UNIVERSITAT POLITÈCNICA
DE CATALUNYA

PhD Thesis

A multi-microphone approach to speech processing in a smart-room environment

Author: Alberto Abad Gareta

Advisor: Dr. Fco. Javier Hernando Pericás

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

Barcelona, February 2007

A mis padres y hermana,

Abstract

Recent advances in computer technology and speech and language processing have made possible that some new ways of person-machine communication and computer assistance to human activities start to appear feasible. Concretely, the interest on the development of new challenging applications in indoor environments equipped with multiple multimodal sensors, also known as smart-rooms, has considerably grown.

In general, it is well-known that the quality of speech signals captured by microphones that can be located several meters away from the speakers is severely distorted by acoustic noise and room reverberation. In the context of the development of hands-free speech applications in smart-room environments, the use of obtrusive sensors like close-talking microphones is usually not allowed, and consequently, speech technologies must operate on the basis of distant-talking recordings. In such conditions, speech technologies that usually perform reasonably well in free of noise and reverberation environments show a dramatically drop of performance.

In this thesis, the use of a multi-microphone approach to solve the problems introduced by far-field microphones in speech applications deployed in smart-rooms is investigated. Concretely, microphone array processing is investigated as a possible way to take advantage of the multi-microphone availability in order to obtain enhanced speech signals. Microphone array beamforming permits targeting concrete desired spatial directions while others are rejected, by means of the appropriate combination of the signals impinging a microphone array.

A new robust beamforming scheme that integrates an adaptive beamformer and a Wiener post-filter in a single stage is proposed for speech enhancement. Experimental results show that the proposed beamformer is an appropriate solution for high noise environments and that it is preferable to conventional post-filtering of the output of an adaptive beamformer. However, the beamformer introduces some distortion to the speech signal that can affect its usefulness for speech recognition applications, particularly in low noise conditions.

Then, the use of microphone arrays for specific speech recognition purposes in smart-room environments is investigated. It is shown that conventional microphone array based speech recognition, consisting on two independent stages, does not provide a significant improvement with respect to single microphone approaches, especially if the recognizer is adapted to the actual acoustic environmental conditions. In the thesis, it is pointed out that speech recognition needs to incorporate information about microphone array beamformers, or otherwise, beamformers need to incorporate speech recognition information. Concretely, it is proposed to use microphone array beamformed data for acoustic model construction in order to take more benefit from microphone arrays. The result obtained with the proposed adaptation scheme with beamformed enrollment data shows a remarkable improvement in a speaker dependent recognition system, while only a limited enhancement is achieved in a speaker independent recognition system, partially due to

the use of simulated microphone array data.

On the other hand, a common limitation of microphone array processing is that a reliable speaker position estimation is needed to correctly steer the beamformer towards the position of interest. Additionally, knowledge about the location of the audio sources present in a room is information that can be exploited by other smart-room services, such as automatic video steering in conference applications. Fortunately, audio source tracking can be solved on the basis of multiple microphone captures by means of several different approaches.

In the thesis, a robust speaker tracking system is developed based on successful state of the art SRP-PHAT algorithm, which computes the likelihood of each potential source position on the basis of the generalized cross-correlation estimations between pairs of microphones. The proposed system mainly incorporates two novelties: firstly, cross-correlations are adaptively computed based on the estimated velocities of the sources. The adaptive computation permits minimizing the influence of the varying dynamics of the speakers present in a room on the overall localization performance. Secondly, an accelerated method for the computation of the source position based on coarse-to-fine search strategies in both spatial and frequency dimensionalities is proposed. It is shown that the relation between spatial resolution and cross-correlation bandwidth is a matter of major importance in this kind of fast search strategies. Experimental assessment shows that the two novelties introduced permit achieving a reasonably good tracking performance in relatively controlled environments with few non-overlapping speakers. Additionally, the remarkable results obtained by the proposed audio tracker in an international evaluation confirm the convenience of the algorithm developed.

Finally, in the context of the development of novel technologies that can provide additional cues of information to the potential services deployed in smart-room environments, acoustic head orientation estimation based on multiple microphones is also investigated in the thesis. Two completely different approaches are proposed and compared: on the one hand, sophisticated methods based on the joint estimation of speaker position and orientation are shown to provide a superior performance in exchange of large computational requirements. On the other hand, simple and computationally cheap approaches based on speech radiation considerations are suitable in some cases, such as when computational complexity is limited or when the source position is known beforehand. In both cases, the results obtained are encouraging for future research on the development of new algorithms addressed to the head orientation estimation problem.

Resumen

Los avances recientes en tecnología informática y procesado del habla y del lenguaje, entre otros, han hecho posible que nuevos modos de comunicación entre las personas y las máquinas empiecen a parecer factibles. Concretamente, el interés en el desarrollo de nuevas aplicaciones en entornos cerrados equipados con múltiples sensores multimodales, también conocidos como salas inteligentes, ha aumentado considerablemente en los últimos tiempos.

En general, es bien sabido que la calidad de las señales de habla capturadas por micrófonos que pueden encontrarse a varios metros de distancia de los locutores se ve severamente degradada por el ruido acústico y por la reverberación de la sala. En el contexto del desarrollo de aplicaciones del habla en entornos de salas inteligentes, el uso de sensores que no sean molestos es un requisito habitual. Es decir, normalmente no está permitido o no es posible usar micrófonos cercanos o de solapa, y por lo tanto, las tecnologías del habla desarrolladas tienen que basarse en las señales capturadas por micrófonos lejanos. En estas situaciones, las tecnologías del habla que habitualmente funcionan razonablemente bien en entornos libres de ruido y reverberación sufren un descenso drástico en sus prestaciones.

En esta tesis se investigan métodos multi-micrófono para solventar los problemas que provoca el uso de micrófonos lejanos en las aplicaciones del habla que habitualmente se desarrollan en salas inteligentes. Concretamente, se estudia el procesado de arrays de micrófonos como un método posible de aprovechar la disponibilidad de múltiples micrófonos para obtener señales de voz mejoradas. Mediante la correcta combinación de las señales que inciden en una agrupación de micrófonos, el procesado de arrays permite apuntar direcciones espaciales concretas a la vez que otras se rechazan.

Para la mejora del habla con arrays de micrófonos, en la tesis se propone el uso de un nuevo esquema robusto de conformación que integra en una sóla etapa un conformador adaptativo y una etapa de post-filtrado de Wiener. Los resultados obtenidos muestran que el conformador propuesto es una solución adecuada para entornos muy ruidosos y que, en general, es preferible al uso convencional de etapas de post-filtrado a la salida de un conformador adaptativo. Sin embargo, el conformador muestra cierta degradación de la señal de voz que puede afectar a su utilidad para aplicaciones de reconocimiento del habla, especialmente cuando el ruido no es demasiado importante.

A continuación se investiga el uso específico de arrays de micrófonos para el reconocimiento del habla en entornos de salas inteligentes. Se demuestra que el uso convencional de arrays de micrófonos para reconocimiento del habla, que consiste en su aplicación en dos etapas independientes, no aporta una mejora significativa respecto al uso de técnicas mono-canal, especialmente, si el reconocedor está adaptado a las condiciones reales del entorno acústico. En la tesis se hace énfasis en la necesidad de que el reconocimiento del habla incorpore información de la

conformación con arrays de micrófonos, o alternativamente, que los conformadores incorporen información del reconocimiento del habla. Más concretamente, se propone el uso de datos capturados por un array de micrófonos y luego procesados por un conformador para la construcción de los modelos acústicos, para de esta manera, obtener un mayor beneficio de los arrays. La aplicación del esquema propuesto de adaptación con datos conformados de un array de micrófonos permite obtener una mejora considerable en un sistema de reconocimiento dependiente de locutor, mientras que en el caso de un sistema independiente de locutor sólo se obtiene una mejora muy limitada, debido en parte al uso de datos de array simulados.

Por otro lado, una limitación habitual del procesado de arrays de micrófonos es que se necesita una estimación verosímil de la posición del locutor para poder apuntar correctamente hacia la posición de interés. Además, el conocimiento de la posición de las fuentes acústicas que puedan estar presentes en una sala es una información que puede ser aprovechada por otros servicios que se desarrollan en las salas inteligentes, como por ejemplo para apuntar automáticamente una cámara en vídeo-conferencias. Afortunadamente, existen numerosos métodos que permiten resolver el problema del seguimiento de fuentes acústicas basándose en las señales capturadas por múltiples micrófonos.

Concretamente, en la tesis se desarrolla un sistema robusto de localización de locutor basado en uno de los algoritmos actuales de mayor éxito consistente en el cómputo de la verosimilitud de cada posible posición basándose en las estimaciones de las correlaciones cruzadas generalizadas entre pares de micrófonos. El sistema propuesto incorpora principalmente dos novedades. Primero, las correlaciones cruzadas se calculan de forma adaptativa basándose en las velocidades estimadas de las fuentes. Este cálculo adaptativo se hace de manera que se minimice el efecto de las diferentes dinámicas de las fuentes presentes en la sala en el resultado de la localización. Segundo, se propone el uso de un método acelerado para el cálculo de la posición basado en estrategias de búsqueda de menor a mayor resolución tanto en el dominio espacial como frecuencial. De hecho, se muestra que la relación entre resolución espacial y el ancho de banda considerado en el cálculo de las correlaciones cruzadas es un aspecto fundamental a tener en cuenta en la aplicación adecuada de este tipo de estrategias rápidas. Las dos novedades comentadas permiten que el sistema propuesto alcance unos resultados razonablemente buenos cuando se evalúa en escenarios relativamente controlados y con pocos locutores que no se solapan. Además, la conveniencia del sistema de localización acústica propuesto queda de manifiesto si se atiende a los destacados resultados que se obtuvieron en una evaluación internacional.

Finalmente, en la tesis también se estudia el problema de la estimación de la orientación del locutor en base a las señales capturadas por múltiples micrófonos en el contexto del desarrollo de nuevas tecnologías que puedan aportar información adicional para los sistemas que potencialmente pueden actuar en salas inteligentes. En concreto, se proponen y comparan dos métodos completamente diferentes. Por un lado, métodos sofisticados basados en la estimación conjunta de la posición y de la orientación que permiten obtener estimaciones aceptables a cambio de un elevado coste computacional. Por otro lado, los métodos más simples que se basan en consideraciones sobre el diagrama de radiación del habla aunque no son capaces de igualar las prestaciones de los métodos sofisticados, también pueden resultar adecuados en algunos casos, como cuando se sabe la posición de antemano o cuando la complejidad computacional está limitada. En ambos casos, los resultados obtenidos permiten ser optimistas de cara al futuro desarrollo de nuevos algoritmos dedicados a la estimación de la orientación del locutor.

Resum

Els avenços recents en tecnologia informàtica i processament de la parla i del llenguatge, entre altres, han fet possible que noves maneres de comunicació entre les persones i les màquines comencin a semblar factibles. Concretament, l'interès en el desenvolupament de noves aplicacions en entorns tancats equipats amb múltiples sensors multimodals, també coneguts com sales intel·ligents, ha augmentat considerablement darrerament.

En general, és ben conegut que la qualitat de les senyals de la parla capturades per micròfons que poden trobar-se a diversos metros de distància dels locutors es veu severament degradada pel soroll acústic i per la reverberació de la sala. En el context del desenvolupament d'aplicacions de la parla en entorns de sales intel·ligents, l'ús de sensors que no siguin molestos és un requeriment habitual. És a dir, no està normalment permès o no és possible fer servir micròfons propers o de solapa, i per tant, les tecnologies de la parla desenvolupades han de basar-se en les senyals capturades per micròfons llunyans. En aquestes situacions, les tecnologies de la parla que habitualment funcionen raonablement bé en entorns lliures de soroll i reverberació pateixen una davallada dràstica en les seves prestacions.

En aquesta tesis s'investiguen mètodes multi-micròfon per a solucionar els problemes que provoca l'ús de micròfons llunyans en les aplicacions de la parla que habitualment es desenvolupen en sales intel·ligents. Concretament, s'estudia el processament d'arrays de micròfons com a un mètode possible d'aprofitar la disponibilitat de múltiples micròfons per a obtenir senyals de veu millorades. Mitjançant la correcta combinació de les senyals que incideixen en una agrupació de micròfons, el processament d'arrays permet apuntar direccions espacials concretes a l'hora que altres es rebutgen.

Per a la millora de la parla amb arrays de micròfons, en la tesis es proposa l'ús d'un nou esquema robust de conformació que integra en només etapa un conformador adaptatiu i una etapa de post-filtrat de Wiener. Els resultats obtinguts mostren que el conformador proposat és una solució adequada per a entorns molt sorollosos i que, en general, és preferible a l'ús convencional d'etapes de post-filtrat a la sortida d'un conformador adaptatiu. No obstant això, el conformador mostra una certa degradació de la senyal de veu que pot afectar a la seva utilitat per a aplicacions de reconeixement de la parla, especialment quan el soroll no és massa important.

A continuació s'investiga l'ús específic d'arrays de micròfons per al reconeixement de la parla en entorns de sales intel·ligents. Es demostra que l'ús convencional d'arrays de micròfons per al reconeixement de la parla, que consisteix en la seva aplicació en dues etapes independents, no aporta una millora significativa respecte de l'ús de tècniques mono-canal, especialment, si el reconeixedor està adaptat a les condicions reals de l'entorn acústic. En la tesis es fa èmfasis en la necessitat de que el reconeixement de la parla incorpori informació de la conformació amb arrays de micròfons, o alternativament, que els conformadors incorporin informació del reconeixement

de la parla. Més concretament, es proposa utilitzar les dades primer capturades per un array de micròfons i després processades per un conformador per a la construcció dels models acústics, per a d'aquesta manera, obtenir un major benefici dels arrays de micròfons. La aplicació del esquema proposat d'adaptació amb dades conformades d'un array, permet obtenir una millora considerable en un sistema de reconeixement depenent de locutor, mentre que en el cas d'un sistema independent de locutor només s'obté una millora molt limitada, degut en part a l'ús de dades d'array simulades.

Per altra banda, una limitació habitual del processament d'arrays de micròfons és que es necessita una estimació versemblant de la posició del locutor per a poder apuntar correctament cap a la posició d'interès. A més, el coneixement de la posició de les fonts acústiques que poden estar presents en una sala és una informació que pot ser aprofitada per altres serveis que es desenvolupen en les sales intel·ligents, com per exemple per a apuntar automàticament una càmara en vídeo-conferencies. Afortunadament, existeixen nombrosos mètodes que permeten sol·lucionar el problema del seguiment de fonts acústiques basant-se en les senyals capturades per múltiples micròfons.

Concretament, a la tesis es desenvolupa un sistema robust de localització de locutor basat en un dels algorismes actuals de major èxit que consisteix en computar la versemblança de cada possible posició basant-se en les estimacions de les correlaciones creuades generalitzades entre parelles de micròfons. El sistema proposat incorpora principalment dues novetats. Primer, les correlacions creuades es calculen de forma adaptativa basant-se en las velocitats estimades de les fonts. Aquest càlcul adaptatiu es realitza de manera que es minimitzi l'efecte de les diferents dinàmiques de les fonts presents en la sala en el resultat de la localització. Segon, es proposa l'ús d'un mètode accelerat per al càlcul de la posició basat en estratègies de cerca de menor a major resolució tant en el domini espacial com en el freqüencial. De fet, es mostra que la relació entre resolució espacial i l'ample de banda considerat en el càlcul de les correlacions creuades és un aspecte fonamental a tenir en compte en l'aplicació adequada d'aquest tipus d'estratègies ràpides. Les dues novetats comentades permeten que el sistema proposat assoleixi uns resultats raonablement bons quan s'evalúa en escenaris relativament controlats i amb pocs locutors que no se solapin. A més, la conveniència del sistema de localització acústica proposat queda de manifest si s'atenen els destacats resultats que es van obtenir en una evaluació internacional.

Finalment, a la tesis també s'estudia el problema de l'estimació de l'orientació del locutor en base a las senyals rebudes per múltiples micròfons, en el context del desenvolupament de noves tecnologies que poden aportar informació addicional per als sistemes que potencialment poden actuar en sales intel·ligents. En concret, es proposen i comparen dos mètodes completament diferents. Por una banda, mètodes sofisticats basats en l'estimació conjunta de la posició i de l'orientació permeten assolir estimacions acceptables a canvi d'un elevat cost computacional. Per altra banda, els mètodes més simples que es basen en consideracions sobre el diagrama de radiació de la parla encara que no són capaços d'assolir les prestacions dels mètodes sofisticats, també poden resultar adequats en alguns casos, como ara quan es coneix la posició amb antelació, o bé quan la despesa computacional està limitada. En tots dos casos, els resultats obtinguts permeten ser optimistes de cara al futur desenvolupament de nous algorismes adreçats a l'estimació de l'orientació del locutor.

Agradecimientos

Buscando inspiración para superar de una manera acertada este delicado apartado y que no deje descontento a nadie, he acabado por leer lo que escribí en mi proyecto final de carrera. Resulta que todos los que aparecen en aquellos agradecimientos, siguen estando a mi lado, lo cual refuerza aún más si cabe, el hecho de que me sienta afortunado y orgulloso por la familia y amigos que tengo. A todos, incluido los que no cito a continuación, les estoy agradecido por dejarme formar parte de sus vidas.

En particular, mi insignificante forma de agradecer todo el amor, cariño y apoyo recibido de mis padres y hermana, es dedicándoles esta tesis. Fer, Pili y Marta: gracias por todo, os quiero mucho.

Me gustaría recordar también a los amigos que han estado más cerca de mi en estos cuatro últimos años y agradecer los buenos momentos compartidos que sin duda me han ayudado a seguir adelante. En especial, a mis compañeros de piso Simón, Aysel y Judith, gracias por aguantarme todos los días. A Ángel, Javi, Mingo y demás "peña micros" por esos grandes momentos delante de la consola (y delante de las barras). A Daniel, Rubén, David, Jaime y Marcos por los ratos pasados en la cochera del Pini. A Adrià, Maribel, Mariella y muchos más sufridores colegas de doctorado, que con su amistad han contribuido a hacer de la tesis una carga un poco menos pesada. A la gente de antes y de ahora de Telecogresca y Taller de Só, que siempre han sido un soplo de aire fresco importantísimo para mi. Por último, y por honrar el hecho de que siempre es el que llega más tarde a todos los sitios, quiero darle las gracias a Luque porque, parafraseándolo, es la persona con la que comparto más nodos.

Finalmente, me queda agradecer a aquellos que han tenido una implicación más profesional (aunque no exenta de lo personal) en que esta tesis se haya realizado. En primer lugar a mi director de tesis Javier, agradecerle que haya sabido llevarme a buen puerto a pesar de lo complicado que a veces puedo llegar a ser. A los numerosos compañeros que han pasado por el despacho por haberme ayudado cuando lo he necesitado, y muy especialmente a Jaume, Pere, Pablo, Jordi y Andrey. También agradecer a Joachim que se haya ofrecido a revisar la tesis. Por último, a las tres personas, al margen de mi director de tesis, con las que más estrechamente he colaborado: a Climent, que siempre ha estado dispuesto a ayudarme, a Dušan, un modelo en el que fijarme, y por supuesto a Carlos, que ha tenido una contribución fundamental en esta tesis.

Alberto Abad Gareta

Febrero 2007

Contents

1 Introduction **1**

1.1 Thesis Context 1

1.2 Objectives Statement 4

1.3 Dissertation Outline 5

2 Multi-microphone Processing **7**

2.1 Problem statement: Far-field environment 8

2.1.1 Acoustic noise 9

2.1.2 Reverberation 10

2.2 Fundamentals of array signal processing 12

2.2.1 Basic concepts 13

2.2.2 Applications: Beamforming and DOA estimation 20

2.2.3 Some microphone array particularities 26

2.3 Alternative multi-microphone approaches 30

2.3.1 Blind source separation 30

2.3.2 Multi-channel dereverberation techniques 31

2.3.3 Binaural processing 32

3 Speech Enhancement and Recognition with Microphone Arrays **33**

3.1 Microphone array processing for speech enhancement 34

3.1.1 Fixed beamforming 34

3.1.2 Adaptive beamforming 38

3.1.3 Post-filtering techniques 41

3.2 Overview of Automatic Speech Recognition 43

3.2.1 Front-End 44

3.2.2 Back-End 48
3.2.3 Approaches to speech recognition with microphone arrays 54

4 Contributions to Microphone Array Speech Enhancement and Recognition 57
4.1 Integrated Wiener-filtering and Adaptive Beamforming 58
4.1.1 Prior work 59
4.1.2 The proposed Integrated Wiener-filtering with Adaptive beamformer . . . 60
4.1.3 Experimental evaluation 63
4.1.4 Conclusions 71
4.2 Development of an ASR system for a smart-room environment 72
4.2.1 Data resources 74
4.2.2 Baseline Automatic Speech Recognition system 76
4.2.3 Speaker adaptation 78
4.2.4 Acoustic matched training and adaptation 80
4.2.5 Impact of beamforming on ASR 83
4.2.6 Conclusions 91

5 Audio Source Tracking and Head Orientation Estimation 93
5.1 TDOA and DOA estimation approaches 94
5.1.1 Direction of Arrival estimation 95
5.1.2 Time Difference of Arrival estimation 97
5.2 Position estimation approaches 100
5.2.1 Direct approaches 101
5.2.2 Closed-form approximations 101
5.2.3 SRP-based approaches 103
5.3 Source Tracking approaches 105
5.4 Head orientation estimation 108

6 Contributions to Speaker Tracking and Head Pose Estimation 111
6.1 Study of head orientation influence in a smart-room environment 112
6.1.1 Talker directivity and reverberation: The effect of orientation 112
6.1.2 Effect of head orientation on the speaker localization performance 114
6.1.3 Conclusions 119

6.2 Person tracking system for smart-room environments 119
6.2.1 Audio person tracking system baseline 119
6.2.2 Adaptive smoothing factor for Cross-Power Spectrum (CPS) estimations . 121
6.2.3 The two-pass search algorithm 124
6.2.4 Comparative evaluation 127
6.2.5 The CLEAR 2006 evaluation campaign 130
6.2.6 Conclusions 134
6.3 Head Orientation estimation 134
6.3.1 The SRP-PHAT based head orientation estimator 135
6.3.2 The HLBR head orientation estimator 138
6.3.3 Experimental evaluation 140
6.3.4 Conclusions 142

7 Conclusions and future work **143**
7.1 Summary and conclusions 143
7.2 Directions for future work 148

Bibliography **151**

List of Figures

2.1 *On the left, speech utterance captured by a close-talking microphone. On the right, the same speech utterance captured by a far-field microphone.* 9

2.2 *Schematic draw of a typical speaker-to-receiver room impulse response. Three parts can be distinguished: the direct wave, the early reflections and the late reflections.* 11

2.3 *Example of spatial diversity in the case of sinusoidal sources impinging an array of sensors. The values observed at a concrete time instant drawn in a spatial axis depend on the direction of arrival of the sources, their temporal frequency and the position of the sensors.* 13

2.4 *Example of a far-field source impinging a uniform linear array from angle φ_s. The delay of the waves arriving to each sensor is related to the inter sensor separation and the direction of arrival of the source.* 16

2.5 *Example of a far-field source impinging a planar array from azimuth φ_s and elevation θ_s. The delay of the waves arriving to each sensor is related to the inter sensor separation and the direction of arrival of the source.* 16

2.6 *Frequency domain processing of broadband signals. The process is split into 3 stages: firstly, the time domain array data is filtered with short time Fourier transform; next, narrow band beamformer or DOA estimation method is individually applied to the vector formed by all the sensors signal at a concrete frequency; and finally, in the case of beamforming the data is re-synthesized in the time domain or in the case of DOA estimation, multiple DOAs must be combined.* 25

2.7 *Energy gain response to different frequencies of a delay-and-sum beamformer steering 0° with a microphone array of 5 sensors separated 5 cm. Low spatial resolution can be observed below $f = (c/2d)$, while undesired grating lobes appear at higher frequencies.* 27

2.8 *The time delay of arrival of the waves impinging a microphone array depends on the wave propagation model. On the left, plane wave propagation model resulting in time delays dependent on the direction of arrival of the source and the inter sensor separation. On the right, spherical wave propagation model resulting in time delays dependent on the position of the source and the microphones.* 28

2.9 *Energy gain response to $400Hz$ and $1000Hz$ of a delay-and-sum beamformer steering 0° with microphone arrays of 2, 5 and 10 sensors separated 5 cm. Increasing the number of microphones results in narrower beams.* 29

2.10 *Energy gain response to* $400Hz$ *and* $7000Hz$ *of a delay-and-sum beamformer steering 0° with microphone arrays of 5 sensors separated 2, 5 and 10 cm. Increasing the microphone separation results in narrower beams at low frequencies and the appearance of undesired gratings lobes at high frequencies.* 30

