimal frequency invariant response for an arbitrary array configuration with a very small number of microphones, and it is capable of nulling several interferent sources simultaneously. The method first decouples the spatial selectivity from the frequency selectivity by replacing the set of real sensors by a set of virtual ones, which are frequency invariant. Then, the same array coefficients can be used for all frequencies. A frequency invariant beamformer in the subbands using least square approach is designed [**70**].

Conventional and null-steering beamformer often requires the knowledge of the directions of target and interference sources, and the output signal to interference noise ratio (SINR) is not maximized by the estimated weights [**71**]. Optimal beamformer overcomes

some limitations based on the adaptive estimation of optimum mean square error (MSE) parameters on the available dataset. The weights are selected by minimizing the mean power of the processor while maintaining the unity response in the look direction of the desired source. The constraint included takes care that the signal remains undistorted. Alternatively, this adaptive array processing can be divided in two categories: element-space processing and beam-space processing.

In element-space processing, the signals derived from the elements are weighted and summed to produce array output. The beam-space processing is a two stage scheme where the first stage takes the array signals as input and produces a set of multiple outputs. These outputs are then weighted and combined to produce the array output. In general, the beam space processing arrays are used in situations where the number of interferences is much less than the number of elements. They offer computational advantage over the element-space processing array. For element-space processing, the constraints on the weights are imposed to prevent the signal that arrives from look direction to be distorted. It makes the array designing more robust. Beam-space processors are reported to produce superior performance in the presence of look direction errors. Beam-space processors have been studied under many different names, including Howell-Applebaum array [**72**], Generalized sidelobe canceller [**73**] [**74**]. One of the advantages of the beam-space processing is that the number of degrees of freedom of an array that are used to achieve adaptivity are proportional to the number of unwanted signals rather than the number of array elements. In the absence of errors, both processing schemes yield identical results. Both Applebaum and Least mean squares (LMS) adaptive array techniques based on statistical methodologies were applied to many applications that iteratively null the jammers. The LMS array relies on a locally generated reference signal to guide feedback loops that generate weights. The Applebaum array uses a steering vector to control the feedback loops and does not require reference signal. The difficulty with this method is that the designer must know where to point the beam. Thus a priori knowledge of arrival angle of the desired signal is required. When this angle is not known, one can choose the Applebaum array in the form of power inversion array, where steering vector turns one element on and the rest off. The turned on element is so chosen that its pattern covers some large sector of space from which desired signals may arrive.

To deal with the problem of non-stationarity, adaptive beamformers are very useful [**75**]. This kind of beamformers optimally extract non-stationary desired signals from the non-stationary interference for a given array geometry. In this context, it is worth to mention that beamformers can partially dereverberate the desired signal, although reverberation is strongly correlated with the desired signal. The beamformer generally increase the power ratio between direct path and reverberation.

2.4.1.5 Some useful beamforming applications in speech technologies

Many microphone array based speech recognition systems have successfully used delay-and-sum processing to improve recognition performance, and because of its simplicity, it remains the method of choice for many array based speech recognition systems [**68**] [**76**] [**77**]. The delay-and-sum beamformer can be generalized to a filter and sum beamformer where rather than a single weight, each microphone signal has an associated filter and the captured signals are filtered before they are combined. Both the delay-and-sum and filter-and-sum methods are examples of fixed beamforming algorithms, as the array processing parameters do not change dynamically over time. If the source moves then the delay values will of course change, but these algorithms are still considered fixed parameter algorithms.

In adaptive beamforming, the array-processing parameters are dynamically adjusted according to some optimization criterion, either on a sample-by-sample or on a frame-by-frame basis. In some cases, the filter parameters can be calibrated to a particular environment or user. For example, a calibration scheme is designed for a hands-free telephone environment in an automobile [**78**]. A series of typical target signals from the speaker, as well as jammer signals from the hands-free loudspeaker, are captured in the car and used for initial calibration of the parameters of a filter-and-sum beamforming system. These parameters are then adapted based on the stored calibration signals and updated noise estimates.

In [**74**], robustness of the adaptive technique (GSC) was improved by reducing the signal cancellation that results from tracking errors. These adaptive methods are very handy and used in many speech processing applications. As an example, in [**79**], efforts have been made to improve the speech recognition task in the environments where close-talking

microphone is not available and the speech is captured by an array of microphones located some distance from the user.

A filter-and-sum array-processing algorithm called Likelihood Maximizing Beamforming (LIMABEAM) in which the filter parameters are optimized in a data-driven fashion using the statistical models of a speech recognition system is developed in [**80**] . The optimization of the filters parameters is formulated as a maximum-likelihood parameter estimation problem. In addition, a subband filtering approach to LIMABEAM, called Subband-Likelihood Maximizing Beamforming (S-LIMABEAM) is also investigated in [**81**] . In this algorithm, processing is performed in the DFT domain, treating each DFT coefficient over time as a time-series.

Noise suppression and echo cancellation in speech technologies are used in different environments like in car, air cockpit etc. A speech reinforcement system for car cabin communication is proposed in [**82**] . The system uses a set of acoustic echo cancellers, echo suppression filters and noise reduction stages based on Wiener filters for its implementation. A combined fixed/adaptive beamforming for speech recognition in car environments had been studied in [**83**].

2.4.2 Blind source separation

Blind source separation is generally a technique for estimating original sources from observed mixed signals without a-priori information about the source signals, hence the term blind. It has been an emerging field of interest due to a number of interesting applications in audio, speech, image signal processing and many other applications. Many different approaches have been attempted by numerous researchers using artificial learning, higher order statistics, minimum mutual information, beamforming based adaptive signal separation and noise cancellations, each claiming various degrees of success.

The blind source separation problem was viewed from an information theoretic framework in [**84**]. They also demonstrated the separation and the deconvolution of mixed sources. But their adaptive methods were different from the cumulant based cost functions [**85**].

But the common factor between most of the BSS techniques is that they rely on statistical independence as its separation criterion. Independence, as the term suggests, is about making the signals statistically independent by reducing mutual information between the signals. The author in [**86**] described that the joint approximate diagonalization of eigen-matrices (JADE) is equivalent to informatic approaches. The contrast function used in their work is effectively a measure of mutual information between the cross-cumulants of signals.

Principal component analysis (PCA) is a second order statistical method that decorrelates the data and reduces the dimension of the problem, but does not achieve full independence [**87**]. In this context, independent component analysis (ICA) has the capacity to make the signals fully independent, since it is a higher order statistical method [**84**]. Sometime, combination of the PCA with ICA is used for BSS [**88**]. The reason for combining them is to reduce the dimension of the problem before implementing the ICA algorithm, thus making it easier for the ICA algorithm to solve the problem. In such cases, the PCA is implemented using a recursive method and the information maximization is implemented for the ICA.

In [**89**], the authors were the first to capture higher order statistics by decorrelating nonlinear transformations of the signals. The authors in [**90**] [**91**], presented a simple algorithmic architecture which in effect performs density estimation and is based on prior knowledge of the cumulative density function of the source signals. The author made modifications to the updated equations to dramatically improve convergence and computational costs [**92**].

In case of convolutive mixtures of the signals, general broadband approach based on second order statistics is presented in [**93**]. The proposed algorithm avoids several known limitations of the conventional narrowband approximation, such as internal permutation problem. The higher order statistics generally produces better separation [**94**] [**95**]. It has been proposed and implemented a higher order cumulant based contrast function for the separation problem. The authors considered the separation problem for the convolutive mixtures in the case where non-Gaussian source signals are not necessarily the filtered version of independent and identically distributed sequence. The separation is done through a deflation method, where sources are extracted one by one.

Some work in this area was inspired from semi-blind approaches. As in [**96**], a new contrast function that makes use of reference signals is proposed. The main advantage of this approach consists in the quadratic form of the criteria; the extraction of one source thus reduces to a simple optimization task, for which comparatively fast and efficient algorithms are available. The separation of the other sources from the mixtures is then carried out by an iterative deflation method.

In case of multichannel blind signal processing for the convolutive mixtures, a 'Triple-N ICA for convolutive mixtures' (TRINICON) framework is presented in [**97**]. It deals with both blind source separation and multichannel blind deconvolution (MCBD). It is based on the use of multivariate pdfs and a compact matrix notation, which considerably simplifies the representation and handling of the algorithms. In that work, the authors exploited three fundamental signal properties: non-whiteness, non-gaussianity and non-stationarity by using a information theoretic cost function.

The blind source separation techniques are expected to be computationally complex. Interestingly, a real-time broadband convolutive BSS has been implemented [**98**]. It uses the same algorithms like in [**97**], which is based on second order statistics and a block-on-line update method for de-mixing filters. The system is tested in a reverberant room with moving speakers. In [**99**], a recursive method for BSS or/and for ICA is proposed. The relation of this recursive method of BSS/ICA with the conventional gradient-based method is quite similar to the relation of the RLS method with the LMS method in adaptive filtering.

2.4.3 Time-frequency sparseness

Time-frequency sparseness based source separation has been very popular research area for source separation. In [**100**], a time-frequency sparseness based technique that uses BSS of a time-varying number of moving sources is presented. They develop two online algorithms based on time-frequency sparseness and are able to track and separate moving sources in mildly reverberant environments in real-time. The advantage of this approach is that it does not require the number of sources or an initialization of DOAs. In [**101**], the authors have proposed a spectrogram image features, based on the visual signature extracted from the

sound's time-frequency representation for sound event classifications in the mismatched conditions.

Overlapping sound event recognition using local spectrogram features with the generalized Hough transform (GHT) is attempted in [**102**] [**103**]. The local features are extracted from the interest-points in the spectrogram. Then the features are clustered for each sub band and a general temporal distribution function for each feature cluster is generated. These are called codebook clusters. While recognizing, matching of the codebook entries are done using GHT which detects shapes in the image through a voting procedure. In the recognition, the matching of the features is done through a temporal distribution function which acts as a voting function to perform GHT.

An underdetermined blind source separation (BSS) using a compressed sensing (CS) approach is proposed in [**104**] . It contains two stages. In the first stage, a modified K-means method is used to estimate the unknown mixing matrix. The second stage is to separate the sources from the mixed signals using the estimated mixing matrix from the first stage. In the second stage a two-layer sparsity model is used. The two-layer sparsity model assumes that the low frequency components of speech signals are sparse on K-SVD dictionary and the high frequency components are sparse on discrete cosine transformation (DCT) dictionary. This model, taking advantage of two dictionaries, can produce effective separation performance even if the sources are not sparse in time-frequency (TF) domain. A framework for separating and reconstructing multichannel speech sources from compressively sensed linear mixtures is explored in [**105**]. The conventional approaches for blind speech separation are almost based on the Nyquist sampling theory. The author proposed an approach which uses the multichannel compressive sensing theory for blind speech separation. The linear programming and gradient based methods are used to separate the sources and has a lower computational complexity. Compared with the conventional blind speech separation, the proposed approach can reduce the requirements of sampling speed and operating rate of the devices.

2.4.4 Non-negative matrix factorization

Among the other popular techniques applied in the noisy or overlapped audio recognition, non-negative matrix factorization (NMF) is the one that produces successful separation and

could assist other tasks like automatic music transcription, object coding, special sound effects, acoustic scene analysis etc.

In [**106**], the author presented a framework featured by two approaches: un-directed and directed NMF model. The un-directed NMF model decomposes the mixing data in an unsupervised manner but requires human interaction for clustering. The directed NMF is performed under the direction of pre-trained models and provided with isolated training data, does not need any user interaction.

In [**107**], the authors used a convolutive NMF based approach for detecting and modeling the acoustic events. As NMF is useful for parts based decompositions of data; in this work authors have used to discover a spectro-temporal patch bases that best describes the data, with the patches corresponding to the event like structures. Then the features are derived from these activations of the patch bases. They performed the AED task with 16 meeting room acoustic events in the presence of added noise.

A sound event detection system that uses a sound source separation in the front end for natural multisource environments is proposed in [**108**]. The audio is pre-processed using NMF and is capable of separating upto four individual signals from the overlapping events. The NMF is calculated by minimizing the Kullback-Leibler divergence between the original spectrogram and a reconstructed spectrogram. The magnitude spectrum of the time domain signal calculated before performing NMF. After factorization, the complex spectrum is reconstructed and then converted back to the time domain signal.

A multichannel high resolution NMF for modelling convolutive mixtures of non-stationary signals in the time-frequency domain is explored in [**109**]. The authors presented a unified probabilistic model called HR-NMF that permits to overcome this limit by taking both phases and local correlations in each frequency band into account. This model is estimated with a recursive implementation of the Expectation-Maximization (EM) algorithm.

In [**110**], author proposed an algorithm for separating monaural audio signals by non-negative tensor factorization of modulation spectrograms. The modulation spectrogram is able to represent redundant patterns across frequency with similar features, and the tensor factorization is able to isolate these patterns in an unsupervised way. The method overcomes the limitation of conventional non-negative matrix factorization algorithms to utilize the

redundancy of sounds in frequency. In the proposed method, separated sounds are synthesized by filtering the mixture signal with a Wiener-like filter generated from the estimated tensor factors.

2.5 Classifiers in the AED task

Any recognition task requires a classification. The task of classification is to provide a label for an unseen input pattern. However, a poor feature processing can hardly be compensated by a good classification. One of the very first works on audio classification used a minimum distance classification model, i.e. a simple distance based classifier with the Euclidean distance between extracted features [**111**]. The minimum distance classifiers choose a class according to the closest training sample. A little more complex algorithms determine the k-nearest neighbours to an unknown input, and then they choose the class that is most represented by them. In that case, classification becomes very complex when a lot of training data is used, as one must measure a distance to all the training samples. By using clustering and storing only the centers of the clusters (class prototypes) the computational efficiency can be improved. The mentioned algorithms and other related optimization steps for audio classification have been reviewed in [**112**].

A rule based classification algorithm that initially also relies on a good feature extraction has been used in [**113**]. In that work, several task-specific features have been proposed with a set of heuristic classification rules.

Among other classification paradigms, a way to classify audio data consists to use already developed and well-tested speech recognition algorithms. In ASR, usually GMMs and HMMs are used. They are well suited to work with time series data, and to deal with the information included in the temporal evolution of the audio signal. Many audio recognition works have exploited the mentioned techniques. GMMs have been used in [**112**] [**114**], and HMMs in [**4**] [**115**] [**116**] [**117**] [**118**] [**119**].

Instead of using generative classification models as GMM, discriminative classification models have been used in a number of works, like ANN in [**87**] [**120**] [**121**], linear vector quantization in [**122**], decision trees in [**123**], support vector machines in [**2**] [**124**]. Most common practice is to use supervised training for event recognition. However, in [**125**],

authors presented a semi-supervised learning which helps in sound event classification. They showed that adding unlabelled sound event data to the training set based on sufficient classifier confidence level after its automatic labeling level could significantly enhance classification performance. Like in speech recognition, the authors in [**126**], developed a technique for detecting signature audio events which is based on identifying patterns of occurrences of automatically learned atomic units of sound called acoustic unit descriptors.

2.5.1 Missing feature techniques

Missing features technique is a method to deal with the problem of noisy or overlapped audio recognition. Speech recognition systems perform poorly in the presence of corrupting noise. Missing feature methods attempt to compensate for the noise by removing noise corrupted components of spectrographic representations of noisy speech and performing recognition with the remaining reliable components.

Conventional classifier-compensation methods modify the recognition system to work with the incomplete representations so obtained. This constrains them to perform recognition using spectrographic features, which are known to be less optimal than cepstra. In [**127**], two missing-feature algorithms that reconstruct complete spectrograms from incomplete noisy ones are presented. Cepstral vectors are then derived from the reconstructed spectrograms for recognition. One algorithm uses MAP procedures to estimate corrupt components from their correlations with reliable components. The other one clusters spectral vectors of clean speech. Corrupt components of noisy speech are estimated from the distribution of the cluster that the analysis frame is identified with.

In [**128**], the authors presented an automatic speech recognition system that uses a missing data approach to compensate for challenging environmental noise containing both additive and convolutive components. The unreliable and noise corrupted “missing” components are identified using a Gaussian mixture model (GMM) classifier based on a diverse range of acoustic features. To perform speech recognition using the partially observed data, the missing components are substituted with clean speech estimates those are computed using both sparse imputation and cluster based GMM imputation. Compared to two reference mask estimation techniques based on interaural level and time difference-pairs, the missing

data approach in that work significantly improved the keyword accuracy rates in all signal-to-noise ratio conditions when evaluated on the CHiME reverberant multisource environment corpus. Of the imputation methods, cluster based imputation was found to outperform sparse imputation.

The conventional missing features techniques require either the substitution of spectrogram elements (imputation) or the classifier modification (marginalization) [**54**]. But in [**129**], a noise robust cepstral feature called missing feature linear-frequency cepstral coefficients (MF-LFCC) is used that transforms both clean and noisy signals into a similar representation. They have used a computer vision technique called blob detection to separate the reliable signal from the noisy or unreliable background in the time-frequency domain.

2.6 Acoustic source localization

Most of the existing indoor acoustic source localization systems in the state of the art are based on the time difference of arrival (TDOA) of the waves captured by microphones placed at different locations. Considering the relative slow speed of sound in air (around 342 m/s) in indoor environments, the localization systems rely on these small differences between signals. This kind of approach faces three problems: detecting the number of sources at a given time (detection), gathering the corresponding position of every detected source (localization), and estimating their trajectories over time (tracking). The small time differences are determined by the source position and the microphone constellation. These differences are usually obtained by the following methods in a nonexclusive form. 1) Time delay: the time needed by the signal to reach every microphone is different. It is possible to reconstruct the source position, assuming the exact location of the microphones. This approach does not require a prior knowledge about the environment, and therefore is the preferred option in most of the practical applications. Usually omnidirectional microphones are employed. 2) Impulse response: for every source position and microphone location, the signal travels across different channels, which has a varying effect in the impulse response of every microphone. A model based approach, which consists of creating a complete map of the impulse response for every location of the meeting room and each microphone, enables the possibility to estimate the source position. The Multiple Inputs Multiple Outputs (MIMO) channel identification task

[**97**] [**130**], tightly related to the BSS task, aims at estimating the impulse responses online without any calibration step. This approach would permit the localization and the separation of the signals of multiple acoustic sources present in the room. This task is still open to research, with preliminary tests published in [**131**]. 3) Microphone channel: recently has been proposed to use directional microphones placed in the same position, but pointing to different directions [**132**]. In this case, it could be possible to take advantage of the knowledge of the microphone direction-dependent response to reconstruct the speaker position. This kind of approach may be combined with the previous ones.

2.6.1 Time difference of arrival estimation

The approaches to TDOA estimation based on the cross-correlation between pairs of microphones are the most common and popular in speech applications. In the particular case of an scenario with multiple acoustic sources and only one pair of microphones available, an alternative technique to estimate multiple time delays is the Adaptive Eigenvalue Decomposition Algorithm [**133**]. However, in the general case of multiple microphone pairs and multiple acoustic sources (or multiple reflections), the task of relating the numerous time-delays provided by each microphone pair with the exact position of every source becomes non-obvious.

2.6.1.1 Techniques based on correlation

Consider a room provided with a set of N microphones from which we choose M microphone pairs. Let p denote a position in space. Considering the speed of sound is v_s, the time delay of arrival $\tau_{i,j}$ of an hypothetical acoustic source located at p between two microphones i, j with position m_i and m_j is:

$$\tau_{i,j} = \frac{\left\| p - m_i \right\| - \left\| p - m_j \right\|}{v_s} \tag{2.6}$$

The cross-correlation function is a measure of the similarity between signals for any given time displacement and ideally it should exhibit a prominent peak in correspondence to

the delay between the pair of signals [**134**]. The presence of disturbing factors such as noise or path differences from the audio source to the microphones can considerably mask this peak and the effect of reverberation is specially harmful for the quality of the recorded speech [**135**]. In order to increase robustness against these factors, the cross-correlation function is usually weighted attending to different optimality criteria, in what is named Generalized Cross-Correlation (GCC) [**136**]. It can be expressed in terms of the inverse Fourier transform as follows:

$$R_{ij}(\tau) = \int_{-\infty}^{\infty} \psi(f) G_{i,j}(f)\, e^{j2\pi f \tau}\, df \tag{2.7}$$

where $G_{i,j}(f) = M_i(f) M_j^*(f)$ is the estimated cross power spectrum and $\psi(f)$ is the weighting function. And the TDOA is estimated:

$$\hat{\tau}_{ij} = \arg\max_{\tau} R_{ij}(\tau) \tag{2.8}$$

GCC has different properties depending on the weighting function. Some weighting functions can improve the TDOA estimation against reverberation, while others offer more immunity against additive noise. As a result, weighting functions should be adapted to the environmental conditions of the corresponding application. In environments that are characterized by the presence of high additive noise, whose spectrum is known, a function widely used in audio processing is the Maximum Likelihood (ML). One of the main drawbacks of this technique is that in many applications it is not possible to know the noise spectrum beforehand. On the other hand, reverberation is the most harming phenomenon present in room applications and the ML function does not provide any special robustness against reverberant conditions.

A widely used weighting function in acoustic source localization is the Phase Transform (PHAT), also known in the literature as cross power-spectrum phase (CSP) [**137**] [**138**] technique that is usually considered useful in reverberant scenarios and defined as:

$$\psi(f) = \frac{1}{\left| M_i(f) M_j^*(f) \right|} \tag{2.9}$$

The magnitude of the PHAT-weighted cross-power spectrum is constant and equal unity for all frequencies, and therefore, all time-delay related information lies within the phase. This has a normalizing effect in the cross-correlation function, being GCC-PHAT directly comparable between microphone pairs independently of the microphone gains, which avoid the need for tedious microphone calibration. On the other hand, the PHAT weighting is very convenient since it does not depend on the noise spectrum and its whitening effect on the cross-power spectrum removes from the cross-correlation function the periodicities produced by the voice autocorrelation. Finally, the main strength of GCC-PHAT is its inherent robustness against reverberation by giving equal importance to every frequency independently of the signal, under the assumption that the signal to reverberant ratio is constant for all the frequencies. For each pair of sensors, the loci of all points, which have the same TDOA, lie in the surface of a hyperboloid. An estimate of the 3D source location is then given as the location which best fits the potential source surfaces across all sensor pairs, which leads to the minimization of an over determined and nonlinear error function. There are many works in the literature for finding the best fit. Some of them are explained in [**139**] [**140**] [**141**] [**142**] [**143**]. A ML approach can be found in [**144**] based on the TDOA estimation from *M* different microphone pairs, defined as:

$$E(p) = \sum_{m=1}^{M} \frac{\left| \hat{\tau}_m - \tau_m(p) \right|^2}{\sigma_{\hat{\tau}_m}^2} \tag{2.10}$$

where $E(p)$ is the ML error-function, $\tau_m(p)$ is the theoretical TDOA from a potential source at position p to the microphone pair m, and $\hat{\tau}_m$ and $\sigma_{\hat{\tau}_m}^2$ are the estimated TDOA and it variance at the *m-th* microphone pair. The most likely position for a given set of TDOA consists of the minimization of:

$$\hat{p} = \arg\min_{p} E(p) \tag{2.11}$$

For real-time applications, closed-form estimators that approximate the minimization problem to a suboptimal solution at a very low computational cost have been developed. Triangulation is the simplest solution [**145**], however, it is difficult to take advantage of multiple sensors and the TDOA redundancy. In general, most closed-form algorithms employ a least-squares approach to construct the error function based on the multiple TDOAs. The minimization of different error functions result in many different estimators with varying performance and complexity. Examples of closed-form estimators are plane intersection [**146**], spherical intersection [**147**], spherical interpolation [**142**], etc. In general, closed-form algorithms make no additional assumption about the distribution of measurement errors, making them very sensitive to errors in the TDOA estimation.

2.6.2 SRP-PHAT technique

In real scenarios dealing with high levels of reverberation and distant sources, the performance of the most of the above-mentioned techniques are very poor and SRP based approaches are better suited. In the last years, much research efforts have been devoted to develop direct approaches, able to estimate the source position in a single step. Algorithms in this class commonly aim to maximize, or minimize a given objective error function either iteratively or exploring over a grid of predefined locations. Direct approaches have not been often considered in past years because of their high computational demands. However, recently there is a growing interest in this type of approaches, since the increasing availability of computing power makes some solutions feasible. Some iterative algorithms like the simplex method presented in [**148**] try to solve the ML error function 2.10. Unfortunately, iterative methods suffer from the drawback of needing an initial estimate close to the real source position in order to avoid instability problems, reaching local minima, or slow convergence.

The simplest example of exploration algorithm is to scan the set of predefined possible locations with a delay-and-sum beamformer [**149**]. Signals received at each sensor are aligned according to the theoretical propagation delays to each location, then compute the energy of the sum of all aligned signals. When the beamformer focuses near the real source position, the energy of the combined output signal presents a maximum peak. This type of exploration method is known as the Steered Response Power (SRP) algorithm. First research works on this

field were developed by Alvarado when he introduced the concept of power field in [**150**]. However, in noisy and reverberant environments the energy is not a reliable feature. Silverman and Kirtman introduced in [**151**] an alternative method to power field which employs the cross-correlation function in a two-stage search algorithm. Later, the Global Coherence Field (GCF) was described in [**137**] and proposes to use a steered beamformer based on a coherence measure instead of power. More recently, a particular implementation of GCF based on GCC-PHAT function and referred to as SRP-PHAT was investigated by DiBiase in [**152**]. Among the all methods based on steered beamforming (SB) [**153**], the most popular is the SRP [**150**] [**154**] [**68**]. The SRP-PHAT algorithm performs very robustly in reverberant environments due to the PHAT weighting, and actually, it has turned out in one of the most successful state-of-the-art approaches to microphone array sound localization [**136**] [**155**].

A localization technique based on ML using a noise model, which is closely related to SRP, has been reported in [**156**]. In [**157**], a source localization method is presented using the generative model based fitting with sparse constraints. The generative model is defined to explain the acoustic power maps obtained by the SRP strategies. An optimization approach is then used to fit the model to real input SRP data and estimate the position of the acoustic source. While optimizing, sparse constraints in the parameters of the model are included, enforcing the number of simultaneous active sources to be limited. In addition, a subspace analysis is used to filter out portions of the input signal that cannot be explained by the model.

Speaker localization and the orientation in a multimodal and smart-room environment is carried out in [**158**]. The work describes the development of a robust speaker tracking system based on the audio signals captured by a set of distributed microphones in UPC's smart-room. The location estimates gathered by the acoustic localization algorithms are usually contaminated by spurious measurements due to noises or reflections of the voice with adjacent objects. Two approaches to filter the noise-corrupted location estimates according to a motion model to obtain a reliable smooth track of the acoustic sources are used based on the Kalman filter and sequential Monte Carlo methods. The acoustic localization and tracking algorithms have been adapted to other speech technologies like speaker identification, speaker diarization and acoustic event detection.