3.1 *On the left, delay-and-sum beamformer in a time domain implementation. On the right, frequency and spatial response of delay-and-sum with 5 microphones separated 5 cm. Low spatial resolution is observed at the low frequency range, while spatial aliasing appears at the high frequency range.* 35

3.2 *On the left, example of an harmonically nested array for four different octaves: the minimum spacing of the microphones is 2cm in the case of the highest frequency dedicated sub-array. On the right, frequency and spatial response of the nested array. A clear reduction of the beampattern variation with respect to the delay-and-sum beamformer is obtained.* 36

3.3 *Filter and sum beamforming in the frequency domain of frequency bin* f. 37

3.4 *On the left, super directive beamformer in frequency domain implementation. On the right, frequency and spatial response of super directive beamformer with 5 microphones separated 5 cm. Compared to delay-and-sum beamformer, a more constant response is obtained for all frequencies.* 38

3.5 *Example of a GSC-like beamformer with its two path structure: the fixed beamformer (FBF) path and the adaptive path, which is composed by a blocking matrix (BM) and an adaptive multiple-input canceller (MC).* 40

3.6 *Multichannel Wiener post-filter of a delay-and-sum beamformer. The received speech signals are previously time-aligned by a time delay compensation (TDC) module.* 42

3.7 *Schematic structure of the basic steps of an automatic speech recognition system.* 44

3.8 *Block diagram of the mel frequency cepstral coefficients (MFCC) feature extraction process.* 47

3.9 *Example of a 10 band mel scaled triangular filter-bank in the range of 0-4 kHz.* . 47

3.10 *Example of generation of a concrete observation sequence by a left-to-right Markov chain of five states: two of them non-emitting -the first and the last- and three emitting states.* 50

4.1 *GSC beamformer with CCAF-LAF structure as proposed in Hoshuyama et al. (1999).* 60

4.2 *On the left side, schema of the common post-filter approach. On the right side, schema of the proposed integrated structure.* 61

4.3 *Detailed structure of the proposed IWAB beamformer.* 63

4.4 *On the left side SNR gain of the beamformers under study for different input SNRs. On the right side SIR gain of the same beamformers for different input SIRs.* 67

4.5 *LAR distance results of the beamformers under study and of the fourth unprocessed microphone array signal for different input SNRs.* 68

4.6 On the left LAR distance results for different input SNR and SIR 10 dBs and on the right results for different input SIRs and SNR 10 dBs. 69

4.7 *LAR distance results for an array with element spacing of 3 and 4 cm of the beamformers under study and the fourth unprocessed microphone array signal.* . . 71

4.8 *LAR distance results for a speaker located at 1 and 3 meters of the beamformers under study and the fourth unprocessed microphone array signal.* 71

4.9 *LAR distance results for a speaker located at 1 meter and a radio jamming signal from 45 degrees competing with the target speaker of the beamformers under study and the fourth unprocessed microphone array signal.* 72

4.10 *Sensors set-up in the UPC smart-room and positions of the recorded speaker..* . . 75

4.11 *Block diagram of the proposed scheme for adaptation with beamformed data.* . . . 88

6.1 *Talker diagram in the horizontal and vertical plane (after Chu and Warnock (2002)).* 113

6.2 *Multi-path example where direct sound wave is 0 and reflective waves are 1, 2 and 3. In this situation, indirect sound waves could reach sensors with higher energy.* 114

6.3 *Sensors set-up in the UPC smart-room.* 116

6.4 *Error in mm for various orientations in point P4 using T-array 1 (up), T-arrays 1&2 (middle) and the 3 T-arrays (bottom).* 117

6.5 *On the left, zenithal camera snapshot. On the right, example of the Spatial Likelihood Function obtained with the SRP-PHAT process. A maximum peak can be clearly observed close to the position of the main speaker.* 120

6.6 *Value of the adaptive smoothing factor depending on the estimated velocity.* . . . 122

6.7 *Blocks diagram of the Kalman based tracker used for velocity estimation.* 123

6.8 *Normalized directivity function depending on the angle differences for the computation of microphone pairs weighting.* 136

List of Tables

4.1 Coherence measures of the noise references given by the ABM ($n_1[n]$-$n_7[n]$), their average (AVG) and the output of MC ($e[n]$) with the speech components ($d_s[n]$) and the noise components ($d_w[n]$) of the corresponding FBF output. 63

4.2 *Word error rate (%) speech recognition results of the beamformers under study and the fourth unprocessed microphone array signal.* 70

4.3 *Speech recognition word error rate with HMMs trained with WSJ0 clean data set.* 78

4.4 *Speech recognition word error rate comparing HMM non-adapted and HMM adapted performances.* 79

4.5 *Speech recognition word error rate with non-adapted, speaker adapted and general non-native English adaptation HMMs.* 80

4.6 *Speech recognition word error rate with robust HMM set, testing with the US native speaker data set and the UPC unmatched speaker data set.* 81

4.7 *Speech recognition word error rate comparing the US native matched noised HMMs system and specific MLLR adaptation of the unmatched speaker stating from baseline clean models and noised models.* 82

4.8 *Speech recognition word error rate of the artificially noised HMM system tested with the US matched speaker, the UPC unmatched speaker, adapting and testing with the UPC speaker, and adapting with the artificially noised TED database and testing with the UPC unmatched speaker.* 83

4.9 *Speech recognition word error rate of the speaker adapted close-talking and distant-talking systems when tested with matched environment data and delay-and-sum beamformed data.* 84

4.10 *Speech recognition word error rate of the two speaker dependent systems adapted to the distant-talking environment when they are tested with beamformed data (delay-and-sum and weighted delay-and-sum).* 87

4.11 *Speech recognition word error rate comparing adaptation with single distant microphone data and testing with both single microphone and beamformed data, in contrast to model adaptation and testing with beamformed data.* 88

4.12 *Speech recognition word error rate of the proposed beamforming and adaptation approach tested with beamformed data based on actual labeled speaker position or estimated speaker position.* 89

4.13 *Speech recognition word error rate of the proposed adaptation scheme compared to single distant-talking microphone adaptation and testing, and to conventional application of beamforming as a pre-processing step in the case of the speaker independent system for non-native US English speakers.* 90

5.1 *Summary of the basic steps of the localization algorithm described in Brandstein et al. (1995).* 103

5.2 *Summary of the basic steps of the SRP-PHAT algorithm.* 104

6.1 *RMSE in mm of SRP and CL techniques (mean of all the points and orientations).* 118

6.2 *A-MOTA results of 8 randomly selected seminars with different forgetting factors for the computation of the cross-power spectrums. Depending on the specific data, high forgetting factor values show a very different behaviour.* 121

6.3 *Summary of the basic steps of the proposed localization system based on SRP-PHAT algorithm with a two-pass search strategy and adaptive forgetting factor for the estimation of the cross-power spectrums.* 127

6.4 *Audio person tracking results of interactive (IBM) seminars.* 129

6.5 *Audio person tracking results of non-interactive (UKA) seminars.* 129

6.6 *Audio person tracking results of interactive (IBM) seminars comparing the proposed two-pass search strategy and the computational expensive exhaustive SRP search.* 130

6.7 *Audio person tracking results of non-interactive (UKA) seminars comparing the proposed two-pass search strategy and the computational expensive exhaustive SRP search.* 131

6.8 *Summary of the recorded data by IBM, Istituto Trentino di Cultura (ITC), Research and Education Society in Information Technology (RESIT), University of Karlsruhe (UKA) and Universitat Politènica de Catalunya (UPC).* 132

6.9 *Audio results of the UPC system in the CLEAR 2006 evaluation for both single and multi-person tracking.* 133

6.10 *Results for acoustic single person tracking task in the CLEAR 2006 evaluation.* . 133

6.11 *Results for acoustic multi-person tracking task in the CLEAR 2006 evaluation.* . 134

6.12 *Summary of the basic steps of the joint source position and orientation estimation method based on SRP-PHAT algorithm (SRPPHAT-J).* 137

6.13 *Summary of the basic steps of the fast orientation estimation method based on SRP-PHAT algorithm (SRPPHAT-F).* 137

6.14 *Head pose orientation results of the four methods evaluated.* 141

6.15 *Audio person tracking results of the SRPPHAT-J and SRPPHAT-F algorithms evaluated with the CLEAR head pose database.* 142

Chapter 1

Introduction

This first chapter introduces the work carried out and overviews the main motivations and problems faced in the development of the present thesis. In Section 1.1 the research topic is contextualized and, particularly, the need for multi-microphone approaches in some highlighted speech applications is justified. Next Section 1.2 defines the objectives of the present dissertation. Last Section 1.3 describes the organization of the rest of the document.

1.1 Thesis Context

Human beings are extraordinary!!!

Lecturers might probably get confused with this starting declaration. What does the author mean with it? In what sense are humans extraordinary? What is new in this sentence? Is this document an engineering PhD thesis?

Questions such as these are completely understandable. However, if we forget for a moment the convenience of this starting sentence in a PhD thesis about multi-microphone processing, and we try to guess the results of a survey about the level of agreement in a wide sense with this sentence, we could surely bet on a positive result.

Motivations for supporting this affirmation could vary enormously and it is not the object of this discussion to list them or to argue in favour or against of some of the reasons. In fact, the intention is to discuss only one answer. One in a million that was appreciated by the author in the process of development of this thesis.

For the author it has been amazing to have the possibility to discover and appreciate the extraordinary capability of humans for the perception of the surrounding medium.

There are thousands of common behaviours related to the human perception of the objects

and particularly of the sounds emitted by them, that we difficultly realize of their complexity until we deeply study them. Indeed, this complexity appears even more evident when some of these human capabilities are intended to be imitated by computers.

A common example that is used for demonstrating the extraordinary abilities of our auditory system is the well-known cocktail party effect. The cocktail party effect describes the ability to focus one's listening attention on a single talker among a mixture of conversations and background noises, ignoring other conversations. Concretely, and this is the reason for the name of the mentioned effect, this phenomenon is easily appreciated in a crowded party. In this context, we are able to maintain a fluent conversation with an other person, even when other interfering conversations are happening around us or when we are close to loudspeakers playing music. In other words, we are able to focus our attention just to the desired "target" and to consider the rest of the interfering sources as background noise that we can ignore. In this case, the hearing reaches an high noise suppression of the background noise and enhances our focus of attention. In such a situation, a microphone recording placed exactly in our position will show a big difference.

Since this effect was firstly noticed by Cherry in 1953, the process in which the auditory system segregates a mixture of sounds derived from natural environment has been the subject of extensive research and is commonly known as "auditory scene analysis" (ASA). More concretely, the aim of the researchers in the areas of acoustics, physiology and psychology devoted to understand the ASA process is to discover the way in which our hearing sense is able to detect, localize, recognize, emphasize and segregate sound sources.

A necessary consequence of the advances in research and understanding of the human auditory system is the development of computational techniques capable of imitating its behaviour or some aspects of it. This research area is known as Computational auditory scene analysis (CASA).

In general, computational techniques that are traditionally classified as CASA are those that try to model the process of the human hearing system. For instance, some binaural processing applications try to imitate our two ears system to solve the problem of localization. However, there are technologies that are not usually considered CASA from the point of view that they are not concerned with the way that humans carry out ASA, but that are basically aimed to solve the same problems. From a practical point of view, the objective is not just to imitate the way in which our auditory system does its function, the general aim is the development of computational solutions related to the abilities of the human auditory system and understanding in order to improve the capabilities of person-machine communication.

Person-machine communication has been one of the most important research milestones of the speech community for decades. Particularly, there is a clear belief that spoken dialogue

would be the most natural and powerful user interface to computers. With recent improvements in computer technology, speech and language processing, and also in other fields such as image processing, new goals for computer communication and assistance are starting to appear feasible. Currently, it becomes evident that computers have to adapt to human requirements, being involved in human communication activities, and requiring the minimal possible awareness from the users. Consequently, there is a need of perceptual user interfaces which are multimodal and robust, and which use unobtrusive sensors.

An interesting example of these new challenging multimodal research efforts can be found in the development of smart-rooms. A smart-room is a closed space provided of multiple microphones and cameras, which is designed to assist and complement human activities. For instance, some of the innovative services offered are lecture summarization, identification of people attending a conference, or composition of a draft about what was said in a meeting.

Such highly sophisticated multimodal services are based on the information provided by many basic technological components of the various modalities. In the case of the audio processing, some of the technologies that are involved in these services are Voice Activity Detection, Automatic Speech Recognition, Speaker Identification and Verification or Acoustic Event Classification.

While most of these technologies perform reasonably well in controlled scenarios, in the context of the development of hands-free speech applications where close-talking microphones are not allowed, they present common problems. In this case, the situation is similar to the one described by the cocktail party phenomenon. Audio signals recorded with microphones that can be located several meters away from the source of interest are severely degraded by noise, reverberation, and also position, orientation and dynamics of multiple concurrent speakers. As a consequence of these sources of degradation, audio technologies show a dramatic lost of performance.

One field of growing interest to reduce problems introduced by distant microphone recordings consists in taking advantage of the multi-microphone availability. More concretely, microphone array processing has been broadly investigated as a pre-processing stage in order to enhance the recorded signal that might be used for any speech application. The basic idea of beamforming is to generate a directive beam that targets the desired direction based on the combination of the signals arriving to an array of microphones.

Unfortunately, most of the speech enhancement techniques based on multi-microphone processing rely on one fundamental cue that is mostly unknown: the source position. The need for a reliable target position estimation in the beamforming applications is one of the reasons for the increasing interest in the acoustic source localization and tracking topic. Furthermore, accurate knowledge of the position of the events or the speakers present in a room is also useful

for other multimodal services like analyzing group dynamics or behaviors, deciding which is the active speaker among all the presents, or providing information for an automatic steering camera system. Furthermore, knowledge about speaker head orientation can be a useful information in such applications.

Hence, in real smart-room environments, a multi-microphone approach to speech processing would permit to enhance multiple captured speech signals on the basis of a estimated speaker position, and eventually these enhanced signals might be used for improving the performance of an automatic speech recognition application, or any other speech based application.

1.2 Objectives Statement

Research on multi-microphone approaches to speech processing has been continuously growing during the last two decades, until currently becoming a remarkable research line in many of the most important speech related investigation laboratories.

The need for taking off multi-microphone processing research in the UPC Speech Processing Group is one of the fundamental motivations of this thesis, which has been further motivated by the recent projects in which the Speech Processing Group has been involved, such as the EU funded CHIL project.

Thus, the work presented in this dissertation is the first effort to open and explore this appealing research line. Consequently, one of the main goals of the thesis is to fully investigate the current state of the art of the multi-microphone processing topic, in order to meet the leading research laboratories and also to establish a consistent basis for future research and development.

More concretely, the thesis is addressed to investigate and propose solutions in two concrete fields of interest strongly related as mentioned above: speech enhancement and recognition with microphone arrays and audio source localization and tracking.

In both cases, the objective is to propose computational solutions able to work reasonably well in a broad range of environments and conditions, but with particular attention to the smart-room case. Particularly, a practical goal of this thesis is to build technological components to be deployed in the smart-room recently built at one of the UPC laboratories.

Regarding microphone array speech enhancement, the objective is to propose practical solutions capable of taking advantage of the microphone network available for enhancing the performance of the various speech technologies that can be developed in smart-room applications. Concretely, research will be principally focused to the development of a robust automatic speech recognition system on the basis of the multiple microphone far-field sensors.

The main objective of the research on acoustic source localization and tracking, will be the

development of an algorithm for the reliable estimation of the position of the acoustic events inside a room, and particularly for the speaker localization.

Finally, and closely related to the source localization work, in this thesis is intended to provide an insight and some solutions to a topic of very recent research interest, that is the problem of head orientation estimation on the basis of multiple microphone recordings.

1.3 Dissertation Outline

This thesis is organized as follows:

Chapter 2 describes the problem that motivates this thesis: the degradation caused by the use of far-field microphones in speech processing applications. Microphone array signal processing is presented as an alternative to solve these problems. Hence, an overview of the fundamentals of array signal processing theory and the main particularities of microphone array signal processing is provided as the basic background for the following chapters. Additionally, some notable alternatives to microphone array processing are also briefly commented.

In Chapter 3 the most representative approaches to the problem of speech enhancement with microphone arrays are introduced. Microphone arrays for automatic speech recognition is also described, providing an overview of HMM based recognition systems and describing some of the most outstanding state of the art proposals.

In Chapter 4 the main contributions and work carried out by the author related to the speech enhancement and recognition with microphone arrays topic are shown. Firstly, an integrated adaptive beamformer with post-filtering is presented, showing convenient performance for speech enhancement applications. Secondly, a robust medium size vocabulary Automatic Speech Recognition system for smart-room environments is developed and described. Based on both robust model construction and microphone array processing, the system achieves a remarkable performance compared to the initial baseline system.

In Chapter 5 the problem of audio source localization and tracking is decomposed into three basic problematics: estimation, positioning and tracking. The most remarkable current state of the art solutions to address these problems are reviewed and classified. Additionally, recent state of the art approaches to acoustic head pose estimation are commented.

In Chapter 6 audio source tracking and head orientation estimation contributions are presented. Firstly, a novel study of the degrading effects of head orientation in selected source localization tasks is described. Secondly, the design process and the experiments carried out to develop a robust audio person tracking system for smart-room environments is explained. Performance achieved by the proposed system in an international competitive evaluation is also

provided. Finally, new proposals on head orientation estimation are discussed and assessed.

Finally, Chapter 7 summarizes the major results and contributions of this thesis, and highlights some directions for future research.

Chapter 2

Multi-microphone Processing

This chapter describes the problem that motivates this thesis: the degradation caused by the use of far-field microphones in speech processing applications. Microphone array signal processing is presented as an alternative to solve these problems. Hence, an overview of the fundamentals of array signal processing theory and the main particularities of microphone array signal processing is provided as the basic background for the following chapters. In addition, some alternative multi-microphone approaches are also commented.

Speech recorded in real environments by distant microphones is dramatically degraded by factors like noise and reverberation. This degradation strongly affects accuracy of the speech systems depending on this input data and as a consequence, the development of robust processing techniques for speech enhancement has been a main topic of research among the speech community along decades (Junqua & Haton, 1996). For this reason, there exist a countless variety of proposed techniques addressing these problems and particularly most of them are oriented to mono-channel applications.

A natural extension to the usage of a single microphone is the development of multi-channel approaches capable of taking benefit of the spatial diversity. One of the most interesting possibilities for this purpose and that it has been broadly exploited in other signal processing fields is the array signal processing theory (Van Veen & Buckley, 1988). Array processing basically consists in the design of spatial filters capable of selecting a desired space direction and rejecting all the others. Concretely, microphone array processing can be used as a pre-processing stage in order to enhance the recorded signal but, unfortunately, as it will be shown in this section, it presents some extra difficulties due to the special characteristics of the speech signal.

However, microphone array processing is not the unique approach to multi-channel processing and other completely different techniques like blind source separation techniques based on the Independent Component Analysis (Hyvärinen et al., 2001) are also possible. Although multi-channel approaches not related to beamforming are initially out of the scope of this thesis, some of them will be briefly commented at the end of this chapter.

2.1 Problem statement: Far-field environment

The communication process, from an engineering point of view, consists in the exchange of information between an emitter and a receiver through a channel. In order to achieve a successful communication it is necessary that the emitter and the receiver use common codifications - in other words, they can understand each other - and the channel must be suitable for transmitting the message.

In any radiated communication system, the channel is affected by several sources of distortion that can seriously threaten the success of the communication. Some of these disturbing effects are noises and interferences, attenuation due to propagation or presence of obstacles, channel distortions or multi-path propagation.

Speech communication is obviously not free of these undesirable effects. In human-to-human speech communication the extraordinary abilities of our auditory system permits successful communications even in extremely adverse conditions, such as the one already described called cocktail party effect.

In the case of human-machine communication the basic machine acquisition system is the microphone. The quality of the signal captured by a microphone depends on two major factors: acoustic noise and reverberation. In the case of applications relaying in close talking microphones the influence of these two factors is relatively low, which permits the development of successful interfaces. As long as microphones are located further from the sources of interest the influence of noise and reverberation becomes critical in the performance of the developed systems.

Hence, the basic problem that motivates this Thesis is the degradation suffered by speech signals collected in real environments, particularly when the microphones are located far away of the sources of interest.

As an example, Figure 2.1 shows the same speech utterance recorded with a close-talking microphone and with a far-field microphone at 4-5 meters in a smart-room. The two main sources of degradation can be distinguished affecting the far-field signal: the additive acoustic noise that can be observed during the silence periods and the reverberation that smears and degrades the speech signal.

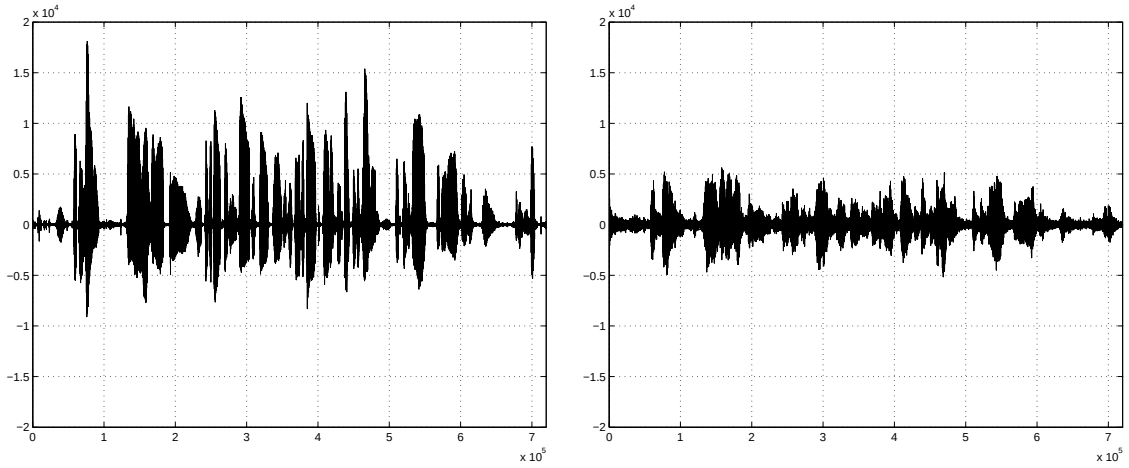

Figure 2.1: *On the left, speech utterance captured by a close-talking microphone. On the right, the same speech utterance captured by a far-field microphone.*

2.1.1 Acoustic noise

Hereinafter, acoustic noise refers to the overall of undesired sound events, that is, any addition of external disturbances to the target information received at a microphone. In this way, acoustic noise - or simply noise - does not refer to concrete statistical, frequency, spatial or propagation characteristics.

For instance, in the case of automatic speech recognition in a smart-room environment everything but the target speaker is considered noise although their characteristics can significantly diverge. This is the case of other persons in the room that eventually speak and the noise produced by computer fans or air conditioning. In the first case, noise has speech-like characteristics, while in the second case noise is highly stationary. However, these two sources show a common characteristic and is the fact that the noise is produced in a very concrete position in space. In other words, the signal captured by two microphones at different positions will be highly correlated. The coherence is a measure of normalized correlation, which is defined as:

$$\Gamma_{ij}(f) = \frac{\Phi_{ij}(f)}{\sqrt{\Phi_{ii}(f)\Phi_{jj}(f)}} \tag{2.1}$$

where Φ_{ij} is the cross-spectral density between signals i and j. Its magnitude is bounded by $0 \leq |\Gamma_{ij}|^2 \leq 1$. In the case of directional sources $|\Gamma_{ij}|^2 \approxeq 1$.

In general, noises are going to be classified into two broad kinds: directional noises as the ones commented and non-directional noise.

Non-directional noise refers to the background noise that is affecting microphone recordings and that in some sense can be considered as noise coming from everywhere. There are different

kinds of background noise basically depending on the coherence of the noise field in two different positions. For instance, background noise in most room applications is usually considered to form a spherical isotropic field also known as diffuse noise. The coherence of a diffuse noise field in two points of space depends on the distance (d_{ij}) - the most separated the less coherence - and the frequency (f) - high coherence for low-frequency components and decreasing for high frequencies - as follows:

$$\Gamma_{ij}(f) = sinc(\frac{2\pi f d_{ij}}{c}) \tag{2.2}$$

Alternatively to diffuse fields other noise fields depending on the scenario can better approximate the real background noise. In some applications, the background noise is considered to be spatially white, that is, the noise at different points in space is completely uncorrelated ($|\Gamma_{ij}|^2 \approxeq 0$). Spatially white noise fields are far from being real in speech applications, however in some derivations of the array signal processing theory is commonly assumed. Electrical noise introduced by each microphone is usually not correlated between microphones and can be considered to be white.

2.1.2 Reverberation

The propagation of acoustic signals in a closed space is generally multi-path, that is, a sound wave propagates from its origin to a point in the space across multiple paths. In this way, the signal recorded by a microphone in addition to the contribution of the direct path from the sound source to the microphone, includes multiple indirect paths due to reflections with the different surfaces of the room or non-straight paths due to diffraction with the objects present or the medium. The number of multiple paths is very large tending to infinite, resulting in a phenomenon very different to echoes which directions that can be distinguished. The multi-path propagation phenomenon of sound waves is commonly known as reverberation.

In practice, the effect of reverberation states as the persistence of sound after the original sound is removed. The effect depends basically on the characteristics of the space where sound is produced, for instance in large chambers the persistence of reverberation is long and can be clearly heard. This phenomenon in a room can be characterized by the reverberation time, which is defined as the time required for a sound in the room to decay by 60 dB.

The reverberation time is controlled primarily by two factors - the surfaces in the room and the size of the room. The surfaces of the room determine how much energy is lost in each reflection. Highly reflective materials increase the reverberation time as they are very rigid. Absorptive materials, reduce the reverberation time. In general, the absorptivity of most materials usually varies with frequency. For a simple room, the Sabine's formula gives a relation

between the volume of the room ($V(m^3)$), the area of each surface ($S_i(m^2)$) and its absorption coefficient a_i:

$$T_{60} = \frac{0.161V}{\sum_i a_i S_i} \tag{2.3}$$

Room reverberation is also characterized by the speaker-to-receiver room impulse response, which is schematically shown in Figure 2.2. Three parts of the room impulse response can be distinguished: direct wave, early reflections and late reflections. The set of early reflections are well defined directional reflections that are directly related to the shape and size of the room, as well as the position of the source and listener in the room. After the early reflections, the rate of the arriving reflections increases greatly. These reflections are more random and difficult to relate to the physical characteristics of the room. These are the late reflections which are also called the diffuse reverberation, since its characteristics are similar to a diffuse field. While the direct wave and early reflections follow the inverse square propagation law, the late reflections remain approximately constant in a room independently on the source and receiver positions. Thus, the ratio of energy in the direct path impulse to the energy in the remainder of the impulse response is inversely proportional to the distance traveled between the source and the receiver. This ratio also characterizes the speaker-to-receiver room impulse response and is named the signal-to-reverberation ratio (SRR).

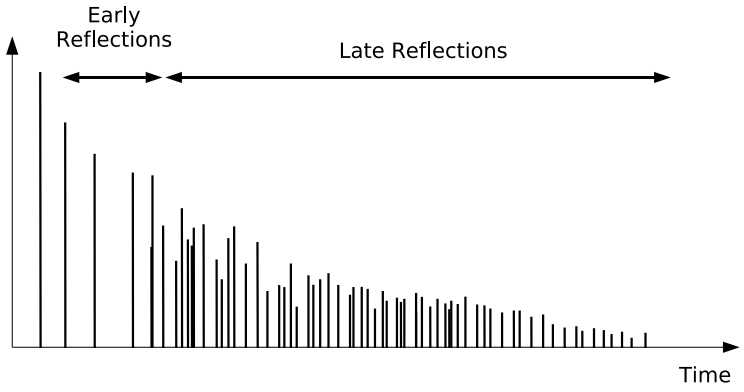

Figure 2.2: *Schematic draw of a typical speaker-to-receiver room impulse response. Three parts can be distinguished: the direct wave, the early reflections and the late reflections.*

In general, the amount of smearing caused by reverberation is a function of both the reverberation time and the SRR. The early reflections cause a distortion called coloration, however it is mostly the late reflections that smear the speech spectra and reduce the intelligibility and quality of speech signals, resulting in a matter of major importance in most robust speech applications. However, it is worth noting that in other fields of acoustics like music, the reverberation provides some desired effects that improves the hearing experience. In this way, reverberation is

not essentially a damaging effect, it depends on the application field.

2.2 Fundamentals of array signal processing

In general, every process of selection, filtering or classification firstly consists in the extraction of particular characteristics of a given signal. These characteristics should permit differentiate various phenomenons in order to select those desired.

In traditional digital signal processing we obtain these differential characteristics or diversity in the time domain. By sampling signals at different time instants we obtain a time diversity that permits, in a second step, designing filters capable of selecting particular characteristics of the given signal. For instance, we can design a simple Finite Impulse Response (FIR) filter devoted to enhance concrete frequencies.

An array is simply a set of sensors that permits obtaining samples of a signal at different spatial positions, thus generating diversity in the space domain in terms of time delays.

It is possible to make a correspondence between classical signal processing based on time diversity and array signal processing based on spatial diversity. Consider a sinusoidal source impinging a uniform linear array, that is, an array with sensors linearly distributed and equi-spaced. Furthermore, the source is assumed to be far enough of the array to consider that the wavefronts arriving to the sensors are plane. In such situation, a source located close to perpendicular direction to the array as the one labeled in Figure 2.3 as a), will arrive approximately at the same time at each sensor. In other words, if we annotate amplitudes received at each sensor of a source located at that position, they will be approximately equal and as a consequence if we draw these amplitudes in an spatial axis they will look like a constant (low spatial frequency). The source in that position is characterized by this concrete spatial signature.

If we now consider the same situation with a source located at approximately 45° (Figure 2.3, source b)), the amplitude received at a time instant or spatial signature at each sensor will depend on the delay of arrival between each sensor, resulting in a higher spatial frequency. Furthermore, this spatial signature will depend not only on the relative positions of the source with respect to the array, it will depend also on the temporal frequency of the signal.

Anyway, the important fact is that if the spatial signature of the signals arriving to an array depend somehow on the position of the source as shown, in the same way that a FIR filter is designed to select concrete frequencies, it is also possible to design spatial filters in order to select and enhance concrete directions and reject others. This process of spatial filtering is also known as beamforming.

Although the problems of temporal filtering and beamforming are quite similar, there exist

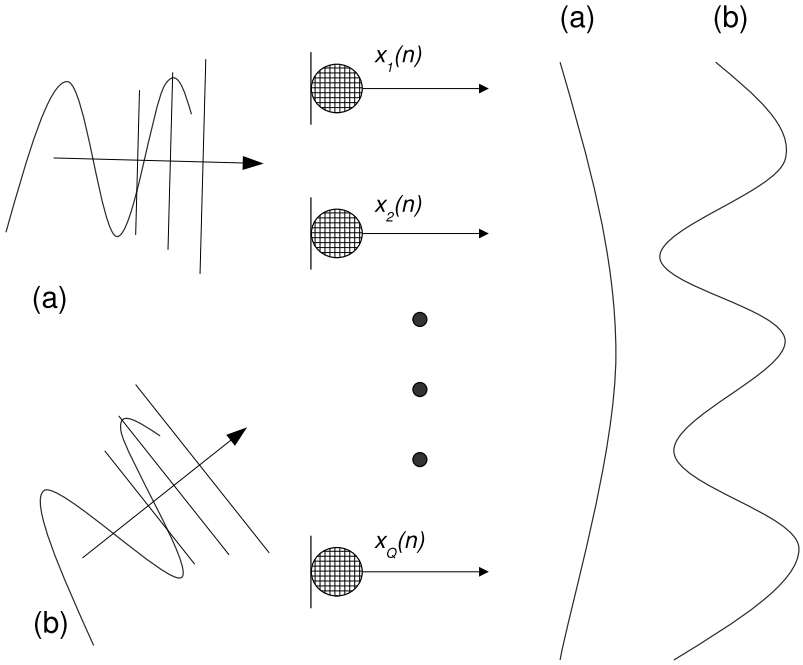

Figure 2.3: *Example of spatial diversity in the case of sinusoidal sources impinging an array of sensors. The values observed at a concrete time instant drawn in a spatial axis depend on the direction of arrival of the sources, their temporal frequency and the position of the sensors.*

some important differences that make array signal processing more complex and sophisticated. For instance, an obvious difference is that arrays do not need to be linear or uniform. Then, this section seeks to develop the principles of array processing first retaining generality for any kind of sensor arrays, to later discuss some of the particularities of microphone array signal processing.

2.2.1 Basic concepts

Array signal processing as already mentioned basically consists on the design of selective spatial filters in order to enhance or reject concrete spatial directions. The basis of these techniques relies on the information about the source position that is included in the phase of the different signals captured in each sensor and in the correct combination of these signals.

The previous example of Figure 2.3 and in general the array signal processing theory assumes that source signals are narrow band and that they are located in the far-field and as a consequence they are received as plane waves.

Wave propagation

Information about distant events is carried to sensors by propagating waves. The physics of propagation is described by the wave equation for the appropriate medium and boundary conditions. The fundamental wave propagation equation is:

$$\nabla^2 s(\mathbf{r},t) = \frac{1}{c^2}\frac{\partial^2 s(\mathbf{r},t)}{\partial^2 t} \tag{2.4}$$

where $s(\mathbf{r},t)$ represents a general scalar field, c is interpreted as the speed of propagation and ∇^2 is the Laplacian operator. The wave equation in the case of electromagnetic fields can be derived from Maxwell's equations where $s(\mathbf{r},t)$ would represent the electric field $\mathbf{E}$. In acoustics, this equation takes the same form lead from the basic laws of physics and the conservation of mass, with $s(\mathbf{r},t)$ representing sound pressure at a point in space and time and c the speed of sound.

Equation 2.4 governs how signals pass from a source radiating energy to an array. The solution to the wave equation for monochromatic plane waves is as follows:

$$s(\mathbf{r},t) = Ae^{j(2\pi f_o t - \mathbf{k}^T \mathbf{r_q})} \tag{2.5}$$

where A is a complex constant, $\mathbf{k}$ is the wave vector which magnitude, usually named wavenumber, is $|\mathbf{k}| = \frac{2\pi}{\lambda} = \frac{2\pi f_o}{c}$, being f_o and λ the carrier frequency and wavelength of the arriving wave.

In array signal processing, a source is considered to be in the far-field when its distance to the array is $r > 2L^2/\lambda$ where L is the size of the array. In this case, the arriving wavefronts are considered to be plane and the solution of Equation 2.5 is valid. Otherwise, a near field source propagates with a spherical wavefront and a different solution to the wave equation in 2.4 must be obtained.

The plane wave solution means that at a concrete time instant, all the points that have identical constant phase form a perpendicular plane to the vector propagation $\mathbf{k}$. Thus, the plane wavefronts propagation model permits defining arriving signals to sensors in terms of delayed versions of the emitted signal depending on its direction of arrival.

Because the wave equation is linear, because complex exponentials solve the wave equation, and because arbitrary functions can be expressed as a weighted superposition of complex exponentials, any signal, no matter what its shape, satisfies the wave equation.

Time delay of arrival

Consider a source emitting a narrow band signal $x_s(t)$, assuming that it is in the far-field region, the signal $x_q(t)$ received at sensor q with $q = 1 \dots Q$ can be expressed as:

$$x_q(t) = x_s(t - t_o - \tau_q) \tag{2.6}$$

where t_o is the propagation delay to a reference point common to each sensor and τ_q is the delay with respect to the reference point that will show each sensor depending on its position and the direction of arrival of the source.

The delay τ_q due to the propagation velocity of the wavefront is obtained from the scalar product of the wave vector $\mathbf{k}$ and the position of the sensor $\mathbf{r_q}$.

$$\mathbf{k}^T \mathbf{r_q} = 2\pi f_o \tau_q \tag{2.7}$$

Figure 2.4 shows the example of a uniform linear array. The time delay with respect to a reference point, for instance the coordinate origin, can be related to the angle of arrival of the source (φ_s) as follows and can be checked with simple trigonometric relations:

$$\mathbf{k}^T \mathbf{r_q} = \frac{2\pi f_o}{c} \begin{bmatrix} \cos(\varphi_s) & \sin(\varphi_s) \end{bmatrix} \begin{bmatrix} d_q \\ 0 \end{bmatrix} = \frac{2\pi f_o}{c} d_q \cos(\varphi_s)$$

$$\tau_q = \frac{d_q}{c} \cos(\varphi_s) \tag{2.8}$$

where d_q is the distance from the reference point to the q sensor. Notice that the reference point was arbitrary chosen, for instance in this case it could be more practical to select the position of the first sensor arriving the signal as the reference point. Thus, signals arriving to all the other sensors will be delayed with respect to this first sensor.

It is also worth to notice that in this case of linear array there is no difference between a signal arriving from φ_s or $-\varphi_s$. Even more, there is no difference for every signal arriving with the same angle with respect to the array axis, showing a cylindrical symmetry. Regarding symmetrical directions with respect to the perpendicular plane to the array (say $\pi - \varphi_s$), it is possible to disambiguate them preserving the sign of d_q, thus generating positive or negative delays with respect to the reference point depending on which sensor the signal arrives first. In general, it can be said that a linear array is only capable of distinguish diversity in one dimension.

Figure 2.5 illustrates the more general case of a planar array. In this case the delay depends on both the azimuth (φ_s) and the elevation (θ_s) as follows:

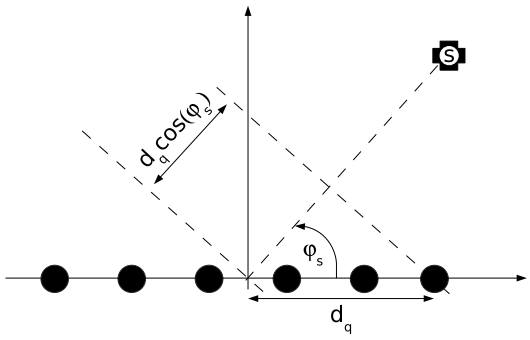

Figure 2.4: *Example of a far-field source impinging a uniform linear array from angle φ_s. The delay of the waves arriving to each sensor is related to the inter sensor separation and the direction of arrival of the source.*

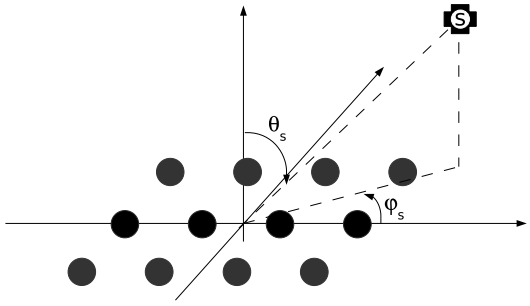

Figure 2.5: *Example of a far-field source impinging a planar array from azimuth φ_s and elevation θ_s. The delay of the waves arriving to each sensor is related to the inter sensor separation and the direction of arrival of the source.*

$$\mathbf{k}^T\mathbf{r_q} = \frac{2\pi f_o}{c}\begin{bmatrix}\cos(\varphi_s)\sin(\theta_s) & \sin(\varphi_s)\sin(\theta_s) & \cos(\theta_s)\end{bmatrix}\begin{bmatrix}d_q\cos(\varphi_q)\\ d_q\sin(\varphi_q)\\ 0\end{bmatrix} = \frac{2\pi f_o}{c}d_q\sin(\theta_s)\cos(\varphi_s-\varphi_q)$$

$$\tau_q = \frac{d_q}{c}\sin(\theta_s)\cos(\varphi_s - \varphi_q) \tag{2.9}$$

In this case of a planar array it is possible to distinguish sources in two dimensions, however ambiguity with respect to array plane still remains.

Narrowband assumption

The second condition that has been considered in all the examples and that the basic array signal theory assumes is that the sources emit narrow band signals.

This fact can be easily understood as follows. Assume there is a pass-band signal $x_s(t)$ with complex envelope $a_s(t)$ and carrier frequency f_o impinging an array in the far-field. The received signal at sensor q of Equation 2.6 - omitting the time constant propagation delay term t_o - can now be written as,

$$x_q(t) = x_s(t-\tau_q) = a_s(t-\tau_q)\exp(j2\pi f_o(t-\tau_q)) \tag{2.10}$$

In order that the spatial signature of a signal such as this is only dependent on its position and the carrier frequency, it becomes necessary that the complex envelope is slow varying. In other words, the group delay must be negligible for any sensor of the array:

$$x_q(t) = x_s(t - \tau_q) \cong a_s(t) \exp(j2\pi f_o(t - \tau_q)) \tag{2.11}$$

$$a_s(t - \tau_q) \cong a_s(t) \tag{2.12}$$

In the frequency domain:

$$A_s(f) \exp(-j2\pi f \tau_q) \cong A_s(f) \tag{2.13}$$

In this situation, only the carrier with its associated phase delay signs in the array with information depending on the angles of arrival. From Equation 2.13, it can be derived that the narrow band condition is accomplished if the product of frequency and delay group is negligible. In the worst case the maximum frequency will be the bandwidth of the signal, say B_o, and the maximum delay will correspond to the maximum sensor distance (L) divided by the constant propagation velocity (c),

$$f\tau_q < B_o \frac{L}{c} = \frac{B_o}{f_o} \frac{L}{\lambda} << 1 \tag{2.14}$$

$$\frac{L}{\lambda} << \frac{f_o}{B_o} \tag{2.15}$$

Hence, this condition is satisfied and we can consider a narrow band scenario if the dimension of the array in terms of wavelengths is much smaller than the ratio between the center frequency and the bandwidth of the sources present.

Spatial aliasing

A particular problem of array processing that must be taken into account is due to the coupling of both temporal and spatial dimensionalities. Regarding Figure 2.3 it was shown that the spatial signature of a source was depending on the position of the emitting source, its carrier frequency and obviously the structure of the receiving array indeed. These dependencies suggest a possible drawback, in some cases sources arriving from different directions are probably signing with an identical spatial signature. Thus, a beamformer can be enhancing two or more simultaneous directions, one of them desired and the others undesired. This problem is known as spatial aliasing.

Spatial aliasing can happen if the sensors are located too far, concretely if the distance between sensors is larger than half the wavelength of the arriving waves. Hence, the condition

for avoiding the appearance of undesired grating lobes is $d \leq \lambda/2$, being d the inter sensor separation. In practice, spatial discrimination suffers as a result of the smaller than necessary array, resulting $d = \lambda/2$ a convenient choice for the sensor separation.

Signal model at the array

The values taken by each sensor of the array at a concrete time instant can be written in a vector form that is usually named snapshot vector. In a free of noise scenario with only one band-pass emitting source s, the snapshot $\mathbf{x_t}$ observed at sampled time instant t can be written as follows

$$\mathbf{x_t} = a_s(t)\mathbf{s_s}\exp(j2\pi f_o t) \tag{2.16}$$

$$\mathbf{s_s} = [\exp(-j2\pi f_o \tau_1), \ldots, \exp(-j2\pi f_o \tau_q), \ldots, \exp(-j2\pi f_o \tau_Q)]^T \tag{2.17}$$

where $\mathbf{s_s}$ is known as the steering vector and contains the information about the phase delay to each sensor. Notice that Equation 2.16 clearly separates the temporal information of the source and the position information. Hereinafter, the carrier term is going to be omitted for shake of simplicity, since it will be only necessary to add it to the following derivations.

In a more general scenario with N_s emitting sources and in presence of non-directional additive noise $\mathbf{n_t}$ the snapshot can be written as follows:

$$\mathbf{x_t} = \sum_{s=1}^{N_s} a_s(t)\mathbf{s_s} + \mathbf{n_t} \tag{2.18}$$

In general, the snapshots captured by the array are obtained at a fixed interval time and, in a digital implementation, Equation 2.18 can be rewritten in digital notation as:

$$\mathbf{x_n} = \sum_{s=1}^{N_s} a_s(n)\mathbf{s_s} + \mathbf{n_n} = \mathbf{S}\mathbf{a_n} + \mathbf{n_n} \tag{2.19}$$

being

$$\mathbf{a_n} = \begin{bmatrix} a_1(n) \\ a_2(n) \\ \vdots \\ a_{N_s}(n) \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} \mathbf{s_1} & \mathbf{s_2} & \ldots & \mathbf{s_{N_s}} \end{bmatrix} \tag{2.20}$$

The statistics of the data observed at the array play an important role to understand and develop the uses of array processing, since evaluation of spatial filters usually involves power or

variance. The vector of data received at the sensors is usually considered to be zero mean and long term stationary. The spatial covariance matrix $\mathbf{R_x}$ formed by the cross-covariance information among the various sensors - assuming that noise and directional sources are statistically independent ($\mathrm{E}\left\{\mathbf{n_n a_n}^H\right\} = \mathbf{0}$) - is defined as follows:

$$\mathbf{R_x} = \mathrm{E}\left\{\mathbf{x_n x_n}^H\right\} = \mathbf{S}\mathrm{E}\left\{\mathbf{a_n a_n}^H\right\}\mathbf{S}^H + \mathrm{E}\left\{\mathbf{n_n n_n}^H\right\} = \mathbf{S R_s S}^H + \mathbf{R_n} \tag{2.21}$$

$\mathbf{R_s}$ and $\mathbf{R_n}$ are the cross-covariance matrices of the sources and the additive noise respectively. In this case of narrow band signals and assuming that sources are uncorrelated between them ($\mathrm{E}\{a_i(n)a_j(n)^*\} = 0$ if $i \neq j$) the sources cross-covariance matrix is a diagonal matrix formed by the variances or power of each source (σ_s^2). Additionally, if the noise is spatially white and of identical variance σ_n^2 at each sensor, the spatial covariance matrix of Equation 2.21 simplifies to

$$\mathbf{R_x} = \sum_{s=1}^{N_s} \sigma_s^2 \mathbf{S}\mathbf{S}^H + \sigma_n^2 \mathbf{I} \tag{2.22}$$

Other considerations

In addition to the plane wave propagation and the narrow band assumptions, there are also some other restrictions that have been implicitly assumed in the derivation of the above theory. In most of the cases these conditions can be assumed without an important impact on the resulting performance of array applications, however in some other situations they can seriously affect the validity of the explained theory. Some of the most important effects that deserve to be mentioned are ideal propagation channel, punctual emitting sources and calibrated and isotropic sensors.

Firstly, the propagating channel has been assumed to be linear and not distorting. In general, real channels will show damaging effects such as dispersion, attenuation, refraction and diffractions. Although the impact of these harmful channel characteristics will not be an object of study, they must be kept in mind to understand why array processing fails in some situations. As an example, a very intuitive case is the impact of lossy media or attenuation: in real environments an array's range is limited, thus, the further the source is located from the array, the more difficult it becomes to extract its radiation from noise.

Secondly, the sources have been assumed to be punctual, that is, the waves impinging the array have been considered to come from a unique point in space. In many situations, the sources are distributed and this model can not be considered.

Finally, the sensors may not be calibrated or be isotropic, that is, different sensors can show different gains and at the same time each sensor can have a direction and frequency dependent

response. In the case that non punctual sources or sensor calibration must be taken into account, their effects can be incorporated to the frequency propagation response or steering vector. In the simplest case of different sensor gain, it can be considered by simply scaling with different weights the corresponding component of the steering vector.

2.2.2 Applications: Beamforming and DOA estimation

The main objective of array processing in classical fields of application such as communications, sonar or radar is the design of spatial filters able to recover enhanced versions of emitted signals and the exploration and localization of these sources.

Beamforming

The spatial filtering operation or beamforming can be written in terms of the snapshot vector of Equation 2.19 and a vector $\mathbf{w}$ that represents the weight applied to each sensor of the array:

$$y(n) = \mathbf{w}^H \mathbf{x_n} = \mathbf{w}^H \mathbf{S} \mathbf{a_n} + \mathbf{w}^H \mathbf{n_n} \tag{2.23}$$

The beamformer response for each angle and for a concrete frequency can be computed as the product of the beamformer weights and the steering vector of every angle at that frequency as follows:

$$|H(f, \varphi)|^2 = |\mathbf{w}^H \mathbf{s}_\varphi|^2 \tag{2.24}$$

As already mentioned, the object of beamforming is to find this set of weights $\mathbf{w}$ that permits enhancing the desired sources, while non-desired components are reduced. There exist a lot of approximations and criteria for the beamformer design, but basically they can be classified into data independent methods or statistically optimum methods.

Data independent methods consist on the design of fixed beamformer patterns independently on the arriving data according to spatial restrictions or other kinds of restrictions.

Statistically optimum beamformers are based on the statistics of the arriving data to optimize some functional that makes the beamformer optimum in some sense under specific conditions. Usually, functionals are based on the minimization of the output power of the beamformer or the output noise power:

$$\mathrm{E}\left\{|y(n)|^2\right\} = \mathbf{w}^H \mathrm{E}\left\{\mathbf{x_n} \mathbf{x_n}^H\right\} \mathbf{w} = \mathbf{w}^H \mathbf{R_x} \mathbf{w} \tag{2.25}$$

The data statistics can be assumed beforehand or estimated. In addition, this kind of optimum beamformers usually incorporate spatial restrictions resulting in conditional optimization problems, which is clearly necessary in the case of minimizing the output beamformer power to avoid cancellation of the desired source.