Generally, the SRP based techniques use computationally intensive grid search methods to find a global maximum. Attempts to reduce the computational load of the exhaustive peak search have also been investigated, like stochastic region contraction [**159**], coarse-to-refine search [**160**]. A computationally viable SRP has been proposed in [**161**], where an inverse mapping is introduced that maps the relative delays to a sets of candidate locations. In [**162**] [**161**] [**163**], the authors discussed the computational issues and proposed optimization methods so that the SRP based techniques can be implemented in real time.

2.6.3 Multiple source localization

Indeed, the localization of simultaneously active sources with the energy or energy-like based approaches has to face a stronger challenge than for the single source case. A method for tracking the positional estimates of multiple talkers in the operating region of a microphone array has been presented in [**164**], where initial talker location-estimates are provided by a time-delay based localization algorithm. These raw estimates are spatially smoothed by a Kalman filter derived from a set of potential source motion models. Data association techniques based on the estimate clustering and source trajectories are incorporated to match location observations with individual talkers.

Multiple source localization in the reverberant environment is attempted in [**165**]. The approach is based on a disturbed harmonics model of time delays in the frequency domain and employs the well-known ROOT-MUSIC algorithm, after a preliminary distributed processing of the received signals. Candidate source positions are then estimated by clustering of raw TDOA estimates.

For single sound source localization, the CSP (cross-power spectrum phase analysis) method has been widely used. However, when localizing multiple sound sources, the CSP method has a problem that the localization accuracy is degraded due to cross-correlation among different sound sources. To solve this problem, in [**166**], the authors propose a new method which suppresses the undesired cross-correlation by synchronous addition of CSP coefficients derived from multiple microphone pairs. Experimented in a real room environment, they have showed that the method improves the localization accuracy when increasing the number of the synchronous addition.

A method for locating multiple sound sources using only a local segment of data from a large-aperture microphone array is proposed in [**167**]. The proposed method employs the proven robust SRP-PHAT as a functional, then an agglomerative clustering, and a low-cost global optimization (stochastic region contraction). Of course, these approaches rely on some kind of measurement of the acoustic energy as a function of space, and require additional methods with some optimization for localizing sources in a multiple source environments. But, the identity of the sound is not used at all.

2.7 Chapter summary

In this Chapter, we have quickly reviewed the work done so far in the area of acoustic event detection from the application point of view. In that context, the temporal overlapping problem is mentioned and how this problem affects the different areas of speech technologies has been presented briefly. Some of the possible solutions of the overlapping problem are also mentioned. Source separation prior to detection could be a reasonable one. Then, a literature review of different source separation techniques and their applications in different speech technologies has been presented. In the later part of this Chapter, a brief state of the art of acoustic source localization has been presented. A short review of ASL in multi source environment is also presented.

Chapter 3. Basic techniques used in this thesis

3.1 Chapter overview

The detection and localization of acoustic events in a room environment has to face the challenge of overlapped sounds, i.e. sounds that occur simultaneously. The detection problem can be tackled by carrying out some kind of source signal separation followed by detection (that includes identification) of each of the overlapped sounds. In this Chapter, either blind source separation based on the deflation method or beamforming is used for signal separation. The beamformers are designed with two or three microphones per array. An alternative detection approach relies on modeling all possible overlapping combinations of acoustic events. Both detection approaches are explored in our work.

An alternative solution to the overlapping problem, which is based on source separation, is presented in Section 2. The Section also describes the different source separation approaches, like BSS and beamforming based technique. A full detection system that is based on the beamforming based source separation is presented in Section 3. The fuzzy integral based fusion of information sources, which could be used in the decision block of the system is presented in Section 4. Acoustic scenario and the databases, which will be used in this thesis work, are described in Section 5. And the Chapter summary is presented in Section 6.

3.2 Source separation based approach

The problem of overlapping can be solved at the signal level by source signal separation, i.e. the signals are separated first, and then detected.The advantage of these methods over the model based one is that they do not require the extra models for the overlapped signals. But, on the other hand, an extra separation block is needed at the first stage of the system. Two very different techniques are considered in this work: 1) a blind source separation (BSS) technique which employs a contrast function based deflation method [**94**], and 2) a computationally simpler array processing based separation technique which employs null steering beamforming (NSB). The first technique is expected to produce better separation than the second, but it can hardly be implemented in real-time. The second one is a partial separation method which, when the number of microphones is very low, as in our case, does not achieve a quality of the separated signals as good as the first, but it still may be useful for detection and localization.

3.2.1 Source separation based on deflation

In blind source separation, a number of signals are separated from the set of mixed signals with or without the aid of information about the source signals or the mixing process. The block diagram of the BSS based AED system is depicted in Figure 2. For our work, we have selected an iterative BSS technique where the source signals are extracted from the mixtures one by one [**94**] [**95**]. The main assumptions are: the signals are stationary and statistically mutually independent, there are more sensors than sources, and the mixing system is a FIR filter. After separation, the output signals correspond in any order to the source signals passed through a scalar filter. If the sources are temporally independent and identically distributed, the scalar filter further reduces to a delay and scaling factor. Here we will use a deflation based BSS approach which consists of using a contrast function to transform the original problem into an optimization problem [**96**]. There are several contrast functions which can be used for this optimization. In this work, to reduce time complexity, we have used a quadratic contrast function with 4^{th} order cumulants, like in [**96**].

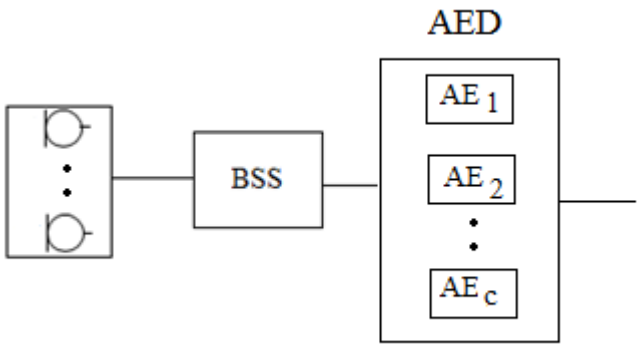

Figure 2 : Blind source separation based AED

3.2.2 Source separation based on beamforming

In this second approach, source separation is based on signal processing using a null steering beamformer (NSB). In the first stage, the NSB adapts the microphone array pattern by steering the main beam towards the desired source and placing nulls in the directions of the interference sources as described in Figure 3. Thus the contribution of one of the simultaneous sounds to the beamformer output is expected to be lower than its contribution to the beamformer input. In the case of two sources, we will have two NSB in parallel, so each of them will nullify a different source signal. Indeed, beamforming is based on the prior knowledge of the direction of the desired and interference sources, which can be provided by an ASL system. Thus, each NSB has two inputs: 1) the multimicrophone signal, and 2) the position coordinates or direction of arrival (DOA) of the sources. In the reported experiments, a linear array of either only two or three microphones is used.

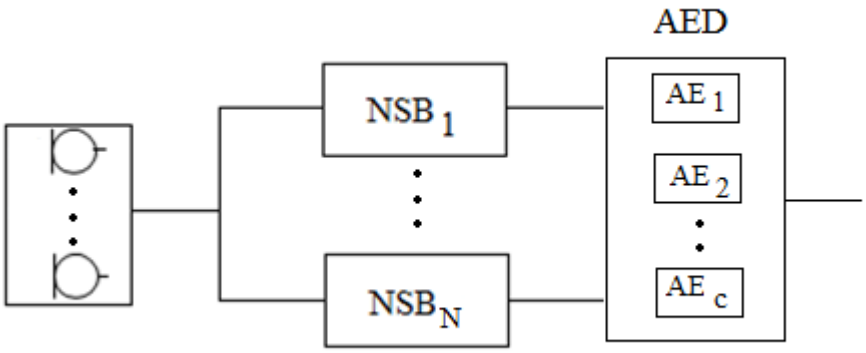

Figure 3 : Beamforming based AED

3.2.3 Null steering beamforming

Null steering beamforming (NSB) is one of the earliest, but potentially very useful, beamforming techniques. It belongs to a class of very popular and widely used beamforming techniques called multiple side lobe cancellers (MSC) [**56**] [**72**]. NSB adapts the sensor array pattern by steering the main beam towards the desired source and placing nulls in the direction of the interference sources [**75**]. Here, beamforming is based on the prior knowledge of the direction of the desired and interference sources. Two types of beamforming are attempted in our work: 1) frequency dependent beamforming, and 2) frequency invariant beamforming. Both are discussed in the following sub-sections.

3.2.3.1 Frequency dependent beamforming

In this type of beamforming, the radiation pattern of the beamformer is dependent on frequency and direction of arrival. Two types of implementations are possible: 1) time domain and 2) frequency domain. In the general scenario, we may have a desired source or target with several other spatially distributed interference sources. So a first-order NSB is needed which uses multiple microphones [**56**] [**64**]. If N is the number of microphones and M is the total number of sources overlapped temporally, the output of a beamformer can be expressed for each time instant as

$$\mathbf{x}^H \mathbf{W} \tag{3.1}$$

where **X** is the *N*x1 input signal vector, and **W** is the *N*x1 weight vector, which is the solution of the equation

$$\mathbf{S}^H \mathbf{W} = \mathbf{e}^H \tag{3.2}$$

where **e** = [1 0….0] is 1x*M matrix*, and **S**=[$S_{m,n}$] is a *M*x*N* matrix whose elements are defined by the equation:

$$s_{m,n} = \exp\left(j2\pi f (n\text{-}1)\left(d \cos\theta_m\right) / v_s \right) \tag{3.3}$$

being *n* the microphone index, $s_{1,n}$ the steering vector for the target source, and $s_{2,n}$ to $s_{m,n}$ is the steering vector for the interfering sources, which are a function of the direction of arrival (DOA) θ_m; *f* is the operating frequency (narrow-band signals are assumed), *d* is the spacing between the consecutive microphones of the linear uniform array, and v_s is the velocity of sound in m/sec. In case of two simultaneous sources, one is the target and other is an interference source that has to be nulled. In that case, steering vectors are:

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} = \begin{pmatrix} 1 & exp(j2\pi f (d \cos\theta_1) / v_S \\ 1 & exp(j2\pi f (d \cos\theta_2) / v_S \end{pmatrix} \tag{3.4}$$

So, we have to use a minimum of two microphone signals to solve Eq. 3.2. Within the each T-shaped array of our room, we have maximum three microphones linearly spaced. So, using the above mentioned technique with three microphone signals, it is possible to separate a target signal from the maximum of two other interference sources. As the steering vectors are dependent on *f* and θ_m, the effective radiation pattern of the beamformer is strongly influenced by the frequency. It imposes some kind of limitations to this time domain beamforming with the narrowband consideration in spite of its simplicity. However, the problem of frequency dependency can be overcome by converting the time domain signal into frequency domain and using the narrowband beamforming for each frequency bin, as indicated in Figure 4 [**57**]. Here each of the microphone signals is converted to a frequency domain representation by using DFT and then applying beamforming for each frequency bin before conversion back to time domain by inverse DFT. Another possibility is to use a frequency invariant beamforming which is discussed in the following sub-section.

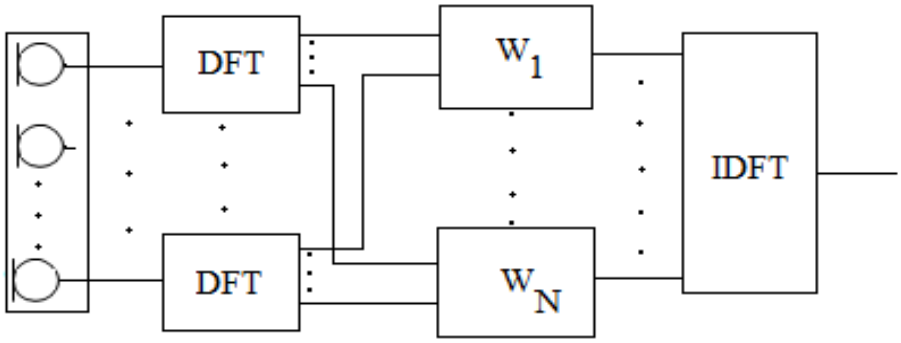

Figure 4 : Frequency domain beamforming

3.2.3.2 Frequency invariant beamforming

The main objective of this method is to introduce a broadband beamforming by decoupling the spatial selectivity from the frequency selectivity through a parameterization of the filter coefficients. Once the decoupling is done, the frequency invariant response is obtained by choosing the same coefficients for multiple frequencies. Generally, the broadband frequency invariant beamforming mostly requires a continuous aperture of sensors with large number of microphones on a specific geometry. However, this method uses a numerical approach to construct an optimal frequency invariant response for an arbitrary array configuration with a very small number of microphones (as much less as 2) [**69**] and capable of nulling many interference sources simultaneously. Moreover, it is possible to steer the resulting frequency invariant response by combining it with the spherical decomposition of the beam pattern. As depicted in Figure 5, the frequency invariant beamforming (FIB) method first decouples the spatial selectivity from the frequency selectivity by replacing the set of real sensors by a set of virtual ones, which are frequency invariant. Then, the same array coefficients can be used for all frequencies.The filter array response in frequency domain is represented as:

$$h(f,\theta) = \sum_{n=1}^{N} c_n(f) g_n(f,\theta) \qquad (3.5)$$

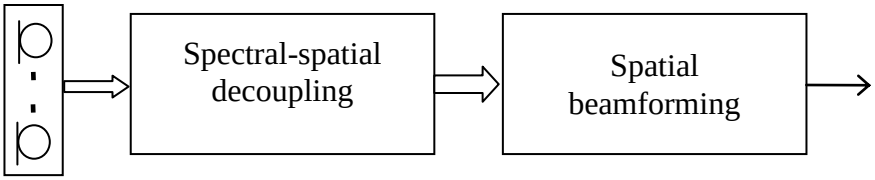

Figure 5 : Frequency invariant beamforming

where N denotes the number of real sensors in any array geometry. This equation can be seen as a parameterization of the filter array response for each frequency f with coefficients $c_n(f)$ and basis function $g_n(f,\theta)$. Modifying coefficients $c_n(f)$ will affect the frequency and spatial response simultaneously because $g_n(f,\theta)$ depends on both frequency f and direction of arrival θ. The goal of this frequency invariant beamforming is to find a new parameterization for $c_n(f)$,

$$c_n(f) = \sum_{l=1}^{L} b_{nl}(f)\hat{c}_l(f) \tag{3.6}$$

such that $b_{nl}(f)$ is the basis transform that converts the array response into frequency invariant array response by replacing N sensors indexed by n by a new set of L virtual sensors indexed by l,

$$\sum_{n=1}^{L} g_n(f,\theta)b_{nl}(f) = \hat{g}_l(\theta) \tag{3.7}$$

These virtual sensors are frequency invariant. Note that the $\hat{c}_l(f)$ coefficients are same for all frequencies. For the calculation of $b_{nl}(f)$ in Eq. 3.7; it is required to use an elegant analytic inversion formula which imposes some limitations like $l<N$ and all the sensors are placed equidistant in a defined geometry. To overcome the problem of the restricted number of sensors and the restrictions on sensor location imposed by analytic inversion, least squares solution is used [**168**]. The arrival angle in this case is discretized by θ_q, $q=1, 2,....,Q$ and the Eq. 3.6 can be rewritten in matrix format as:

$$\mathbf{G}(f)\mathbf{B}(f) = \hat{\mathbf{G}} \tag{3.8}$$

where $[\mathbf{G}(f)]_{qn} = g_n(f,\theta_q)$, $[\mathbf{B}(f)]_{nl} = b_{nl}(f)$ and $[\hat{\mathrm{G}}]_{ql} = \hat{g}_l(\theta_q)$. The least squares solution to this equation is defined by,

$$\mathbf{B}(f) = \mathbf{G}^{\dagger}(f)\hat{\mathbf{G}} \tag{3.9}$$

Where $\mathbf{G}^{\dagger}(f)$ is the pseudo inverse of $\mathbf{G}(f)$. After calculation of this basis transform $\mathbf{B}(f)$, it is made steerable by spherical harmonics. To elaborate the whole methodology in a compact form, we can write the array response in the following format,

$$h(f,\theta) = \mathbf{g}(f,\theta)\mathbf{B}(f)\mathbf{D}(\alpha,\beta,\gamma)\hat{c}(f) \tag{3.10}$$

where $\mathbf{g}(f,\theta)$ is a vector of N sensor responses to a plane wave, $\mathbf{B}(f)$ uncouples the spatial response from spectral response, $\mathbf{D}(\alpha,\beta,\gamma)$ steers the spatial response into an arbitrary direction and is the rotation matrix commonly known as Wigner rotation matrix which is a function of azimuth α, elevation β and the spin angle about z-axis γ. Therefore the basis transform has a rotations to any directions specified by (α, β, γ) and $\hat{c}(f)$ defines it's spatial profile for each frequency separately. Choosing the same coefficients for all frequencies yields an array response that is frequency invariant. When computing the basis $\mathbf{B}$ by Eq. 3.9, it is required that matrix $\mathbf{G}$ to be of full rank. Generally, the matrix is ill conditioned for lowest frequencies and therefore cannot be accurately inverted when computing pseudo inverse $\mathbf{G}^{\dagger}(f)$. This is to be expected that at lower frequencies, the less separation between sensors prevents effective spatial resolution. Similarly at higher frequencies, finite and comparative more separation between sensors generates side lobes and thus generates non invertible $\mathbf{G}^{\dagger}(f)$. In this situation of numerical inversion, the instability leads unwanted gains for the noise content of the signal which affect the desired signal destructively. So, for the calculation of the optimum basis transform $\mathbf{B}$, an added term $\mathbf{\Sigma}(f)$ is introduced in Eq. 3.9 considering the presence of an uncorrelated and homogenous noise in space.

$$\mathbf{B} = (\mathbf{G}^{H}\mathbf{G} + \mathbf{\Sigma})^{-1}\mathbf{G}^{H}\hat{\mathbf{G}} \tag{3.11}$$

where $\mathbf{\Sigma}$ is a NxN square diagonal matrix with powers $\boldsymbol{\sigma}^2(f)$ on the diagonal which result in the conventional regularization of the pseudo inverse and thus reduce the effect of the unwanted amplification of the noise signals.

3.3 Scheme for the whole detection system based on beamforming

The basic schematic diagram of the whole system that could be used either for detection, or for localization, or for joint detection and localization is presented in Figure 6. The system at its front end uses a beamforming based source separation. The overall system is an integration of several subsystems. These sub-systems are: 1) beamforming based signal processing, placed at the front end of the system after the audio acquisition, 2) feature extraction, 3) classifiers, and 4) decision block.

The proposed system at its front end consists of a set of null steering beamformers. These beamformers are designed to work with as much less as two microphones and allows us to design a sensor array pattern that steers the main beam towards the desired source, and places nulls in the direction of interferent sources. Given the broadband characteristics of the audio signals, in order to determine the beamformer coefficients we can use frequency invariant beamforming (FIB). The method, explained in the previous Section, uses a numerical approach to construct an optimal frequency invariant response for an arbitrary array configuration with a very small number of microphones, and it is capable of nulling several interferent sources simultaneously [**69**]. As depicted in Figure 5, the FIB method first decouples the spatial selectivity from the frequency selectivity by replacing the set of real sensors by a set of virtual ones, which are frequency invariant. Then, the same array coefficients can be used for all frequencies.

Scheme for the whole detection system based on beamforming

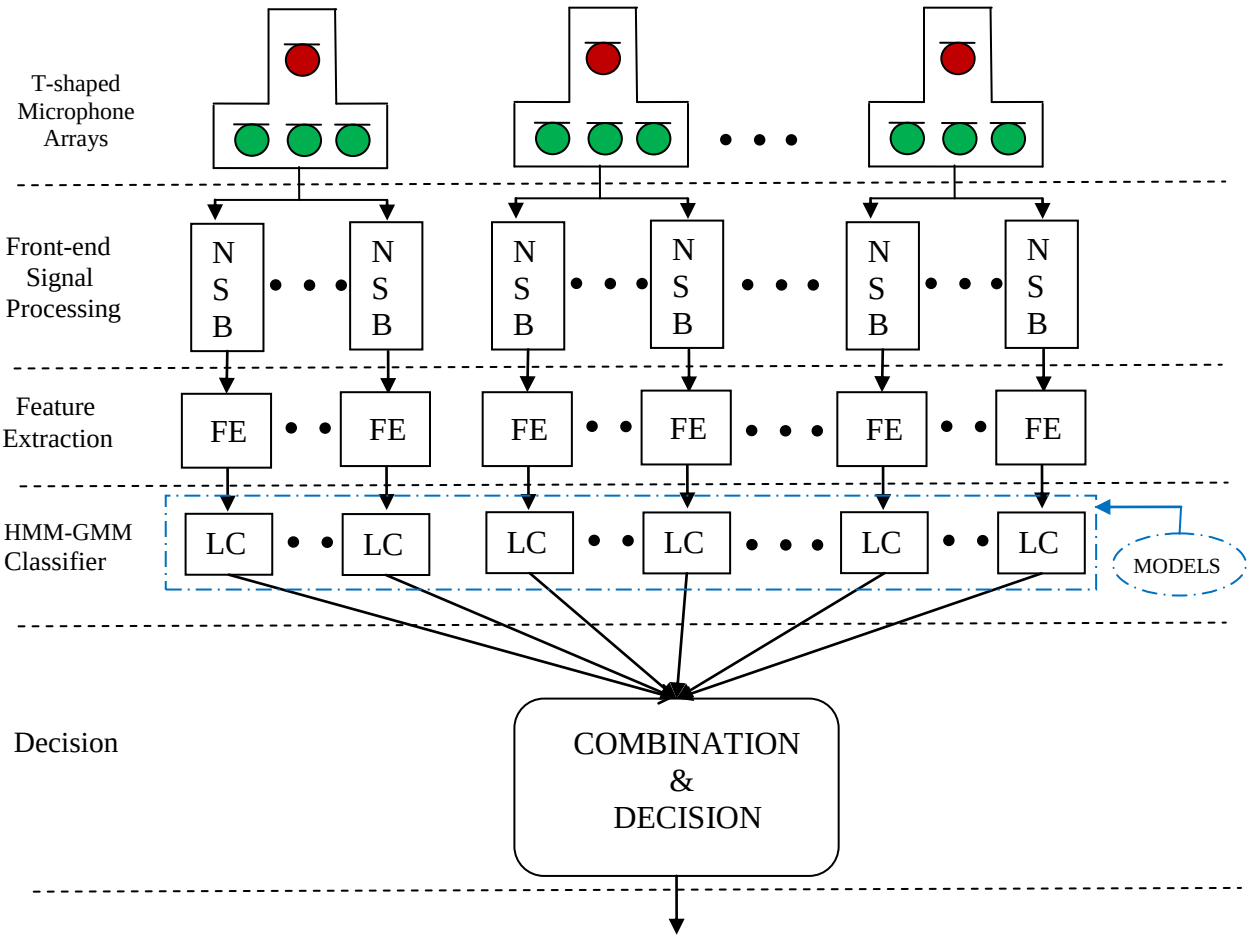

Figure 6 : Block diagram of a source separation (beamforming) based detection system

An illustrative example is shown in Figure 7; note how the null beams are rather constant along frequency. Indeed, in our case we cannot expect with this approach a perfect separation of the different mixed signals at the output of the NSB, since we use a small number of microphones per array, and also because of echoes and room reverberation. Actually, with such few microphones, it is expected that the beamformers have wider lobes and the sources are less well separated. But on the other hand, it facilitates a computationally efficient working environment.

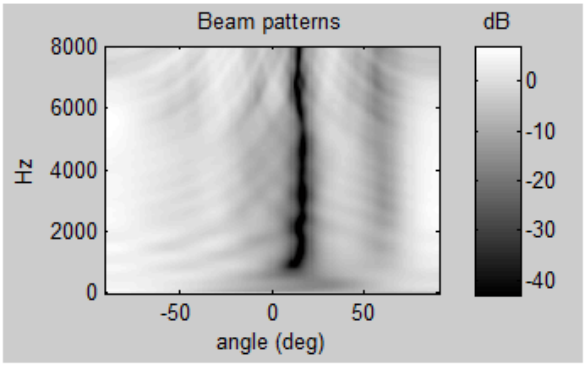

Figure 7 : Beam patterns for null steering beamformer

In the second stage of the system, feature extraction block extracts a set of audio spectro–temporal features for each signal frame. In all the applications, we have considered the frame length of 30 ms with 20 ms shift, and a Hamming window is applied. We have used frequency-filtered log filter-bank energies (FF-LFBE) for the parametric representation of the spectral envelope of the audio signal [**169**]. For each frame, a short-length FIR filter with a transfer function $z\text{-}z^{-1}$ is applied to the log filter-bank energy vectors and end-points are taken into account. Here, we have used 16 FF-LFBEs along with their 16 first temporal derivatives, where the latter represents the temporal evolution of the envelope. Therefore, the dimension of the feature vector is 32.

In the next stage, we have used a HMM-GMM classifier. The HTK toolkit is used for developing the HMM-GMM based classifier [**170**]. There is one left-to-right HMM with three emitting states for each AE and silence. The observation distribution of these states is Gaussian mixture with continuous density, and consists of 32 components with diagonal covariance matrix. Each HMM is trained with the signal segments belonging to the corresponding event class using the standard Baum–Welch training algorithm [**171**].

In the final stage, the decision block takes decision depending on the application of the system with the scores and hypothesized class from the previous HMM-GMM classification stage. In such type of system based on multiple arrays and working in a complex scenario with multiple simultaneous acoustic sources, it may require to combine the scores and class hypothesis in the decision block [**172**]. In case of combining classifier scores, add or product

rule could be adopted. It is also possible to use some machine learning approach for these kinds of combinations for the decision. Also the classifiers working in parallel could be considered as an individual information sources for the decision block. In such case, fuzzy integral (FI) based fusion of these information sources often beneficial for the performance of the overall system [**173**].

3.4 Fuzzy integral fusion of information sources

In the proposed scheme described in the last Section, the information fusion is done at the decision levels. Here, we use a specific classifier at each path from the output of beamformers and then combine the output scores of these classifiers. Each such classifier or some specific combination of classifiers acts as an independent "expert", giving its opinion about the unknown class. The fusion rule then combines the individual experts. In the presented work, fusion is carried out using fuzzy integral (FI) [**172**] [**173**] fusion approach at the decision level. Unlike non-trainable fusion operators (mean, product), the statistical FI avoid the assumption of equal importance of information sources. Moreover the FI fusion operator also takes into account the interdependences among the information sources. The main motivation here is to compensate possible misclassification errors of a certain classifier with other available classifiers and to end up with a more reliable overall decision.