Furthermore, since data statistics may change, there are adaptive approximations to the beamforming problem.

An excellent overview of the main proposals can be found in (Van Veen & Buckley, 1988). As an example, the derivation of a simple beamformer with spatial restrictions by minimization of a data driven criterion will be shown.

In general, spatial restrictions can be forced to the beamformer by simply imposing a desired response to concrete directions, that is, a concrete value is forced for a given steering vector. Particularly, it is commonly desired a unity response to the position of the desired source and the cancellation of all the other directional interferences. Additionally, it is also possible to add some other restrictions not only related to spatial response. Anyway, the maximum number of restrictions is determined by the number of sensors of the array. Spatial restrictions can be written as follows:

$$\mathbf{w}^H\mathbf{S} = \mathbf{c}^T \qquad \mathbf{w}^H \begin{bmatrix} \mathbf{s_1} & \mathbf{s_2} & \dots & \mathbf{s_{N_s}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix} \tag{2.26}$$

The design of the sample beamformer would be completed demanding minimum response to non-directional additive noise in addition to targeting the desired source. Assuming white noise, that is the noise cross-covariance matrix is $\mathbf{R_n} = \sigma^2\mathbf{I}$, the solution is then given by the beamformer that minimizes its norm at the same time that the spatial restriction is accomplished, which can be obtained by simple conditional minimization and results in $\mathbf{w} = \mathbf{S}(\mathbf{S}^H\mathbf{S})^{-1}\mathbf{c}$.

In the particular case of forcing only the target source condition, the solution is a scaled version of the steering vector of the desired source ($\mathbf{s_d}$) as follows: $\mathbf{w} = \frac{\mathbf{s_d}}{||\mathbf{s_d}||}$. That is, the beamformer is simply a phase delay vector. Under a time domain interpretation, the beamformer would consist in the correct alignment of the different sensor signals according to the expected inter sensor delay and averaging of these aligned signals. This beamformer has several names depending on the field of application. Hereinafter, it will be referred as the delay-and-sum beamformer. It is worth to notice that in the scenario of only existing the desired source and in presence of non-directional white noise this beamformer is optimal.

The delay-and-sum beamformer is a very particular and simple beamformer and has been selected as an example since it will appear continuously in the following sections and chapters. However, other statistically optimum beamformers deserve to be highlighted at this moment.

When the characteristics of the noise cross-covariance matrix are not assumed beforehand,

the general solution that minimizes the noise variance at the beamformer output is known as the linear constraint minimum variance beamformer and is $\mathbf{w} = {\mathbf{R_n}}^{-1}\mathbf{S}(\mathbf{S}^H{\mathbf{R_n}}^{-1}\mathbf{S})^{-1}\mathbf{c}$. Again the particular case of only assuming the existence of non-directional noise and the source of interest results in $\mathbf{w} = {\mathbf{R_n}}^{-1}\mathbf{s_d}({\mathbf{s_d}}^H{\mathbf{R_n}}^{-1}\mathbf{s_d})^{-1}$ and is usually named the minimum variance distortionless response (MVDR) beamformer.

The previous beamformer is useful from a theoretical point of view to design optimum beamformers in presence of concrete noise characteristics, for instance it is very easy to derive the delay-and-sum beamformer assuming white noise. However, when the noise cross-covariance matrix must be estimated from real observed data, this beamformer can become unpractical since it is often difficult to obtain noise alone data. Alternatively, it is in general more practical to estimate the array data cross-covariance matrix and minimize the total power output subject to restrictions. This solution is $\mathbf{w} = {\mathbf{R_x}}^{-1}\mathbf{S}(\mathbf{S}^H{\mathbf{R_x}}^{-1}\mathbf{S})^{-1}\mathbf{c}$ and it is commonly known as the linear constraint minimum variance beamformer. Particularly, when the unique spatial restriction is steering to the desired source, the solution is the minimum power distortionless response (MPDR) beamformer: $\mathbf{w} = {\mathbf{R_x}}^{-1}\mathbf{s_d}({\mathbf{s_d}}^H{\mathbf{R_x}}^{-1}\mathbf{s_d})^{-1}$.

DOA estimation

Source localization and concretely estimation of directions of arrival is the other main field of application of array signal processing. As in the case of beamforming, the development of direction of arrival estimators has been object of extensive research resulting in a countless number of proposals.

For completeness of this section, a brief overview of the main proposals of classical array processing to direction of arrival estimation will be provided without entering in detail. An excellent introductory review of the main contributions can be found in (Krim & Viberg, 1996). Following this work, estimation techniques can be classified into two main categories, namely spectral-based and parametric approaches.

Spectral-based techniques form a depending DOA function based on the received data, thus resulting in a spatio-spectral functional, where the directions of the present sources are obtained as the highest peaks of the functional. These techniques are classified into to broad categories: beamforming techniques and subspace-based methods.

Beamforming techniques are also known as exploration or steered-beamformer techniques. The reason of these names is the way in which the spatio-spectrum functional is obtained. It basically consists on the exploration of all the possible directions with an array and measuring the output power obtained, that is, an estimation of the power across the space is obtained steering a beamformer to the different directions as follows:

$$\hat{\varphi} = \arg\max_{\varphi} \mathbf{w}(\varphi)^H \mathbf{R}_{\mathbf{x}} \mathbf{w}(\varphi) \tag{2.27}$$

The difference between the various proposals lies in the concrete beamformer selected. For instance, the well-known Capon estimator (Capon, 1969) is the particular case of DOA estimation with the MPDR beamformer.

Subspace methods are high-resolution spectral estimators based on the decomposition and exploitation of the properties of the cross-covariance matrix. Usually these algorithms are considered as cancellation or detection methods because, instead of directly estimating the power coming for all the possible positions, they locate sources by finding the directions that are orthogonal to the noise subspace.

The main representing technique of this group is the MUSIC algorithm (sch,). Again assuming that the various sources are narrow band and uncorrelated between them and that the noise is spatially white and of identical variance at each sensor, then the spatial covariance matrix $\mathbf{R}_{\mathbf{x}}$ is the one of Equation 2.22 and it can be decomposed in eigenvalues (λ_q) and eigenvectors ($\mathbf{e}_q$) as follows:

$$\mathbf{R}_{\mathbf{x}} = \sum_{s=1}^{N_s} \sigma_s^2 \mathbf{S}\mathbf{S}^H + \sigma_n^2 \mathbf{I} = \sum_{q=1}^{Q} \lambda_q \mathbf{e}_q \tag{2.28}$$

In this case, the $Q - N_s$ eigenvectors corresponding to the lowest eigenvalues form the noise space orthogonal to the source steering vectors. The MUSIC algorithm finds the DOA of the present sources as the peaks of the inverse projection of the explored steering vectors to the noise space in the following way:

$$\hat{\varphi} = \arg\max_{\varphi} \frac{1}{\sum_{q=N_s+1}^{Q} |\mathbf{s}^H(\varphi)\mathbf{e}_q|^2} \tag{2.29}$$

Finally, more recent approaches to DOA estimation propose a fully parametric estimation of the data model. The methodology requires a statistical framework for the data generation. Then, with some model-based approach such as the maximum likelihood (ML) technique, the various parameters are estimated. Thus, not only the DOA can be estimated, also the original source data or the noise are parameters that can be obtained. These methods are in general computationally quite more complex than the previous ones. It is for that reason that some subspace-based approximations or particular solutions for uniform linear arrays have been also proposed. For a complete review of these methods see (Krim & Viberg, 1996).

Broadband processing

Beamformers and DOA estimation methods presented here, and in general all the exposed array theory, are essentially limited to processing narrow band data. This is a realistic assumption in some fields of processing such as radar or communications, however in other fields like sonar and acoustic applications the received signals are broadband.

The extension of the narrow band methods to broadband scenarios is naturally done by filtering the array data in several sub-bands that could be individually considered narrow band. Then, each of these sub-bands can be processed by a beamformer or a DOA estimator independently, and in this way the broadband problem is decomposed into several narrow band problems. The usual way of filtering the array data is applying a short-time Fourier transform, which permits converting time domain broadband signals into a bank of complex narrow band signals. Each one of these narrow band signals has a central frequency corresponding to the frequency bin of the Fourier transform. This broadband to narrow band process is represented in Figure 2.6 for the particular case of a beamforming application.

The formulation in the frequency domain can be expressed in an equivalent way to previous narrow band theory. Assume N_s broadband sources impinging an array in presence of noise, the signal observed in sensor q at time instant t in the time domain is:

$$x_q(t) = \sum_{s=1}^{N_s} x_s(t - \tau_{sq}) + n(t) \tag{2.30}$$

And in the frequency domain:

$$X_q(t,f) = \sum_{s=1}^{N_s} X_s(t,f) exp(-j2\pi f \tau_{sq}) + N(t,f) \tag{2.31}$$

Again we can define the vector of frequency f observed by the array in terms of the steering vector:

$$\mathbf{X}(t,f) = [X_1(t,f), \ldots, X_q(t,f), \ldots, X_Q(t,f)]^T = \sum_{s=1}^{N_s} X_s(t,f)\mathbf{s_s} + \mathbf{N}(t,f) = \mathbf{S}\mathbf{X_s}(t,f) + \mathbf{N}(t,f) \tag{2.32}$$

And in the same way that the cross-covariance matrix was previously derived, it is also possible to define the cross-covariance matrix for each frequency f or more precisely, the cross-power spectral density matrix for each frequency f:

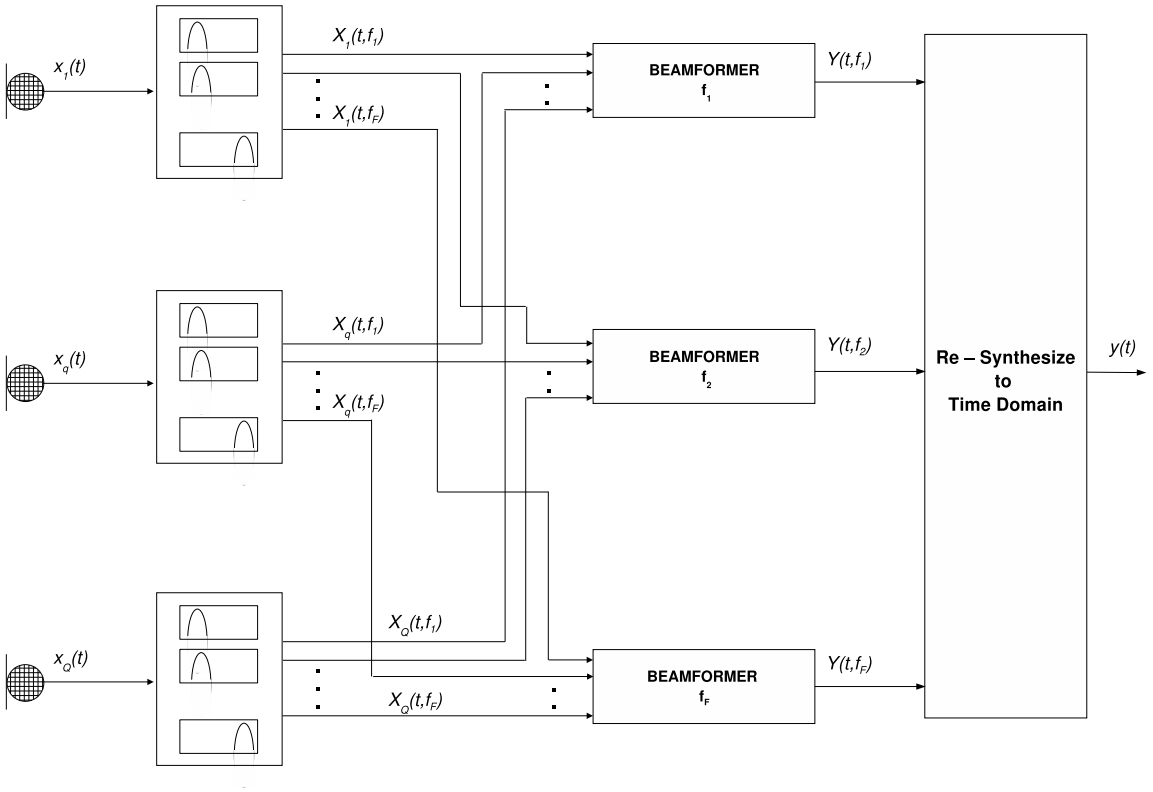

Figure 2.6: *Frequency domain processing of broadband signals. The process is split into 3 stages: firstly, the time domain array data is filtered with short time Fourier transform; next, narrow band beamformer or DOA estimation method is individually applied to the vector formed by all the sensors signal at a concrete frequency; and finally, in the case of beamforming the data is re-synthesized in the time domain or in the case of DOA estimation, multiple DOAs must be combined.*

$$\mathbf{\Phi_x}(f) = \mathrm{E}\left\{\mathbf{X}(t,f)\mathbf{X}(t,f)^H\right\} = \mathbf{S}\mathrm{E}\left\{\mathbf{X_s}(t,f)\mathbf{X_s}(t,f)^H\right\}\mathbf{S}^H + \mathrm{E}\left\{\mathbf{N}(t,f)\mathbf{N}(t,f)^H\right\}$$

$$\mathbf{\Phi_x}(f) = \mathbf{S}\mathbf{\Phi_s}(f)\mathbf{S}^H + \mathbf{\Phi_n}(f) \tag{2.33}$$

In the case of beamforming applications, this frequency domain processing results in a problem of beamforming F sub-bands to later re-synthesize them into the time domain by means of inverse filtering, as it is shown in Figure 2.6. In general, the filters are estimated independently for each frequency, for instance applying one of the beamformers described previously:

$$Y(t,f) = \mathbf{W}(\mathbf{f})^H\mathbf{X}(\mathbf{t},\mathbf{f}) \tag{2.34}$$

The frequency domain beamforming can lead to problems of circular convolution that can be avoided with a convenient selection of the processing frame size, the frame shift and the length of the Fourier transform.

Regarding the DOA estimation, one narrow band method is applied to each frequency band. In this case, it is not necessary to re-synthesize data to the time domain since the object is just to estimate DOA. The problem associated to this frequency domain DOA estimation process is that a different estimation is obtained for each frequency bin.

2.2.3 Some microphone array particularities

Throughout this section, the array signal processing theory has been described keeping generality with independence on the field of application. Although the state of the art in the fields of speech beamforming and audio source localization is provided in next chapters 3 and 5 respectively, a brief insight in some of the particular characteristics of microphone arrays is required at this point to understand what makes microphone array processing a very specific signal processing task. More reading about microphone array particularities can be found in (Van Compernolle & Van Gerven, 1995; Brandstein & Ward, 2001; McCowan, 2001; Sanchez-Bote, 2004).

In addition to the common problems derived from the assumed model such as sensibility to steering errors, difficult calibration, different sensors radiation diagrams and, of course, noise and non ideal channel propagation, the speech signal shows some particularities that makes the use of microphone arrays for speech applications a more difficult task than in other traditional fields of application. Some of these harmful characteristics are for instance the low-stationarity of the speech signal and its burst nature that can affect the design of data-driven beamformers and source localization algorithms. Furthermore, microphone array applications are usually developed in environments where there may be more power in the reverberant signal than in the direct one, and the interferences and in general the noise can vary a lot, even some of them can be non-stationary and have similar spectral characteristics and angles of arrival to the ones of the desired signal.

However the main drawback of the application of microphone arrays is due the fact that the two basic assumptions, that is the narrow band and the far-field assumption, are not true for the speech signal in most of the cases. Thus, the design and development of beamformers for speech is strongly conditioned. To better understand how these two problems affect, a delay-and-sum beamformer will be again considered as an example.

Hence, consider a microphone linear array of 5 microphones with 5 cm of separation between each sensor. The desired source is assumed to be located in front of the array and there can exist other directional and non-directional undesired sources. Assume at this moment that the waves of the desired source are arriving exactly at the same time to each microphone of the array, that

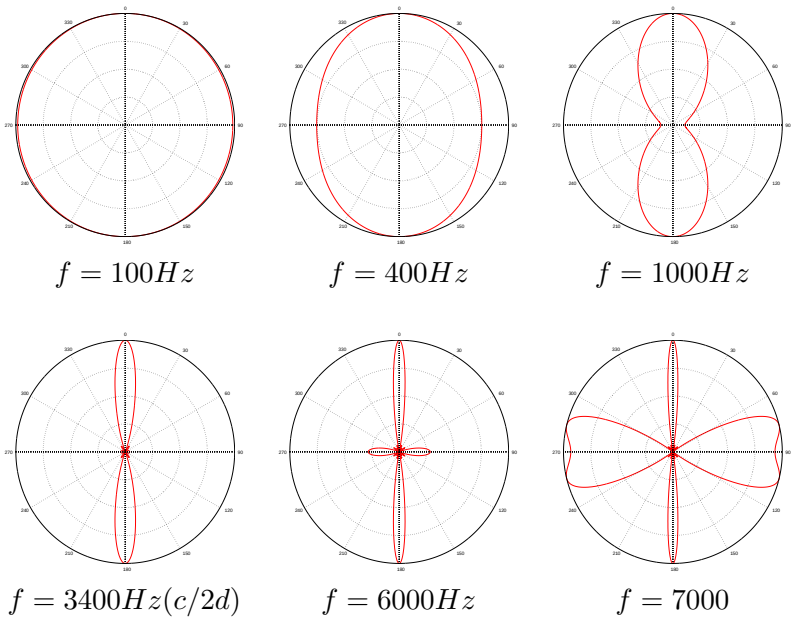

Figure 2.7: *Energy gain response to different frequencies of a delay-and-sum beamformer steering 0^{o} with a microphone array of 5 sensors separated 5 cm. Low spatial resolution can be observed below $f = (c/2d)$, while undesired grating lobes appear at higher frequencies.*

is plane wave propagation is accepted. A delay-and-sum would simply consist in the average of the microphone signals in order to steer the front direction, since they are already aligned.

Figure 2.7 shows the response (see Equation 2.24) of the sample beamformer to different frequencies. From these energy gain diagrams, it is firstly clear that for a concrete frequency the array response to non steered directions is not constant and that the attenuation of an undesired directional source varies depending on its direction of arrival. Secondly, the array response depends on the frequency of the impinging waves. On the one hand, the beamformer becomes less selective, for frequencies below the one corresponding to wavelength equal to twice the distance between sensors (in this case $f < 3400Hz$), resulting almost useless at very low frequencies. On the other hand, frequencies above $3400Hz$ are affected by spatial aliasing and grating lobes appear enhancing undesired directions.

As the speech signal is a broadband signal the application of narrow band beamforming solutions as the one of the example results in poor performance at low frequencies, possible spatial aliasing effects at the high frequency range and frequency distortion of undesired directional sources due to different response to a concrete direction depending on the frequency.

Regarding the far-field condition, typical distances in speech applications from speakers to the array in a meeting room can go from less than 1 meter to 5 or 6 meters. In most of these situations the arriving wavefronts can not be considered plane.

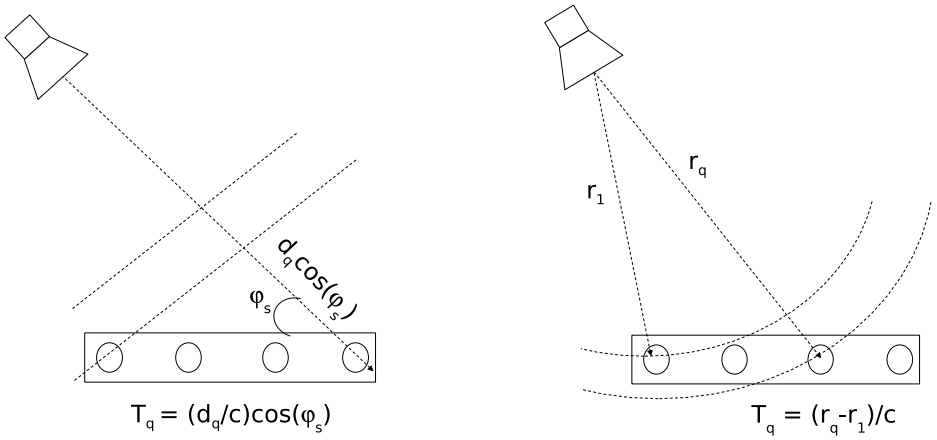

Figure 2.8: *The time delay of arrival of the waves impinging a microphone array depends on the wave propagation model. On the left, plane wave propagation model resulting in time delays dependent on the direction of arrival of the source and the inter sensor separation. On the right, spherical wave propagation model resulting in time delays dependent on the position of the source and the microphones.*

Figure 2.8 shows a comparison of plane wave propagation versus spherical propagation. As it has been already pointed out, in the case of plane wave propagation the frequency response or steering vector of the waves arriving to an array can be expressed in terms of the inter-senor separation and the angle of a arrival of the source.

In the case of spherical propagation, the delays of arrival to each microphone depends on the different distances traveled by the wave to each microphone. Furthermore, the near-field model can be completed taking into account attenuation due to different propagation to each microphone. In this way, considering the first microphone as the reference one, Equation 2.17 can now be rewritten for the near-field scenario as follows:

$$\mathbf{s_s} = [1, g_2 \exp(-j2\pi f_o \tau_2), \ldots, g_q \exp(-j2\pi f_o \tau_q), \ldots, g_Q \exp(-j2\pi f_o \tau_Q)]^T \tag{2.35}$$

where $\tau_q = \frac{r_q - r_1}{c}$, and g_q is the attenuation factor with respect to the first microphone due to spherical propagation.

Additionally, some details and relations about the number of microphones and the distance between them deserve to be mentioned to further show the complexity of microphone array design.

In Figure 2.9 is shown the response to different low frequencies ($400Hz$ and $1000Hz$) with an array of two, five and ten microphones with 5 cm of separation. The increase of the number of microphones results in narrower beams, thus improving performance in low frequency regions, in exchange of more complicated gain patterns and a larger array size, that can affect to the

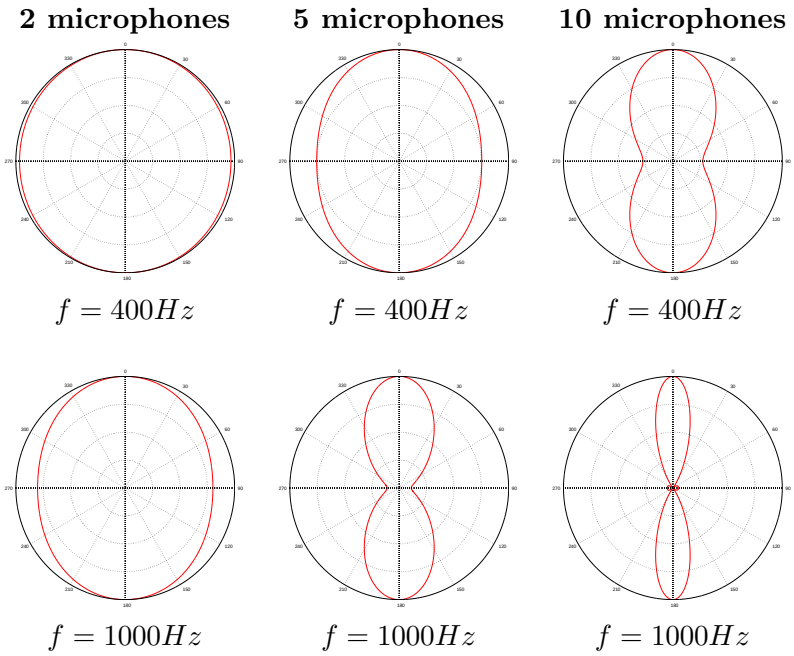

Figure 2.9: *Energy gain response to* $400Hz$ *and* $1000Hz$ *of a delay-and-sum beamformer steering* $0°$ *with microphone arrays of 2, 5 and 10 sensors separated 5 cm. Increasing the number of microphones results in narrower beams.*

plane wave propagation assumption in the case it was considered.

Concerning the problem of spatial aliasing at higher frequencies, it can be solved as already mentioned if $d \leq \lambda_{min}/2$. For instance, to avoid spatial aliasing in a range of $0 - 8kHz$ it is necessary an inter-sensor separation of about 2 cm, resulting in a very poor response to low frequencies for typical arrays of not more than 10 microphones. Figure 2.10 shows the response to low frequency ($400Hz$) and to high frequency ($7000Hz$) of the array of five microphones with 2 , 5 and 10 cm of separation, thus it is clearly seen the coupling effect of low spatial resolution at low frequencies and spatial aliasing at high frequencies.