We are searching for a suitable fusion operator to combine a finite set of information sources$Z$= {1,….,z}. Let D= {D_1, D_2, ….,D_z} be a set of trained classification systems and Ω = {c_1, c_2,…c_n} be a set of class labels. Each classification system takes as input a data point $x\epsilon$ $\Re^n$ and assigns it to a class label from Ω. Alternatively, each classifier output can be formed as an N-dimensional vector that represents the degree of support of a classification system to each of N classes. It is convenient to organize the output of all classification systems in a decision profile:

$$DP(x) = \begin{bmatrix} d_{1,1}(x)....d_{1,n}(x).....d_{1,N}(x) \\ ..... \\ d_{j,1}(x)....d_{j,n}(x).....d_{j,N}(x) \\ ..... \\ d_{z,1}(x)....d_{z,n}(x).....d_{z,N}(x) \end{bmatrix} \tag{3.12}$$

where a row is classifier output and a column is a support of all classifiers for a class. We suppose these classifier outputs are commensurable, i.e. defined on the same measurement scale (most often they are posterior probability-like).

Let's denote by h_i, $i=1, 2,..., z,$ the set of output scores of the z classification systems for the class c_n (the supports for class c_n, i.e. a column for decision profile) and before defining how FI combines information sources, let's look to the conventional WAM fusion operator. A final support measure for the class c_n using WAM can be defined as:

$$M_{WAM} = \sum_{i \in Z} \mu(i) h_i \tag{3.13}$$

where $\sum_{i \in Z} \mu(i) = 1$, for $\mu(i) \geq 0$ for all $i \in Z$

The WAM operator combines the score of z competent information sources through the weights of importance expressed by $\mu(i)$. For the weights in WAM operator we use uniform class noise model with the weights computed as $\mu_i = E_i^{E_i}(1 - E_i)^{1-E_i}$, where E_i is the training error of class c_i [**172**]. The main disadvantage of the WAM operator is that it implies preferential independence of the information sources.

Let's denote with $\mu(i, j) = \mu(\{i, j\})$ the weight of importance corresponding to the couple of information sources i and j from Z. If μ is not additive, i.e. $\mu(i, j) \neq [\mu(i) + \mu(j)]$ for a given couple $\{i, j\} \subseteq Z$, we must take into account some interaction among the information sources. Therefore, we can build an aggregation operator starting from the weighted arithmetic mean, adding the term of "second order" that involves the corrective coefficients $\mu(i, j) - [\mu(i) + \mu(j)]$, then the term of "third order", etc. Finally, we arrive to the definition of

the FI: assuming the sequence h_i, $i=1,..,z$, is ordered in such a way that $h_1 \le ... \le h_z$, the Choquet *fuzzy integral* can be computed as

$$M_{FI}(\mu, h) = \sum_{i=1}^{z} \left[\mu(i,..., z) - \mu(i+1,..., z)\right] \; h_i \tag{3.14}$$

where $\mu(z+1) = \mu(\text{ø}) = 0$. The value $\mu(S)$ can be viewed as a weight related to a subset S of the set Z information sources. It is called fuzzy measure (FM) for $S, T \subseteq Z$ it has to meet the following conditions:

$$\mu(\text{ø}) = 0, \mu(Z) = 1, \qquad \text{Boundary}$$

$$S \subseteq T \Rightarrow \mu(S) \le \; \mu(T), \qquad \text{Monotonicity}$$

For instance, as an illustrative example let's consider the case of 2 information sources with unordered system outputs h_1=0.4 and h_2=0.3, and corresponding fuzzy measures μ(1)=0.6 and μ(2)=0.8. Note that μ(0)=0 and μ(1,2)=1. In that case, the Choquet *fuzzy integral* is computed as $M_{FI}(\mu,h) = (\mu(1,2)- \mu(1))h_2 + \mu(1)h_1$=0.36.

A large flexibility of the FI aggregation operator is due to the use of FM that can model importance and interaction among criteria. And although the FM $\mu(i)$ provides an initial view about the importance of information source i, all possible subsets of Z that include that information source should be analysed to give a final score. For instance, we may have $\mu(i) = 0$, suggesting that element i, $i \notin T$, is not important; but if, at the same time, $\mu(T \cup i) >> \mu(T)$, this actually indicates i is an important element for the decision. For calculating the importance of the information source i, the Shapley score [**174**] is used. It is defined as:

$$\phi(\mu, i) = \sum_{T \subseteq Z \setminus i} \frac{(|Z| - |T| - 1)!|T|!}{|Z|!} [\mu(T \cup i) - \mu(T)] \tag{3.15}$$

Generally, Eq. 3.15 calculates a weighted average value of the marginal contribution $\mu(T \cup i) - \mu(T)$ of the element i over all possible combinations. It can be easily shown that the information source importance sums to one.

3.5 Acoustic scenario and databases

State-of-the-art empirical and statistical data driven methods in audio recognition depend to a large extend on sufficient and appropriate sample data, often covering a particular domain, acoustic environment, recording channel or modality. One of the problems when dealing with multimodal AED task in the meeting-room environment is lack of the annotated data to evaluate the performance of the proposed techniques. There exists a relatively large database of sounds, like RWCP sound scene database [**175**], but only a small part of the sounds included in that database can be considered as usual or at least possible in a meeting room and only audio modality is available for those sounds. Another relatively large and multimodal AMI corpus [**176**] contains only a limited number of AE instances that is not appropriate to develop AED technologies.

For meeting-room environments, the task of AED is relatively new; however, it has already been adopted as a semantically relevant technology in CHIL European project (2004-2007) and two international evaluation campaigns: in CLEAR (Classification of Events, Activities, and Relationships evaluation campaigns) 2006 [**38**], by three participants, and in CLEAR 2007 [**39**], by six participants. To support these evaluations a large multimodal and multi-site corpus for AED in meeting-room environment has been created.

Since the employed cameras in CLEAR'07 evaluation corpus do not provide a close view of the subjects under study, a new database has been recorded at UPC smart-room with 5 calibrated cameras and 6 T-shaped 4-microphone clusters. This database includes two kinds of datasets: 8 recorded sessions of isolated AEs, where 6 different participants performed 10 times each AE, and a spontaneously generated dataset which consists of 9 scenes about 5 minutes long with 2 participants that interact with each other in a natural way: discuss certain subject, drink coffee, speak on the mobile phone, etc. Although the interactive scenes were recorded according to a previously elaborated scenario, we call this type of recordings "spontaneous" since the AEs were produced in a realistic seminar style with possible overlap with speech. Manual annotation of the data has been done to get an objective performance evaluation. This database is publicly available from the author and the detailed description of this database is presented in.

The above mentioned databases include 13 semantic classes (classes of interest), i.e. types of AEs that are: “door knock”, “door open/slam”, “steps”, “chair moving”, “spoon/cup jingle”, “paper work”, “key jingle”, “keyboard typing”, “phone ring”, “applause”, “cough”, “speech”, “silence”. Among them, there is one AE, “silence” which is never evaluated. The details of the database in terms of the number of occurrences per AE class are shown in Table 1.

Table 1 : Number of occurrences per acoustic event class

Event Type	Label	Number of Occurrences in audio-visual database	
		UPC iso multimodal (S-recordings)	**UPC spontaneously generated (T-recordings)**
Door knock	[kn]	79	27
Door open/slam	[ds]	256	82
Steps	[st]	205	153
Chair moving	[cm]	242	183
Spoon/cup jingle	[cl]	96	48
Paper work	[pw]	91	146
Key jingle	[kj]	82	41
Keyboard typing	[kt]	89	81
Phone ring	[pr]	101	29
Applause	[ap]	88	9
Cough	[co]	90	24
Speech	[sp]	74	255
Silence	[si]	Not annotated explicitly	

Figure 8 shows the Universitat Politècnica de Catalunya (UPC)'s smart-room, with the position of its six T-shaped 4-microphone arrays on the walls. For training, development and testing of the system in this thesis, we have used part of a publicly available multimodal database recorded in the UPC's smart-room. Concretely, we use 8 recording sessions of audio data which contain isolated acoustic events (S-recordings). The approximate source positions of the acoustic events (AE) are shown in Figure 8. Each session was recorded with all the six T-shaped microphone arrays. The overlapped signals used for development and testing of the systems were generated adding those AE signals recorded in the room with a speech signal, also recorded in the room, both from all the 24 microphones. To do that, for each AE instance, a segment with the same length was extracted from the speech signal starting from a random position, and added to the AE signal. The mean power of speech was made equivalent to the mean power of the overlapping AE. That addition of signals produces an increment of the background noise level, since it is included twice in the overlapped signals; however, going from isolated to overlapped signals the SNR reduction is slight: from 18.7dB to 17.5dB. Although in our real meeting-room scenario the speaker may be placed at any point in the

room, in the experimental dataset its position is fixed at a point at the left side (SP, in Figure 8).

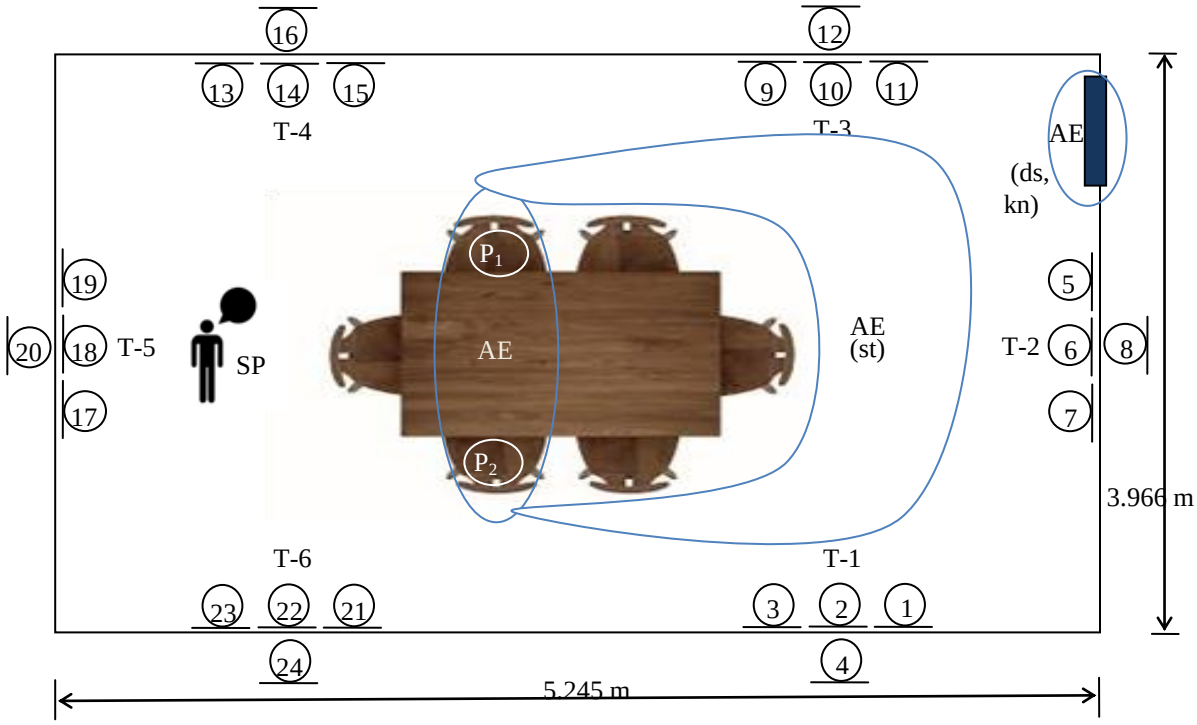

Figure 8 : Smart-room layout, with the positions of microphone arrays (T-i), acoustic events (AE) and speaker (SP)

In case of spontaneous recordings (T-recordings), two participants take the positions P_1 and P_2. They interact with each other through a conversation and the all AEs are produced from the positions as specified in Figure 8.

All signals were recorded at 44,1 kHz sampling frequency, and further converted to 16 kHz.

3.6 Chapter conclusions

In this Chapter, we have discussed the problem of source overlapping. This problem not only affects the performance of the AED system, but also severely affects the other associated system. Several solutions to overcome this problem are also discussed here. Model based approach is one of the baseline systems that performs well in a limited scenario but suffers from a scalability problem. It means that the model based approach is hardly feasible in multi-source scenarios where either the number of events or the number of simultaneous sources is large, since all the possible combinations of events have to be modelled. Alternatively, the problem can be tackled at levels other than the model one, like at the signal level. In such case, some kind of source signal separation is carried out and followed by detection (that includes identification) of each of the overlapped sounds. A beamforming based source separation technique which is computationally less demanding than a BSS based technique is desirable due to its possibility of being implemented in the real-time system. A system of such kind has been proposed, which consists of a null steering beamforming based partial signal separation technique, followed by a likelihood ratio based classifier and a decision block. Moreover, the similar structure of this kind of system could be used for different applications, like detection, localization, resolving the permutation problem in multisource scenario etc. Different types of beamforming techniques and their application possibilities in the proposed system are also discussed in this Chapter. It is also discussed the acoustic scenario associated with the multimodal database that will be used in this thesis work.

Chapter 4. Source ambiguity resolution of overlapped sounds in a multi-microphone room environment

4.1 Chapter Overview

When several acoustic sources are simultaneously active in a meeting-room scenario, and both the position of the sources and the identity of the time-overlapped sound classes have been estimated, it remains the problem of assigning each source position to one of the sound classes. To solve this source ambiguity problem, we present in this Chapter a position assignment (PA) system that performs a one-to-one correspondence between the set of source positions and the set of class labels. Both frequency dependent and frequency invariant beamformers are designed and used in the experiments. Moreover, both signal based fusion and score level fusion of pairs of microphones are also proposed. Inclusion of speech model in the classifier with the acoustic event models and fusing the scores would be a possibility for improving the performance of the overall system.

The source ambiguity problem is described in Section 2. The position assignment system with different possibilities is presented in Section 3. Experiments are reported in Section 4, along with some practical issues about the system implementation and the used metrics. A conclusion is presented in Section 5.

4.2 Problem of source ambiguity

Sound is a rich source of information. For that reason, machine audition [**177**] plays an important role in many applications. In particular, in meeting room scenarios, acoustic event detection (AED) systems try to determine the identity of an occurring sound and the time interval when it is produced. Acoustic source localization (ASL) systems estimate its position in space. Both tasks become much more challenging when there exists sound simultaneity, i.e. several sounds overlapping in time in a given space.

In a typical meeting-room acoustic scene where a person is speaking at a given time, and other non-speech sounds may happen simultaneously with the speaker's voice, the problem is dealt by detecting and localizing the acoustic event that may be temporally overlapped with speech. The detection of overlapping events may be dealt with different approaches, either at the signal level, at the model level, or at the decision level. In [**3**], a model based approach was adopted for detection of events in our meeting-room scenario with two sources, one of which is always speech, and the other one is a different acoustic event from a list of 11 pre-defined events. The same approach is used in the current real-time system implemented in our UPC's smart-room that includes both AED and ASL [**48**].

In that model based approach, we face a permutation problem. In fact, the AED system gives the hypothesized identities of the overlapped sounds, but does not associate each of them to one of the available source positions that are provided by the ASL system. The same problem may be encountered by using other AED approaches; for instance, if a blind source separation technique is used prior to the detection of each of the isolated events.

To have an unambiguous spatial analysis in any one of the detection approaches, each of the detected acoustic events has to be assigned to one of the given or estimated source positions. Therefore, to solve the source ambiguity problem, we present a position assignment (PA) system that performs a one-to-one correspondence between the set of source positions and the set of class labels. It is based on partial source separation achieved using beamforming, followed by a log-likelihood ratio based classifier.

4.3 Source position assignment

The block diagram of the whole system that performs position assignment from the outputs of the acoustic event detection and localization systems is depicted in Figure 9. The AED and ASL techniques used in our system are described in [**48**]. The AED system employs only one microphone and uses a model based approach. The ASL system, which employs all 24 microphones, is based on the SRP-PHAT localization method. The model based AED system outputs either one or two AE hypothesis (if they are two, one of them is speech). On the other hand, in the online implementation at the UPC's smart-room, the ASL system provides either one or two source positions. Hence, there are 4 different cases for mapping the detected events into the detected positions: 1-1, 1-2, 2-1, and 2-2. As can easily be seen, there exists an ambiguity in the last three cases. In this work, we focus on the most general 2-2 case, where we have two detected events, i.e. E (one of the 11 possible AE) and "sp", and two source positions: P_1 and P_2. If the problem of assigning the two events to the two positions is solved, the other two cases with ambiguity (1-2 and 2-1) can be solved using the same approach. In this Section we aim to design a system that can be deployed in real time in the room to resolve that ambiguity in the correspondence between detected AEs and acoustic source positions.

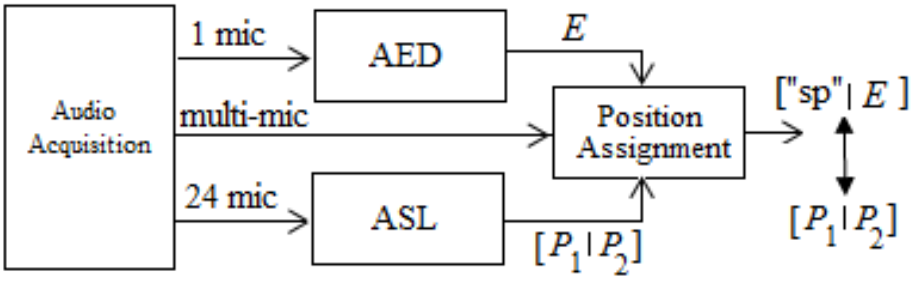

Figure 9 : Block diagram of the whole system

4.3.1 Scheme of the PA system

Here we assume there are two simultaneous events and one of them is always speech, so at the output of the AED system we need only the hypothesized identity of the non-speech AE, as

indicated in Figure 10 by *E*. The ASL system provides an estimate of the two source positions: P_1 and P_2. Hence, the position assignment (PA) block actually is a binary classifier that assigns *E* to either P_1 or P_2.The PA system, which is shown in Figure 10 for one array, has at its front-end two null-steering beamformers (NSB), which work in parallel. The main beam of each NSB is steered towards the desired source and a null is placed in the direction of the interferent source, so each NSB will nullify a different source signal. Thus the contribution of one of the simultaneous sounds to the beamformer output is expected to be lower than its contribution to the beamformer input. Indeed, beamforming is based on the prior knowledge of the direction of the desired source and the interferent source, which can be provided by an ASL system. Thus, each NSB requires two inputs: 1) the multimicrophone signal, and 2) the position coordinates or direction of arrival (DOA) of the sources.

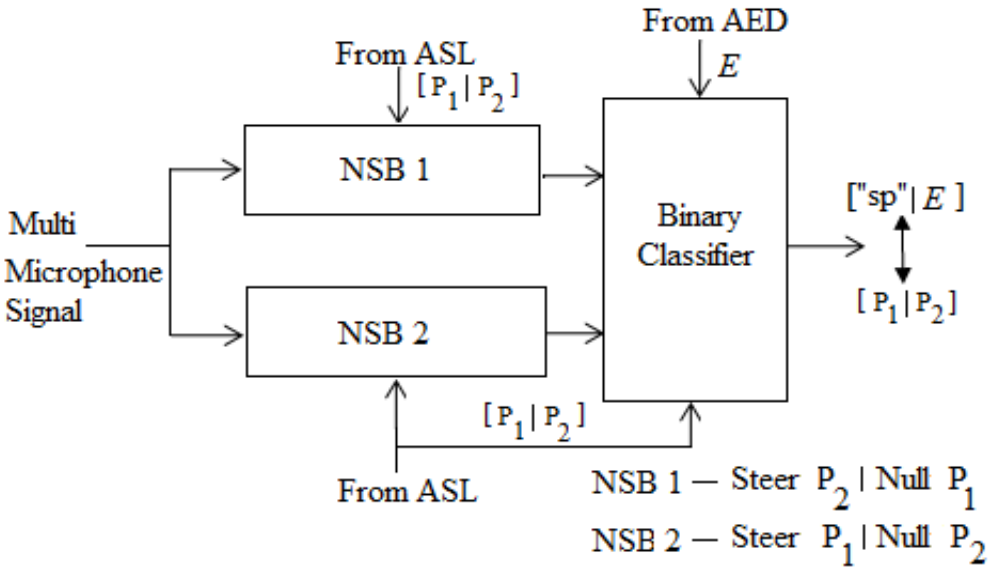

Figure 10 : Position assignment system

Each of the beamformers is followed by feature extraction (FE) and likelihood computation (LC), which uses the HMM model corresponding to the acoustic event *E* as shown in Figure 11. Finally, a decision block makes the assignment based on the computed log likelihoods LL_i, i=1,2.The two log-likelihood scores indicated in Figure 11 are combined

in the decision block to compute the following single-array score S in terms of the log-likelihood-ratio LLR_1.

$$S=LLR_1= (LL_1 - LL_2) \quad (4.1)$$

If S is positive, the AE E is associated to the position P_2, and if S is negative, it is associated to P_1.

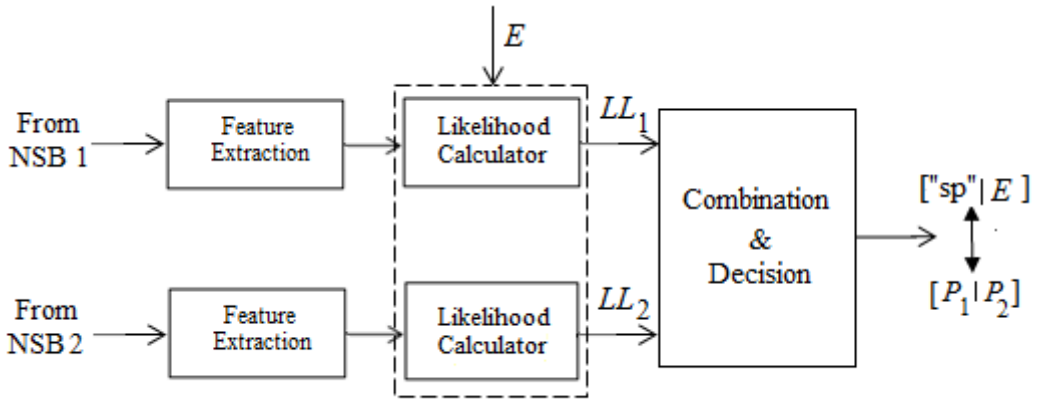

Figure 11 : Classifier and decision block of PA system

4.3.2 Null steering beamforming

Null steering beamforming adapts the sensor array pattern by steering the main beam towards the desired source and placing nulls in the direction of the interference sources [**75**]. In a general scenario, we may have a desired source with several other spatially distributed interference sources. So a first-order NSB is needed which uses multiple microphones [**56**] [**64**]. The solution for the weight matrix in this type of beamformer is achieved by setting to unity the desired response at the direction of the target sound, and setting it to zero at the direction of the interferent sources. In our particular scenario, as we have only two sources, one is the target and the other is the interference that has to be nulled. Therefore, we need to use a minimum of two microphone signals to get a solution for the weight matrix. As in each T-shaped array from our room there are three linearly spaced microphones, with three microphone signals it is possible to separate a target signal from two interferences.

Two types of beamforming are tried in our work: frequency dependent beamforming, and frequency invariant beamforming. In frequency dependent beamforming (FDB), the

radiation pattern of the beamformer is dependent on the direction of arrival and also on the frequency, since a narrow-band signal is assumed. Hence, in spite of its simplicity, this approach requires tuning the frequency to the particular conditions of the scenario. Given the broadband characteristics of the audio signals, another possibility is to determine the beamformer coefficients using a technique called FIB. The method, proposed in [**69**] [**168**], uses a numerical approach to construct an optimal frequency invariant response for an arbitrary array configuration with a very small number of microphones, and it is capable of nulling several interferent sources simultaneously. The FIB method first decouples the spatial selectivity from the frequency selectivity by replacing the set of real sensors by a set of virtual ones, which are frequency invariant. Then, the same array coefficients can be used for all frequencies.

4.3.3 Combination of pairs of microphones at the signal level and at the decision level

At least two microphone signals are required to implement a 1^{st} order beamformer [**64**]. Each T-shape array in our smart room has maximum of three linearly spaced microphones. In this work, we consider the use of either 2 or 3 microphones, linearly arranged, from those T-shape arrays. When using the 3-microphone array, we can work with two pairs of microphone signals [**178**]. The outputs of the two 1^{st}-order beamformers are combined either at the signal level or at the classifier level. The signal level combination is shown in Figure 12. In the first stage of the system, we have two 1^{st} order beamformers and both of them are designed with two linearly spaced and consecutive microphones. Here we choose to work with the consecutive microphones because the spacing between them is smaller compared to the alternately selected microphones. In fact, the microphone spacing in the earlier case is half of the spacing in the later one. Then these two pairs of microphones are combined at the second stage. In addition, the combination at the classifier level is the option taken in the scheme of Figure 13. The weight vectors $w_{i,j}$ in each of the NSB are calculated from the positions provided by the ASL block, where i indicates the number of the beamformer, and j is the number of the pair of microphone for each beamformer.

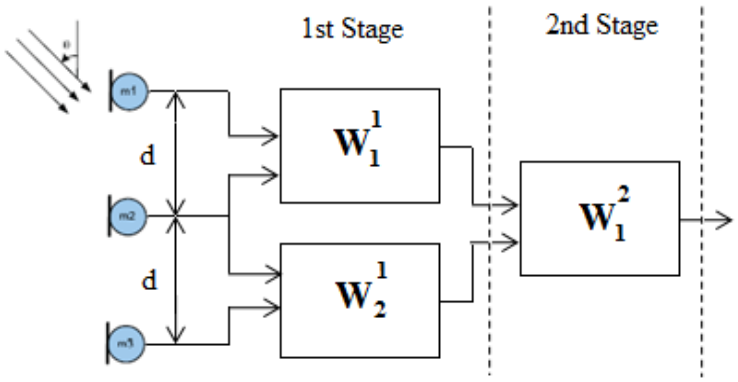

Figure 12 : Signal level combination

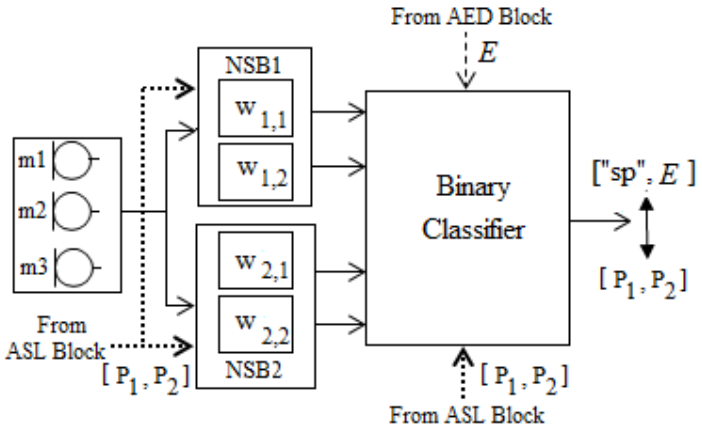

Figure 13 : Classifier level combination of microphone pairs

4.3.4 PA system with AE and speech model based classifier

In order to get the most from the available information, the classifier can include a speech model besides the AE model. Indeed, we can work separately with only either the speech model based classifier or the AE model based classifier. All those options have been tested and the results are reported in Section 4.4. As shown in Figure 14, two different NSBs are needed, in general, for each parallel path of the system; one will be steered to the AE source and the other to the speech source. Moreover, when time-domain FDB is used, a different beamformer is needed for each acoustic event class, since the frequency f is tuned to a particular event. In case of FIB, one beamformer is sufficient at each parallel path.