To conclude this section, it is interesting to get a brief insight on the possible theoretical effects of reverberation in microphone array applications. The reverberation model of Section 2.1 distinguishes between early and late reflections. Ideally the latter can be assumed to have an equivalent impact on beamforming and localization to the one of a diffuse noise field. The former are well-defined non-direct paths from source to microphones and from the point of view of microphone array processing can be considered as directional undesired interferences arriving from different directions. In this section it has been shown how a classical beamformer as the delay-and-sum responds to directional interferent sources depending on their direction of arrival. For localization algorithms the reflections appear as competitors. The additional problem that introduces reverberation is that these interferent reflections are highly correlated with the source of interest. In many DOA estimation approaches the problem of coherent interferences results

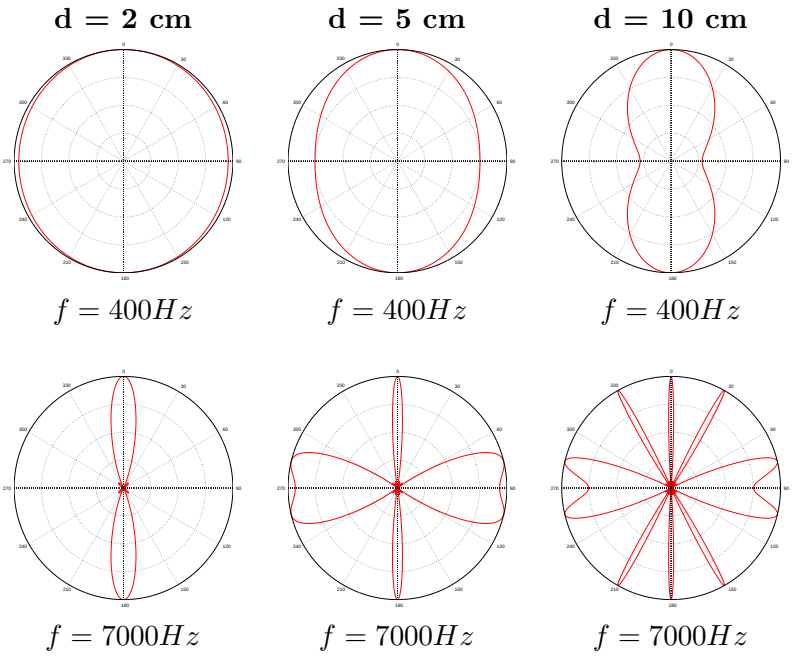

Figure 2.10: *Energy gain response to* $400Hz$ *and* $7000Hz$ *of a delay-and-sum beamformer steering* $0°$ *with microphone arrays of 5 sensors separated 2, 5 and 10 cm. Increasing the microphone separation results in narrower beams at low frequencies and the appearance of undesired gratings lobes at high frequencies.*

in poor performance of the algorithms. Considering beamforming, data-driven beamformers and particularly adaptive beamformers can be dramatically affected by directional coherent interferences, in some cases resulting in cancellation of the source of interest.

2.3 Alternative multi-microphone approaches

Alternatively to multi-microphone processing based on the array signal processing framework, other approaches to speech enhancement, source separation and localization can be found in the speech related literature. Some of the most outstanding are blind source separation techniques, multi-channel dereverberation techniques and binaural processing techniques.

2.3.1 Blind source separation

The blind source separation (BSS) task consists on the estimation of a demixing matrix, that permits the separation of the various sources present in multiple microphone signals without any a priori information about the sensors placement or the sources characteristics. This is accomplished by means of independent component analysis (ICA) techniques. In few words, ICA techniques generally assume that the number of sources is known and that it is equal

or less than the number of microphone channels. Considering that all the source signals are statistically independent, then the demixing matrix can be iteratively estimated based on the maximization of an appropriate independence criteria measuring the degree of independency of the demixed output signals. Depending on the concrete approach, the independence criteria varies including second order statistics, higher order statistics and information theory based measures.

While simple ICA methods show impressive results when trying to separate instant mixtures, that is linear additive combinations of several sources, their performance in more realistic convolutive mixtures is in general unsatisfactory. A possible approach to face the convolutive mixture problem based on ICA techniques consists in the estimation in the frequency domain of different demixing matrices for each frequency bin independently. Thus, a time domain convolutive problem is converted into several frequency domain instant mixture problems. However, one general limitation of ICA techniques is the problem of permutation, that is, there is no certainty about which demixed output corresponds to which source signal. This problem when processing several frequency bins means that there is not guarantee that the separated frequency components at a particular output will always correspond to the same source.

Blind source separation based on independent component analysis with application to the speech separation problem has actually become one of the most appealing and active research fields in the speech processing community. As a consequence, many efforts have been recently devoted to solve their current limitations and new promising theoretical advances and approaches are continuously appearing. One can find detailed description on BSS and ICA principles and methods in (Cardoso, 1998; Hyvärinen & Oja, 2000; Hyvärinen et al., 2001).

2.3.2 Multi-channel dereverberation techniques

It has been already commented that reverberation is a significant cause of loss of quality and intelligibility of speech in far-field applications. As a consequence, research and development of algorithms aimed to reduce the effect of reverberation, generally named dereverberation techniques, has been a main topic of interest among the speech research community.

Perhaps the most straightforward idea for combating reverberation is equalization of the speaker-to-receiver impulse response. If the acoustic impulse response is known (from calculations or measurements), reverberation can be removed by using the inverse filter. Unfortunately, inverting an acoustic impulse response is not readily possible. Approximations to the inverse filter can be constructed , however some of these equalizer approaches spread speaker-to-receiver impulse response energy along the equalizer time span, which actually results in more reverberant speech signals. When considering multiple channels, it is shown in (Miyoshi & Kaneda, 1988) that the exact inverse is possible to obtain if the transfer functions have no common zeros.

However, sensitive to the accuracy of the measured/estimated acoustic impulse responses and concerns about the numerical stability render in not practical solutions.

Another possible approach is to face reverberation as a matched filtering problem (Flanagan et al., 1993). In this way, the microphone signals are filtered with the time-reversed acoustic impulse responses. In the case of multiple microphones the matched filters are used in a kind of filter-and-sum beamformer. This technique is less sensitive to the accuracy of the impulse responses, but no perfect dereverberation can be obtained and, although being theoretically superior to classical delay-and-sum beamformer, it has not been shown to provide significant improvements in some outstanding speech applications such as ASR.

In general, practicality of these inverse filter or matched filtering approaches is further complicated by the difficulty of estimating and tracking the acoustic impulse response in real-time applications.

2.3.3 Binaural processing

A human subject localizes, recognizes and focuses different sound sources based on binaural signals, that is, the signals arriving to their two ears. Consequently, the imitation of these abilities on the basis of only two observed signals has traditionally been an important field of interest of the research community.

Based on binaural signals, one can find audio source localization techniques. Most of binaural localization approaches rely in two fundamental cues that are the interaural level differences (ILD) and the interaural time differences (ITD), that is, the amplitude level and time differences between the two observed signals. Also based on binaural signals, the separation of overlapping sound sources has been extensively investigated. A complete review on binaural processing can be found in (Viste, 2004).

Chapter 3

Speech Enhancement and Recognition with Microphone Arrays

This chapter shows the main contributions in the field of microphone array processing for speech enhancement with the intention of providing a significant review of the current state of the art. Also, automatic speech recognition is introduced and the most significant works on speech recognition with microphone arrays are commented.

By means of beamforming is possible to spatially filter signals arriving to a microphone array enhancing concrete desired directions while others are rejected. As a result of this multi-microphone processing, enhanced signals are generally less reverberated and both background and directional noises are reduced compared to single microphone recordings. However, the particular characteristics of microphone array processing prompt speech beamforming to be still a challenging topic of research. In the recent literature one can find many approximations to speech enhancement. In general, most of them use a microphone array as a pre-processing step to capture signals that can be used for any speech application and particularly for speech recognition.

Next section highlights the most significant proposals in speech enhancement with microphone arrays and details some of the most relevant beamformers used in speech applications. The different works in this field are aimed to obtain practical solutions and in some cases are particularly designed to solve some of the drawbacks of speech beamforming, for instance proposing solutions to deal with the broad band nature of speech signals or to avoid unrealistic assumptions about the far-field wave propagation model.

Last section presents state of the art Hidden Markov Model based speech recognition. An introductory overview to feature extraction, acoustic modelling and some robustness issues is

provided. Then, the most relevant recent works in the field of automatic speech recognition with microphone arrays are higlighted.

3.1 Microphone array processing for speech enhancement

Microphone array research has been addressed by many different approaches, but must of them can be basically summarized into two major trends: fixed and adaptive beamforming. Moreover, there is an extensive literature related to microphone array post-filtering and in general to the application or generalization of mono-channel techniques to the multi-channel case.

3.1.1 Fixed beamforming

Most conventional approaches consist of fixed beamformers as the delay-and-sum (Johnson & Dudgeon, 1993), filter-and-sum (Johnson & Dudgeon, 1993) or the super directive beamformer (SDB) (Cox et al., 1986). The common characteristic of the fixed approaches is that the array processing parameters do not change dynamically over time.

Delay-and-sum beamformer

In a time-domain implementation as the one shown in Figure 3.1, the delay-and-sum beamformer (Johnson & Dudgeon, 1993) basically consists on the alignment of the different microphone signals to compensate for the different path lengths from the source to the various microphones, and the combination of these aligned signals together. It can be expressed mathematically as follows:

$$y(n) = \sum_{q=1}^{Q} \alpha_q x_q(n - \tau_q) \tag{3.1}$$

where α_q is the weight given to each different microphone and τ_q is the delay that compensates the different propagation delays. Usually, the weight α_q is equal to $1/Q$ resulting in the average of the aligned signals, however it is possible to select other criteria for microphone weight for instance depending on the propagation model (see far-field and near-field discussion in 2.2.3), compensation of different sensor gain or even different signal to noise ratio. Obtaining τ_q is a problem of time delay estimation or more generally of speaker localization that will be tackled in Chapter 5. The simplicity of the delay-and-sum beamformer is the most important strength, resulting in many cases a convenient and practical choice for many microphone array applications. Thus, delay-and-sum is widely used despite its frequency depending response, the

impossibility of reducing highly directive noise sources and all the limitations already commented in the previous chapter.

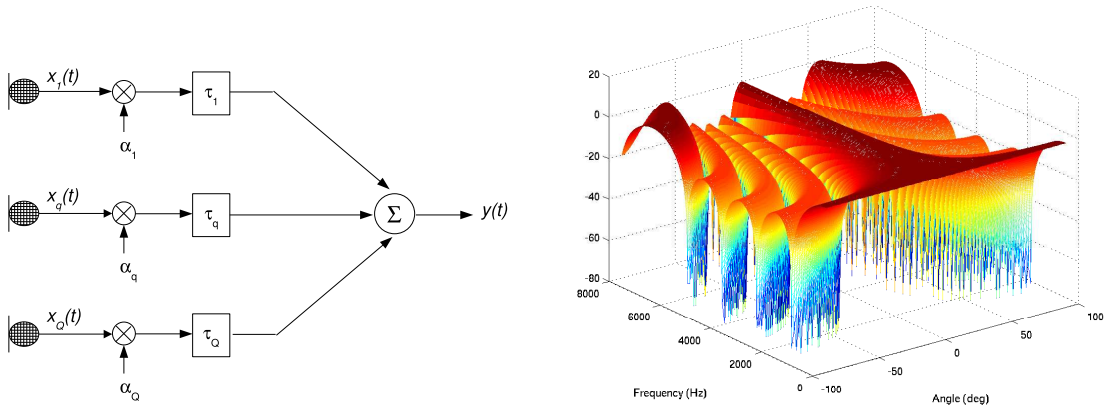

Figure 3.1: *On the left, delay-and-sum beamformer in a time domain implementation. On the right, frequency and spatial response of delay-and-sum with 5 microphones separated 5 cm. Low spatial resolution is observed at the low frequency range, while spatial aliasing appears at the high frequency range.*

Particularly interesting are those proposals devoted to reduce the frequency dependent response of the delay-and-sum beamformer, but that can eventually be applied to any other beamformer structure. The basic idea is to design a beamformer such that the spatial response is the same over a wide frequency band. A possible simple approximation for speech applications consists of processing different frequency bands in different harmonically nested sub-arrays (Kellermann, 1991; Flanagan et al., 1991; Sanchez-Bote et al., 2003) with different inter-microphone separations, as the one shown in Figure 3.2. 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.

Filter-and-sum beamformer

A generalization of the delay-and-sum beamformer is known as the filter-and-sum beamformer (Johnson & Dudgeon, 1993). In a time domain implementation, instead of simply combining aligned and weighted signals, each microphone signal is filtered with an associated filter dependent on the channel before being combined. This beamformer can be expressed as follows:

$$y(n) = \sum_{q=1}^{Q} h_q(n) * x_q(n - \tau_q) = \sum_{q=1}^{Q} \sum_{l=0}^{L-1} h_q(l) x_q(n - l - \tau_q) \tag{3.2}$$

where $*$ denotes convolution operation as developed in the right side of the identity and

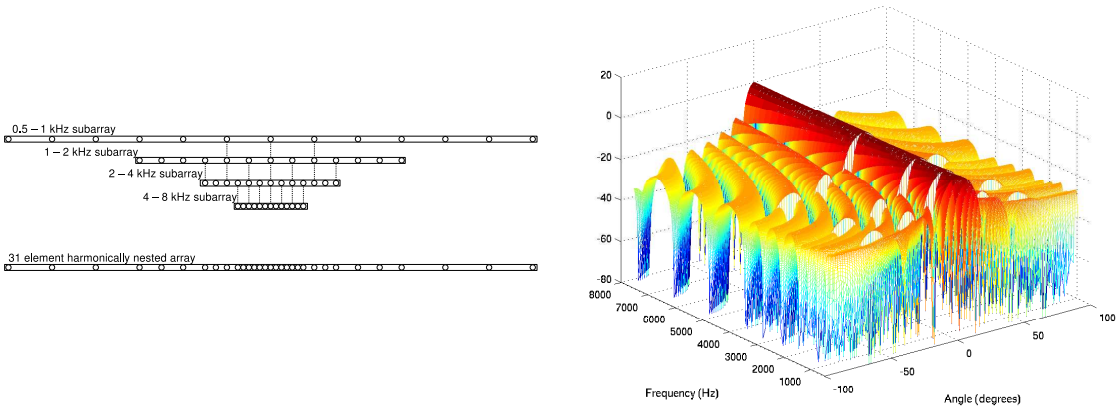

Figure 3.2: *On the left, example of an harmonically nested array for four different octaves: the minimum spacing of the microphones is 2cm in the case of the highest frequency dedicated sub-array. On the right, frequency and spatial response of the nested array. A clear reduction of the beampattern variation with respect to the delay-and-sum beamformer is obtained.*

$h_q[n]$ is the filter associated to microphone q of length L. In general, a filter-and-sum approach would permit more sophisticated and specific array responses, that are not possible with simple delay-and-sum approaches, for instance enhancing or rejecting particular directions in addition to the desired one.

While designing these filters is not always easy and computationally efficient in the time domain, it was already shown in Section 2.2.2 how a broad band signal like speech can be beamformed in the frequency domain: the speech signal is decomposed in several frequency bands, a beamformer is designed and applied to each band and the resulting beamformed frequency bands are re-synthesized. Considering $\mathbf{X}(t,f) = [X_1(t,f)...X_Q(t,f)]^T$ the vector formed by the microphone signals at frequency f and time instant t, and $\mathbf{W}(f)$ the beamformer weights vector for that frequency, then the output $Y(t,f)$ at this frequency f is given by:

$$Y(t,f) = \sum_{q=1}^{Q} W_q(f) X_q(t,f) = \mathbf{W}(f)^H \mathbf{X}(t,f) \tag{3.3}$$

The interpretation of the filter-and sum beamformer as a frequency domain processing, permits applying spatial restrictions and different optimization criteria, as shown in Section 2.2.2.

Again, among the filter-and-sum literature for speech applications, the approaches devoted to reduce the frequency variant response are of particular interest. These are generally known as frequency invariant beamformers. Frequency invariant beamforming aims to parameterize array filter coefficients such that the spectral and spatial response profiles of the array can be adjusted independently. Solutions to this problem have been presented for far-field (Ward et al.,

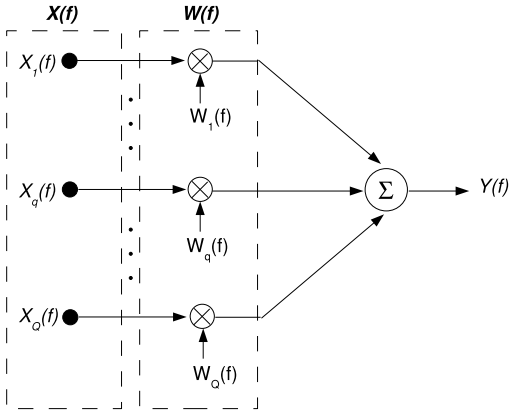

Figure 3.3: *Filter and sum beamforming in the frequency domain of frequency bin f.*

1995) and near-field (Abhayapala et al., 2000). They develop frequency invariance property for theoretical continuous sensor, and then they approximate this sensor by an array of discrete sensors. Unfortunately, to assure frequency invariance for low speech frequencies, say 300 Hz , the method may require an array of several meters long, resulting in a practical solution only for the mid and high frequency range. A recent appealing approach based on least-squares inversion has been proposed for arbitrary sensor configurations of few microphones in (Parra, 2006).

Super directive beamformer

The super directive beamformer (Cox et al., 1986) is a particular case of filter-and-sum beamformer designed to maximize the directivity in the direction of the speech source for a diffuse (spherically isotropic) noise field. Background noise in room environments is usually well approximated by a diffuse noise field, thus the super directive beamformer is a common choice in microphone array applications. The super directive beamformer in the frequency domain can be derived from the minimum variance distortionless response beamformer (MVDR) shown in the last section for the particular case of diffuse noise field as follows:

$$\mathbf{W(f)} = \mathbf{\Gamma}^{-1}(f)\mathbf{s_d}(\mathbf{s_d}^H\mathbf{\Gamma}^{-1}(f)\mathbf{s_d})^{-1} \tag{3.4}$$

$\mathbf{\Gamma}(f)$ represents the cross-coherence matrix at frequency f of the diffuse noise between sensors, and $\mathbf{s_d}$ is the steering vector of the desired direction for that frequency. The value of each $\Gamma_{ij}(f)$ depends on the distance between microphones i and j and the frequency f as shown in Equation 2.2.

The above solution is known to lead to undesirable gain of incoherent noise due to sensitive to microphone and channel mismatch, particularly at low frequencies. Therefore, an additional

constraint on the white noise gain is generally added in order to obtain a more robust solution, yielding

$$\mathbf{W}(f) = (\mathbf{\Gamma}(f) + \epsilon\mathbf{I})^{-1}\mathbf{s_d}(\mathbf{s_d}^H(\mathbf{\Gamma}(f) + \epsilon\mathbf{I})^{-1}\mathbf{s_d})^{-1} \tag{3.5}$$

An small amount is added to each diagonal component of $\mathbf{\Gamma}(f)$ solving the problem of incoherent noise amplification.

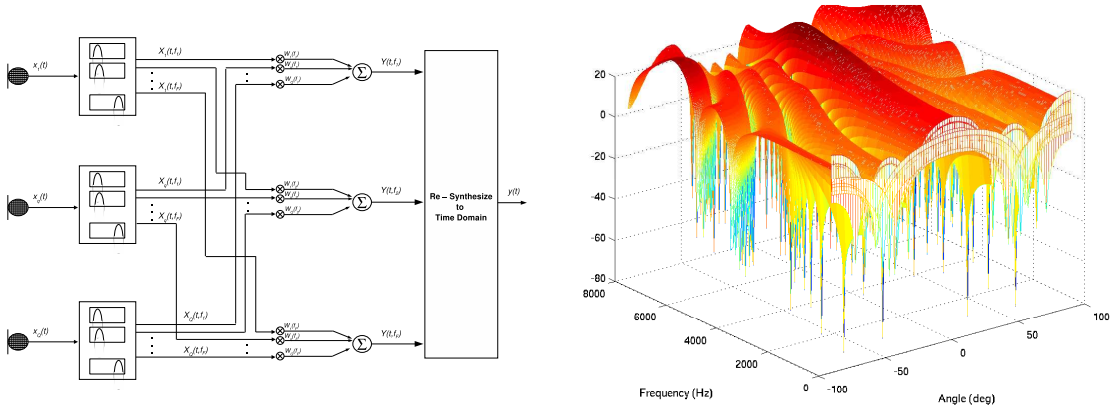

Figure 3.4: *On the left, super directive beamformer in frequency domain implementation. On the right, frequency and spatial response of super directive beamformer with 5 microphones separated 5 cm. Compared to delay-and-sum beamformer, a more constant response is obtained for all frequencies.*

As commented super directive beamformer is widely used in microphone array applications and several works for instance proposing near-field solutions (Täger, 1998) or applying it to hearing aids (Kates, 1993) applications can be found. Further detail can be found in some nice reviews of super directive microphone array applications in (Bitzer & Simmer, 2001) and (Elko, 2000).

3.1.2 Adaptive beamforming

The noise present in real environments changes both spatially and spectrally along time. Thus, data-driven beamforming methods (see Section 2.2.2) must deal with adaptive estimation of noise characteristics. In fact, most adaptive algorithms instead of explicitly estimating the noise characteristics indeed, compute the beamformer weights in a manner that the actual noise characteristics are taken into account.

The well-known Frost Algorithm (Frost, 1972) is an adaptive solution to the MPDR beamformer. By means of a constrained LMS algorithm (Haykin, 2001), the filter weights applied to each sensor are adaptively adjusted to minimize the output power of the beamformer at the same time that maintains a constant response at the direction of the target source.

The Generalized Sidelobe Canceller (GSC) (Griffiths & Jim, 1982) is an alternative elegant solution to the same problem tackled by Frost. The GSC beamformer converts the constrained minimization problem in an unconstrained minimization problem thanks to a two path structure processing.

In general, adaptive beamforming compared to fixed beamforming permits higher capability of noise reduction, particularly of a priori unknown directional noises, thanks to its ability to adapt to the actual noise scenario. However, as a consequence of the minimization process, adaptive beamforming algorithms are very sensitive to steering errors and might suffer from signal leakage and degradation. As a result, conventional adaptive filtering approaches have not gained widespread acceptance for speech applications, and the development of robust modifications to avoid signal leakage and cancellation has been an important matter of interest in microphone array applications.

Alternatively to the GSC based approaches, other adaptive beamforming solutions have been also proposed for microphone array processing. In (Masgrau et al., 1999), several adaptive schemes are compared for different microphone array configurations in presence of directional noise, ambient noise and reverberation.

Generalized Sidelobe Canceller

Most common and successful approaches in microphone array applications are based on modifications of the well-known Generalized Sidelobe Canceller (GSC). A GSC beamformer basically consists of a fixed and an adaptive path. The adaptive path tries to estimate the non-desired components through a spatial blocking matrix that blocks target signal direction and allows all the other directions. These non-desired components are used for adaptively reducing the correlated components of the output of the fixed beamformer in order to obtain an enhanced output. This reduction noise stage can be done with a simple unconstrained LMS algorithm, since the desired signal is assumed to be blocked by the blocking matrix. A diagram of the GSC beamformer is shown in Figure 3.5.

In the simplest case the fixed beamformer of the GSC is a delay-and-sum beamformer. If we assume that microphone signals have been time aligned the output of the fixed beamformer (y_q) can be written in terms of the delay-and-sum weights and the captured snapshot as follows:

$$y_q(n) = \mathbf{w}^T \mathbf{x}(n) \qquad \mathbf{w} = \begin{bmatrix} 1/Q & 1/Q & \cdots & 1/Q \end{bmatrix}^T \tag{3.6}$$

In this case the blocking stage is achieved by simply subtracting pairs of sensors. Then, the blocking matrix and the output of the blocking matrix is:

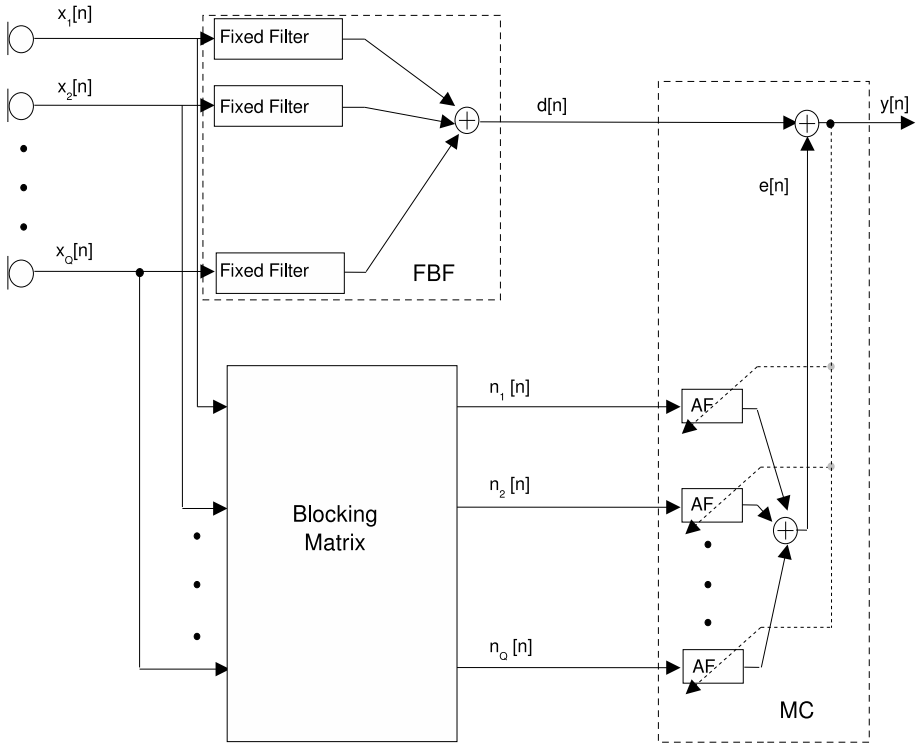

Figure 3.5: *Example of a GSC-like beamformer with its two path structure: the fixed beamformer (FBF) path and the adaptive path, which is composed by a blocking matrix (BM) and an adaptive multiple-input canceller (MC).*

$$\mathbf{x_o}(n) = \mathbf{B}\mathbf{x}(n) \qquad \mathbf{B} = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & \dots & \ddots & \ddots & \dots & \vdots \\ 0 & \cdots & 0 & 1 & -1 & 0 \\ 0 & \cdots & 0 & 0 & 1 & -1 \end{bmatrix} \tag{3.7}$$

The output of the adaptive path can now be written in terms of $\mathbf{x_o}$ and an adaptive filter $\mathbf{a}$

$$y_a(n) = \mathbf{a}^T\mathbf{x_o}(n) \tag{3.8}$$

and the total output of the GSC beamformer as:

$$y(n) = y_q(n) - y_a(n) = \mathbf{w}^T\mathbf{x}(n) - \mathbf{a}^T\mathbf{x_o}(n) \tag{3.9}$$

Since the output of the delay-and-sum contains the desired source (in addition to residual noise and interference terms) and the output of the blocking matrix only contains noise and interference terms, finding the filter $\mathbf{a}$ which minimizes the output power in $y(n)$ is equivalent to the MPDR beamformer and can be obtained by means of an unconstrained LMS algorithm:

$$\mathbf{a}_{n+1} = \mathbf{a}_n + \mu \mathbf{y}(n)\mathbf{x_o}(n) \tag{3.10}$$

where μ is the step size.

It is worth noting that the adaptive path of this structure is able to reduce only coherent noise. In other words, in presence of completely incoherent noise the adaptive path will not improve the performance of the fixed beamformer. It is for that reason, that GSC beamformers are particularly useful for rejection of unknown directional interferences.

As commented, steering errors due for instance to inaccurate propagation model, misadjustment in the microphone positions, and wrong estimated or assumed source position, causes signal leakage into the blocking matrix output which results in target signal cancellation at the microphone array output. Additionally, source signal reflections and particularly in microphone array processing the reverberation is an important reason of signal leakage, since it can be interpreted as interference signals correlated with the target source and arriving from different directions. A variety of techniques to increase robustness of GSC and make it practical exist, some of them devoted to improve the blocking matrix stage (Claesson & Nordholm, 1992; Hoshuyama et al., 1999) and others to restraint excess growth of adaptive coefficients (Cox et al., 1987; Claesson & Nordholm, 1992). A good overview of robust adaptive beamforming based on GSC can be found in (Hoshuyama & Sugiyama, 2001).

In practice, in many works based on GSC beamformers with microphone arrays, the filter adaptation of the noise cancellation stage is done in an intermittent fashion to avoid signal cancellation (Van Compernolle, 1990). In this way, adaptive filters are only updated when there is only presence of noise and the desired speaker is not active. As a consequence, some efforts must be addressed to develop a reliable speaker activity detector.

In addition to works related to reduce the signal leakage, other relevant proposals for speech signals can be found, for instance in (Gannot et al., 2001) GSC is generalized for the case of arbitrary transfer functions instead of assuming a delay-only propagation channel. In (McCowan et al., 2002) explicit model for the near-field is introduced.

3.1.3 Post-filtering techniques

In many speech applications conventional beamforming approaches as the delay-and-sum or the filter-and-sum, exhibit an insufficient improvement in terms of noise reduction. In order to obtain better free of noise speech signals, one method that increases the performance of beamformers is to add a post-processing stage at the beamformer output.

Wiener filtering for microphone arrays

In (Zelinski, 1988), it is proposed an adaptive Wiener post-filter with delay-and-sum beamformer as the one shown in Figure 3.6. It is shown that incorporating a post-filter with the beamformer allows use of knowledge obtained with spatial filtering to also allow effective frequency filtering of the signal, resulting in a both spatial and frequency enhancement.

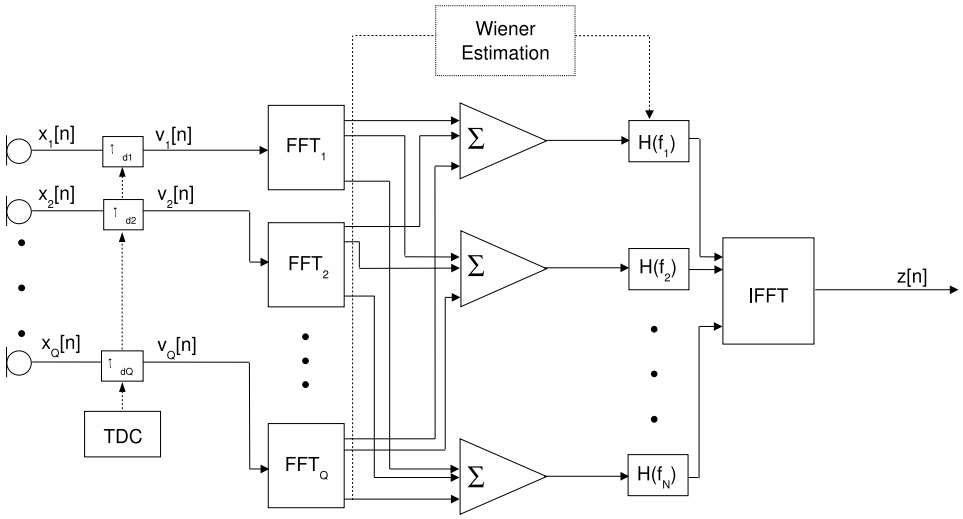

Figure 3.6: *Multichannel Wiener post-filter of a delay-and-sum beamformer. The received speech signals are previously time-aligned by a time delay compensation (TDC) module.*

The general Wiener post-filter is formulated in terms of the cross-spectral densities of noise at the beamformer output and the desired source as:

$$H(f) = \frac{\Phi_{ss}(f)}{\Phi_{ss}(f) + \Phi_{nn}(f)} \tag{3.11}$$

The Wiener filter can be estimated thanks to the availability of multiple inputs that permits computing the power spectral density of the target signal and the one of the noise combining the cross-power spectral densities and the power spectral density of the different microphones of the array. Assuming that the received signal is an additive mixture of the desired signal and noise, that they are uncorrelated and that noise is uncorrelated also between microphones and have an equal power spectral density, then:

$$\Phi_{v_i v_j}(f) = \Phi_{ss}(f) \qquad \Phi_{v_i v_i}(f) = \Phi_{ss}(f) + \Phi_{nn}(f) \tag{3.12}$$

In this case, the Wiener filter equation can be estimated by averaging as

$$H(f) = \frac{\frac{2}{Q(Q-1)} \Re \left\{ \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \hat{\Phi}_{v_i v_j}(f) \right\}}{\frac{1}{Q} \sum_{i=1}^{Q} \hat{\Phi}_{v_i v_i}(f)} \tag{3.13}$$

It is clear that given the above assumptions the post-filter is particularly convenient in presence of spatially white noise, however it is also useful in diffuse noise fields which reasonably approximate these conditions. In fact, the Zelinski post-filter was deeply studied in (Marro et al., 1998) for the general case of a filter-and-sum beamformer and was studied for different non-ideal conditions in terms of beamformer characteristics, such as noise reduction and array gain. In this work, it was shown that the post-filter is effectively able to cancel any incoherent noise, that the rejection to coherent noise correlated and uncorrelated with the desired signal is improved if they are not arriving from the same direction and that it is robust to minor steering errors. The general expression of the Wiener post-filter for any beamformer is:

$$H(f) = \frac{\sum_{i=1}^{Q} |w_i(f)|^2}{\sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} w_i(f) w_j^*(f)} \frac{\Re \left\{ \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \hat{\Phi}_{v_i v_j}(f) \right\}}{\sum_{i=1}^{Q} \hat{\Phi}_{v_i v_i}(f)} \tag{3.14}$$

The Zelinski post-filter has been extensively used in microphone array works, for instance as part of a GSC-like beamformer (Fischer & Simmer, 1995) or in combination with a speech dereverberation technique based in the separate processing of the minimum-phase and all-pass components of the input speech signal (Gonzalez-Rodriguez et al., 2000).

3.2 Overview of Automatic Speech Recognition

Automatic speech recognition stands out as one of the most interesting, practical and challenging applications of speech processing. An automatic speech recognizer is basically a pattern classification system with different sounds or sequences of sounds, such as phonemes, groups of phonemes or even words, modelled as distinct acoustic classes. The object of the speech recognizer is to provide the hypothesized chain of classes that better match the input speech utterance. Obtaining the best sequence of classes means obtaining the best sequence of sounds, and thus, an hypothesized text transcription.

This thesis deals with the use of microphone arrays for different speech applications, consequently a fundamental field of application consists of processing the signals received by an array of microphones to be used as the input of a speech recognition system. In fact, this is exactly the most common way of incorporating microphone arrays to speech recognition, that is, beamforming is used as a pre-processing speech enhancement algorithm applied to the received signals to enhance them previous to being recognized. Alternatively to the use of microphone

arrays as a pre-processing speech enhancement stage, there are some recent works specifically aimed at improving speech recognition accuracy.

In order to understand the works carried out in microphone array processing for speech recognition and the work of the next chapter, the manner in which the recognizers operate must be reviewed. The process involves different levels of knowledge such as physical, acoustic and language information, which interact between them. This multiple knowledge level processing is still not well understood and this is the main reason for automatic speech recognition becoming a complex and still not solved problem. However, a practical and feasible way of tackling the problem is shown in Figure 3.7.

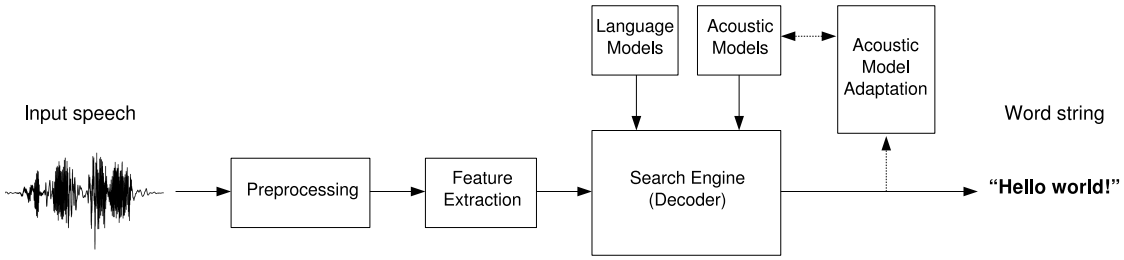

Figure 3.7: *Schematic structure of the basic steps of an automatic speech recognition system.*

First, the waveform received is pre-processed to enhance the signal by means of simple pre-emphasis filtering, or more sophisticated noise reduction or dereverberation algorithms. Then, the processed speech waveform is converted into a sequence of meaningful feature vectors, which are used to feed the speech decoder. The decoder seeks for the best match between a sequence of features and every possible sequence of acoustic classes, using the available information from the acoustic and language models, which are typically obtained in a training phase prior to the recognition step. Additionally, the recognition result can be used as feed-back information to better adjust the acoustic models to the actual speaker or acoustic environment in a process generally named adaptation.

In the next sections, the operation of the different stages of speech recognition systems is described, focusing our attention to the description of those systems that are based on Hidden Markov Models (HMM). Then, some outstanding selected approaches to speech recognition with microphone arrays will be commented.

3.2.1 Front-End

Speech recognition systems do not actually perform the recognition or decoding step directly on the speech signal. Rather, the speech waveform is divided into short frames of samples, which

are converted to a meaningful set of features. The duration of the frames is selected so that the speech waveform can be regarded as being stationary. In addition to this transformation, some pre-processing techniques are applied to the waveform signal and also some robust techniques can be applied later at the feature domain. The whole set of processing steps applied to the speech signal in order to obtain the final feature vector is commonly known as the front-end of the speech recognizer.

Pre-processing

Pre-processing is understood to be applied at the speech signal level in order to enhance it and to better prepare it for the speech recognition. There are several of pre-processing possibilities to be applied to the speech signal but they can be coarsely classified into three kinds: general pre-processing, noise compensation techniques and dereverberation techniques.

By general pre-processing I mean the processing applied in almost all the state of the art systems that can be considered a kind of standard. Particularly, pre-emphasis filtering is the very first enhancement step for speech signals. It is aimed to compensate for attenuation of high frequencies, thus, high frequency components are emphasized while low frequency components are attenuated. It is typically performed by means of a simple high-pass FIR filter, such as $H(z) = 1 - az^{-1}$, being a close to 1. Additionally, DC offset removing is also typically applied to the input speech.

Regarding noise suppression techniques, the most popular are the so-called spectral methods, mostly because of their simplicity and effectiveness. These approaches, first estimate both the short-time noise and noisy speech spectrums, to later, according to a suppression rule, apply a spectral data-dependant time-varying gain function (filter) to the noisy speech amplitude spectrum. The enhanced magnitude and noisy phase spectrums are then combined to produce a clean short-time spectrum estimate. If time-domain re-synthesis is needed, overlap-add (OLA) methods are typically used. A wide range of suppression rules exist. The most popular are spectral substraction (Boll, 1979) and the well-known Wiener filtering technique. The former estimates the clean speech spectrum on the basis of a linear substraction of the noise spectrum to the noisy speech spectrum. The latter, estimates a gain function depending on the SNR, so that the mean square error between the noisy and the estimated speech is minimized. In general, spectral methods rely on a good estimate of the noise spectrum that is typically obtained from non-speech segments of the speech signal.

With respect to dereverberation techniques some basic ideas where already presented in previous chapter. It was shown that some mono-channel dereverberation techniques assume some a priori knowledge about the channel, concretely the speaker-to-receiver impulse response. Usually these techniques are quite sensitive to accurate impulse response knowledge and can result in

strong degradation of the processed signal in the case of wrong estimations. Alternatively, there are some approaches that try to blindly dereverberate the received signal without any explicit estimation of the speaker-to-receiver response (Gillespie et al., 2001). These techniques are usually high computational consuming and are not very well matched with the problem of speech recognition. In fact, most of the approaches for dereverberation are evaluated on simulated reverberant speech (Allen & Berkley, 1979) for better algorithm analysis, but also due to a lack of robustness for processing real reverberant speech. This lack of robustness has traditionally result in low widespread of mono-channel dereverberation techniques for real speech recognition applications. Consequently, and also due to the increasing of computational capability, multichannel dereverberation techniques (Miyoshi & Kaneda, 1988) have become an important field of interest.

Feature extraction

The object of the feature extraction step is to obtain a small set of parameters from each frame of the (pre-processed) input speech data containing the relevant speech information and being the most robust to variations as possible. Mel-frequency cepstral coefficients (MFCC) (Davis & Mermelstein, 1980) and linear predictive cepstral coefficients (LPCC) (Rabiner & Juang, 1993) are probably the most extended, although other speech representations have been investigated and robust feature extraction is still a matter of interest in ASR research. The features used in this Thesis in Chapter 4 are the MFCCs, and the extraction process will be described next. Additionally, some widely used robust feature post-processing techniques are going to be commented. For more information about alternative feature representations, one can refer to (Huang et al., 2001).

Mel-frequency cepstral coefficients The complete feature extraction process for the obtainment of the mel-frequency cepstral coefficients is depicted in Figure 3.8. First,to extract the MFCC features, the discrete Fourier transform (DFT) is applied to windowed overlapped frames of the input signal to obtain the short term energy spectrum as follows:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi \frac{kn}{N}} \tag{3.15}$$

$$|X(k)|^2 = X(k)X(k)^* \tag{3.16}$$

where $x(n)$ is the input speech signal, $X(k)$ its resulting DFT, $|X(k)|^2$ the short time energy spectrum , and N is the DFT size.

The obtained spectrum is mapped into a mel scale multiplying it by a series of triangular weighting functions called mel filters. These triangular filters are equally distributed along

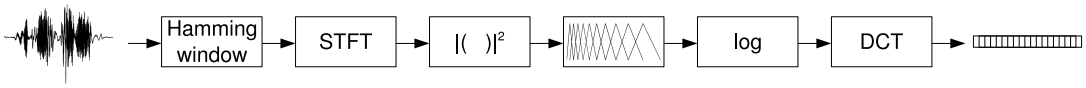

Figure 3.8: *Block diagram of the mel frequency cepstral coefficients (MFCC) feature extraction process.*

the mel frequency scale with a 50% overlap between consecutive triangles. As a result, the mel spectrum vector of the frame is obtained with each component representing the energy in each of the mel filters. The object of this mel scale warping is to approximate to the human auditory system's response more closely than the linearly-spaced frequency. Figure 3.9 shows an example of a 10-band mel scaled triangular filter-bank.

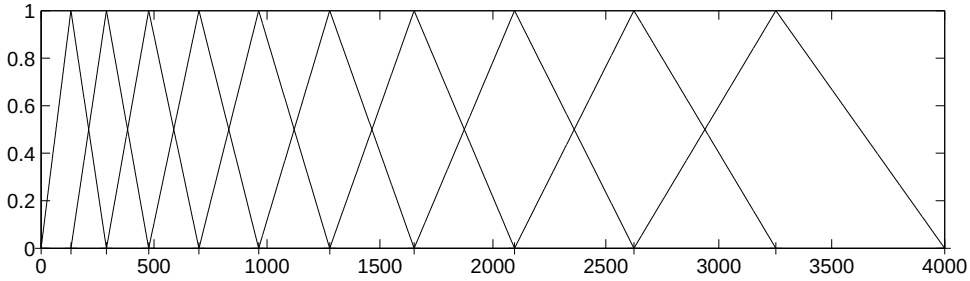

Figure 3.9: *Example of a 10 band mel scaled triangular filter-bank in the range of 0-4 kHz.*

The vector formed by these outputs is generally known as filter-bank energies (FBE). In fact, to better approximate the human auditory processing, the natural logarithm of the FBE vector is taken as follows:

$$S(m) = \log_e \sum_{k=0}^{N-1} H_m(k)|X(k)|^2 \tag{3.17}$$

where $S(m)$ is the logarithmic filter bank energy (logFBE) and $H_m(k)$ represents the response of the m-th mel filter to frequency bin k.

Finally, discrete cosine transformation is applied to the logFBE vector to convert it into the cepstral domain, besides to achieving an high degree of decorrelation of the output features.

$$c(n) = \sum_{m=0}^{M-1} S(m) \cos\left(\frac{\pi n m}{M}\frac{m+1}{2}\right) \tag{3.18}$$

These output features are generally truncated (liftered) in order to obtain the final mel frequency cepstral coefficient vector. Additionally, the input to the speech recognizer is usually a sequence of vectors composed by the mel-frequency cepstral coefficients, as well

as its first and second temporal derivatives, which are approximated by differences with the neighboring feature vectors.

Feature post-processing Some additional post-processing techniques are usually applied to the feature vector to improve speech recognition. In general, these techniques show an excellent trade-off between performance improvement and computational cost, as a result some of them have became part of almost all state of the art front-ends.

Cepstral mean substraction (CMS) (Atal, 1974) is a widely used method for removing short-term invariant linear channel distortion in speech signals. Since the cepstral transformation described above is an homomorphic transformation, a linear distortion channel affects feature vectors in an additive manner. Consequently, it can be removed by means of the substraction of the overall average of the cepstral coefficients over a speech segment.

Cepstral mean and variance normalization (CMVN) (Viikki & Laurila, 1998) technique is a generalization of the CMS technique. The mean and variance of each component vector is estimated over a speech segment and, in addition to substract the mean to each feature component, the variance of each component is normalized to be equal to unity. This simple correction permits acoustic classes to have more invariant position and size in the feature space, thus, reducing mismatch between training and test conditions, and becoming more robust to both noisy and reverberant speech.

3.2.2 Back-End

The object of the back-end stage of a recognizer is to effect a mapping of the speech vectors provided by the front-end and the wanted underlying sequence of acoustic classes modelling concrete symbols (phonemes, letters, words...). The main problem of the recognition process is related to the fact that the mapping from speech to acoustic classes is not one-to-one. The reasons are that different classes can give rise to similar speech sounds, and additionally that different speakers or environments, causes a large variability in speech realizations. In this way, in frame-based statistical speech recognition systems the speech production mechanism is characterized as a random process which generates a sequence of feature vectors. In this case, the isolated word recognition problem (each acoustic class represents a whole word and only isolated words are recognized) can be regarded as:

$$\hat{w} = \arg\max_{w \in W} P(w|\mathbf{O}) \tag{3.19}$$

where w represents a word of the vocabulary W and $\mathbf{O}$ is the sequence of speech vectors provided by the front-end or observations. The sequence of observations is defined as $\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T$, where $\mathbf{o}_t$ is the observed feature vector at time instant t. Hence, the problem is

that of finding the word that has more likely been said given a concrete sequence of observation vectors. This probability can be computed following the Bayes' Rule:

$$P(w|\mathbf{O}) = \frac{P(\mathbf{O}|w)P(w)}{P(\mathbf{O})} \tag{3.20}$$

Because the maximization is with respect to the word w for a given (and therefore fixed) sequence of observation $\mathbf{O}$, the denominator term can be ignored in the maximization, resulting Equation 3.19 in

$$\hat{w} = \arg\max_{w \in W} P(\mathbf{O}|w)P(w) \tag{3.21}$$

Thus, the most probable spoken word depends only on a set of prior word probabilities $P(w)$ known as the language score which are obtained from a language model, and the class conditional observation densities $P(\mathbf{O}|w)$ known as acoustic likelihood or acoustic score. In general, the problem of directly estimating $P(\mathbf{o}_1, \mathbf{o}_t, \ldots|w)$ from examples of speech is not feasible, but if a parametric model of word production is assumed the problem results in the estimation of a short set of parameters. In practice, the most extended and successful statistical parametric approach to speech recognition is the Hidden Markov Model (HMM) paradigm (Rabiner, 1989; Rabiner & Juang, 1993), that supports both acoustic and temporal modeling. Alternatively, artificial neural networks (ANN) have been proposed as an efficient approach to acoustic modeling, but it is not clear how to use them to model the temporal evolution of speech (Tebelskis, 1995). As a consequence, ANN-HMM hybrid systems have been recent focus of research in order to combine the strengths of the two approaches.

HMM-based modeling

In Hidden Markov Model (HMM) speech recognition systems, the random process which corresponds to the generation of a particular word is modeled by a Markov model. An HMM is a finite state machine characterized by the number of states. A transition between two states i and j is produced at each time instant t with an associated state-transition probability $a(i, j)$. Each time a state j is entered, a speech vector $\mathbf{o}_t$ is generated with a certain probability $b(\mathbf{o}_t, j)$, known as the output probability distribution. The state output probability distribution functions are usually modeled as Gaussians or mixtures of Gaussians. In order to improve computational efficiency, the Gaussians are typically assumed to have diagonal covariance matrices. Thus, the output probability of a observation vector $\mathbf{o}_t$ belonging to the state j of an HMM is represented in function of M_s Gaussian mixtures as

$$b(\mathbf{o}_t, j) = \sum_{m=1}^{M_s} c_{jm} N(\mathbf{o}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}) \tag{3.22}$$

where c_{jm}, $\boldsymbol{\mu}_{jm}$ and $\boldsymbol{\Sigma}_{jm}$ are the mixture weight, mean vector and covariance matrix associated with the m-th Gaussian in the mixture density of state j of the HMM. $N(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multi-variate Gaussian:

$$N(\mathbf{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{o}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{o}-\boldsymbol{\mu})} \tag{3.23}$$

where n is the dimensionality of $\mathbf{o}$.