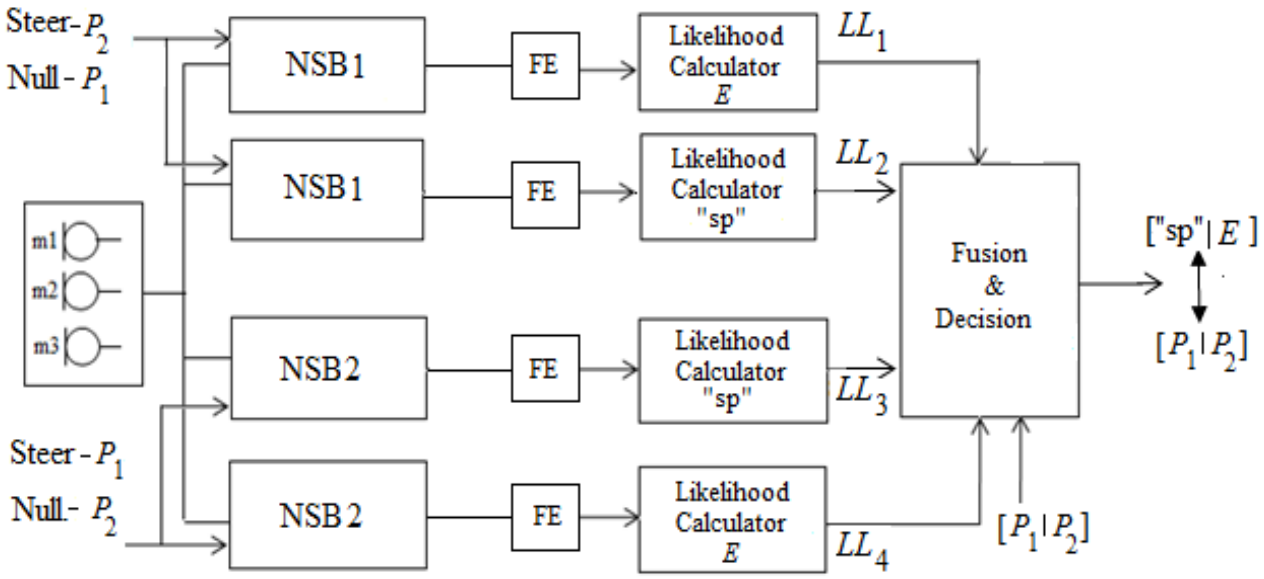

Figure 14 : PA system based on AE and speech models for one array

4.3.4.1 Single-array classification stage

As shown in Figure 14, the classification stage of the PA system with a single array consists of feature extraction, followed by log-likelihood calculation, and a binary decision block. Features are extracted from the audio signals with frame length of 30ms, and 20ms frame shift. As features, we use frequency-filtered log filter-bank energies (FF-LFBE). In our experiments, we consider a 32 dimension feature vector (16 FF-LFBE and their first temporal derivatives). As shown in the scheme of Figure 14, there is a set of two likelihood calculators for each parallel channel, one of them to calculate the model based log-likelihood for the AE label (*E*), provided by the AED system, and the other to calculate it for speech. Here we employ Hidden Markov Models (HMM), where Gaussian Mixture Models (GMM) is used to calculate the emission probabilities. 32 Gaussian components with diagonal covariance matrices are used per model. 11 HMM are trained with isolated events using the Baum-Welch algorithm.

The four log-likelihood scores (LL_i, i=1,2,3,4) indicated in Figure 14 are combined in the decision block to compute the following single-array score *S*in terms of the log-likelihood-rations LLR_1 and LLR_2 :

$$S = LLR_1 + LLR_2 = (LL_1 - LL_2) + (LL_3 - LL_4) \quad (4.2)$$

If S is positive, the AE E is associated to the position P_2, and if S is negative, it is associated to P_1. Let's illustrate it with a particular case. Assume P_1 truly corresponds to speech and P_2 to the acoustic event E. When using the AE model, it is expected to get comparatively higher log-likelihood from the output of NSB1 (LL_1) than from the output of NSB2 (LL_4). For the clean speech model, we expect to get comparatively higher log-likelihood from the output of NSB2 (LL_3) than from the output of NSB1 (LL_2). If that is the case, the decision is taken that speech is at P_1 and E is at P_2, which is the correct decision. Note that with this type of combination, the decision block gives equal importance to all the four likelihood calculator outputs.

As we have already mentioned that in order to get the most from the available information, the classification stage includes a speech model besides the AE model. Indeed, the system could also work with only either the speech model based classifier or the AE model based classifier. To study the contribution of each one of the models, all those options have been tested and the results are reported in Section 4.4. For the decision, if only either the AE or the speech based classifier is used, just either LL_1-LL_4 or LL_3-LL_2, respectively, is needed.

4.3.4.2 Multi-array fusion

As it was mentioned above, in our position assignment system we use all the six 3-microphone linear arrays deployed in the room. For taking the assignment decision, the six sets of scores LLR_1 and LLR_2, computed as indicated in Eq. 4.2 are combined either with a uniformly-weighted average of the 12 values or by fuzzy-integral based fusion. The scores at the output of the classification stage can be linearly combined by using an optimal fusion approach that assigns an individual weight to each of them. However, in this work we are going to consider a more sophisticated weighting technique that considers all subsets of information sources: the fuzzy integral (FI) approach [**173**].

The scores at the output of the classification stage can be linearly combined by using an optimal fusion approach that assigns an individual weight to each of them. However, in this

work we are going to consider a more sophisticated weighting technique that considers all subsets of information sources: the fuzzy integral (FI) approach [**173**].

Let's denote by h_i, $i=1,2,...,z$, the set of output scores (LLR_1 and LLR_2) of the $z/2$ single-array systems. Assuming the sequence h_i, $i=1,2,...,z$, is ordered in such a way that $h_1 \leq .... \leq h_z$, the Choquet fuzzy integral can be computed as

$$M_{FI}(\mu,h) = \sum_{i=1}^{z} [\mu(i,....,z) - \mu(i+1,...,z)]h_i \qquad (4.3)$$

where $\mu(z+1)=0$. The value $\mu(M)$ can be viewed as a weight related to a subset M of the set Z of z information sources. It is called fuzzy measure and, if M and T are subsets of Z, it has to meet the following conditions:

Boundary*:*	$\mu(\varnothing)=0$, $\mu(Z)=1$
Monotonicity:	$M \subseteq T \Rightarrow \mu(M) \leq \mu(T)$

In this work, we have used a supervised gradient based training algorithm for learning the fuzzy measures from the training data with cross-validation [**179**].

4.4 Experiments

The PA experiments are done under the assumption that there is always an AE overlapped with speech. We also assume that the identity of the AE event is known, to avoid the propagation of the AED errors to the PA system. Additionally, we assume that the approximate position in the room of the AE source and the speaker are known. Thus, the PA system only has to make a binary decision (the AE is from position P_1 or position P_2), that will be either correct or incorrect. And in the latter case it will be counted as an error.

4.4.1 Evaluation metrics

To design and evaluate the performance of the system, we define the position assignment rate (PAR) metric for a given AE class as the quotient between the number of correct decisions and the total number of occurrences of that class in the testing database. Then, the PAR will be averaged over the classes to have the final evaluation measure. For reference, we also consider a second metric, called Diff_LL, which is the value of the S score provided that the assignment

is correct (LL_1-LL_4 for the AE based system, or LL_3-LL_2 for the speech based one, or S when both the AE model and the speech model are used, according to the Figure 14). Actually, that score can be considered as an estimate of the degree of source separation carried out by the beamformers when a correct assignment is made. While maximization of the PAR is our main criterion for evaluation, Diff_LL has been used with a second level priority. When tuning the frequency *f* for the FDB case, sometimes occurs that, the PAR for an AE is the same for several frequency values, since the number of AE occurrences is not high enough. Then, the frequency *f* that maximizes Diff_LL is chosen.

4.4.2 System implementations

In this paper, we consider two types of beamforming at the front end of the PA system: FDB and FIB. Their design requires the DOA angles corresponding to the target and the null, i.e. the DOA from the source positions P_1 and P_2. For both of them, we have worked with approximate angles from visual inspection during recording. Alternatively, for the AE position we have used the output of a one-source ASL system. Regarding the speech source position, in our all experiments, we have used the speaker's position specified during recording.

The formulation of the FDB approach is comparatively simpler than the FIB one. As the audio signals are wideband, two types of implementations are considered for the FDB approach: 1) in the time domain, and 2) in the frequency domain. The former approach, as it assumes a narrow-band signal, requires frequency tuning to get optimal results. Therefore, the beam patterns for this type of NSB are heavily dependent on both the DOA and the frequency *f*. As we are looking to design our system for different types of acoustic events with diverse spectra, the choice of a frequency value for each specific event becomes necessary. Here, we adopted an exhaustive search technique for that. We varied the frequency from 100 Hz to 8 KHz with intervals of 100 Hz, and observed the performance of the system for each acoustic event separately.

In the frequency domain implementation, beamforming is applied to each frequency bin. Moreover, as the separation between the microphones (20cm) is suited to an operating frequency smaller than 1 KHz [**61**], we have low-pass filtered the input signal with 1 KHz cut-

off frequency. The main advantage of this frequency domain implementation is that it does not require any frequency tuning.

4.4.3 Experiments with the FDB based system

To assess the performance of the PA system, several experiments have been conducted. In our offline experiments, we use data from seven sessions (S02-S08) for training the system and tuning the frequency *f* in the FDB case, and one session (S01) is left for testing. Six of the seven training sessions are used for training the AE models, and the remaining session is used for testing each frequency value. Cross-validation is used to have a better statistical meaning. The frequency that shows the best average behavior according to the PAR metric is chosen. An initial experimental results obtained from the training database (sessions S02-S08), with two microphones (first-order FDB) from array T6, are presented in Table 2. It is worth mentioning that the AEs source positions and the speech source position are physically rather well separated from the viewpoint of the array T6. We assume in these first tests that only the speaker position is known, and the AE position is estimated with a one-source ASL system. We also assume that the identity of the AE event is known, to avoid the propagation of the AED errors to the PA system.The first column indicates the label of the acoustic events overlapped with speech. The second, third and fourth columns indicate, respectively, the position assignment rate (PAR), the difference of log-likelihoods (Diff_LL), and the selected operating frequencies. Those frequencies are estimated, like the HMM models, from the seven training sessions of the database.

Table 2 : PA metrics values and optimized frequencies for the FDB case

AEs	PAR (%)	Diff_LL	Operating Frequency (Hz)
ap	87	1.15	600
cl	80	1.34	600
cm	94	1.68	700
co	70	0.97	1000
ds	83	0.91	700
kj	95	1.79	700
kn	87	1.01	600
kt	99	3.17	800
pr	94	1.49	600
pw	100	1.14	200
st	95	1.22	500
Average	**89**	**1.44**	

The frequency domain implementation with two microphones is compared with the time domain implementation for FDB. The results obtained with only the AE models, are shown in Table 3, for both metrics, and averaging over all acoustic event occurrences in the testing dataset. Like the previous test, here also we assume that only the speaker position is known, and the AE position is estimated with a one-source ASL system. To have a better statistical meaning, the testing results are obtained with all 8 sessions with a leave-one-out criterion, i.e. we recursively keep one session for testing, while all the other 7 sessions are used for training. Notice that the PAR score of the frequency domain system is lower than that of the time domain system. However, its performance is much higher in terms of Diff_LL, which indicates that it achieves a better NSB based signal separation when the position assignment decision is correct.

Table 3 : PA performance comparison of the time-domain and frequency-domain 2-microphone based FDB

	Time-domain	Frequency-domain with LPF
PAR (%)	87	81.6
Diff_LL	1.38	3.6

FDB is implemented and tested initially using two microphones and then extending it to three. While using the maximum of 3 microphones in an array, we combine the two elemental beamformers which are designed with as much less as 2 microphones. The combination of the two beamformers at both the signal level and at the classifier level has been tested. For comparing these two alternatives, we used beamformers implemented in the time domain. The results obtained with only the AE models, for both metrics, and averaging over all acoustic event occurrences in the testing dataset are shown in Table 4. The classifier level likelihood combination proves to be preferable to the signal level one.

Table 4 : FDB based PA performances of signal level and classifier level combination

	Signal level comb.	**Classifier level comb.**
PAR (%)	88.4	89.3
Diff_LL	2.47	2.28

All the previous experiments are done using only AE models. As explained in the previous sub-Section 4.3.4, we also have the possibility of using speech models along with AE models. To check the performance of the PA system when either only the AE model or only the speech model is used, we have performed experiments for the array T6. Four types of configurations are considered: 1) the classifier is based only on AE models, 2) the classifier is based only on the speech model, 3) the likelihood (LL) combination introduced in sub-secction 4.3.4.1 that uses the outputs of both the AE-model based and the speech-model based classifiers, and 4) the fusion of likelihood scores with the FI technique. In the FI based fusion, we have used a supervised gradient-based training algorithm for learning the fuzzy measures from the training data with cross-validation [**179**]. And for that, we have used a 5-fold cross validation on the training data to stop the training process so that it is not over trained. The results obtained with both metrics, averaging over all the testing dataset, are presented in Table 5. In that table, we observe that the LL combination and the FI fusion improve the performance of the system with respect to the use of only one type of model. Between both types of combination/fusion, the FI one shows a better performance.

Table 5 : PA rate and Diff_LL for the PA system with the T6 array alone: only the AE model, only the speech model, the combination of models with the S scores, and their FI based fusion for FDB

	AE model	**Speech model**	**Average of LLR scores**	**FI based fusion**
PAR (%)	87	88.3	89.4	90.2
Diff_LL	1.38	1.23	1.48	1.62

4.4.4 Experiments with the FIB based system

Similar to the FDB based system, the design of the frequency invariant beamformers also require the DOA angles corresponding to the target and the null, i.e. the DOAs from the source positions P_1 and P_2. We have considered two different options regarding the approximate positions of the acoustic events from which the DOAs are extracted. First, we have used the position of the event estimated by an ASL system based on the SRP-PHAT technique. So, in that case, the beam steers to the direction of the specific event position. The beampttern of that type of FIB is shown in Figure 15. Note that the right side beampattern has narrower main lobe. And, in the second option, we have considered, for each array, the same approximate DOA for the whole set of acoustic events. It is obtained as a DOA average over the AE source positions, which are known from visual inspection during recording. In the latter case, we have designed a beampattern with a broader main lobe (as shown in the right side of Figure 16) to approximately encompass all the positions of the acoustic events.

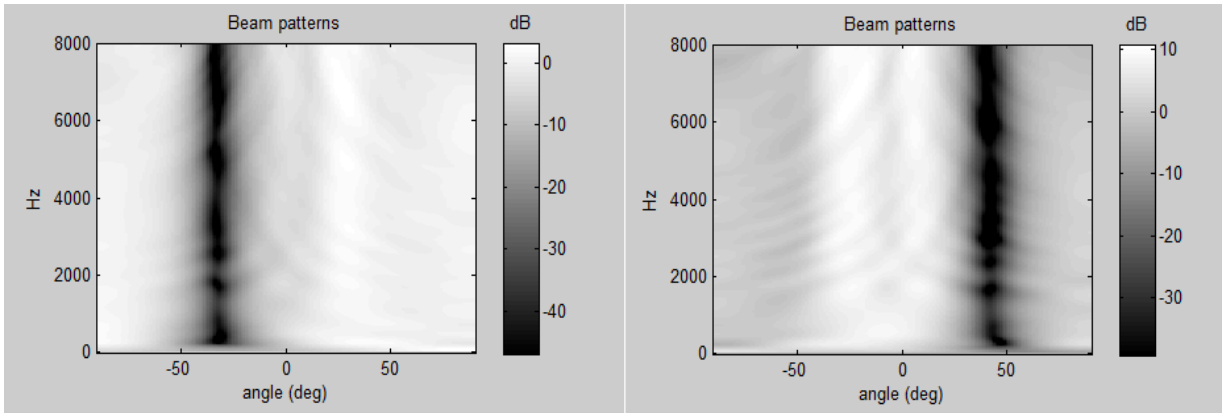

Figure 15 : Beampattern of the FIB. Left: null towards speech; right: null towards an AE, a narrow beam

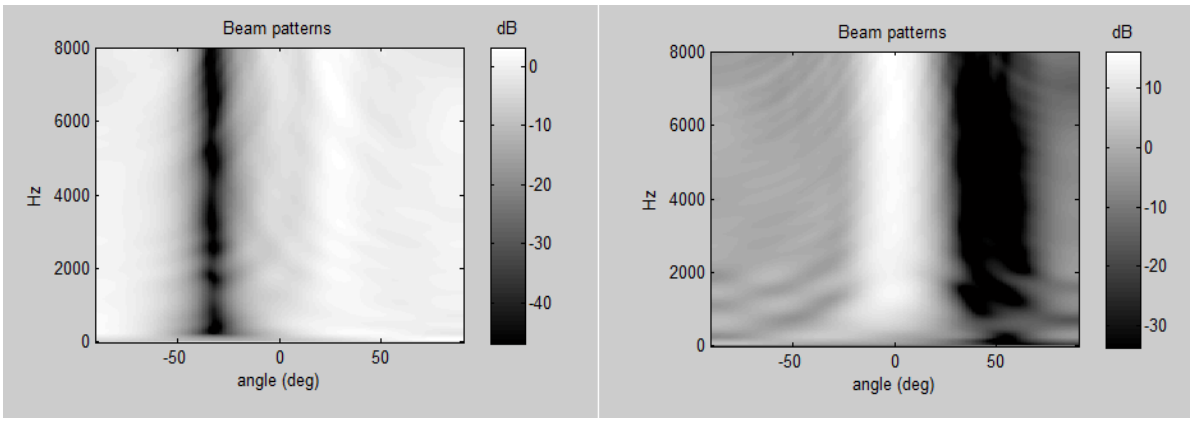

Figure 16 : Beampattern of the FIB. Left: null towards speech; right: null towards the AEs in the case of a wide beam encompassing the DOAs of all the AEs

To assess the performance of the FIB based PA system that is depicted in Figure 14, several experiments have been conducted. The testing results are obtained with all the 8 recording sessions (S01-S08), using a leave-one-out criterion, and averaging over all the testing dataset. In all the FI-based fusions, we have used a 5-fold cross validation on the training data to stop the training process and avoid over-fitting. To check the performance of

the PA system when either only the AE model or only the speech model is used, we have performed experiments for the array T6, using visually inspected positions for AEs and a broad beam. The results are shown in Table 6.

Table 6 : PA rate and Diff_LL for the PA system with the T6 array alone: only the AE model, only the speech model, the combination of models with the S scores, and their FI based fusion for FIB

	AE model	**Speech model**	**Average of LLR scores**	**FI based fusion**
PAR (%)	86.2	81	87.1	91.2
Diff_LL	1.44	1.07	2.53	2.88

We can observe, from the results in Table 6 that the combination of the two models with the *S* score, that averages the scores LLR_1 and LLR_2, improves the performance of the system with respect to the use of only one type of model. The improvement is much more noticeable using the FI based fusion of the two scores. Notice that the AE-model based system works much better than the speech model based one. In fact, the former uses a more specific model, because the speech model is obtained from the whole set of speech sounds. In that table, we also show the Diff_LL score. Notice that, in general, it is well correlated with the PAR one. However, there is a large difference between the values of Diff_LL for the AE-based case and the LL combination case, in contrast with the very small difference there is in terms of PAR. It means that the use of both models allows to achieve a much stronger confidence on the PA decision when it is correct.

Table 7 shows the PAR scores when all six-microphone arrays are employed, either alone or in combination. The results are given for the two types of DOA settings mentioned above. Also, two types of intra-array combinations are considered, as in Table 6 (second column of Table 7): the average of LLR scores given by (1), and the FI based fusion.

Table 7 : PA performance (in %) for each single array and for the two considered combinations of the six arrays

DOA setting	Intra-array combination	T1	T2	T3	T4	T5	T6	Average of LLR scores	FI based fusion
Broad-beam nulling angle	Average of LLR scores	83.1	77.3	81.3	88.9	88.2	87.1	89.8	93.5
	FI based fusion	83.5	77.1	82.3	92.8	92.7	91.2	-	93.6
ASL-estimated AE positions	Average of LLR scores	88.2	85.6	89.8	91.2	92.1	91	93.6	95.4
	FI based fusion	88.3	84.9	90.2	92.7	93	92.2	-	95.7

In the first columns of Table 7 we have the PAR (in %) when each one of the arrays is used alone. Notice that the PAR scores of the upper half of the array numbers (T4, T5, and T6) are higher than the ones of the lower half, what is in agreement with the reason given above to choose the array T6 for the experiments presented in Table 6.

Note from the two last columns in Table 7 that the accuracy obtained from either the average of LLR scores or the FI based fusion of the whole set of arrays is higher than the accuracy obtained from any of the single arrays. Comparing both types of fusion, the FI one shows a noticeable better performance for both DOA setting cases, arriving to a PA error of only 4.3%.

The use of intra-array FI based fusion improves significantly the PAR scores with respect to using a uniformly-weighted average of LLR scores, for the upper-half arrays. So, though by employing the FI approach there is the cost of having to learn the fuzzy measures from data, it may be a good choice when the quality of the signal separation is not too low, like it presumably happens with the upper-half arrays.

Regarding the type of DOA setting, the ASL-estimated AE position based system works always better than the one that uses an average DOA based on visual inspection (i.e. a broad-beam nulling angle). That could be expected, since the beam pattern is specific of each event class, whereas the broad beam encompasses all the angles of the AE source positions. While

the latter design simplifies the overall system, as it does not require a precise source position and may avoid an additional external ASL block, it is specific of the given scenario, so it has to be re-designed when the scenario changes.

4.5 Chapter conclusions

In this Chapter, an attempt to resolve the source identification ambiguity that appears when an acoustic event overlapped with speech is detected is carried out. A position assignment system consists firstly of a set of null-steering beamformers to carry out different partial signal separations for each microphone array has been proposed and tested. The beamformers are followed by model-based likelihood calculations, using both the acoustic event model and the speech model, to obtain a couple of likelihood ratios, which are multiplied to get a final score per array.

Two types of beamforming techniques are compared. Both acoustic event models and speech models are used for the likelihood computation with a fusion technique in the decision block. A product rule based combination of their likelihood calculators improves the performance of the system. The formulation of frequency dependent beamformer is simpler than the frequency invariant one, but it is more dependent on the particular conditions of the production of acoustic events in the room. On the other hand, the alternative FIB technique does not require frequency tuning and thus it is less dependent on the concrete scenario.

Moreover, the FIB based position assignment performance has been observed with all the microphone arrays distributed in the room, and the inter-array level fuzzy integral based likelihood fusion shows a further improvement. The FI based fusion of the six scores, one per array, yields the best assignment error that is smaller than 5%. Although the position assignment system has been developed for the problem encountered in our current scenario with two acoustic sources, its new formalism can be extended to more sources, as it will be done in the next Chapters for AED and ASL.

Chapter 5. Real time multi-microphone classification and detection of simultaneous acoustic events

5.1 Chapter overview

Time overlapping of acoustic signals, which so often occurs in real life, is a challenge for current state-of-the-art sound recognition systems. In this Chapter, we propose efficient approaches for classifying, detecting, identifying, and positioning a set of simultaneous acoustic events in a room environment, using multiple arbitrarily located microphone arrays, and working in real time.

Assuming both the end-points of the events and a set of acoustic source positions, the use of a frequency invariant null-steering beamformer for each position and each array yields a set of signals, which show different balances among the various acoustic sources. For each signal, a model based likelihood computation is carried out to obtain a matrix of likelihood scores. Then a MAP criterion is used to jointly classify the event and assign each of them to a given source position.

In this Chapter, we also propose an acoustic event detection method that uses all the available microphone arrays. This approach does not require any assumptions about the start and end time stamps of the occurred events. Using the same framework, the system is also able to assign each detected event to each source position.

The system for classification and position assignment of simultaneous acoustic events is described in Section 2. The detection approach is presented in Section 3. Experimental work on event classification and detection is reported in Section 4, and a conclusion is given in Section 5.

5.2 MAP based overlapped event classification and position assignment

A computationally efficient approach, which is based on signal separation by using multiple linear microphone arrays that are composed of a small number of microphones, is proposed. Let us assume that several acoustic source positions are provided. They may have been either provided from the visually inspected positions during recording or estimated by an external localization system that uses the available set of microphone arrays. As mentioned in the previous Chapters, in the proposed approach, the arrays can be located arbitrarily. For deployment, this is an advantage with respect to using spatially structured array configurations. Assuming a set of P hypothesized source positions, a set of P beamformers is used to separate up to some extent each hypothesized source from the others. Using those (partially) separated signals, acoustic event classification is carried out using a maximum-a-posteriori (MAP) criterion. Moreover, each hypothesized event is assigned to a given source position using the same framework. The beamformers are based on a frequency invariant null steering approach.

5.2.1 Scheme for a multiple array based classification system

As shown in Figure 17, in the proposed system, firstly the multi-channel signal collected by each of the microphone arrays is driven to a set of null-steering beamformers (NSB). Each beamformer is steering to a different source position while placing nulls at the direction of the other source positions.

Feature extraction (FE) is applied at the output of each beamformer, to subsequently compute a likelihood score (LC), by using previously trained models of the acoustic event classes. At last, a decision module carries out the classification of the event identities by integrating the likelihood scores using a MAP criterion. Both the beamformer design and the MAP classification are presented in the two following sub-sections.