The set of parameters $B = \{c_{jm}, \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}\}$ for all mixture components for all states in the HMM, together with the matrix of state transition probabilities $A = \{a(i,j)\}$, is the complete set $\lambda = \{A, B\}$ of statistical parameters that define an HMM.

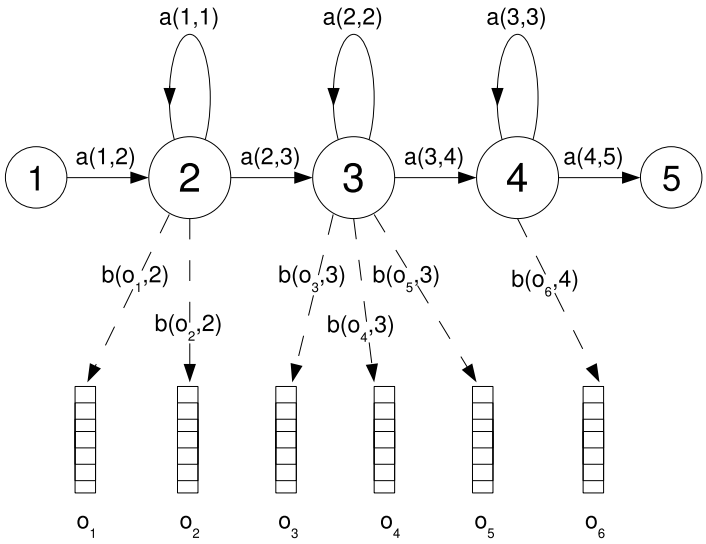

Figure 3.10: *Example of generation of a concrete observation sequence by a left-to-right Markov chain of five states: two of them non-emitting -the first and the last- and three emitting states.*

Speech is generally modelled by left-to-right chains as the one shown in Figure 3.10, where the only allowable transitions are back to the current state or to the state of the right. Additionally, in most cases the initial and final states are non-emitting states, that is, no observations are generated, thus these states do not have associated an output probability distribution. This is usually done to facilitate the construction of composite models for more general applications different to isolated word recognition.

Given a certain model M characterized by its statistical parameters $\lambda_M = \{A_M, B_M\}$, the joint probability that $\mathbf{O} = \mathbf{o}_1, \mathbf{o}_t, \ldots, \mathbf{o}_T$ is generated by the model M transiting thorough a

state sequence $\mathbf{s} = s_1, s_2, \ldots, s_T$ can be simply calculated as the product of the transition and output probabilities as follows:

$$P(\mathbf{O}, \mathbf{s}|M) = a_M(s_0, s_1)b_M(\mathbf{o}_1, s_1)a_M(s_1, s_2)b_M(\mathbf{o}_2, s_2)\ldots \tag{3.24}$$

where s_0 and s_{T+1} are forced to be the input and the exit non-emitting states respectively. The problem is that only the observation sequence $\mathbf{O}$ is known and the state sequence $\mathbf{s}$ remains hidden. Hence, the complete probability $P(\mathbf{O}|M)$ is computed as the sum over the space of all the possible state sequences S:

$$P(\mathbf{O}|M) = \sum_S a_M(s_0, s_1)\prod_{t=1}^{T} b_M(\mathbf{o}_t, s_t)a_M(s_t, s_{t+1}) \tag{3.25}$$

In practice, this probability is usually approximated only by the most likely state sequence:

$$P(\mathbf{O}|M) \approxeq \max_{\mathbf{s}\in S}\left\{a_M(s_0, s_1)\prod_{t=1}^{T} b_M(\mathbf{o}_t, s_t)a_M(s_t, s_{t+1})\right\} \tag{3.26}$$

Hence, the problem of isolated word recognition can now be tackled if Hidden Markov models are available for each word w with their corresponding transition probabilities and output probabilities parameters ($\lambda_w = \{A_w, B_w\}$). If we replace the acoustic likelihood by the expression of 3.26, then the Equation 3.21 results

$$\hat{w} = \arg\max_{w\in W, \mathbf{s}\in S}\left\{a_w(s_0, s_1)\prod_{t=1}^{T} b_w(\mathbf{o}_t, s_t)a_w(s_t, s_{t+1})P(w)\right\} \tag{3.27}$$

The HMM framework can easily be expanded to model strings of words. The HMM of a sequence of words $\mathbf{w} = [w_1, w_2, \ldots, w_T]$ can be built concatenating the HMMs of the individual words if they have been trained independently. In fact, the idea of concatenating individual models of words to construct models of sentences can be generalized to the case of concatenating any kind of sub-word unit. That is, HMMs of for instance phonemes can be used to built an HMM of the constituent words of a sentence. In practice, this generalization of continuous recognition of any kind of acoustic class implies an high computational demand in the evaluation of all the possible sequence of units, resulting unpractical its direct computation and becoming necessary efficient strategies for decoding that will be commented next.

HMM training

One of the main appealing characteristics of the HMM approach to statistical speech recognition relies in the possibility of automatically determining the parameters of a particular model provided speech realizations of that acoustic class by efficient and robust re-estimation procedures. Thus, the problem initially commented of covering the variabilities of speech realization is elegantly solved providing a sufficient representative amount of training data of each word (or recognition unit), since a HMM will implicitly model all the sources of variability in real speech present in the data. In fact, the amount and the kind of training data affect the resulting HMM system and can determine its usability, for instance, it is possible to develop speaker dependent systems with few data of a single speaker, while much more data is needed for a speaker independent system.

For detailed information on HMM training, one can refer to (Rabiner, 1989; Rabiner & Juang, 1993). In few words, the procedure can be reduced to a fist step of initialization, which is usually based on the equal separation of the observations of a concrete model between all the states and the computation of an initial mean an variance estimation of each state. Then, the maximum likelihood state sequence using the Viterbi algorithm (Forney Jr, 1973) is used to reassign the observation vectors to states and then mean and variances are updated with the resulting reassigned observations. Once the estimations do not change, a second step of more accurate parameters search is applied by means of the Baum-Welch re-estimation formulae. It is an estimation-maximization algorithm that basically lies in the assignment of each observation vector to every state in proportion to the probability of the model of being in that state when the vector was observed, instead of assigning the observation to a single state. This is efficiently done using the Forward-Backward algorithm (Devijver, 1985). The Baum-Welch re-estimation is iteratively applied until no improvements in the likelihood of the observation vectors are obtained given the previous and the new re-estimated model.

In order to train independent models for each word or sub-word unit following this procedure, the boundaries of the realizations of each unit are needed. In sub-word recognition systems hand-labeling the boundaries of the units of the training data can be unfeasible and inaccurate. The usual way to tackle the training of continuous word or sub-word recognition systems is to first initialize all the models assigning the global speech mean and variance. Then, the Baum-Welch procedure is applied similarly as for the isolated case but rather than training each model individually a composed model is built concatenating the sub-word models for each training utterance an in this way all models are trained in parallel. To start with this procedure during the first cycle of re-estimation it is necessary to uniformly segment each training utterance. The hope then is that enough of the phone models align with actual realizations of that phone so that on the second and subsequent iterations, the models align as intended.

In practice, most of the medium and large vocabulary recognition systems are based on HMMs of phonemes or context dependent phonemes, more concretely, the so-called triphones are widely used. Triphones are phonemes with left and right context, that is, a different recognition unit is considered for each phoneme depending on its predecessor and posterior phonemes. Usually, to reduce the total number of parameters needed by the HMMs and to increase the robustness of the estimations for all the phonemes or triphones modeled, the parameters of the Gaussian distributions are shared across states of various models. This process is commonly called state tying and the states which are sharing parameters are known as tied states. It is also usual in triphone systems to tie the transition matrices of all the set of triphones of a same phoneme.

Recognition and Viterbi decoding

In order to build an speech recognition system able to obtain the most likely word or sequence of words spoken, in addition to the set of acoustic models for all the possible recognition units it is also necessary, as shown in Equation 3.27, to construct a language model that contains the information of all the possible sequence of words that the system is expected to recognize, that is, the prior probability $P(w)$ must be modeled. Depending on the object of the system, this language model can be simply constructed by means of a hand-made word network if few constrained sentences need to be recognized or also it can be built based on stochastic approaches, estimating it from transcribed speech corpora. Among statistical approaches, N-grams (Jurafsky & Martin, 2000) are the most widely used, both because of its simplicity and recognition performance. Additionally, in conjunction to the language model it is necessary to have a dictionary of pronunciations that permits constructing a recognition network specifying what is allowed to be spoken and how each word is pronounced in terms of recognition units.

Hence, given a recognition network and its associated set of HMMs, the task of the decoder is to find the path thorough the network which is more likely for a given unknown speech utterance. This search problem is efficiently solve by the Viterbi algorithm (Forney Jr, 1973), which is extensively used in most recognizers. For large vocabulary ASR, sub-optimal solutions and optimized search strategies are usually taken due to the enormous computational cost related to optimal search.

Acoustic model adaptation

Acoustic model adaptation can be eventually used to adapt a set of well-trained generic HMMs to the specific characteristics of a small amount of enrollment data. The most typical application of model adaptation is addressed to modify speaker independent models to a set of models specialized for a particular speaker on the basis of utterances from this speaker. Thus, an improved

performance is expected when this new adapted models are used to recognize speech from that particular speaker, since the variability of the adapted models is reduced. In general, most of the model adaptation techniques can be applied both in a supervised manner, where the transcriptions of speech utterances are used for adapting the generic models, or in an unsupervised manner, where no transcriptions are needed. Furthermore, adaptation can be applied incrementally for each adaptation utterance as it becomes available or in a batch or static manner if the adaptation data is available as a block of data. It is worth noting that although most adaptation techniques have been mainly proposed for speaker adaptation and as a consequence they have been widely used for that purpose, the generic aim of these techniques is to approximate the acoustic space of a set of HMMs to the acoustic characteristics of the adaptation data. Thus, if the adaptation data is representative of a specific acoustic environment where several utterances from different speakers were collected, the result of applying an adaptation technique will be the adaptation to that specific acoustic environment. For instance, acoustic models for specific use in telephone applications can be obtained by means of adaptation from a set of HMMs trained with clean data collected with high quality microphones in a studio, rather than training a complete new set of models specific for telephone applications.

Several approaches exist for model adaptation, such as maximum-likelihood linear regression (MLLR) (Leggetter & Woodland, 1995; Gales & Woodland, 1996), maximum a posteriori (MAP) (Gauvain & Lee, 1994), Parallel Model Combination (PMC) (Gales & Young, 1993) or Jacobian Adaptation (JA) (Sagayama et al., 1997; Abad et al., 2003). MLLR is one of the most successful and widely used adaptation technique and it will be recalled in Chapter 4. It estimates a set of linear transformations for the mean and variance parameters of a Gaussian mixture HMM system. If few adaptation data is available, it is possible to simply estimate a global adaptation that is applied to every mixture component. As more adaptation data is available, more specific transformations can be estimated to be applied to a specific grouping of Gaussians, for instance, they could be grouped into the broad phone classes: silence, vowels, stops, glides, nasals, fricatives, etc. The set of Gaussians which share a transform is usually referred to as a regression class.

3.2.3 Approaches to speech recognition with microphone arrays

The significant advances in ASR technology, some of them commented above, have result in the development of high performance systems for the case where there is a good match between the HMM models, that is the training corpora used for constructing them, and the real scenario of application, that is the testing data. However, most of these systems suffer from lack of robustness to strong mismatches between the training and the testing data. For instance, a system developed with data collected with a close talking microphone in a quite environment would suffer an important drop of performance when tested with data collected with a far-field

microphone in a noisy environment.

As a consequence, many research efforts have been addressed to increase robustness to the data mismatch problem, and to the variability of the acoustic conditions, particularly when the objective is to develop a distant-talking system. Some of the possibilities for robust distant-talking ASR systems have been already commented. For instance, pre-processing techniques, or robust feature processing techniques, are addressed to improve the performance of these systems in the front-end stage. At the back-end level it is possible to increase robustness of the HMM set by means of acoustic model adaptation. Another possibility is the use of large training corporas covering various environment conditions or the introduction of noisy speech material in the training corpora, which is known in the literature as training data contamination (Gong, 1995), to increase the robustness of the constructed models.

In the first section of this chapter, it has been shown the potential use of microphone arrays as a speech enhancement technique able to emphasize or reject concrete spatial directions on the basis of multiple microphone availability. As a consequence, one can also think in the application of microphone array beamforming as a pre-processing stage of an ASR system. Thus, enhanced samples of the captured speech can be obtained, eventually reducing the mismatch between the data collected by far-field microphones and the one collected by a close-talking microphone, and improving the performance of distant-talking ASR systems.

Hence, as suggested, the first immediate application of any one of the beamforming techniques described above to the speech recognition problem is their use as a pre-processing stage. Many examples can be found in the literature of microphone array based speech recognition systems in this way. For instance in (Giuliani et al., 1995) and (Yamada et al., 1996), the delay-and-sum beamformer is used in ASR experiments with different tracking strategies. In fact, despite the limitations of the delay-and-sum approach, it is probably the most widely used beamformer for ASR applications. Also, adaptative beamformers have been used for automatic speech recognition, particularly in applications where the speaker remains almost static as for instance in car applications (Grenier, 1992). In general, the use of microphone arrays as a pre-processing stage has been reported to increase the performance of ASR systems, but the improvement observed is not in well accordance with the expected increase based on the SNR gains observed and the subjective enhancement of the quality of the speech signals. In other words, microphone array processing methods are designed according to various waveform-level criteria, such as maximum SNR, maximum array gain, or minimum mean squared error, that does not necessarily result in improved recognition accuracy.

To better match the use of microphone arrays for speech recognition some early approaches proposed the use of well-known successful robust techniques in combination with beamforming. For instance, in (Giuliani et al., 1996; Kleban & Gong, 2000) the use of microphone arrays and different acoustic model adaptation techniques is explored.

More recently, a new appealing technique for microphone array processing have been specially proposed for speech recognition in (Seltzer, 2003; Seltzer et al., 2004). Since speech recognition is a pattern classification problem, rather than tackling beamforming as an independent signal processing problem, the beamformer is designed to maximize the likelihood of the correct hypothesis. On the basis of a filter-and-sum beamformer, the likelihood maximizing beamforming (LIMABEAM) (Seltzer et al., 2004) uses the information from the speech recognition system itself to find the array parameters that improve speech recognition performance in an unsupervised or supervised way. Recent works have appeared as an extension of the idea of this ASR dedicated beamformer, for instance optimizing the filter-and-sum coefficients in the MFCC domain rather than in the logFBE domain (Raub et al., 2004) or incorporating an N-best search strategy to improve results of the unsupervised LIMABEAM (Brayda et al., 2006).

Chapter 4

Contributions to Microphone Array Speech Enhancement and Recognition

This chapter shows the contributions and experiments carried out by the author in the field of speech enhancement and speech recognition with microphone arrays. First, a robust beamformer based on the integration of a GSC-like beamformer with multi-channel Wiener post-filtering is presented mainly for speech enhancement purposes. Then, a medium size vocabulary automatic speech recognition system for practical application in the UPC smart room is developed.

Microphone array processing techniques for speech enhancement and speech recognition have grown in the last times to currently become a considerably important research topic in speech processing. The widespread gained by microphone array processing is stated in the previous state of the art of Chapter 3, where only some of the many contributions have been possible to be emphasized.

Although the intensive dedication, the problem of enhancing far-field speech signals and more concretely distant speech recognition are still open problems. Consequently, microphone array processing techniques for speech enhancement and more particularly for speech recognition in the context of smart room environment applications is one of the two main pillars of this Thesis and the work carried out by the author in this field is addressed in this chapter.

Concretely, first it is described early work presented in (Abad & Hernando, 2004b, 2004a) related to the design of a robust beamformer that integrates in a single stage, alternatively to most of the previous reported works, the directional noise reduction capability of a modification of the Generalized Sidelobe Canceller with adaptive blocking matrix, and the great performance of the Wiener post-filtering techniques against low-correlated noises. Experimental results show

that the proposed integrated Wiener-filtering with adaptive beamformer is more robust than conventional post-filtering of the output to correlated and uncorrelated noise, and presents a lower level of degradation of the speech signal besides a better speech recognition performance.

Furthermore, the development process of an automatic speech recognition system for medium size vocabulary tasks for particular use in the UPC smart-room environment is described. In this work, speaker mismatch and environmental mismatch problems are addressed by means of robust acoustic model building and remarkable improved results are obtained. In fact, an on-line implementation of the ASR system derived from this work is currently being used for demonstration purposes in the context of the EU funded CHIL project research activities carried out at the UPC. Additionally, several microphone array alternative approaches are investigated in order to incorporate them into the proposed ASR system. As a result of this recent work, some preliminary experiments and proposals are provided. Particularly, microphone array beamforming is assessed to be used as an integral part of an acoustic model adaptation scheme, rather than as a conventional pre-processing stage prior to feature extraction.

4.1 Integrated Wiener-filtering and Adaptive Beamforming

Many of the most popular approaches to microphone array beamforming shown in the previous chapter are developed under the assumption of some specific static noise characteristics. That is, a fixed beamformer is designed to operate in concrete environment conditions providing in this way a simple, stable and robust solution when these conditions are fulfilled. However, in most real scenarios the acoustic environment is continuously changing. For instance, in smart-room applications, such as meeting assistance, new participants can appear, change their position, make noises, turn on their laptops or ask questions. In this context, it would be preferable to use beamformers capable of adapting to the ever changing conditions.

Generalized Sidelobe Canceller like beamformers (see Section 3.1.2) are the most successful and extended adaptive approaches. Thanks to its two path structure, shown in Figure 3.5, it is able to reject and cancel highly directive noise sources, without prior information about their spatial or spectral characteristics. Unfortunately, some problems arise with this kind of beamformers. First, their poor performance when low correlated noises affect microphone signals, particularly at the low frequency range. Second and most importantly, the blocking step is generally far from being perfect in real conditions due to steering errors or small error adjustments in the microphones of the array. Consequently, signal leakage appears at the output of the blocking matrix, resulting in practice in high signal cancellation at the beamformer output.

To overcome the problem of poor performance in presence of low correlated noise, a practical solution is the use of the post-processing Wiener filtering shown in Section 3.1.3. To solve the problems of the signal cancellation of adaptive beamforming, some solutions where already

commented in Section 3.1.2. Concretely, in (Hoshuyama et al., 1999) is proposed designing an adaptive blocking matrix (ABM) with control of the allowable target error region using the output of the fixed part of the GSC structure and the microphone signals, rather than a fixed blocking matrix.

The motivation of the work presented here is to design a robust beamformer in presence of both correlated and uncorrelated noise by means of the combination of adaptive beamforming with post-filtering. Some previous works (Bitzer et al., 1999; McCowan et al., 2000) tried to do it applying post-filter at the output of a GSC-like beamformer. However, in these works the post-filter considered was optimally designed for the delay-and-sum beamformer and the problems associated with target error tracking were not addressed. In this section, adaptive GSC-like beamformer with ABM is integrated with Wiener filtering by modifying the FBF part. Wiener filtering is applied to generate a cleaner output of the FBF that is used to better estimate the ABM and as the reference for the MC. In this way, the proposed beamformer is expected to be a higher directive and low correlated noise cancellation structure with less distortion of the speech signal in a single stage.

Experimental evaluation in next Section 4.1.3 shows that the proposed integrated Wiener-filtering with adaptive beamformer (IWAB) results in an appropriate technique for speech applications, particularly in high noise scenarios. Moreover in the experiments it is shown that integrating Wiener filtering in the structure of a robust beamformer outperforms traditional post-filtering of the output of the beamformer.

4.1.1 Prior work

Robust Generalized Sidelobe Canceller

The IWAB beamformer is based on the robust GSC structure proposed in (Hoshuyama et al., 1999), shown in Figure 4.1. It basically consists on a robust modification of the GSC, where the blocking matrix is adaptively designed to allow a concrete target-looking error region and to minimize the leakage of the desired signal at the beamformer output. CCAF filters (Coefficient Constrained Adaptive Filters) in the adaptive blocking matrix (ABM) minimize the output of the ABM resulting in target tracking. Thus, CCAFs processing consists on a NLMS (Normalized LMS) algorithm constrained to a maximum and a minimum bound for the values of the coefficients, according to the maximum allowable look-direction error desired. LAF filters (Leaky Adaptive Filters) are used in the multiple-input canceller (MC) to minimize the components of the fixed beamformer output (FBF) correlated with the outputs of the ABM enhancing the robustness of the system. It consists on a NLMS updating process with a small leakage constant that avoids excess growth of tap coefficients and that prevents the signal target cancellation when minimization at the ABM is incomplete.

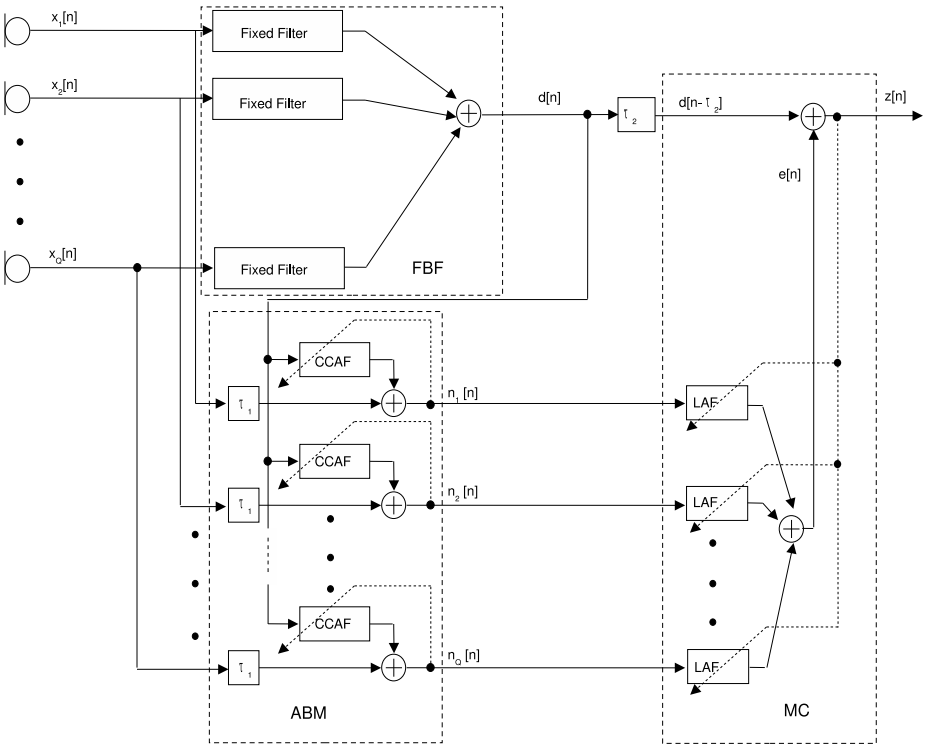

Figure 4.1: *GSC beamformer with CCAF-LAF structure as proposed in Hoshuyama et al. (1999).*

Wiener post-filtering

The Wiener post-filtering described in Section 3.1.3 is the one considered in this work. The optimal Wiener post-filter for the case of the delay-and-sum beamformer under the assumptions previously detailed is the one of Equation 3.13.

4.1.2 The proposed Integrated Wiener-filtering with Adaptive beamformer

In order to design a beamformer that can perform correctly in a wide range of applications, robust beamformers based on modifications of a GSC-like structure have been used in combination with post-filtering in some previous works. In (Bitzer et al., 1999) a super directive beamformer in a GSC-like structure in combination with post-filter is presented. A structure consisting of a fixed near-field super directive beamformer and an adaptive sidelobe cancelling path adding a post-filter is reported in (McCowan et al., 2000). The beamformer proposed in (Fischer & Simmer, 1995), similarly to the new proposed approach, uses Wiener Filter as the fixed part and includes fixed temporal low-pass filters in a fixed BM to reduce signal cancellation. Thanks to this combination these techniques can obtain a good performance against both directional and diffuse noises.

The novelty in this work (Abad & Hernando, 2004b, 2004a) is the integration of the stage of adaptive beamforming with adaptive blocking matrix and the stage of post-filtering in a single one. The way in which it is achieved consists in applying Wiener filtering to generate the output of the FBF instead of applying it at the final output of the beamformer. This filtered output of the fixed path of the beamformer is used by the ABM and by the MC.