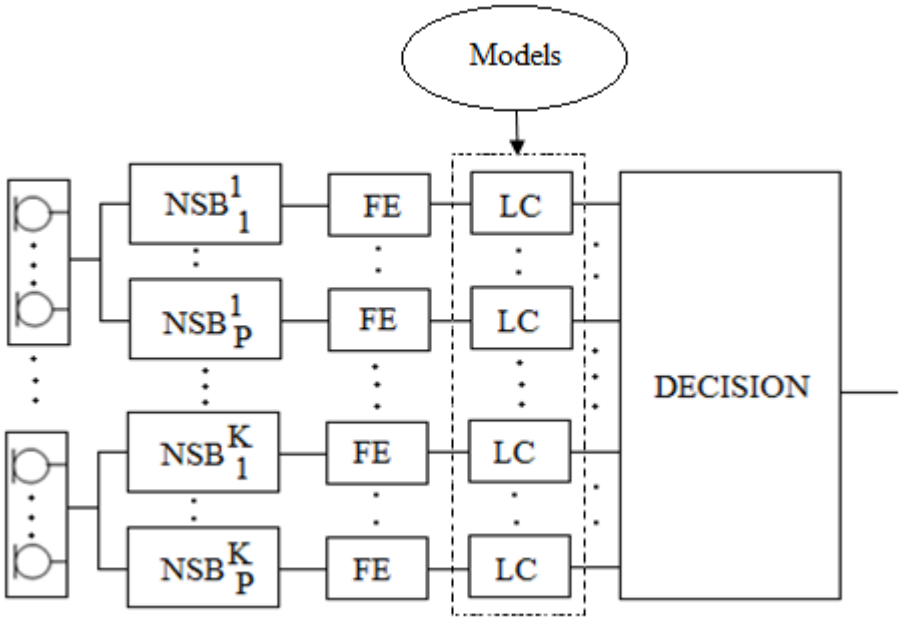

Figure 17 : Scheme of the whole classification system using K arrays

A null-steering beamformer (NSB) is capable of placing nulls at different positions in the sensor array patterns. The beamformers are based on a frequency invariant null steering approach as described in Chapter 3, which uses a numerical approach to construct an optimal frequency invariant response for an arbitrary array configuration with a very small number of microphones, and it is capable of nulling several interferent sources simultaneously. Here, we have designed the beamformer using 3 linearly spaced microphones for each array. An illustrative example is shown in Figure 18; note how the null beams are rather constant along frequency.

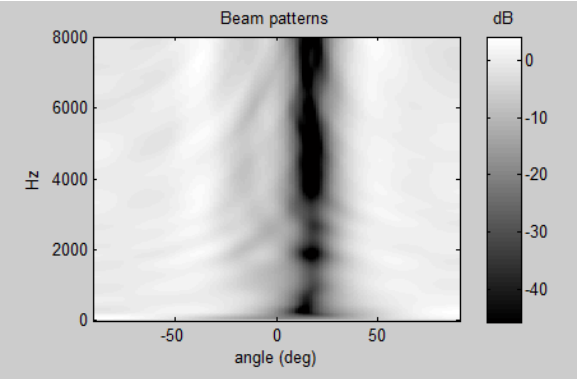

Figure 18 : FIB beam pattern; there is a null at 15 degrees

Indeed, in our case we cannot expect with this approach a perfect separation of the different mixed signals at the output of the NSB, since we use a small number of microphones per array, and because of echoes and room reverberation.

A MAP criterion is used in our system. To determine the likelihoods, the acoustic events are modeled with Hidden Markov models (HMM), and the state emission probabilities are computed with continuous density Gaussian mixture models (GMM).

Let's assume we have a set of N simultaneous events E_i, $1 \leq i \leq N$, that belong to a set of C classes, a set of P acoustic source positions, and a set of K microphone arrays. Each array steers a NSB to each of the source positions while nulling the others. So from array processing, we have a set of PK output signals, and after likelihood computations, we have a PxK-dimensional matrix of likelihood scores. We will assume also that each class c_i has a prior probability $p(c_i)$, and each estimated source position s_j has an associated probability $p(s_j)$. The latter may be provided by the ASL system.

Performing null steering beamforming with the k-th microphone array, which has at its input the multi-channel signal X_k (notice that, to simplify notation, we do not consider time indices), P output signals will be obtained, one for each NSB pattern. Let us denote with s_j the NSB that has the position s_j as target and the other positions as nulls. We want to determine the posterior probability of a given class c_i for that k-th array through all the P NSBs (note that our NSBs only separate the signals partially, so a class actually produced at position s_j may still be observed in all the NSBs that do not steer at s_j). By using the product combination rule [**172**] (i.e. assuming the output signals of the beamformers are independent), we have

$$p(c_i \mid X_k) = \prod_{j=1}^{P} p(c_i|s_j,X_k)p(s_j) = \prod_{j=1}^{P} p(X_k|c_i,s_j)p(c_i)p(s_j) / p(X_k) \quad (5.1)$$

where $p(X_k \mid c_i, s_j)$ is the likelihood of class c_i obtained from its corresponding HMM-GMM model.

For combining the posterior probabilities from the various microphone arrays, we will use again the product combination rule, so the optimal class c_o will be obtained with

$$c_o = \underset{c_i}{\operatorname{argmax}} \prod_{k=1}^{K} p(c_i|X_k) \quad (5.2)$$

In the case of *N* simultaneous sources and assuming they correspond to *N* different classes, the recognized identities of those classes are obtained by applying Eq. 5.2 *N* consecutive times and leaving each time the recognized class out.

This beamforming based approach for AED allows to easily assign the optimal class to one of the given source positions. In fact, the optimal position $s_o^{\,i}$ of the *i*-th event source out of the *N* simultaneous sources will be chosen as the one steered by the beamformers whose outputs show a maximum product of posteriors over all arrays given the optimal class:

$$s_o^{\,i} = \underset{s_j}{\operatorname{argmax}} \prod_{k=1}^{K} p(s_j|c_o,X_k) = \underset{s_j}{\operatorname{argmax}} \prod_{k=1}^{K} p(X_k|c_o,s_j)p(s_j) / p(X_k) \quad (5.3)$$

5.3 MAP based detection of overlapped acoustic events

Acoustic event detection requires both segmentation of the audio stream, and classification of the segments. We perform simultaneous segmentation and classification using state-of-the-art methods similar to the continuous speech recognition ones [**180**].

The goal of AED for a single channel can be formulated as follows: find the event sequence that maximizes the posterior probability of the event sequence $\Omega = (c_1, c_{1,\ldots,}c_M)$ given the observations $O = (o_1, o_2, \ldots, o_T)$:

$$\hat{\Omega} = \arg\max_{\Omega} P(\Omega \mid O) = \arg\max_{\Omega} P(O \mid \Omega)P(\Omega) \tag{5.4}$$

The acoustic model $P(O|\Omega)$ for each AE uses HMM-GMM. $P(\Omega)$ is a prior probability of AE sequence Ω. In order to avoid the dependence of AE sequence to the particular recording scenario we assume that all sequences of AEs are equally probable. Feature set consists of 16 FF LFBE coefficients with their first time derivatives as described in Chapter 3. The observation distributions of the states are incrementally-trained Gaussian mixtures with continuous densities. Each HMM is trained with the signal segments belonging to the corresponding event class from development data, using the standard Baum–Welch training algorithm [**171**]. The HTK toolkit [**170**] is used for training and testing the HMM–GMM system. The HMM topology for each AE is determined during a cross-validation procedure on the development data. The number of emitting states and Gaussian mixtures per state depends much on the amount of available training data. Usually the number of emitting states for each meeting-room AE ranges from 1 to 5 and the number of Gaussian mixtures ranges from 2 to 16. For testing, the Viterbi algorithm is used to find the sequence of states with the highest probability, resulting in a sequence of detected AEs [**171**] [**53**].

For a set of K microphone arrays and P beamformers per array, we need to perform the detection at each path and we have to find the event sequence that maximizes the posterior probability of the event sequence $\Omega = (c_1, \ldots..,c_M)$ given observations $X^{k,j}=(X_1, \ldots\ldots, X_T)$ for channel j of array k:

$$\begin{aligned} \Omega^{k,j} &= \arg\max_{\Omega} \; p(\Omega \mid X^{k,j}) \\ &= \arg\max_{\Omega} \; p(X^{k,j} \mid \Omega)\, p(\Omega) \end{aligned} \tag{5.5}$$

By using the Viterbi decoding separately at each channel of each array for a given segment of observation, we have a PxK set of hypothesis, each having a sequence of detected events and their corresponding end-points. In the decision stage, the best channel for deciding

the start and end time-stamps of the event is chosen by maximizing the posterior probability among all the channels from all the arrays:

$$\hat{\Omega}_O = \underset{k,j}{\operatorname{argmax}} \; p(\Omega \mid X^{k,j}) \tag{5.6}$$

Only the segmentation information about the events (the end-points) is used obtained from the Eq. 5.6. However, to decide the class identity, we follow the classification methods described in Eq. 5.1 and 5.2. This detection methodology has an advantage of utilizing information from all the channels and from all the arrays.

Finally, the assignment of each optimal class to one of the given source positions is performed by Eq. 5.3.

5.4 Experiments

In our experimental work, we consider a meeting room scenario with a predefined set of 11 acoustic events plus speech [**3**] [**48**]. Like in [**48**], there may exist either 0, 1 or 2 simultaneous events, and, in the last case, one of the events is always speech. However, the reported experiments correspond to the case of two overlapped events, since it is the most general one. For training, development and testing of the system, we have used, as in [**48**], part of a publicly available multimodal database recorded in the same smart-room.

As the number of sources is P=2 in our scenario, two frequency invariant beamformers are used per array: NSB1 and NSB2. This beamformers are specific for each array. In our experiments, to set the directions of arrival of the beamformers (for both the target source and the null source) we use those positions, knowledge about which was gathered during the recording. One angle corresponds to the speech source (the speaker's position is static), and the other one is for the acoustic event. In this way, for a particular array, NSB1 steers to the AEs and nulls speech, and NSB2 steers to speech and nulls AEs.

Here, we have considered two different options regarding the positions of the acoustic events from which the DOAs are extracted. First, we have considered, for each array, the approximate DOA for the whole set of acoustic events. It is obtained as a DOA average over the AE source positions, which are known from visual inspection during recording. In this

case, we have designed a beampattern with a broader main lobe to approximately encompass all the positions of the acoustic events.So the beamformers are not adapted to each particular AE instance.And, in the second option, we have used the position of each specific event and therefore the beamformers are adapted to each particular AE instances.In this latter case, the beam steers to the direction of the specific event position and thus beampattern has narrower main lobe.

5.4.1 Metrics

To evaluate the performance of the systems, we have used two metrics: one for the evaluation of the classification system and the other for detection system.

The metric used for the evaluation of the classification system is the quotient between the number of correctly classified and the total number of occurrences of that class in the testing database, i.e. the classification accuracy (CA).

To evaluate the detection system, we use the metric AED-ACC. In support of CHIL evaluation campaigns a specific metric for AED technology evaluation has been defined. The metric referred to AED-ACC is employed to assess the accuracy AED systems. This metric is defined as the F-score (the harmonic mean between precision and recall):

$$AED-ACC = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5.7}$$

where

$$Precision = \frac{number\ of\ correct\ system\ output\ AEs}{number\ of\ all\ system\ output\ AEs},$$

$$Recall = \frac{number\ of\ correctly\ \det ected\ reference\ AEs}{number\ of\ all\ reference\ AEs}.$$

A system output AE is considered correct if at least one of two conditions is met: 1) There exists at least one reference AE whose temporal centre is situated between the

timestamps of the system output AE, and the labels of the system output AE and the reference AE are the same. 2) Its temporal centre lies between the timestamps of at least one reference AE, and the labels of both the system output AE and the reference AE are the same. Similarly, a reference AE is considered correctly detected if at least one of two conditions is met: 1) There exists at least one system output AE whose temporal centre is situated between the timestamps of the reference AE, and the labels of both the system output AE and the reference AE are the same. 2) Its temporal centre lies between the timestamps of at least one system output AE, and the labels of the system output AE and the reference AE are the same.

The AED-ACC metric was used in the last CLEAR'2007 [**39**] international evaluation, supported by the European Integrated project CHIL [**37**] and the US National Institute of Standards and Technology (NIST).

5.4.2 Classification and detection results

The HTK toolkit is used for training and testing the HMM-GMM based classification system [**170**]. As described in Chapter 3, there is one left-to-right HMM with three emitting states for each AE and silence. 32 Gaussian components with diagonal covariance matrix are used per state. Initially, each HMM is trained with the standard Baum-Welch algorithm using signals that have been processed with the beamformer NSB1 of a particular array. Indeed, when testing, the knowledge about the relative position (left/right) of AEs and speech is not used. For each array, the likelihoods are computed by using the same set of AE+silence models for the two beamformer outputs.

The testing results are obtained with all the 8 sessions (S01-S08) with a leave-one-out criterion, i.e. we recursively keep one session for testing, while all the other 7 sessions are used for training. As one of the purposes is to compare the new signal separation based approach with other methods (e.g. model based and BSS based methods as explained in Chapter 3), in the reported experiments, classification is performed uniformly for all the methods, i.e. the annotated time marks of the events are used. For the new technique, classification is carried out by combining the likelihood scores as indicated in Eq. 5.1. Both classes and positions are assigned flat prior probabilities in the reported tests. When using two

arrays, the optimal class is obtained by integrating the posterior probabilities according to Eq. 5.2.

Table 8 shows the classification results obtained with the proposed system, averaging over all the eight testing datasets, for all arrays, and their combination using a product rule. We have performed experiments for all the arrays, using visually inspected positions for AEs and a broad beam angle that approximately encompasses all the AE positions. It can be observed that a better result is obtained for the combination of all arrays than using any individual array.

Table 8 : Classification results obtained using a beamformer having a broad beam angle for all arrays individually and their combinations

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**Product rule based combination**
CA(%)	73.6	71.8	75.1	79.2	78.1	77.8	81.5

Similarly, we present the detection accuracy in Table 9. In detection, we observed that the product rule based combination of scores from all six arrays could not perform the best in terms of detection accuracy, but it is close to the highest accuracy that we get with individual array T5 and much better than the lowest one.

Table 9 : Detection results obtained using a beamformer having a broad beam angle for all arrays individually and their combinations

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**Product rule based combination**
AED-ACC (%)	69.1	64.4	70.2	73.2	76.3	73.1	75.1

We have also performed experiments in order to check the performance of the system when the beamformer is used for each occurrence of event instead of using one beamformer with a broad beam angle that encompasses all the AE positions. The experiments are conducted for the two types of DOA settings: 1. visually inspected AE positions during the recording of the database 2. ASL estimated AE position. As the ASL system used for estimating positions in 1-source scenario, i.e. when a single event occurs at any given time, the

second setting is presented here only as a reference. In addition, to check the performance of the system using the models generated from the signals only processed either by NSB1 or only by NSB2, we have performed experiments. Table 10 shows the classification scores when all six-microphone arrays are employed, either alone or in combination for the all above-mentioned situations. Two types of combinations are considered in these experiments: 1. Product rule based combination of likelihood scores 2. FI based fusion.

Table 10 : Classification performances of individual array and their combination

DOA setting	**Models generated by processing signals from**	**Array**						**Product rule based combination**	**FI Fusion**
		T1	**T2**	**T3**	**T4**	**T5**	**T6**		
Visually inspected AE positions	NSB1	87.5	82.9	82.9	80.2	86.3	86.7	87.6	87.6
	NSB2	85.2	86.4	85.3	84.5	78.1	87.3	89.2	88.8
	NSB1 and NSB2	87.5	86.8	85.4	83.7	83.7	86.1	89.1	88.5
ASL-estimated AE positions	NSB1	84.9	83.5	84.6	82.8	83.8	83.6	88.3	88.6
	NSB2	86.2	88.3	88	87.2	76.9	88.7	90.8	90.4
	NSB1 and NSB2	88.2	88	89.2	86.9	82.4	88.1	90.7	89.8

Note from the last two columns in Table 10 that the accuracy obtained from either the product rule based combination or the FI based fusion of the whole set of arrays is always higher than the accuracy obtained from any of the single arrays. Among the two types of combination, product rule based combination works slightly better than the FI based one. The use of both the models generated from the signals processed by NSB1 and NSB2 improves significantly the CA with respect to using either one of them. It is also observed that most of the time we get higher CA while using the models generated from the NSB2 processed signals than using the models generated from signals processed by NSB1. Regarding the type of DOA setting, the ASL-estimated AE position based system works better than the one that uses the DOA from the visually inspected and approximate AE position during the recording of the database. On the other hand, designing the beamformer for each AE proves improvement than using an average DOA based on visual inspection (i.e. a broad-beam nulling angle presented

in Table 8). That could be expected, since the beam pattern is specific of each event class, whereas the broad beam encompasses all the angles of the AE source positions. While the latter design simplifies the overall system, as it does not require a precise source position (hence a single beamformer is used for all AEs) and may also avoid an additional external ASL block, it is specific of the given scenario, so it has to be re-designed when the scenario changes. The detection results for all the possibilities discussed above are presented in the following Table 11. As usual, most of the time the accuracy obtained from either the product rule based combination or the FI based fusion of the whole set of arrays is higher than the accuracy obtained from any of the single arrays. In some cases, the combinations of arrays do not give better accuracies, but always it is very close to the highest one and much better than the lowest one produced by an array.

Table 11 : Detection performance (AED-ACC in %) for each single array and their combination

DOA setting	**Models generated by processing signals from**	**Array**						**Product rule based combination**	**FI Fusion**
		T1	**T2**	**T3**	**T4**	**T5**	**T6**		
Visually inspected AE positions	NSB1	86.8	81	81.5	78.9	85.5	82.4	85.8	85.9
	NSB2	83.4	85.2	82.9	82.8	77.7	85.4	86.3	86.1
	NSB1 and NSB2	85.6	84.8	82.9	82.6	82.8	85.1	87.2	86.9
ASL-estimated AE positions	NSB1	84.3	81.9	83.3	82.3	83	82.5	86.1	86.2
	NSB2	85.8	86.5	86.1	84.7	77	86.8	88.1	87.7
	NSB1 and NSB2	87.2	86.4	87.6	85.7	81.8	87.1	88.7	88.7

In addition, for comparison purposes, classification results obtained with two very different techniques are reported in Table 12. In coherence with how we train the models for our proposed technique, separated signals have been used for training the corresponding blind source separation (BSS)based system, which is based on the deflation method [**94**]. The model based technique uses a set of 11 models (plus silence) for the acoustic events overlapped with speech. The BSS based system also uses three microphones (T6), and the model based system

uses signal from only one microphone from the same array. The technique presented in this Chapter is using a frequency invariant beamforming at its front end. So we represent it as FIB from now onwards. From the RA results presented in the Table 12, it is clear that the FIB based AE classification system that uses all the six microphone arrays produces superior performance than the other two systems. The model based system shows a higher accuracy than the more computationally demanding BSS based system when both arrays are used.

Table 12 : Comparison of classification results for three different systems

	Source separation based		Model based
	FIB	BSS	
CA(%)	89.1	80.8	83.8

The AED results of the above mentioned 3 systems are presented in the following Table 13. Our proposed FIB based system again outperforms the other two by a significant margin.

Table 13 : Detection accuracies of the three compared systems

	Source separation based		Model based
	FIB	BSS	
AED-ACC (%)	87.2	77.6	80.1

To verify the reliability of the proposed method, we have also performed experiments with the other database (T recordings) where AEs and speech are already overlapped in the recordings from a real scenario. In this database, as explained in Chapter 3, two persons interact spontaneously and the AEs are produced along with speech. The classification results of the proposed system when using either the models generated from the signals processed by both NSB1 and NSB2, for all the arrays, or their combinations are presented in Table 14. The testing results are obtained with all the 9 sessions (T01-T09) with a leave-one-out criterion, i.e. we recursively keep one session for testing, while all the other 8 sessions are used for training. The classification results shown in Table 14 are obtained by averaging over all the classes and all the nine testing datasets, for all arrays and their combination. Like in the previous experiments, it is observed that a better result is obtained for the combination of all arrays than using any individual array. The FI based fusion slightly improved the classification performance with respect to the product rule based combination. The detection results are

presented in Table 15. As a comparison, the detection result for the model based AED system is also shown in the same table. The model based system that uses only one microphone is evaluated with the same experimental setup that the proposed FIB based AED system. It is evident from the experimental results that the FIB based AED performance is little superior to the model based AED. Here the combination could not produce the best result but it is very close to the highest accuracy that we get using individual array and far better than the lowest one.

Table 14 : Classification results with T-recording

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**Product rule based combination of arrays**	**FI fusion**
CA (%)	82.9	82.2	82.9	80.6	80	82.7	83.7	84.3

Table 15 : Detection accuracy comparison of the proposed FIB based and the model based AED using T-recording

	FIB based AED								**Model based AED**
	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**Product rule based combination of arrays**	**FI fusion**	
AED-ACC (%)	79.1	79.2	77.1	77	75.5	79.9	78.7	78.2	65.6

5.4.3 PA result

It is also noticeable that, among the different evaluated techniques, only the proposed FIB based system is able to assign the hypothesized classes to the given source positions without requiring an extra system for that. Its position assignment capability is evaluated with the following experiments.

To have a complete description of the acoustic scene in the smart room, there is still the need of assigning each one of the two positions to each one of the two detected events. The position assignment is done at the decision block according to Eq. 5.3, after the optimal class

is chosen through event classification. In our particular scenario, the optimal event class (1 of 11 AEs) is assigned to one of the source positions, and the other position is assigned to speech.

A position assignment rate (PAR) and difference in log-likelihood (Diff_LL) metrics are used for the PA evaluation and described in the previous Chapter 4. The results with those metrics, averaging over all AEs in the 8 S-recording testing datasets, are presented in the Table 16, for all individual arrays and for the product rule based combination of likelihood ratios. The PA results with the T-recordings are presented in Table 17. Again, the combination of arrays produces the best result. In this case, the tests have been carried out by using two models, which are generated from the processed signals by NSB1 and NSB2. In the decision block, instead of maximizing the product of likelihoods, we maximize the product of the likelihood ratios corresponding to each beamformer. As presented in the Table 16, this system produces better PAR than the best PA result reported in the previous Chapter 4. This improvement may be due to the models used for calculating the likelihood scores. Here, we have used the separated signals after the beamforming for generating those models. However, the system explained in the previous Chapter 4 uses the models generated from the mono-event signals, not from the separated signals. Therefore, in that case, there is a mismatch between the training and testing dataset, which may be the reason for the reduction of PA accuracies. As usual, with the proposed system, a significant PAR improvement is obtained when using the whole set of arrays with respect to the case of using only a single array.

Table 16 : Position assignment results with S recordings

	PA after	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**Product rule based combination of LR**
PAR (%)	Classification	87.2	87.3	91.1	93.6	94.4	94.6	98.2
	Detection	87.4	86.1	90.5	93.4	93.5	93.4	97.1
Diff_LL	Classification	0.78	1.41	1.12	1.5	1.2	1.1	6.6
	Detection	0.7	1.37	1.04	1.3	1.1	0.9	6.4

Table 17 : Position assignment results with T recordings

	PA after	T1	T2	T3	T4	T5	T6	Product rule based combination of LR
PAR (%)	Classification	92.6	93.8	93.1	92.9	95.2	92.4	98.5
	Detection	92.2	93	93.1	93.1	94.6	91.1	97.9
Diff_LL	Classification	0.96	2.31	4.6	0.76	5.71	0.85	15.8
	Detection	0.82	2.22	4.49	0.8	5.99	0.66	15.3

5.5 Chapter conclusions

In this Chapter, a new approach for computationally efficient classification/detection and the positioning of acoustic events that result from the combination of a beamforming based partial signal separation and a MAP based decision has been presented. Assuming both the source positions and the start/end timestamps of the events, the proposed system has been tested for classification and position assignment in a scenario with two sources. In addition, assuming only the source positions, a system is proposed that performs the detection task and outputs both the class identity and the start/end timestamps of the detected events. The core system consists firstly of a set of frequency invariant null steering beamformers to carry out different partial signal separations for each microphone array. The beamformers are followed by model based likelihood calculations, using the models generated from the signals processed by the beamformer from each of the parallel path in order to get the maximum information from the scenario to obtain the posteriors. These scores are then combined using product rule in the decision block, a MAP is used to get the optimal class, and the optimal position for PA.

The proposed system performance is compared with a model based and BSS based systems. The model based system, which is the previous baseline, has a problem of limited scalability regarding both the number of event classes and the number of simultaneous sources. However, the proposed system can overcome this problem. On the other side, the BSS technique is computationally demanding and not suitable for real-time applications. The proposed technique is computationally efficient and can overcome this problem as well. In spite of having these advantages, the proposed FIB based system that uses several small, distributed microphone arrays performs significantly better than the other two methods mentioned above. As the used beamforming technique only partially separates the different sources, we have tried to use as much information as possible from the given scenario by using the models generated from the signals processed by all the beamformers. We found a noticeable improvement in the performance of the system while combining the microphone arrays using a product rule based combination in the decision block before the MAP criterion is applied.

Additionally, both the model based and the BSS based approaches suffer from the permutation problem. They require a separate PA system to solve it. However, the proposed technique includes the posterior assignment of event hypothesis to source positions that the other two techniques do not have.

Chapter 6. Acoustic event localization in a meeting-room scenario

6.1 Chapter overview

Automatic acoustic scene analysis in a room environment using distributed microphone arrays is a very challenging task. The task is even more difficult in a multiple source scenario. In this context, acoustic source localization and sound recognition are common acoustic scene analysis tasks that are usually considered separately. Most source localization techniques rely on some kind of measurement of the acoustic energy as a function of space In this Chapter, a new source localization technique is proposed that works jointly with a sound recognition system. Once the identities and the endpoints of the simultaneous sounds are known, the proposed technique uses the statistical models of those sounds to compute a likelihood score for each model and each position in the room space. Those scores are subsequently combined to find the MAP-optimal positions in the room.

The experimentation of this technique is carried out in the scenario considered in this thesis, consisting of meeting-room acoustic events, either isolated or overlapped with speech.

A brief description about the existing localization technique is presented in Section 2. In Section 3, we present our proposed sound-model-based localization technique in detail. Experimental work on acoustic event localization is reported in Section 4, and a conclusion is given in Section 5.

6.2 SRP-PHAT based acoustic source localization

The task of source localization in smart environments itself is a very challenging due to the presence of influencing factors like noise, reverberations, directivity of the acoustic sources. The task is even more difficult in the situation when multiple acoustic sources are active simultaneously.

In this Section, we describe a standard source localization technique based on SRP-PHAT algorithm. The reason for choosing this algorithm for the baseline system is due to its robustness against reverberation and its relative independence on speaker orientation [**155**] [**181**].

The basic operation of the SRP-PHAT algorithms consists of exploring the 3-dimensional (3D) space, searching for the maximum of the global contribution of the PHAT-weighted cross-correlations from all the microphone pairs. The 3D room space is quantized into a set of positions with typical separation of 5-10 cm. The theoretical TDOA $\tau_{,p,i,j}$ from each exploration position to each microphone pair are pre computed and stored. The estimated acoustic source location is the position of the quantized space that maximizes the contribution of the cross-correlation of all microphone pairs:

$$\hat{p} = \underset{p}{\operatorname{argmax}} \sum_{i,j \in S} R_{m_i m_j}(\tau_{p,i,j}) \tag{6.1}$$

where S is the set of microphone pairs and the cross-correlation with PHAT weighting is computed as:

$$R_{m_i m_j}(\tau) = \int_{-\infty}^{\infty} \frac{G_{m_i m_j}(f)}{\left| G_{m_i m_j}(f) \right|} e^{j2\pi f \tau} \, df \tag{6.2}$$

The set of cross-correlation functions can also be combined to create a Spatial Likelihood Function (SLF) $\phi(p)$, which gives a score for each position x in space by means of the following equation:

$$\phi(p) = \sum_{m=1}^{M} R_m(\tau_{p,m}) \tag{6.3}$$

The phase transform (PHAT) is an especially effective weighting of a GCC for finding a TDOA from acoustic signals in highly reverberant environment. The process is thus to explore over the whole volume and ultimately find the set of one or more distinct maxima. The calculation of any particular point is called a functional evaluation. For the SRP-PHAT functional, a point-source location in the room that gives the maximum value of cross-correlation is determined. Instead of a grid-search, which requires functional evaluation on a fine grid throughout the room, a stochastic region contraction (SRC) is used to find the global maximum as presented in [**162**]. The basic idea of the SRC algorithm is, given an initial rectangular search volume containing the desired global optimum and perhaps many local maxima or minima, gradually, in an iterative process, contract the original volume until a sufficiently small sub-volume is reached in which the global optimum is trapped. The contraction operation on iteration is based on a stochastic exploration of the SRP-PHAT function in the current sub-volume. The main advantages of using the fast optimization technique like SRC are: 1) a more robust procedure against an early wrong decision, and 2) an allowance of the optimum being on the continuum.

6.3 Sound-model-based acoustic source localization

In this Section, we propose that the information about the content of the signals that are captured by distant and distributed microphones may be effectively used for localization of the sources in the space domain. In fact, instead of relying only on energy-like based measures, we propose a similarity measure delivered by a classifier. As the classifier uses models for the different sound classes, we have named this method as sound-model-based (SMB) localization. The identity of the simultaneous sounds and their time positioning are given by an acoustic detection system (AED) as the one presented in previous Chapter 5. Therefore, the sound models are shared by both AED and ASL systems.

In this work, we assume that detection has already been carried out, so the identities and the end points of a set of acoustic events are known. The acoustic events may occur either isolatedly or simultaneously. The proposed system is shown in Figure 19. Let us assume a room with a set of *K* microphone arrays which can be located arbitrarily; for deployment, this is an advantage with respect to using spatially-structured array configurations. The 2-D room space is divided into a set of *P* pre-defined small-area cells.

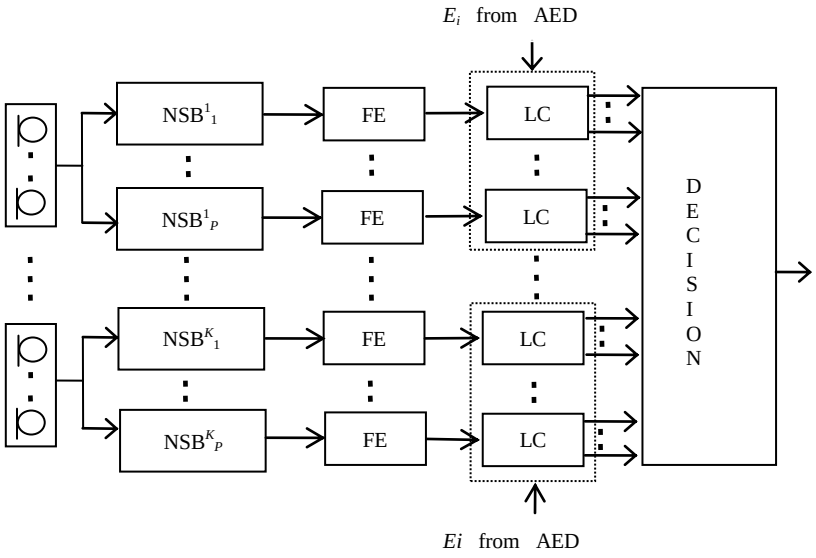

Figure 19 : Sound model-based acoustic source localization system

Note that the vertical coordinate is not considered in this study, but it could also be included. For each microphone array, there is a set of *P* beamformers (NSB), each one attenuating the signals from all the directions except the center of one of the cells. This beamformer is capable of attenuating signals for the different positions in the sensor array patterns. To determine the beamformer coefficients for the wide band audio signals, we use a technique called frequency invariant beamforming (FIB) which is already described in Chapter 3.

6.3.1 Use of the MAP criterion

The output signal of each beamformer enters a classification system. After the feature extraction (FE), a likelihood score (LC) is computed for each of the event classes (hypothesized from an external AED system), by using previously trained acoustic event models (that may be the same models used by the AED system). Let us assume a room with K microphone arrays, and a set of N (possibly simultaneous) events E_i, $1 \le i \le N$, that belong to a set of C different classes. Given a grid of positions s_j, $1 \le j \le P$, in the room, for each array, there is a set of P NSBs, so that the j-th NSB is placing nulls in the directions of the P positions except that of position s_j. Therefore, there is a set of P output signals from each array processor. For a given event E_i, a set of P likelihood scores are obtained from the NSB outputs and using the model of the class E_i. The optimal position $s_o^{\,i}$ of that i-th event out of the N events is chosen to maximize a product of posterior probabilities, i.e.

$$\begin{aligned} s_o^{\,i} &= \underset{s_j}{\operatorname{argmax}} \prod_{k=1}^{K} p(s_j | E_i, X_k) \\ &= \underset{s_j}{\operatorname{argmax}} \prod_{k=1}^{K} p(X_k | E_i, s_j) p(s_j | E_i) / p(X_k) \end{aligned} \tag{6.4}$$

To determine the likelihoods, the acoustic events are modeled with Hidden Markov models (HMM), and the state emission probabilities are computed with continuous density Gaussian mixture models (GMM). An illustration of how the log-likelihood score varies along the angles is shown in Figure 20; there is a minimum for a specific angle, which is the true DOA of the given class.

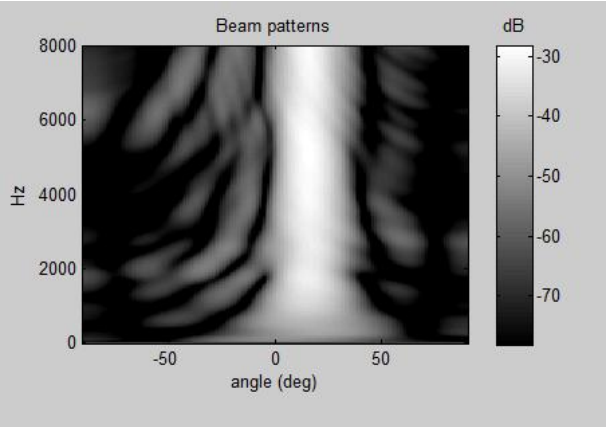

Figure 20 : FIB beam pattern; nulls of the beamformer are placed in other directions except the center of a pre-defined cell

6.4 Experiments

In our experimental work, we consider a meeting room scenario with a predefined set of 11 acoustic events. Like in our previous experiments, we assume that there may simultaneously exist 0, 1 or 2 events, and, in the last case, one of the events is always speech. In the reported experiments, we localize the isolated events (1 source) and overlapped events (2 sources).

The steered-beamforming sound-model-based (SB-SMB) system in Figure 19 is used to localize the acoustic event sources in the room environment. The nulls of the beamformers are placed in other directions except the centers of the pre-defined cells. To facilitate real time processing, we have considered a relatively large cell: 0.6x0.8m. Though a larger cell reduces the resolution of the ASL in the room, it indeed reduces the number of beamformers, which in turn ensures less computational load. In the proposed system, the beamformers are designed to work with the horizontal row of 3 microphones each array in the smart-room has.

In the feature extraction block of the system depicted in Figure 19, like the experiments in the previous Chapters, a set of audio spectro–temporal features is computed for each signal frame. The frames are 30 ms long with 20 ms shift, and a Hamming window is applied. We have used frequency-filtered log filter-bank energies (FF-LFBE) for the parametric representation of the spectral envelope of the audio signal. For each frame, a short-length FIR filter with a transfer function $z\text{-}z^{-1}$is applied to the log filter-bank energy vectors and end-

points are taken into account. Here, we have used 16 FF-LFBEs along with their 16 first temporal derivatives, where the latter represents the temporal evolution of the envelope. Therefore, the dimension of the feature vector is 32.The HTK toolkit is used for developing the HMM-GMM based classifier. There is one left-to-right HMM with three emitting states for each AE and silence. 32 Gaussian components with diagonal covariance matrix are used per state. Initially, each HMM is trained with the standard Baum-Welch algorithm using the beamformed signals for a particular array. For each array, the likelihoods are computed by using the same set of acoustic event models for all the beamformer outputs.

The optimal source position is obtained by maximizing the integrated posterior probabilities over all microphone-arrays for the given optimal class according to 6.4. All the positions are assigned flat prior probabilities in the reported tests. The number of position that we need to estimate is assumed and could be provided by the AED system. For 0 sources, as silence is detected, therefore the system does not need any output position. In the experiments with both isolated and overlapped acoustic events, the optimized position is found by maximizing the integrated posterior over all the arrays, given the optimal class.

6.4.1 Proposed metrics

To test the performance of the model based localization system, we will use two metrics. 1) Acoustic source localization cell error (Cell error): it is defined as the quotient between the number of localization errors and the total number of event occurrences in the testing database. For an event *i*, a localization error occurs when the cell assigned to the true position is not the same as the one estimated by the ASL system. The true position for each event was obtained from visual inspection during the recording of the signal. 2) Root-mean-squared error for localization (RMSE), which is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_e}\sum_{i=1}^{N_e}\left(\frac{\left|x_i^{test} - x_i^{ref}\right|}{\Delta x} + \frac{\left|y_i^{test} - y_i^{ref}\right|}{\Delta y}\right)^2} \tag{6.5}$$

where (x_i^{test} , y_i^{test}) is the 2-D estimated position of the test event, and (x_i^{ref} , y_i^{ref}) is its corresponding reference (true) position. Δx and Δy are the separations along x and y axis of the

pre-defined positions which are considered by quantizing the room space. N_e is the total number of event samples in the testing session.

6.4.2 Results

The testing results are obtained with all the 8 sessions (S01-S08) with a leave-one-out criterion, i.e. we recursively keep one session for testing, while all the other 7 sessions are used for training. Table 18 shows the results obtained with two metrics for the evaluation of the proposed SB-SMB system. As a comparison, in the same table we have also put the result for a SRP-PHAT localization system. The results with the two metrics, averaging over all AEs in the 8 testing datasets, are obtained using all the six arrays (T1 to T6) available in the room.

Table 18 : Performance comparison of isolated (1-source) ASL systems

	SB-SMB	SRP-PHAT
Cell error (%)	13.4	13.8
RMSE	0.41	0.44

The results obtained with the SB-SMB system consider flat values for both $p(s_j|E_i)$ and $p(X_k)$. It is worth noticing that, in the proposed method, we follow an event based approach, which means we localize the whole event. Due to that event based approach, we have to assume that during the whole event the acoustic source is not moving along space. For that reason, we exclude ‘steps’ and ‘chair moving’ from the evaluation. In our approach, the source position is estimated assuming the identity of the event class, and also the start and end points of its occurrence are known. In addition, instead of using the AED system output to set the AE model used by the likelihood calculators in the classifier, the ground truth was used, so the errors from the AED system are not affecting in our tests the measure of localization performance.

In the experiments, the proposed system shows a slightly lower cell error rate than the conventional SRP-PHAT system.

In Table 19, we present the performance scores for the acoustic events in the overlapped case (when an acoustic event is overlapped with speech). In this case, the proposed system

clearly outperformed the SRP-PHAT based system. Also notice that the cell error rate for the proposed technique is only around 1.2% higher than the case of isolated events for the proposed system. Note also that the proposed method may take advantage of the knowledge about the a-priori probabilities of the pre-defined positions for each event class.

Table 19 : Sound-model-based ASL system performance for overlapped (2- source) acoustic events

	SB-SMB	SRP-PHAT
Cell error (%)	14.5	44.2
RMSE	0.53	2.6

6.5 Chapter conclusions

In this Chapter, a novel approach for acoustic source localization based on models of the sounds has been presented which combines a set of beamformers and a MAP based decision. When tested in a meeting-room scenario, the one-source localization performance of the proposed system is slightly better than that of the widely used SRP-PHAT based system, while it is significantly better in the more complex two-source scenario, provided that the exact information about classes and time end-points are available. Note that, unlike the SRP-PHAT system, the SBSMB system requires the identities and the time end-points of the events. However, it may take advantage of the a-priori probabilities of the pre-defined positions for each event class, though they were not used in the experiments. In summary, the presented SBSMB localization technique can be an alternative for localization in a multiple source scenario when it works together with an acoustic event detection system, with the additional advantage that both use the same framework.

Chapter 7. Joint classification and localization of meeting-room acoustic events using distributed microphone arrays

7.1 Chapter overview

Acoustic scene analysis usually requires several sub-systems working in parallel for carrying out the various required functionalities. Focusing to a more integrated approach, in this Chapter we present an attempt to jointly recognize and localize several simultaneous acoustic events that take place in a meeting room environment, by developing a computationally efficient technique that employs multiple arbitrarily located small microphone arrays. First, a joint technique for classification and direction of arrival estimation (which is localization from just one array) will be considered. Assuming a set of simultaneous sounds, for each array a matrix is computed whose elements are likelihoods along the set of classes and a set of discretized directions of arrival. MAP estimation is used then to decide about both the recognized events and the estimated directions. In a second step, 2D localization that integrates information from all the arrays in the room environment will be considered.

The classification plus direction-of-arrival estimation technique is presented in Section 2. In Section 3, classification plus sound-model-based localization technique will be presented in detail. In the same section, a joint, parallel classification and localization technique is described. Experimental work on both of these methods is reported in Section 4, and a conclusion is given in Section 5.

7.2 Classification plus direction-of-arrival estimation

Although classification and localization may work separately for the acoustic scene analysis task, it can be expected that an integrated approach offer advantages in terms of system performances. In Chapter 5, we have implemented a classification system based on signal separation that can easily work in real time by using multiple distributed linear microphone arrays composed of a small number of microphones. The signal separation was done using a frequency independent beamforming at the front end, and assuming the source positions. In this Section, we aim to take a step further in the direction of the integrated approach, by avoiding the assumption that the various acoustic source positions are known, so avoiding the need of an external and specific ASL sub-system. In fact, we present here a new technique as an attempt of jointly recognizing and estimating DOAs, in an unambiguous way, the classes, and the positions of the N simultaneous sounds. Assuming only the x-y plane is needed; this is done by discretizing, for each microphone array, the direction of arrival (DOA) with M angles, and building for each event class a sequence of posterior probabilities along the angle axis. In this way, for each array, i.e. for each multi-channel signal, we have a matrix, where each element of that matrix is the likelihood for a given class and a given angle. The hypothesized event classes are determined from that likelihood matrix by applying the MAP criterion. The angle for which the posterior of a given hypothesized class shows a minimum is taken as the estimated localization angle.

7.2.1 Methodology

In this approach, we aim to build for each microphone array a posterior matrix that contains information about both the identity of the acoustic events that are simultaneously present in the room and the direction of arrival of their acoustic waves to the array. Then, both the identities of the sounds and their DOAs will be estimated with a MAP criterion. As shown in Figure 21, at the front end of the proposed system, the multi-channel signal collected by each of the microphone arrays is driven to a set of null-steering beamformers (NSB). Each NSB is placing a null to a different value of the angular variable θ, which is discretized in M values that uniformly span the angle interval $(-\pi, \pi)$. Note that the vertical coordinate is not

considered in this study. Feature extraction (FE) is then applied at the output of the beamformer, to subsequently compute a set of likelihood scores (LC), by using previously trained HMM-GMM models for the set of *C* acoustic event classes.

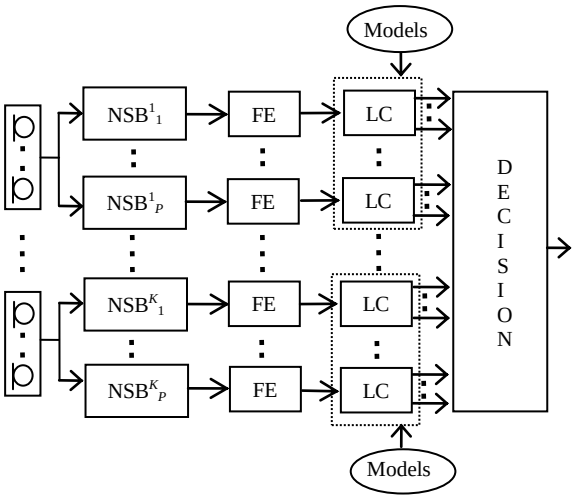

Figure 21 : Joint event classification and localization system

Consequently, when *K* arrays are used, *K*x*M*x*C* likelihood scores are fed to the last block, where a MAP criterion is used to take the decision about the identity of the acoustic events $E_1 \ldots E_N$, and their directions of arrival $\theta_1 \ldots \theta_N$. Note that the number *N* of acoustic sources is hypothesized in this work.

7.2.1.1 Beamformer placing null at the direction of interest

Null steering beamforming (NSB) allows us to design a sensor array pattern that steers the main beam towards a specific direction, places null in that direction and allows the signals to pass from the rest of the directions. The detail of this technique is described in Chapter 3. Indeed, in our case we cannot expect with this approach a perfect separation of the different mixed signals at the output of the NSB, since we use a small number of microphones per array, and because of echoes and room reverberation.

Assuming the end-points of the events are known, it becomes a classification problem. To determine the likelihoods, the acoustic events are modeled with Hidden Markov models (HMM), and the state emission probabilities are computed with continuous density Gaussian mixture models (GMM), like the previous systems in the last Chapters.

Let us assume we have a set of N simultaneous events E_i, $1 \leq i \leq N$, that belong to a set of C classes. For each of the K microphone arrays, there is a set of M beamformers, each one having a null to a different angle θ_j. So there is a set of M output signals for each array, and, after likelihood computations with the HMM-GMM models, we have a MxC-dimensional matrix of likelihood scores, that can be seen as a set of C patterns along the angle variable. An example of such patterns for two different events is shown in Figure 22.

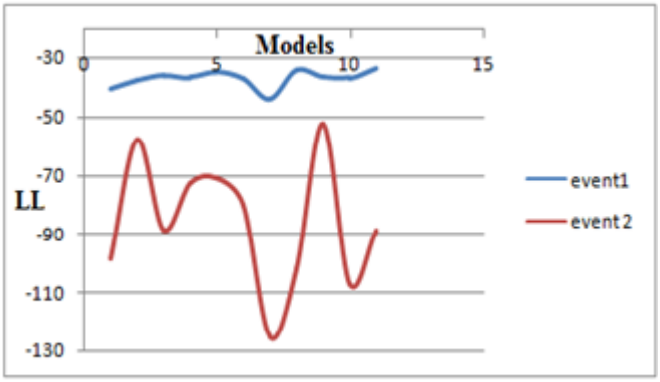

Figure 22 : Patterns of log-likelihood along the 11 models for two different events

Let's denote with X_k the multi-channel signal corresponding to the k-th array (notice that, to simplify notation, we do not consider time indices). We want to determine the posterior probability of a given class c_i for the k-th array through all the NSBs. Note that our NSBs only separate the signals partially, so a class actually produced at the angle θ_j may still be observed in all the NSBs that do not place nulls at θ_j. We will assume that each angle θ_j has an associated prior probability $p(\theta_j)$. By using the product combination rule [**172**] (i.e. assuming the output signals of the beamformers are independent), we have:

$$p(c_i \mid X_k) = \prod_{j=1}^{M} p(c_i|\theta_j, X_k) p(\theta_j) \\ = \prod_{j=1}^{M} p(X_k|c_i,\theta_j) p(c_i)\, p(\theta_j) / p(X_k) \tag{7.1}$$

where $p(X_k|c_i,\theta_j)$ is the likelihood of class c_i for the multichannel signal X_k after it goes through beamformer j, which is obtained from the corresponding HMM-GMM model.

For combining the posterior probabilities from the various microphone arrays, we will use again the product combination rule, so the optimal class c_o will be obtained with

$$c_o = \underset{c_i}{\operatorname{argmax}} \prod_{k=1}^{K} p(c_i|X_k) \tag{7.2}$$

In the case of N simultaneous sources, and assuming they correspond to N different classes, the recognized identities of those classes are obtained by applying Eq. 7.2 N consecutive times and leaving each time the recognized class out. In this work we use a data-dependent likelihood-to-posterior transformation to compute the probabilities $p(c_i|\theta_j,X_k)$ involved in the first line of Eq. 7.1.

The optimal DOA $\theta_o^{\,i}$ of the i-th event source out of the N simultaneous sources is chosen according to:

$$\theta_o^{\,i} = \underset{\theta_j}{\operatorname{argmin}}\ p(\theta_j \,|\, c_i, X_k) \\ = \underset{\theta_j}{\operatorname{argmin}}\ p(X_k \,|\, c_i, \theta_j)\, p(\theta_j) \tag{7.3}$$

where the minimum is taken because a null is placed by the beamformers in the direction of the position, not a maximum. Figure 23 shows an illustration of the variation of the likelihood scores along the angles; and there is a minimum for a specific angle, which actually is the true DOA of the given class.

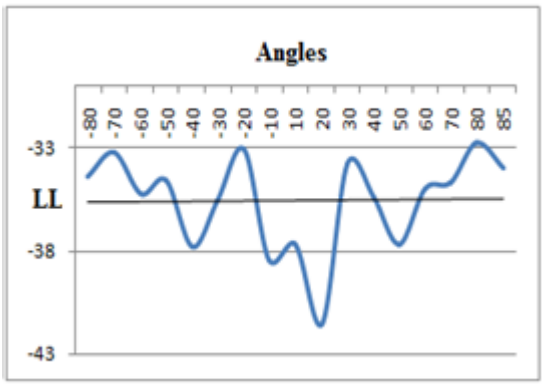

Figure 23 : Log-likelihoods along angles

7.2.1.2 Beamformer attenuates signal from all directions other than the direction of interest

This type of beamforming allows us to design a sensor array pattern that steers the main beam towards the desired source, and attenuates the signal from rest of the directions. Unlike the previous case, it allows to pass the signals from a very narrow direction in the given space. Here also, we use Eq. 7.1 and 7.2 for the classification decision. We have explored two options. 1) Classification decision is made from the likelihoods generated from the HMM-GMM classifiers using the second line of Eq. 7.1 and then applying Eq. 7.2. 2) A data-dependent likelihood-to-posterior transformation is used to compute the probabilities $p(c_i|\theta_j,X_k)$ involved in the first line of Eq. 7.1 and then applying Eq. 7.2, just like the system explained in the previous sub-section.

The optimal DOA $\theta_o^{\,i}$ of the *i*-th event source out of the *N* simultaneous sources is chosen according to:

$$\begin{aligned}\theta_o^{\,i} &= \underset{\theta_j}{\operatorname{argmax}}\ p(\theta_j \,|\, c_i, X_k) \\ &= \underset{\theta_j}{\operatorname{argmax}}\ p(X_k \,|\, c_i, \theta_j)\, p(\theta_j)\end{aligned} \qquad (7.4)$$

where the maximum is taken because the beamformer allows the signal to pass from the direction of of interest, not attenuating.

7.3 Joint classification and localization

In this Section, we present a new technique as an attempt of jointly recognizing and localizing, in an unambiguous way, the classes, and the positions of the *N* simultaneous sounds. Although it is possible to localize the sources, after estimating DOAs for each array and then combining them using intersection [**139**] [**143**], but the sound-model-based (SB-SMB) localization technique from Chapter 6 is considered to be used with the classification system for this joint approach. It has been seen that the SB-SMB localization technique that performs well even in multiple source scenarios and competes favorably with a standard SRP-PHAT based localization technique.

The block diagram of the methodology is similar to the one depicted in Figure 21. Let us assume a set of *K* microphone arrays, which can be located arbitrarily. The 2-D room space is divided into a set of *P* pre-defined small-area cells. Note that the vertical coordinate is not considered in this study, but it could also be included. For each microphone array, there is a set of *P* beamformers (NSB), each one attenuating signals from all the directions except of the centre of one of these pre-defined cells. The output signal of each beamformer enters a classification system. After feature extraction (FE), a likelihood score (LC) is computed for all the event classes, by using previously trained acoustic event models generated from the signals processed by the beamformer. It means that the beamformers allow the signal to pass from the true source direction and attenuating signals from all other directions. Finally, a decision module carries out the localization of the events by combining the likelihood scores using a MAP criterion.

7.3.1 Classification-plus-localization

Given a room with *K* microphone arrays, let us assume we have a set of *N* (possibly simultaneous) events c_i, $1 \leq i \leq N$, which belong to a set of *C* different classes. Given a grid of positions s_j, $1 \leq j \leq P$, in the room, for each array, there is a set of *P* NSBs, so that the *j*-th NSB is placing nulls in the directions of the *P* positions except that of position s_j. So from array processing, we have a set of *P* output signals for each array, and after likelihood computations with the models of all classes, we have a *P*x*C*x*K*-dimensional vector of likelihood scores. We

want to determine the posterior probability of a given class c_i for the k-th array through all the beamformers. Note that our beamformers only separate the signals partially, so a class actually produced at the position s_j may still be observed in all the other beamformers that allow the signal to pass other than s_j. We will assume that each angle s_j has an associated prior probability $p(s_j)$. By using the product combination rule (i.e. assuming the output signals of the beamformers are independent), we have:

$$\begin{aligned} p(c_i \mid X_k) &= \prod_{j=1}^{S} p(c_i|s_j, X_k) p(s_j) \\ &= \prod_{j=1}^{S} p(X_k|c_i, s_j) p(c_i) p(s_j) / p(X_k) \end{aligned} \tag{7.5}$$

where $p(X_k|c_i,s_j)$ is the likelihood of class c_i for the multichannel signal X_k after it goes through beamformer j, which is obtained from the corresponding HMM-GMM model.

For combining the posterior probabilities from the various microphone arrays, we will use again the product combination rule, so the optimal class c_o will be obtained with:

$$c_o = \underset{c_i}{\operatorname{argmax}} \prod_{k=1}^{K} p(c_i|X_k) \tag{7.6}$$

In the case of N simultaneous sources, and assuming they correspond to N different classes, the recognized identities of those classes are obtained by applying Eq. 7.6 N consecutive times and leaving each time the recognized class out. For classification, we have explored two options like we did in the previous Section: 1) Classification decision is made from the likelihoods generated from the HMM-GMM classifiers using the second line of Eq. 7.5 and then applying Eq. 7.6. 2) A data-dependent likelihood-to-posterior transformation is used to compute the probabilities $p(c_i|s_j,X_k)$ involved in the first line of Eq. 7.5 and then applying 7.6.

Once the classification decision is made, the optimal position s_o^i of the i-th event source out of the N simultaneous sources is chosen to maximize a product of posterior probabilities, i.e.

$$s_o^i = \underset{s_j}{\operatorname{argmax}} \prod_{k=1}^{K} p(s_j|c_i,X_k)$$
$$= \underset{s_j}{\operatorname{argmax}} \prod_{k=1}^{K} p(X_k|c_i,s_j)p(s_j|c_i) / p(X_k) \quad (7.7)$$

To determine the likelihoods, the acoustic events are modeled with HMM, and the state emission probabilities are computed with continuous density GMM like in the previous cases.

7.3.2 Joint, parallel classification and localization

In this approach, both classification and localization decisions are made jointly in parallel. The whole system is similar to the one that is depicted in Figure 21. We have a set of *P* output signals for each array, and after likelihood computations with the models of all classes, we have a *P*x*C*x*K*-dimensional vector of likelihood scores. We want to determine the posterior probability of a given class c_i and position s_j for the *k*-th array,

$$p(c_i, s_j \mid X_k) = p(X_k \mid c_i, s_j)p(c_i)p(s_j) / p(X_k) \quad (7.8)$$

For combining the posterior probabilities from the various microphone arrays, we will use the product combination rule, so the optimal class c_o and the optimal position s_j is chosen to maximize a product of posterior probabilities, i.e.

$$c_o, s_o = \underset{c_i,s_j}{\operatorname{argmax}} \prod_{k=1}^{K} p(c_i,s_j|X_k) \quad (7.9)$$

In the case of *N* simultaneous sources, and assuming they correspond to *N* different classes, the recognized identities of those classes and the corresponding positions are obtained by applying Eq. 7.9 *N* consecutive times and leaving each time the recognized class and its corresponding position out.

7.4 Experiments

7.4.1 Joint classification and DOA estimation results

The proposed event classification system at its front end consists of a set of frequency invariant beamformers that span all the angles in the room. The beamformers are designed to work with the horizontal row of 3 microphones each array in the smart-room has. With such a small number of microphones, it is expected that the beamformers have wide lobes and the sources are not well separated. On the other hand, it facilitates a computationally efficient working environment.

In the feature extraction block of the multi-array signal separation based system depicted in Figure 21, a set of audio spectro–temporal features is computed for each signal frame like we did in our all previous experiments. The HTK toolkit is used for training and testing the HMM-GMM system. For working with the null steering beamforming based system, each HMM is trained with the standard Baum-Welch algorithm using mono-event signals from a microphone and for a particular array. This approach actually introduces a mismatch between training and testing conditions, which is a source of classification errors.

Therefore, to compensate for that mismatch, in the decision block we have employed a machine learning based non-linear transformation technique that is unique for all classes. It is trained, in a supervised way, with the likelihoods obtained from the separated signals (the NSB outputs). We have used a multi-layer feed-forward neural network (NN) and a back-propagation training algorithm. The NN consists of three layers: input, hidden and output. We have optimized the number of hidden nodes in the NN through cross-validation. The tan-sigmoid transfer function is used at the output stages of the hidden and the output layers. A fast-scaled conjugate-gradient-based training algorithm is used [**182**]. At the output of the NN, we apply the MAP criterion according to Eq. 7.1 and Eq. 7.2. In the experiments, all the angles are assigned flat prior probabilities.

The testing results are obtained with all the 8 sessions (S01-S08) with a leave-one-out criterion, i.e. we recursively keep one session for testing, while all the other 7 sessions are used for training. Table 20 shows a performance comparison of the proposed system,

averaging over all the eight testing datasets, for six different arrays (T1-T6) and their combination. Here, we have used a machine learning based technique to transform the likelihoods generated from the HMM-GMM based classifier to the posteriors and then a MAP. In comparison with the classification results obtained when the positions of the acoustic sources are known (already presented in Chapter 5), it is observed that CA is degraded. This degradation might be due to the uniform combination of the scores from all the DOAs, also including the unknown desired directions, which are estimated later using the classification hypothesis. Moreover, the number of undesired directions are much larger than the desired directions (in this case, desired directions are only two) that possibly affect the classification performance of the system.

Table 20 : Classification performance of the system with the beamformer placing a null at the direction of interest

	Arrays						Product rule based combination of arrays
	T1	T2	T3	T4	T5	T6	
CA (%)	77.2	75.4	79.1	83.1	81.6	81.4	83.3

The hypothesized classes from the recognizer are used to localize the sources in terms of DOA estimation. It is performed at the decision block with the likelihood scores from HMM-GMM likelihood calculators. Using the mono-event signals instead of the separated signals for generating the HMM-GMM models, it is expected to get more variations in the likelihood scores along the angle and consequently, that choice should help to produce a better estimation. The optimal DOAs of the events for each array are obtained by using Eq. 7.3. Here also, we consider flat prior probabilities $p(\theta_j)$ for all angles.

To test the performance of the localization system, we will use the normalized root means squared error for direction of arrival (RMSE_DOA) given by the following equation:

$$\text{RMSE\_DOA} = \sqrt{\frac{1}{N_e}\sum_{i=1}^{N_e}\left|\frac{\theta_i^{test} - \theta_i^{ref}}{\Delta\theta}\right|^2} \tag{7.8}$$

where θ_i^{test} is the estimated DOA for an event i, θ_i^{ref} is its reference DOA, and N_e is the total number of event samples in the testing session. The reference DOA for each event class is taken from visual inspection during the recording of the signal. In our experiments, the null beam width $\Delta\theta$ is always kept constant (9 degrees).

The testing results for DOA estimation are obtained using all the 8 sessions (S01-S08) with a leave-one-out criterion. Table 21 shows the DOA estimation results obtained for the proposed metric Eq. 7.8, averaging over all the 8 testing datasets, for all different arrays (T1-T6). Since the DOA estimations are made using the individual arrays, and not combining them together, the errors are large.

Table 21 : DOA estimation performance of the system with the beamformer placing a null at the direction of interest

	Arrays					
	T1	T2	T3	T4	T5	T6
RMSE_DOA	3.2	3.7	2.9	2.1	2.9	2.6

The classification results for the system that uses beamformers, which place nulls at all the directions other than the direction of interest, are presented in Table 22. As mentioned earlier, two options regarding the classification decision are evaluated with this system: 1) Only MAP is used with the scores from HMM-GMM classifier (referred as MAP in the table), 2) MAP after transforming the likelihoods from HMM-GMM classifier to the posteriors by a machine learning technique (referred as ML-MAP in the table). For the experiments with the first option, we consider flat prior probabilities $p(\theta_j)$ for all angles and $p(X_k)$ for all the arrays. And the experiments with the second option, flat prior probabilities $p(\theta_j)$ are considered. Here, the classification accuracy (presented in the second row of the table) is slightly reduced from the previous case (reported in Table 20). This might be the due the difference in the type of beamforming used in these two cases. Note in this case, it is also possible to classify the events, only using the MAP and without likelihood to posterior transformation, with 1.8% reduction in CA. However, it was not possible in the previous case (reported result in Table 20), which requires this transformation (ML) before using the MAP.

Table 22 : Classification performance of the system with the beamformer placing nulls at all the directions other than the direction of interest

		Arrays						Product rule based combination of arrays
		T1	T2	T3	T4	T5	T6	
CA (%)	MAP	75.6	74.4	77.3	81.7	79.3	80.1	80.7
	ML-MAP	76.8	75	78.1	82.1	80.8	80.9	82.5

The corresponding testing results for DOA estimation are obtained and presented in Table 23 for the proposed metric Eq. 7.8, averaging over all the 8 testing datasets, for all different arrays (T1-T6). Since the DOAs are estimated using the classification hypothesis, the classification performance has an impact on DOA estimations. It is observed from the result in the Table 23, that for the better CA, the lesser DOA estimation errors are obtained. The same trend is observed while comparing this result with the previous case (presented in Table 21).

Table 23 : DOA estimation performance of the system with the beamformer placing nulls at all the directions other than the direction of interest

		Arrays					
		T1	T2	T3	T4	T5	T6
RMSE_DOA	MAP	3.9	3.8	3.1	2.8	3.2	2.9
	ML-MAP	3.4	3.5	3.1	2.3	3.1	2.7

7.4.2 Joint classification and 2D localization results

The testing results are obtained with all the 8 sessions (S01-S08) with a leave-one-out criterion, i.e. we recursively keep one session for testing, while all the other 7 sessions are used for training. The classification results for the system that uses beamformers, which place nulls at all the directions other than the direction of interest are presented in Table 24. Like in the later system described in the previous sub-section, two options regarding the classification decision are evaluated: 1) only MAP is used with the scores from HMM-GMM classifier (referred as MAP in the table), 2) MAP is employed after transforming the likelihoods from the HMM-GMM classifier to posteriors by a machine learning technique (referred as ML-MAP in the table). For the experiments with the first option, we consider flat prior probabilities $p(s_j)$ for all angles and $p(X_k)$ for all the arrays. And, for the experiments with the

second option, flat prior probabilities $p(s_j)$ are considered. From the presented result in Table 24, it is clear that the likelihood to posterior transformation before the MAP produces better results than using only MAP. Again, in comparison with the classification results obtained when the positions of the acoustic sources are known (presented in Chapter 5), it is observed that CA is degraded. This degradation might be due to the uniform combination of the scores from all the positions, also including the unknown desired positions, which are estimated later using the classification hypothesis. Moreover, the number of undesired positions is much larger than the number of desired ones (two, in the experimental scenario), what may possibly affect the classification accuracy of the system.

Table 24 : Classification results of the system with the beamformer attenuating signals from all the directions other than the direction of interest

		Product rule based combination of arrays
CA (%)	MAP	78.1
	ML-MAP	82.8

Table 25 shows the localization result obtained with two metrics for the proposed joint classification and localization system. The results with the two metrics, averaging over all AEs in the 8 testing datasets, are obtained using all the six arrays (T1 to T6) available in the room. The localization results are obtained by considering flat values for both $p(s_j|C_i)$ and $p(X_k)$.

Table 25 : 2D localization performance of classification-plus-localization system

	MAP	ML-MAP
Cell error (%)	23	19.1
RMSE	0.72	0.65

It is worth remembering that, like in Chapter 6, in the proposed method, an event based approach is followed, which means the localization is performed from a whole event.

Due to that event based approach, we have to assume that during the whole event the acoustic source is not moving along space. For that reason, we exclude ‘steps’ and ‘chair

moving' from the evaluation. In our approach, the source position is estimated using the identity of the event class from the classification hypothesis, but assuming the start and end points of the event are known.

Note that the errors are reduced in the case when likelihood to posterior transformation is applied before the MAP. It is also noticeable that these errors are increased around 5% in comparison with the error that is produced when the localization decision is made assuming the true label of the event class (Table 19 in Chapter 6). Indeed, the proposed method might take advantage of the knowledge about the a-priori probabilities of the pre-defined positions for each event class.

The classification results for the joint parallel system are presented in Table 26. Only MAP is used with the scores from the HMM-GMM classifier according to Eq. 7.9. For that, we consider flat prior probabilities $p(X_k)$ for all the arrays.

Table 26: CA of the joint and parallel classification and localization system

		Product rule based combination of arrays
CA (%)	MAP	84.4

In comparison with the results obtained with the classification-plus-localization system (first row of Table 24), it is observed that CA is improved in this joint parallel approach. However, it could not achieve the CA obtained when the positions of the acoustic sources are previously known (presented in Chapter 5).

Table 27 shows the localization result obtained with two metrics for the proposed joint parallel classification and localization system. The results with the two metrics, averaging over all AEs in the 8 testing datasets, are obtained again using all the six arrays (T1 to T6) available in the room. The localization results are obtained by considering flat values for $p(X_k)$. As the event based approach is followed, again we exclude 'steps' and 'chair moving' from the evaluation, since they are moving sources.

Table 27: 2D localization performance of the joint parallel classification and localization system

	MAP
Cell error (%)	15.8
RMSE	0.58

Note that, the errors are reduced around absolute 3.3% in comparison with the classification-plus-localization system (first column of Table 25). Indeed, the proposed method might take advantage of the knowledge about the a-priori probabilities of the pre-defined positions for each event class.

7.5 Chapter conclusions

In this Chapter, a combined approach for classification and localization of simultaneously occurring meeting room acoustic events is presented. For classification, a computationally efficient beamforming based source separation technique followed by a HMM-GMM based likelihood computation has been presented, either where the estimation is done with a MAP criterion alone or a MAP criterion after applying a data-dependent non-linear transformation. Contrarily to what it was assumed in Chapter 5, here the system does not require any information about the event source positions, since, by using the hypothesized outputs of the recognizer, the system is also able to localize the acoustic events in terms of either DOA or 2D position estimation, so avoiding the need of an external localization system. It is observed that the classification accuracy suffers degradation in the classification-plus-localization approach with respect to the case where the source positions are known. It is also observed that in the classification-plus-localization method, the degradation of classification accuracy increases the localization errors. However, with a joint parallel approach, it is observed that both the classification and localization performances are improved significantly in comparison with the classification-plus-localization approach. Alternatively, the proposed method might take the advantage of the knowledge about the a-priori probabilities of the pre-defined positions for each event class for a better localization performance.

Chapter conclusions

Chapter 8. Conclusion and future work

8.1 Summary of conclusions

This thesis presents the work done by the author in the area of acoustic event detection and localization focusing on the problem of signal overlapping using distributed microphone arrays. The work done could be broadly categorized into two: 1) Event detection that consists of two stages. First, a set of null steering beamformers is used to carry out partial signal separations, by using multiple arbitrarily located linear microphone arrays that are composed of a small number of microphones. Second, acoustic event classification and detection are carried out from the beamformer outputs using a maximum-a-posteriori criterion. 2) A novel sound-model-based event localization technique that could be used either separately or jointly with the event detection system. The HMM-GMM classifier is chosen as the basic technique for both detection and localization.

There are several contributions of this thesis work. To solve the problem of source overlapping, source separation is carried out prior to the detection (that includes identification) of each of the overlapped sounds. This solution is reasonable because it could also be applicable to the case where either the number of events or the number of simultaneous sources is large. This is not feasible with the previous baseline model based approach that suffers from the scalability problem. Beamforming based source separation technique that is computationally less demanding than usual BSS based techniques is desirable due to its possibility of being implemented in real-time system. A system of such kind has been proposed, which consists of a null steering beamforming based partial signal separation technique, followed by a likelihood ratio based classifier and a decision block. Moreover, a similar system structure can be used for different applications, like detection, localization, resolving the permutation problem in a multisource scenario, etc.

In this thesis, a new approach for computationally efficient classification/detection and the positioning of acoustic events that result from the combination of a beamforming based partial signal separation and a MAP based decision has been presented. Assuming both the source positions and the end-points of the events are known, the proposed system has been tested for classification and position assignment in a scenario with two sources. In addition,

assuming only the source positions, a system is proposed that performs the detection task and outputs both the class identities and the endpoints of the events. The core system consists firstly of a set of frequency invariant null steering beamformers to carry out different partial signal separations for each microphone array. The beamformers are followed by model based likelihood calculations. The scores from the arrays are then combined in the decision block, and finally a MAP criterion is used to get the optimal class, and also the optimal position.

The proposed system is compared with a model based system and a BSS system in the case of two sources, one of which is speech. Unlike the model based system, it does not have the problem of limited scalability. On the other hand, unlike the BSS technique, which is computationally demanding, the proposed technique is computationally efficient. So, it is more suitable to real-time applications. Besides having these advantages, the proposed FIB based system that uses several small, distributed microphone arrays, performs significantly better than the other two methods mentioned above. As the used beamforming technique only partially separates the different sources, we have tried to use as much information as possible from the given scenario by using the models generated from the signals processed by all the beamformers. We found a noticeable improvement in the performance of the system both in terms of classification, detection, and position assignment while combining the microphone arrays using a product rule based combination in the decision block before the MAP criterion is applied.

Additionally, both the model based and the BSS based approaches suffer from the permutation problem. In fact, they require a separate position assignment system to solve it. However, the proposed technique includes the posterior assignment of event hypothesis to source positions that the other two mentioned techniques do not have. In this thesis, an attempt is presented to resolve the source identification ambiguity that appears when an acoustic event overlapped with speech is detected. A position assignment system has been proposed and tested in the smart room environment. The already mentioned set of beamformers is followed by model based likelihood calculations, using both the acoustic event model and the speech model, to obtain a couple of likelihood ratio per beamformer, which are multiplied to get a final score. Finally, either a product rule or FI based fusion is used to integrate the scores from all the arrays. Two types of beamforming techniques, a frequency dependent one, and a

frequency invariant one are compared. While, a careful frequency tuning is required in the former case, the alternative frequency invariant technique does not require frequency tuning and thus it is less dependent on the concrete scenario. Important enough, the fusion of the scores from the arrays, yields the best assignment error that is smaller than 5%.

Another significant contribution of this thesis is source localization. A novel approach for acoustic source localization based on the models of sounds has been presented which combines a set of beamformers and a MAP based decision. It has been tested in a limited scenario with one and two sources. The 1-source localization performance of the proposed system is slightly better than the standard SRP-PHAT based system, and it performs significantly better in the more complex 2-source scenario. Therefore, the presented sound-model-based localization technique may be an alternative for event-level localization in a multiple source scenario. Moreover, the proposed technique would obviously find a place at least when it has to work together with a recognition system, since both use the same framework. Interestingly enough, in the case of two simultaneous events, one of which is speech, the acoustic events are localized with a not much worse accuracy than when they occur alone.

In the thesis work, a combined approach for classification and localization of simultaneously occurring meeting room acoustic events is also presented. For classification, a computationally efficient beamforming based source separation technique followed by a HMM-GMM based likelihood computation has been presented, either where the estimation is done with a MAP criterion alone or a MAP criterion after applying a data-dependent non-linear transformation. Contrarily to what it was assumed in the previously described classification work, here the system does not require any information about the event source positions. By using the hypothesized outputs of the recognizer, the system is also able to localize the acoustic events in terms of either DOA or 2D position estimation, so avoiding the need of an external localization system. It is observed that the classification and localization accuracies suffer degradation in this classification based localization approach. On the other hand, working with a joint and parallel approach, it is observed that both the classification and localization performances are improved significantly in comparison with the classification-plus-localization approach. Additionally, it is also observed that the proposed SB-SMB based

localization technique can work independently and without using the class identities from the classification system. Moreover, the proposed method might take advantage of the knowledge about the a-priori probabilities of the pre-defined positions for each event class for a better localization performance. Although classification and localization may work separately for the acoustic scene analysis task, the integrated approach can offer advantages in terms of resource and design.

8.2 Future work

The following list contains the most important points requiring improvements as well as a few directions for future work.

8.2.1 Joint detection and localization

In Chapter 5 of this thesis, both acoustic event classification and detection systems are presented. In addition, in Chapter 7, a joint classification and localization method is proposed and tested. In the future work, a joint detection and localization system can be tested.

8.2.2 Event level to frame level localization, source tracking

In this thesis, a sound-model-based event localization system that works at the event level (not at the frame level), since the localization is performed for the whole event, is proposed. Performing localization with the proposed method, the system has to assume that for the duration of the whole event, the sources do not change their spatial position, which means that they are static sources. Because of that assumption, we had to keep some events out from the evaluation: ‘steps’ and ‘chair moving’. It could be interesting to see the performance of the system when the same proposed technique is used at the frame level instead of localizing for the whole event. If the system is designed to work at the frame level, it can be possible to evaluate it with those moving events. Moreover, source tracking could be thought of when the system works at the frame level.

8.2.3 Inclusion of vertical axis for 3D localization and better separation

In all the reported experiments of this thesis work, three linearly placed microphones for each of the T-shaped arrays are used. It means that the work has been done in the horizontal (i.e. XY plane) plane, so the vertical axis has never been considered. The fourth microphone of each array is aligned vertically with the other three linear, horizontally-placed microphones. Inclusion of the fourth microphone will enable the localization system to work in the vertical axis as well. Therefore, it will be possible to localize the event in the 3D space. In addition, it will be interesting to see that how the inclusion of this fourth microphone affects the amount of signal separation with these beamformers, and thus influences the recognition accuracy.

8.2.4 Overlapping of more than 2 sources

In this thesis work, the systems are proposed to work in a multisource environment. They are tested in a scenario where a maximum of up to two sources are active simultaneously. Future work can be carried out to test the proposed system in a situation that has more than two simultaneous acoustic sources.

8.2.5 Working with other databases, cross site event detection and localization

For the experiments in this thesis, two types of signals are used: 1) artificially overlapped signals by superposition of different signals recorded separately in a real room environment, 2) signals overlapped in a natural way through the interaction of two persons in the room environment (both AEs and speech are produced and overlapped in a natural way). The recognition accuracy is decreased while working with the second type of signals than the first one. One possibility for further research could be to work with different types of overlapped signals and under different noise conditions. However, creating the database with signal overlaps produced in a natural way is difficult problem.

A possible direction of further research could be the cross-site event detection, i.e. the case when acoustic models are created using the database from one site, and testing is performed using the database from another site. This is a natural requirement for many practical applications to work equally well in different conditions and environments. In fact, within the CHIL European project different databases with AEs were recorded from UPC, UKA, IBM, AIT and ITC sites [**37**], which can be used.

Own Publications

- R. Chakraborty, C. Nadeu, and T. Butko, “Detection and positioning of overlapped sounds in a room environment”, Proc. *Interspeech*, Portland, USA, 2012.
- R. Chakraborty, C. Nadeu, and T. Butko, “Binary position assignment of two known simultaneous acoustic sources”, Proc. *IberSPEECH*, Madrid, Spain, 2012.
- R. Chakraborty and C. Nadeu, “Real-time multi-microphone recognition of simultaneous sounds in a room environment”, *Proc. ICASSP*, Vancouver, Canada, 2013.
- R. Chakraborty, and C. Nadeu, “Joint recognition and direction-of-arrival estimation of simultaneous meeting-room acoustic events”, Proc. *Interspeech*, Lyon, France, 2013.
- R. Chakraborty, C. Nadeu, and T. Butko, “Source ambiguity resolution of overlapped sounds in a multi-microphone room environment”, submitted, Eurasip Journal on Audio, Speech and Music Processing.

Bibliography

[1] A. Temko et al., "Acoustic event detection and classification," in *Computers in the Human Interaction Loop*, A. Waibel and R. Stiefelhagen, Eds.: Springer, 2009, pp. 61-73.

[2] A. Temko, "Acoustic event detection and classification," UPC, Barcelona, PhD Thesis 2007.

[3] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters, Elsevier*, vol. 30, no. 14, pp. 1281-1288, 2009.

[4] T. Butko, "Feature selection for multimodal acoustic event detection," UPC, Barcelona, PhD Thesis 2011.

[5] T. Butko et al., "Acoustic event detection based on feature-level fusion of audio and video modlities," *EURASIP Journal on Advances on Signal Processing*, vol. 2011, 2011.

[6] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297-336, October 1994.

[7] D. P. W. Ellis, "Prediction–driven computational auditory scene analysis," MIT, Cambridge, USA, PhD Thesis 1996.

[8] J. Pinquier, "Robust speech/ music classification in audio document," in *ICSLP*, Denver, USA, 2002, pp. 2005-2008.

[9] L. Lu, H. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transaction on Speech and Audio Processing*, vol. 10, no. 7, pp. 504-516, 2002.

[10] K. -H. Lin et al., "Improving faster-than-real-time human acoustic event detection by saliency-maximized audio visualization ," in *ICASSP*, kyoto, Japan, 2012, pp. 2277-2280.

[11] D. Castan and M. Akbacak, "Indexing multimedia documents with acoustic concept recognition lattices," in *Interspeech*, Lyon, France, 2013.

[12] T. Nishiura, S. Nakamura, K. Miki, and K. Shikano, "Environmental sound source identification based on hidden markov models for robust speech recognition," in *Eurospeech*, 2003, pp. 2157-2160.

[13] T. Nishiura and S. Nakamura, "Study of environmental sound source identification based on hidden Markov model for robust speech recognition," *Journal of the Acoustical Society of America*, vol. 114, no. 4, p. 2399, 2003.

[14] R. Cai, L. Lu, H. Hanjalic, and L.-H. Cai, "A flexible framework for key audio effects detection and auditory context inference," *IEEE Transaction Audio, Speech, Language Processing*, vol. 14, no. 3, pp. 1026-1039, 2006.

[15] S. Moncrieff, C. Dorai, and S. Venkatesh, "Detecting indexical signs in film audio for scene interpretation," in *IEEE ICME*, 2001, pp. 989-992.

[16] S. Liu, M. Xu, H. Yi, L.-T. Chia, and D. Rajan, "Multimodal semantic analysis and annotation for basketball video," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1-13, 2006.

[17] Y. Rui, A. Gupta, and A. Acero, "Automatically Extracting Highlights for TV Baseball Programs," in *ACM Multimedia*, 2000, pp. 105-115.

[18] D. S. Pallet, "A look at NIST's benchmark ASR tests: past, present, and future," National Institute of Standards and Technology (NIST), Gaithersburg, USA, Technical Report 2003.

[19] J. Saunders, "Real-time discrimination of broadcast speech/music," in *ICASSP*, Atlanta, USA, 1996.

[20] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature," in *ICASSP*, Munich, Germany, 1997.

[21] W. Chou and L. Gu, "Robust singing detection in speech/music discriminator design," in *ICASSP*, Salt Lake City, USA, 2001, pp. 1331-1334.

[22] T. Izumitani, R. Mukai, and K. Kashino, "A background music detection method based on robust feature extraction," in *ICASSP*, Las Vegas, USA, 2008, pp. 13-16.

[23] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam, "Classification of audio signals using SVM and

RBFNN," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6069-6075, April 2009.

[24] S. Lefèvre and N. Vincent, "A two level strategy for audio segmentation," *Digital Signal Processing*, vol. 21, no. 2, pp. 270-277, March 2011.

[25] T. Butko and C. Nadeu, "Audio segmentation of broadcast news: A hierarchical system with feature selection for the Albayzin-2010 evaluation," in *ICASSP*, Prague, Czech Republic, 2011, pp. 357-360.

[26] D. Castan, A. Ortega, A. Miguel, and E. Lleida, "Broadcast News Segmentation with Factor Analysis System," in *First Workshop on Speech, Language and Audio in Multimedia (SLAM)*, Marseille, France, 2013.

[27] C. N. Doukas and I. Maglogiannis, "Emergency fall incidents detection in assisted living environments utilizing motion, sound, and visual perceptual components," *IEEE Transaction of Information Technology in Biomedicine*, vol. 15, no. 2, 2011.

[28] L. Vuegen, B. Van Den Broek, P. Karsmakers, H. Van hamme, and B. Vansumste, "Automatic monitoring of activities of daily living based on real-life acoustic sensor data: a preliminary study," in *Workshop on speech and language processing for assistive technologies - SLPAT-2013 edition:4*, Grenoble, France, 2013, pp. 1-6.

[29] M. Vacher, D. Istrate, L. Besacier, E. Castelli, and J. Serignat, "Smart audio sensor for telimedicina," in *Smart Object Conference*, 2006.

[30] M. Stäger et al., "Sound button:design of a low power wearable audio classification system," in *IEEE Int. Symp. on Wearable Computers*, 2003, pp. 12-17.

[31] C. Jianfeng, Z. Jianmin, A. Kam, and L. Shue, "An automatic acoustic bathroom monitoring system," in *IEEE International Symposium on Circuits and Systems*, 2005.

[32] J. Chen, A. H. Kam, J. Zhang, N. Liu, and L. Shue, "Bathroom Activity Monitoring Based on Sound," in *Pervasive Computing*, H-W Gellersen, R. Want, and A. Schmidt, Eds.: Springer Berlin Heidelberg, 2005, pp. 47-61.

[33] K.-Y. Huang, C. -C Hsia, M. -S. Tsai, Y.-H. Chiu, and G.-L. Yan, "Activity Recognition by Detecting Acoustic Events for Eldercare," in *6th World Congress of Biomechanics (WCB 2010)*, C. T. Lim and J. C. H. Goh, Eds.: Springer Berlin Heidelberg, 2010, pp. 1522-1525.

[34] J. Schroeder, S. Wabnik, P. W. J. van Hengel, and S. Goetze, "Detection and Classification of Acoustic Events for In-Home Care," in *Ambient Assisted Living*, R. Wichert and B. Ederhardt, Eds.: Springer Berlin Heidelberg, 2011, pp. 181-195.

[35] C. Clavel, T. Ehrette, and G. Richard, "Events Detection for an Audio-Based Surveillance System," in *IEEE International Conference on Multimedia and Expo.*, Amsterdam, Netherlands, 2005, pp. 1306-1309.

[36] J. Kotus, K. Lopatka, and A. Czyzewski, "Detection and localization of selected acoustic events in acoustic field for smart surveillance applications," *Multimedia Tools and Applications*, July 2012.

[37] A. Waibel and R. Stiefelhagen, *Computers in Human Interaction Loop*. New York, USA: Springer, 2009.

[38] CLEAR, 2006. Classification of Events, Activities and Relationships. Evaluation and Workshop. [Online]. http://isl.ira.uka.de/clear06

[39] CLEAR, 2007. Classifications of Events, Activities and Relationships. Evaluation and Workshop. [Online]. http://www.clear-evaluation.org

[40] D. Wang and G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*.: Wiley-IEEE Press, 2006.

[41] NIST. (2009), The NIST Rich Transcription evaluation project website. [Online]. http://www.itl.nist.gov/iad/mig/tests/rt/

[42] R. Vipeperla et al., "Speech overlap detection and attribution using convolutive non-negative sparse coding," in *ICASSP*, Kyoto, Japan, 2012, pp. 4181-4184.

[43] M. Wöllmer, F. Weninger, J. Geiger, B. Schuller, and G. Rigoll, "Noise robust ASR in reverberated multisource environments applying convolutive NMF and Long Short-Term Memory," *Computer Speech & Language*, vol. 27, no. 3, pp. 780-797, May 2013.

[44] S. Wrigley, G. Brown, V. Van, and S. Renals, "Speech and crosstalk detection in multi-channel audio," *IEEE Transactions on Speech Audio Processing*, vol. 13, pp. 84-91, 2005.

[45] M. Zelenak, C. Segura, J. Luque, and J. Hernando, "Simultaneous speech detection with spatial features for speaker diarization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 436-446, 2012.

[46] M. Zelenak, C. Segura, J. Luque, and J. Hernando, "Simultaneous Speech Detection With Spatial Features for Speaker Diarization," *IEEE Transactions on Audio, Speech and Language processing*, vol. 20, no. 2, pp. 436-446, 2012.

[47] J. Geiger et al., "Convolutive Non-Negative Sparse Coding and New Features for Speech Overlap Handling in Speaker Diarization," in *Interspeech*, Portland, USA, 2012.

[48] T. Butko, F. Gonzalez Pla, C. Segura, C. Nadeu, and H. Hernando, "Two-source acoustic event detection and localization: online implementation in a smart.room," in *EUSIPCO*, Barcelona, 2011.

[49] D. Templeton and D. Saunders, *Acoustic Design*. London, UK: Architectural Press, 1987.

[50] R. L. Bouquin and G. Faucon, "Using the coherence function for noise reduction," *IEE Proceedings*, vol. 139, pp. 276-280, June 1992.

[51] J. Bitzer, K. Kammeyer, and K.U. Simmer, "An alternative implementation of superdirective beamforemr," in *IEEE Workshop on Applications of Signal Processing on Audio, Speech and Acoustics*, New York, USA, 1999, pp. 991-994.

[52] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed.: CRC Press, 2007.

[53] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*.: Prentice Hall PTR, 2001.

[54] T. Virtanen, R. Singh, and V. Raj, *Techniques for Noise Robua¡stness in Automatic Speech Recognition*.: John Wiley & Sons, 2012.

[55] S.V. Vaseghi, "Spectral Subtraction," in *Advanced Digital Signal Processing and Noise Reduction*.: John Wiley & Sons, 2000, ch. 11, pp. 333-354.

[56] B. D. Van Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4-24, April 1988.

[57] S. Simanapalli and M. Kaveh, "Broadband focusing for partially adaptive beamforming," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 30, no. 1, pp. 68-80, January 1994.

[58] D. B. Ward, R. A. Kennedy, and R. C. Williamson, "Theory and design of broadband sensor arrays with frequency invariant far-field beam patterns," *Acoustical Society of America*, vol. 97, no. 2, pp. 1023-1034, 1995.

[59] T. D. Abhayapala, R. A. Kennedy, and R. C. Williamson, "Nearfield broadband array design using a radially invariant modal expansion," *Acoustical Society of America*, vol. 107, no. 1, pp. 392-403, 2000.

[60] D. B. Ward, "Theory and Application of Broadband Frequency Invariant Beamforming," Australian National University, PhD Thesis 1996.

[61] Y. R. Zheng, R. A. Goubran, and M. El-Tanany, "Experimental evaluation of a nested microphone array with adaptive noise cancellers," *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 3, pp. 777-786, June 2004.

[62] J. Sanchez-Bote, J. Gonzalez-Rodriguez, and J. Ortega-Gracia, "A real-time auditory based microphone array assessed with E-RASTI evaluation proposal," in *ICASSP*, Hong Kong, China, 2003.

[63] A. Abad, "A multi-microphone approach to speech processing in a smart-room environment," UPC, Barcelona, PhD Thesis 2007.

[64] H. Teutsch and G. W. Elko, "First and second order adaptive differential microphone arrays," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Darrmstadt, Germany, 2001.

[65] M. Wolf and C. Nadeu, "On the potential of channel selection for recognition of reverberated speech with multiple microphones," in *Interspeech*, Makuhari, Japan, 2010.

[66] M. Wolf and C. Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Communication*, vol. 57, pp. 170-180, February 2014.

[67] M. Wolf and C. Nadeu, "Channel Selection Using N-Best Hypothesis for Multi-Microphone ASR," in *Interspeech*, Lyon, France, 2013.

[68] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*.: Springer, 2001.

[69] L. C. Parra, "Steerable Frequency-Invarant Beamforming for Arbitrary Arrays," *Acoustical Society of America*, vol. 119, no. 6, pp. 3839-3847, June 2006.

[70] Y. Zhao, W. Liu, and R. J. Langley, "Design of frequency invariant beamformers in subbands," in *IEEE/SP 15th Woekshop on Statistical Signal Processing*, Cardiff, UK, 2009.

[71] L. C. Godara, "Applications of Antenna Arrays to Mobile Communications, part II: Beamforming and Direction-of-Arrival Considerations," *Proceedings of the IEEE*, vol. 85, no. 8, pp. 1195-1245, August 1997.

[72] S. P. Applebaum and D. J. Chapman, "Adaptive Arrays with Main Beam Constraints," *IEEE Transactions on Antennas and Propagation*, vol. 24, no. 5, pp. 650-662, September 1976.

[73] J. Griffiths and C. W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beamforming," *IEEE Transactions on Antenna and Propagation*, vol. 30, no. 1, pp. 27-34, January 1982.

[74] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A Robust Adaptive Beamformer for Microphone Arrays with a Blocking Matrix Using Constrained Adaptive Filters ," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2677-2684, October 1999.

[75] O. Hoshuyama and A. Sugiyama, "Robust Adaptive Beamforming," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. New York: Springer, 2001, ch. 5, pp. 87-109.

[76] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, "Experiments of hands-free connected digit recognition using microphone array," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, CA, 1997, pp. 490-497.

[77] T. B. Hughes, H. Kim, J. H. DiBiase, and H. F. Silverman, "Performance of HMM speech recognizer using a real-time tracking microphone array as input," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 346-349, May 1999.

[78] S. Nordholm, I. Claesson, and M. Dahl, "Adaptive microphone array employing calibration signals," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 241-252, 1999.

[79] M. L. Seltzer, "Microphone array processing for robust speech recognition," CMU, Pittsburgh PA, PhD Thesis 2003.

[80] M. L. Seltzer and B. Raj, "Calibration of microphone arrays for improved speech recognition," in *Eurospeech*, Aalborg, Denmark, 2001, pp. 1005-1008.

[81] M. L. Seltzer and R. Stern, "Subband Likelihood-Maximizing Beamforming for Speech Recognition in Reverberant Environments," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2109-2121, November 2006.

[82] A. Ortega, E. Lleida, and E. Masgrau, "Speech Reinforcement System for Car Cabin Communications," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 917-929, 2005.

[83] J. H. L. Hansen and X. Zhang, "Analysis of CFA-BF: Novel combined fixed/adaptive beamforming for robust speech recognition in real car environment," *Speech Communication*, vol. 52, no. 2, pp. 134-149, February 2010.

[84] P. Comon, "Independent Component Analysis: A New Concept?," *Signal Processing*, vol. 36, no. 3, pp. 287-314, 1994.

[85] J. F. Cardoso, "Blind Signal Separation: Statistical Principles," *IEEE Proc.*, vol. 86, no. 10, pp. 2009-2025, October 1998.

[86] K. B. Christensen, "The application of digital signal processing to large scale simulation of room acoustics:

frequency response modeling and optimization software for a multichannel dsp engine," *Journal on Audio Engineering Society*, vol. 40, no. 4, pp. 260-276, 1992.

[87] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed.: Wiley-Interscience, 2000.

[88] M. Berg, E. Bondesson, S. Y. Low, S. Nordholm, and I. Claesson, "A combined on-line PCA-ICA algorithm for blind source separaton," in *Asia Pacific Conference on Communications*, Perth, Australia, 2005.

[89] C. Jutten and J. Herault, "Blind separation of sources part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1-10, 1991.

[90] D. T. Pham, P. Garrat, and C. Jutten, "Separation of mixture of independent sources through a maximum likelihood approach," in *EUSIPCO*, Brussels, Belgium, 1992, pp. 771-774.

[91] A. Bell and T. Sejnowski, "An information maximization approaach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129-1159, November 1995.

[92] S. Amari, A. Cichocki, and A. A. Yang, "A new learning algorithm for blind source separation," in *NIPS*, Denver, USA, 1996, pp. 752-763.

[93] H. Buchner, R. Aichner, and W. Kellermann, "A generalized of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 120-134, January 2005.

[94] C. Simon, P. Loubaton, and C. Jutten, "Separation of a class of convolutive mixtures: a contrast function approach," *Signal Processing*, vol. 81, no. 4, pp. 883-887, April 2001.

[95] M. Castella, S. Rhioui, E. Moreau, and J- C Pasquet, "Quadratic Higher-Order Criteria for Iterative Blind Separation of a MIMO Convolutive Mixture of Sources," *IEEE Transactions on Signal Processing*, vol. 55, no. 1, pp. 218-232, January 2007.

[96] M. Castella and E. Moreau, "A new optimization method for reference-based quadratic contrast functions in a deflation scenario," in *ICASSP*, Taiwan, R.O.C, 2009, pp. 3161-3164.

[97] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *ICASSP*, Montreal, Cananda, 2004, pp. 889-892.

[98] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "Real-time convolutive blind source separation based on a broadband approach," in *International Symposium ICA*, Granada, Spain, 2004, pp. 833-840.

[99] S. Ding, J. Huang, and D. Wei, "Real-time blind source separation of acoustic signals with a recursive approach," *International Journal of Computational Intelligence and Applications*, vol. 4, no. 2, pp. 193-206, June 2004.

[100] B. Loesch and B. Yang, "Online blind source separation based on time-frequency sparseness," in *ICASSP*, Taiwan, R.O.C., 2009, pp. 117-120.

[101] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 130-133, 2011.

[102] J. Dennis, H. D. Tran, and E. Chng, "Image feature representation of the subband power distribution for robust sound event classification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 2, pp. 367-377, 2013.

[103] J. Dennis, H. D. Tran, and E. S. Chng, "Overlapping sound event recognition using local spectrogram features with the generalized Hough Transform," in *Interspeech*, Portlan, Oregon, USA, 2012.

[104] G. Bao, Zhogfu Ye, X. Xu, and Y. Zhou, "A Compressed Sensing Approach to Blind Separation of Speech Mixture Based on a Two-Layer Sparsity Model," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 899-906, May 2013.

[105] Q. Li-yan, C. Yin, H. Xu, and H. Li, "Blind separation of speech sources in multichannel compressed sensing," in *IEEE International Instrumentaion and Measurement Technology Conference*, Minneapolis, USA, 2013, pp. 1771-1774.

[106] B. Wang and M. D. Plumbley, "Investigating single-channel audio source separation methods based on non-negative matrix factorization," in *ICA Research Network International Workshop*, 2006, pp. 17-20.

[107] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, 2011.

[108] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound Event detection in multisource environment using source separation," in *CHIME Workshop, satellite event of Interspeech* , Florence, Italy, 2011.

[109] R. Badeau and M. D. Plumbley, "Multichannel HR-NMF for modelling convolutive mixtures of non-stationary signals in the time-frequncy domain," in *Accepted for IEEE Workshop on Applications pf Signal processing to Audio and Acoustics* , New York, USA, 2013.

[110] T. Barker and T. Virtanen, "Non-negative tensor factorization of modulation spectrograms for monaural sound source separation," in *Interspeech*, Lyon, France, 2013.

[111] E. Wold, T. Blum, T. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio ," *IEEE Multimedia*, vol. 3, no. 3, pp. 27-36, 1996.

[112] B. Lukic, "Activity Detection in Public Spaces," Kungliga Tekniska Högskolan (KTH), Stockholm, Master Degree Project 2004.

[113] J. Pinquier and R. André-Obrecht, "Jingle detection and identification in audio documents," in *ICASSP*, Montreal, Canada, 2004.

[114] G. Iyenger, H. Nock, C. Neti, and M. Franz, "Semantic indexing of multimedia using audio, text and visual cues," in *IEEE ICME*, Lausanne, Switzerland, 2002, pp. 369-372.

[115] G. Xu, Y. -F. Ma, H. -J. Zhang, and S. Yang, "Motion based event recognition using HMM," in *ICPR*, Quebec, Canada, 2002, pp. 831-834.

[116] M. Zobl, F. Wallhoff, and G. Rigoll, "Action regognition in meeting scenarios using global motion features," in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2003, pp. 32-36.

[117] R. Wang, Z. Liu, and J. -C. Huang, "Multimedia Content Analysis," *IEEE Signal Processing Magazine*, vol. 17, no. 6, pp. 12-36, 2000.

[118] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. K. Wong, "Integration of multimodal features for video scene classification based on HMM," in *IEEE Workshop on Multimedia Signal Processing*, Copenhagen, Denmark , 1999, pp. 53-58.

[119] M. T. Chan, Y. Zhang, and T. S. Huang, "Real-time lip tracking and bi-modal continuous speech recognition," in *IEEE Workshop on Multimedia Signal Processing*, 1998, pp. 65-70.

[120] S. Petridis and M. Pantic, "Audiovisual discrimination between laughter and speech," in *ICASSP*, Las Vegas, USA, 2008, pp. 5117-5120.

[121] Z. Li and Y.-P. Tan, "Event detection using multimodal feature analysis," in *ISCAS*, Kobe, Japan, 2005, pp. 3845-3848.

[122] M. Cowling and R. Sitte, "Analysis of speech recognition techniques for use in a non-speech sound recognition system," in *6th International Symposium on Digital Signal Proceesing for Communication Systems*, 2002, pp. 16-20.

[123] D. Hoiem, Y. Ke, and R. Sukthankar, "SOLAR: Sound object localization and retrieval in complex audio environments," in *ICASSP*, Philadelphia, USA, 2005, pp. 429-432.

[124] M. Xu, L. -Y. Duan, C. -S. Xu, and Q. Tian, "A fusion scheme of visual and auditory modalities for event detection in sports video," in *ICASSP*, Hong Kong, China, 2003, pp. 189-192.

[125] Z. Zhang and B. Schuller, "Semi-supervised learning helps in sound event classification," in *ICASSP*, Kyoto, Japan, 2012.

[126] A. Kumar, P. Dighe, R. Singh, S. Choudhuri, and B. Raj, "Audio event detection from acoustic unit occurrence patterns," in *ICASSP*, Kyoto, Japan, 2012.

[127] B. Raj, M. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 275-296, September 2004.

[128] S. Keronen et al., "Mask estimation and inputation methods for missing data speech recognition in a

multisource reverberant environment," *Computer Speech & Language*, vol. 27, no. 3, pp. 798-819, May 2013.

[129] Y. R. Leng and H. D. Tran, "Using blob detection in missing feature linear-frequency cepstral coefficients for robust sound event recognition," in *Interspeech*, Portland, Oregon, USA, 2012.

[130] J. Benesty, J. Chen, Y. A. Huang, and J. Dmochowski, "On Microphone-Array Beamforming from a MIMO Acoustic Signal Processing Perspective," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1053-1065, March 2007.

[131] H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, and W. Kellermann, "Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering," in *ICASSP*, Philadelphia, USA, 2005.

[132] M. Matsumoto and S. Hashimoto, "Multiple signal classification by aggregated microphones," *IEICE Transactions on Fundamentals of Electronics, Communication and Computer Sciences*, vol. 88, no. 7, pp. 1701-1707, July 2005.

[133] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, p. 384, 2000.

[134] P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *ICASSP* , Munich, Germany, 1997.

[135] B. Champagne, S. Bedard, A. Stephenne, I. Telecommun, and Q. Verdun, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Transactions on Acoustics, Speech and Audio processing*, vol. 4, no. 2, pp. 148-152, 1996.

[136] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 24, no. 4, pp. 320-327, 1976.

[137] M. Omologo and P. Svaizer, "Acoustic event localization using crosspower-spectrum phase based technique," in *ICASSP* , Adelaide, Australia, 1994.

[138] M. Omologo and P. Svaizer, "Use of crosspower-spectrum phase in acoustic event detection," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 5, no. 3, pp. 288-292, 1997.

[139] E. E. Jan and J. Flanagan, "Sound source localization in reverberant environments using an outlier elimination algorithm," in *IEEE ICSLP*, 1996, pp. 1321-1324.

[140] J. C. Chen, K. Yao, T. L. Tung, C. W. Reed, and D. Chen, "Source localization and tracking of a wideband source using a randomly distributed beamforming sensor array," *International Journal of High Performance Computing Applications*, vol. 16, no. 3, p. 259, 2002.

[141] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943-956, 2002.

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

[143] E. E. Jan, "Parallel processing of large scale microphone arrays for sound capture," Rutgers University, NJ, USA, PhD Thesis 1995.

[144] M. S. Brandstein, "A Framework for Speech Source Localization Using Sensor Arrays," Brown University, PhD Thesis 1995.

[145] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *ICASSP*, Munich, Germany, 1997, pp. 187-190.

[146] R. O. Schmidt, "A new approach to geometry of range difference location," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-8, no. 6, pp. 821-835, 1972.

[147] H. Schau and A. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 8, pp. 1223-1225, August 1987.

[148] Y. Yu and H. F. Silverman, "An improved TDOA-based location estimation algorithm for large aperture

microphone arrays," in *ICASSP*, Montreal, Canada, 2004.

[149] N. Strobel, T. Meier, and R. Rabenstein, "Speaker localization using a steered filter-and-sum beamformer," in *Erlangen Workshop on Vision, Modeling and Visualization*, 1999.

[150] V. M. Alvarado, "Talker Localization and Optimal Placement of Microphones for a Linear Microphone Array Using Stochastic Region Contraction," Brown University, PhD Thesis 1990.

[151] H. F. Silverman and S. E. Kirtman, "A two-stage algorithm for determining talker location from linear microphone array data*," vol. 6, no. 2, pp. Computer, Speech & Language, April 1992.

[152] J. DiBiase, H. F. Silverman, and M. Brandstein, *Microphone Arrays. Robust Localization in Reverberant Rooms*.: Springer, 2001.

[153] J. Dmochowski and J. Benesty, "Steered Beamforming Approaches for Acoustic Source Localization," in *Speech Processing in Modern Communication*, I. Cohen, J. Benesty, and S. Gannot, Eds. Berlin Heidelberg: Springer-Verlag, 2010, ch. 12, pp. 307-337.

[154] J. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Brown University, Providence, RI, USA, PhD Thesis 2000.

[155] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in low noise, reverberative environments?," in *ICASSP* , Las Vegas, USA, 2008, pp. 2565-2568.

[156] C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538-548, April 2008.

[157] J. Velasco, D. Pizarro, and J. Macias-Guarasa, "Source Localization with Acoustic Sensor Arrays Using Generative Model Based Fitting with Sparse Constraints," *Sensors*, vol. 12, no. 10, pp. 13781-13812, October 2012.

[158] C. Segura, "Speaker Localization and Orientation in Multimodal Smart Environments," UPC, Barcelona, PhD Thesis 2011.

[159] M. F. Berger and H. F. Silverman, "Microphone array optimization by stochastic region contraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 39, no. 11, pp. 2377-2386, 2002.

[160] R. Duraiswami, D. Zotkin, and L. S. Davis, "Active speech source localization by a dual coarse-to-fine search," in *ICASSP*, Salt Lake City, USA, 2001.

[161] J. Dmochowski, J. Benesty, and S. Affes, "A Generalized Steered Response Power Method for Computationally Viable Source Localization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2510-2526, November 2007.

[162] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone arrays," in *ICASSP*, Hawaii, USA, 2007.

[163] H. Do and H. F. Silverman, "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC)," in *IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, New York, USA, 2007.

[164] D. E. Sturim, M. S. Brandstein, and H. F. Silverman, "Tracking multiple talkers using microphone-array measurements ," in *ICASSP*, Munich, Germany, 1997, pp. 371-374.

[165] E. D. Di Claudio, R. Parisi, and G. Orlandi, "Multi-source localization in reverberant environments by ROOT-MUSIC and clustering," in *ICASSP*, Istanbul, Turkey, 2000.

[166] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, "Localization of multiple sound sources based on a CSP analysis with a microphone array," in *ICASSP*, Istanbul, Turkey, 2000, pp. 1053-1056.

[167] H. Do and H. F. Silverman, "A method for locating multiple sources using a frame of a large-aperture microphone array data without tracking," in *ICASSP* , Las Vegas, USA, 2008, pp. 301-304.

[168] L. C. Parra, "Least squares frequency invariant beamforming," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, 2005.

[169] C. Nadeu, D. Macho, and J. Hernando, "Frequency & time filtering of filter-bank energies for robust HMM speech recognition," *Speech Communication*, vol. 34, no. 1-2, pp. 93-114, April 2001.

[170] S. J Young et al., "The HTK book (for HTK version 3.2)," Cambridge University, 2002.

[171] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*.: Prentice Hall, 1993.

[172] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*.: Wiley-Interscience, 2004.

[173] A. Temko, D. Macho, and C. Nadeu, "Fuzzy integral based information fusion for classification of highly confusable non-speech sounds," *Pattern Recognition*, vol. 41, no. 5, pp. 1814-1823, May 2008.

[174] M. Grabisch, "Fuzzy integral in multi-criteria decision-making," *Fuzzy Sets & Systems*, vol. 69, no. 3, pp. 279-298, February 1995.

[175] S. Nakamura et al., "Design and collection of acoustic sound data for hands-free speech recognition and sound scene understanding," in *Multimedia and Expo.*, 2002.

[176] AMI Corpus. [Online]. http://corpus.amiproject.org/

[177] W. Wang, Ed., *Machine Audition: Principles, Algorithms and Systems*.: IGI Global Press, 2010.

[178] G. W. Elko and J. Meyer, "Second-order differential adaptive microphone array," in *ICASSP*, Taiwan, R.O.C, 2009, pp. 73-76.

[179] S. Chang and S. Greenberg, "Syllable-proximity evaluation in automatic speech recognition using fuzzy measures and a fuzzy integral," in *IEEE Fuzzy Systems*, St. Louis, USA, 2003, pp. 828-833.

[180] H. Ney and S. Ortmanns, "Dynamic programming search for continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 16, no. 5, pp. 64-83, September 1999.

[181] A. Abad, D. Macho, C. Segura, J. Hernando, and C. Nadeu, "Effect of head orientation on the speaker localization performance in smart-room environment," in *Interspeech*, Lisbon, Portugal, 2005.

[182] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525-533, 1993.