Although, in practice the Wiener filter is applied to the output of the FBF, it is equivalent to the application of identical filters to each single microphone channel. Thus, the proposed beamformer can be understood as a filter-and-sum beamformer nested in a GSC-like robust structure, where the filter is the one given in Equation 3.13. As stated in Section 3.1.3, the output of the delay-and-sum with post-filter is optimal in the case that the noise is uncorrelated between microphones, that noise have equal power spectral density and that the received signal is uncorrelated with the additive noise. Even in the case that some of these conditions are not completely fulfilled, the Wiener post-filter has been shown to provide still better performance than simple delay-and-sum beamformer.

Therefore, the underlying idea behind the entire work is that the better the output of the FBF is, the better the ABM is estimated. Consequently, the MC will have better estimations of the reference target signal and of the reference noises provided by the ABM. Figure 4.2 shows a schematic comparison between the common post-filter approach and the proposed integration in the FBF path.

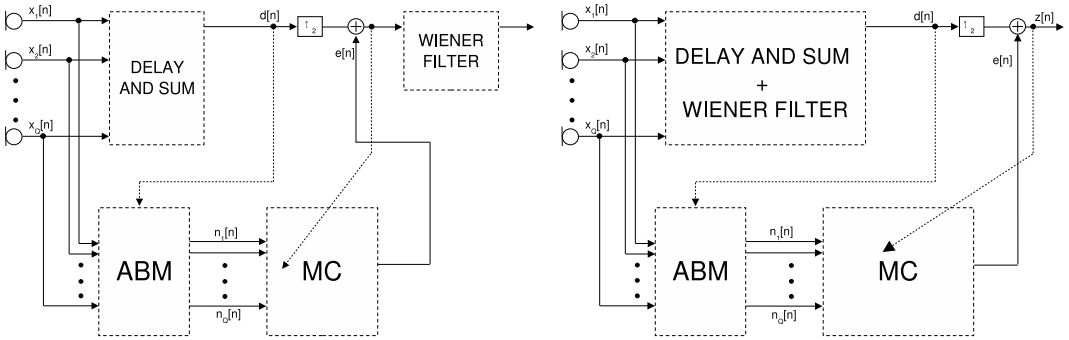

Figure 4.2: *On the left side, schema of the common post-filter approach. On the right side, schema of the proposed integrated structure.*

Furthermore, it is expected that the integration of the Wiener filter inside a GSC-like structure may provide some additional benefits. Usually, when noise is very high, post-filtering techniques can reduce it very well but in exchange for a more degraded signal. The proposed structure is expected to reduce this degradation in two ways.

First, the Wiener filtering is applied to the output of the FBF, that is, the signal for which the optimal Wiener is computed, and not to the output of the complete beamformer. The design of the Wiener post-filtering stage for a concrete beamformer depends on the weights of

the beamformer used as shown in Equation 3.14. For this reason, it would not be an optimal solution to apply the same Wiener filter at the output of the adaptive beamformer, resulting into a lost of the low-correlated noise reduction capability of the overall system.

Second, the MC can compensate some of the artifacts or degradations introduced to the signal by the filtering process. These degradations, besides being disturbing in speech enhancement applications, can also importantly affect the performance of an automatic speech recognition engine trained with clean speech. Distortion reduction can be obtained based on the idea that degradations introduced in the Wiener filtering process will remain at the output of the ABM since they were not originally present in the signals arriving to the array. Therefore, if some of these degradations are present in the noise references the MC will use them to reduce their presence in the reference signal resulting in a less distorted output.

Finally, one fundamental aspect to consider regarding the correct functionality of the MC stage in a GSC-like beamformer is the degree of correlation between the residual noise of the FBF output and the noise references given by the blocking matrix. If the degree of correlation drops in a significant manner, the capability of the adaptive path for reducing noise at the FBF output would be affected, resulting useless depending on the degree of this drop of coherence. With respect to the modifications of the fixed path introduced in the proposed beamformer, the time varying nature of the Wiener filter could result in a drop on the noise components coherence, due to possible inability of the CCAFs of the ABM to track this time variability. Consequently, the noise cancellation capability of the overall system would be seriously affected.

To clarify this key point, it becomes necessary to measure the coherence of the outputs of the ABM and MC ($n_1[n]$-$n_7[n]$ and $e[n]$) with the speech and noise components of the FBF output ($d_s[n]$ and $d_w[n]$ respectively). Thus, GSC and the proposed IWAB beamformers are compared in presence of 0 dBs diffuse noise. More details about experimental set-up with simulated array data are described in next Section 4.1.3. Results on Table 4.1 show a slight drop of the average coherence of the noise references with the noise components at the output of the Wiener filter, while a more important lost can be observed in the coherence of the MC output. Despite this observation, it is true that the coherence with the noise components is still stronger than the coherence with the speech components for both the proposed beamformer and the conventional GSC-like beamformer. In this way, it is not expected an important lost of performance due to low correlation. Moreover, attending to experimental results in Section 4.1.3 it can be confirmed the utility of the adaptive path in the proposed beamformer and a minimum impact of this effect. Otherwise, in the case of a wrong adaptation of the ABM, the overall output of the proposed structure would present an important leakage of the desired signal that has not been observed. The smoothing process carried out in the estimation of the power spectrum and cross-power spectrum densities reduces the time variability of the filter and, in this way, it helps to reduce the commented negative effect.

		$n_1[n]$	$n_2[n]$	$n_3[n]$	$n_4[n]$	$n_5[n]$	$n_6[n]$	$n_7[n]$	AVG	$e[n]$
GSC	$d_s[n]$	0,11	0,09	0,10	0,11	0,02	0,03	0,06	0,07	0,05
	$d_w[n]$	0,28	0,28	0,18	0,32	0,26	0,41	0,31	0,29	0,58
IWAB	$d_s[n]$	0,06	0,04	0,03	0,04	0,04	0,05	0,06	0,05	0,07
	$d_w[n]$	0,19	0,18	0,17	0,13	0,45	0,45	0,31	0,27	0,31

Table 4.1: Coherence measures of the noise references given by the ABM ($n_1[n]$-$n_7[n]$), their average (AVG) and the output of MC ($e[n]$) with the speech components ($d_s[n]$) and the noise components ($d_w[n]$) of the corresponding FBF output.

In Figure 4.3 a detailed diagram of the proposed structure is shown. It can be seen that the Time Delay Compensation (TDC) block has been added to simulate target signal coming always from broadside. The structure is the same of the Figure 4.1, where the FBF is replaced by the delay-and-sum with Wiener filter of Equation 3.13.

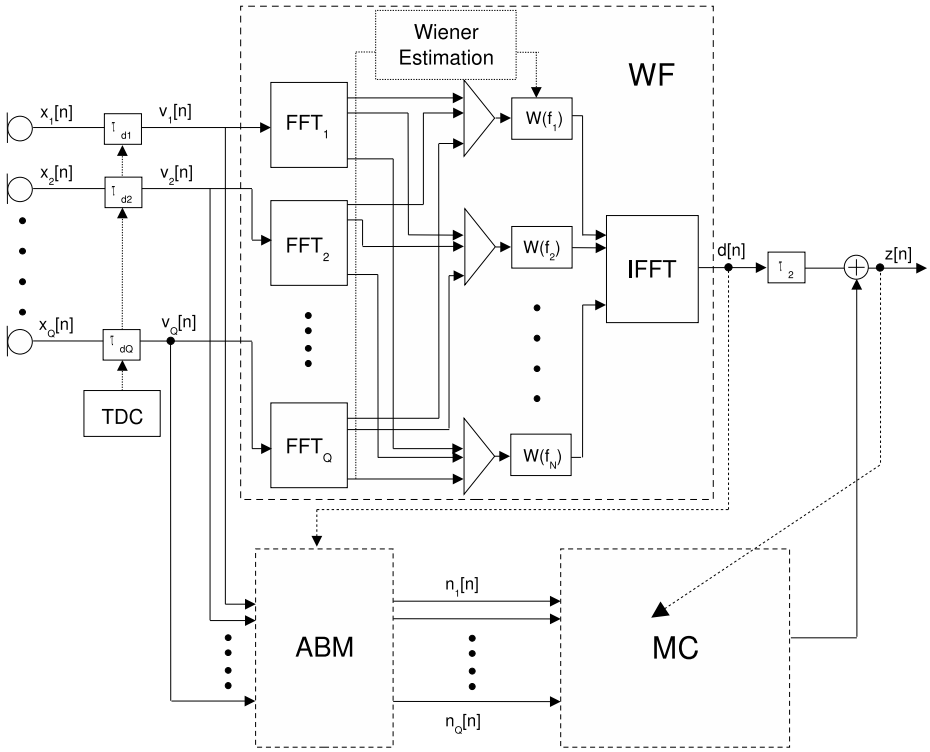

Figure 4.3: *Detailed structure of the proposed IWAB beamformer.*

4.1.3 Experimental evaluation

This section shows the experiments carried out to test the performance of the proposed integrated Wiener-filtering with adaptive beamformer (IWAB) in comparison with the delay-and-sum (DS),

the post-filtered DS (PDS), the GSC with CCAF-LAF structure (GSC) and the same adaptive beamformer with post-filtering (PGSC). Two sets of experiments are conducted. The first one tests the different beamformers for speech enhancement and speech recognition with simulated microphone array data, while the second one compares the beamformers with real microphone array data in speech enhancement experiments. Additionally, some implementation details of the IWAB beamformer are provided.

Implementation details of the IWAB beamformer

Characteristics of the filters and some other details related with the implementation have been selected according to experimental observations. CCAF are 8 taps length, the step size of the NLMS corresponds to 0.1 and the maximum allowable target error region is fixed to 5 degrees. 116 taps compose the LAF filters, the step size of the NLMS is 0.2 and a constant value of 10^{-5} has been selected for the leaky coefficient. Delays τ_1 and τ_2 are fixed to 3 and 10 samples respectively (see Figure 4.1). To obtain the output of the Wiener filtering stage, speech is processed in windows of 64 samples with an overlap of 50 % and FFT of N=128 samples is computed. As a consequence of this processing for the computation of the output of the fixed part of the beamformer, a delay of N/4 samples is introduced that must be compensated. In order to compute the coefficients of the Wiener filter, short term power spectrum and cross-power spectrum densities are recursively estimated with a first order filter (the τ of the forgetting factor is 0.128 s).

The adaptation process of the filters of the ABM and the MC is a key aspect of the implementation of the proposed IWAB beamformer. During the first 375 ms of each speech utterance only CCAFs are adapted, that is, the ABM is initialized. Then, the filters of the ABM are adapted only when speech of the target source is present, and the LAFs of the MC are adapted only in absence of target speech. Hence, ABM and MC are alternatively adapted depending on whether the target speaker is active (adaptation of ABM) or not (adaptation of MC). Periods of vocal activity have been marked with a coarse energy-based voice activity detector applied to the clean reference speech signal. This adaptation process is identically applied to the other adaptive beamformers considered in this study, that is the GSC and the PGSC beamformers.

Microphone array data

Simulated microphone array data Simulated microphone array data is obtained using measured real IRs and recorded ambient noise. The signal model at each microphone is

$$x_q[n] = s_q[n] + i_q[n] + w_q[n] = s[n] * h_{sq}[n] + i[n] * h_{iq}[n] + w_q[n] \tag{4.1}$$

where $x_q[n]$ is the simulated received signal at microphone q, $s[n]$ and $i[n]$ are the original clean target and interfering speech signals, $h_{sq}[n]$ and $h_{iq}[n]$ are the impulse responses from the desired target and the interferer speaker to microphone q and $w_q[n]$ is the ambient noise captured at microphone q.

Real measured IRs and real ambient noise are obtained from the RWCP database (Satoshi et al., 2000). The IRs and the noise used were recorded in a high reverberating meeting room (approximately 780 ms of reverberation time) with a 7-element microphone linear array of 5,66 cm of separation between microphones.

In the speech enhancement experiments, target speaker (male) is situated in the broadside and an interfering speaker (female) is at about 40 degrees. Both the target and the interference microphone array data is generated convolving the corresponding IR with close-talking speech signals of Spanish sentences of about 5 s length, randomly selected from the Speecon database (Iskra et al., 2002). Different signal to noise ratios (SNR) and signal to interfering ratios (SIR) are simulated.

Array data for speech recognition experiments is obtained from part of the Aurora1 test set (Pearce, 1998) convolving it with the IRs and adding noise at different SNRs. In the speech recognition experiments only the target speaker is present.

Real microphone array data Real microphone array data is obtained from the CMU array database (Sullivan, 1996). Speech signals were recorded in both a noisy computer laboratory and a conference room with a 7-element microphone array linearly distributed.

Evaluation metrics

Noise and interference reduction, speech degradation in different situations and automatic speech recognition performance using the different beamformers are evaluated in terms of the signal to noise ratio (SNR) gain and target signal to interference ratio (SIR) gain, log area ratio (LAR) distance and word error rate (WER) respectively.

For this purpose, the beamformer processing is simultaneously applied to the simulated received signals ($x_q[n]$), to the speech target components ($s_q[n]$), to the interferer components ($i_q[n]$) and to noise components ($w_q[n]$). Hence, considering beamforming as a multi-channel time varying filtering process where $h_{BFq}[n]$ is the filter applied to channel q we can generally write

$$z[n] = \sum_{q=1}^{Q} h_{BFq}[n] * x_q[n] = \sum_{q=1}^{Q} h_{BFq}[n] * s_q[n] + \sum_{q=1}^{Q} h_{BFq}[n] * i_q[n] + \sum_{q=1}^{Q} h_{BFq}[n] * w_q[n] \tag{4.2a}$$

$$z[n] = z_s[n] + z_i[n] + z_w[n] \tag{4.2b}$$

where $z[n]$ is the beamformer output and $z_s[n]$, $z_i[n]$ and $z_w[n]$ are the only target, interference and noise processed components respectively.

In this way, SNR and SIR gain can be obtained as the differences of SNR and SIR between the input (SNR_{in} / SIR_{in}) and the output (SNR_{out} / SIR_{out}) of the system.

$$SNR_{in} = \frac{E_{s_{in}}}{E_{w_{in}}} \qquad SIR_{in} = \frac{E_{s_{in}}}{E_{i_{in}}} \tag{4.3a}$$

$$SNR_{out} = \frac{E_{z_s}}{E_{z_w}} \qquad SIR_{out} = \frac{E_{z_s}}{E_{z_i}} \tag{4.3b}$$

where $E_{s_{in}}$ and E_{z_s} are the average energy of the speech components at the input and the output of the beamformer respectively, $E_{i_{in}}$ and E_{z_i} the energy of the interference components present in the input and the output, and $E_{w_{in}}$ and E_{z_w} the energy of noise components. Energies of both desired and interference components are averaged only along activity periods, while noise energy is averaged across all the signal duration. Moreover, the energies of target speech, interfering signal and noise at the input of the array ($E_{s_{in}}$, $E_{i_{in}}$, and $E_{w_{in}}$) are averaged across all the microphones.

LAR distances are computed for a processing signal as the average frame distance of the LAR coefficients vector of the output of the only speech components ($z_s[n]$) of each beamformer under study and the clean reference ($s[n]$). With k being the frame index it can be written as follows

$$LAR_{distance} = \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{LAR}\{\mathbf{s}(k)\} - \mathbf{LAR}\{\mathbf{z_s}(k)\}\|^2 \tag{4.4}$$

The i-th coefficient LAR_i of the LAR vector is obtained as the decimal log area ratio of the i-th parcor coefficient α_i as follows

$$LAR_i = \log_{10} \frac{1 + \alpha_i}{1 - \alpha_i} \tag{4.5}$$

The word error rate (WER) is the metric used to evaluate speech recognition. It is a measure of the average number of word errors taking into account three error types: substitution (the

reference word is replaced by another word), insertion (a word is hypothesized that was not in the reference) and deletion (a word in the reference transcription is missed). The word error rate is defined as the sum of these errors divided by the number of reference words. Notice that given this definition the word error can be more than 100%.

Speech enhancement experiments with simulated microphone array data

Figure 4.4 shows on the left side SNR gain obtained for each technique for different input SNRs and with input SIR fixed to 10 dBs.

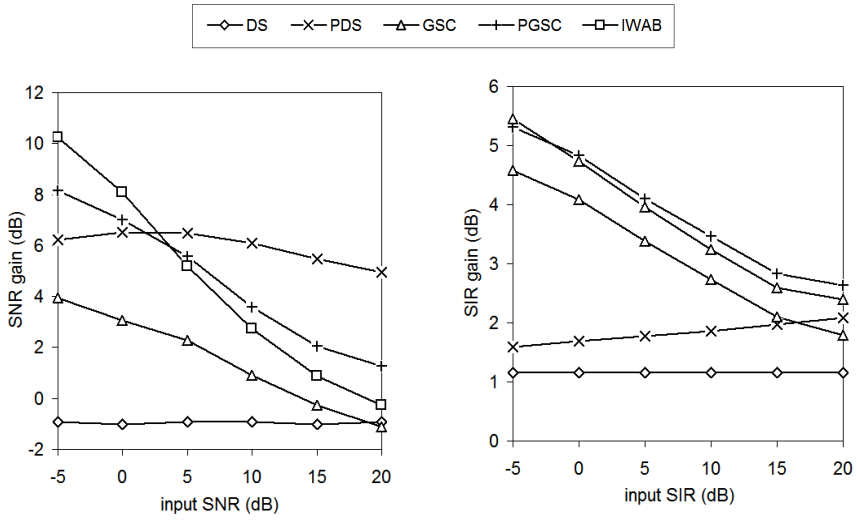

Figure 4.4: *On the left side SNR gain of the beamformers under study for different input SNRs. On the right side SIR gain of the same beamformers for different input SIRs.*

Attending to the SNR gain figure it can be clearly observed that DS is not capable of reducing noise while GSC beamformer improves when SNR is very low, this is due to the fact that ambient noise is not completely uncorrelated and in this way adaptive beamformers can reduce disturbances at the output thanks to their cancelling structure. In general, Wiener filtering based techniques (PDS, PGSC and IWAB) obtain a great noise cancellation in comparison to the DS and GSC beamformers. Concretely, PGSC and IWAB perform similarly in the sense that they obtain better results as long as input SNR decreases, and in particular PGSC obtains better results than IWAB in low noise scenarios and IWAB outperforms PGSC (and the other beamformers) in highly noisy situations, which is a positive characteristic of the proposed beamformer. On the right side, SIR gain obtained for different input SIRs and SNR equal to 10 dBs is shown. Regarding the SIR gain, it can be observed that only adaptive techniques (GSC, PGSC and IWAB) present a relevant interference reduction capability that increases when input SIR decreases. More concretely, PGSC and IWAB show a very similar gain for all the input SIRs clearly outperforming GSC without post-filtering. Although the Wiener filter is

designed to reduce the uncorrelated components of the received signal, the beneficial impact of post-filtering also with coherent noises was already documented in (Marro et al., 1998). This fact can be related to the reduction of the reverberated contribution of the interference that is usually considered to be low correlated for high frequencies.

Figure 4.5 shows speech distortion in terms of the LAR distance when only the target speaker is present and for different input SNRs. Results of the unprocessed fourth microphone of the array (MIC) are also provided. It can be first noticed that Wiener filtering based techniques importantly harm speech signal and especially when the SNR is very low since the filtering process is more aggressive in order to obtain a better noise reduction. However, although DS and GSC do not introduce such degradation, they neither show a relevant reverberation reduction capability; in fact DS shows a slightly higher distortion than the non-processed signal. Comparing the Wiener filtering techniques it can be observed a better performance of the proposed IWAB beamformer in front of PGSC and PDS beamformers. Therefore, the IWAB beamformer, like the other Wiener filtering based techniques, show an high degradation of the signal in exchange of a best performance against ambient noise and interference, but it shows a clear advantage in front of classical post-filtering of the output of a beamformer in terms of speech distortion.

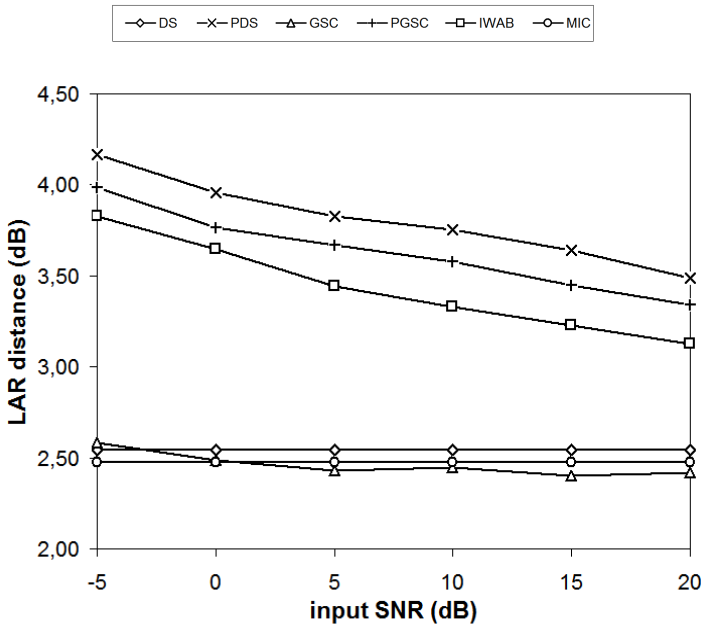

Figure 4.5: *LAR distance results of the beamformers under study and of the fourth unprocessed microphone array signal for different input SNRs.*

Figure 4.6 shows LAR distance results in presence of an interferer speaker and background noise. On the left side SIR is fixed to 10 dBs and different input SNRs are simulated, while on the right side SNR is fixed to 10 dBs for different input SIRs. Similar conclusions to the only target speaker experiment can be drawn for this case.

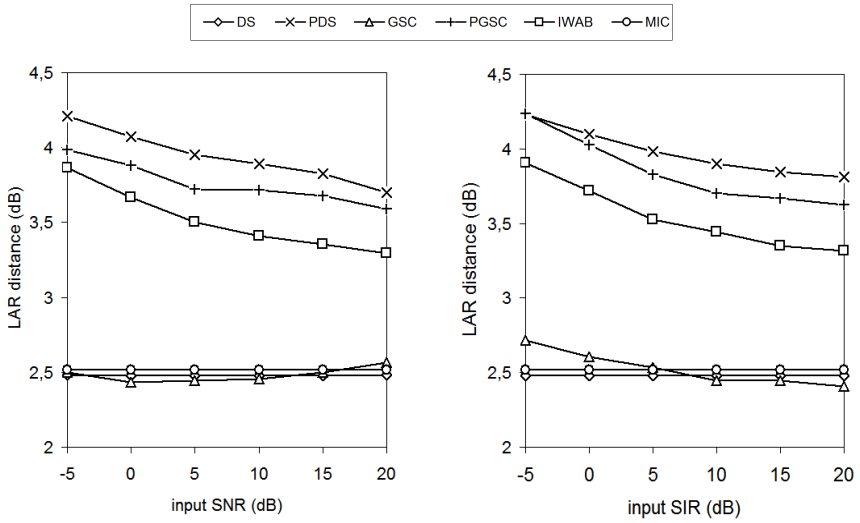

Figure 4.6: On the left LAR distance results for different input SNR and SIR 10 dBs and on the right results for different input SIRs and SNR 10 dBs.

Automatic Speech Recognition tests with simulated microphone array data

Speech recognition tests consist on the usage of the microphone array processing techniques under study as a front-end for a continuous digit recognizer. Recognition system was implemented with HTKv3.2.1 toolkit (Young et al., 2002). A 39 feature vector was used composed by the static cepstrum representation and the delta and acceleration parameters. Continuous density HMM of digits with 18 states and 3 mixtures for each state was trained with the clean speech training set of the Aurora1 database (Pearce, 1998).

In Table 4.2, columns with WER results for various SNRs and a column with the average results (AVER) are shown for all the tested beamformers and also for the fourth microphone of the array (MIC). Although the proposed beamformer is the one that obtains the highest average results, only a marginal improvement is obtained for speech recognition. In fact, in high SNR situations the proposed technique performs worse than others, but this poor performance is extensive to all Wiener filtering based techniques. Hence, it seems again that distortion introduced affects more negatively the performance of the speech recognition system than the influence of the noise reduction obtained when SNR is not low enough. In fact, these speech recognition results are well correlated with the observations of previous section about the superiority of the IWAB beamformer in very low signal to noise ratio conditions and its poorer performance in high SNR cases.

	20	15	10	5	0	AVER
MIC	42.00	44.34	50.45	62.73	82.04	56.31
DS	36.44	39.05	45.47	56.86	75.96	50.76
PDS	41.79	42.98	46.24	52.47	63.28	49.35
GSC	32.02	34.69	39.88	51.61	71.26	45.89
PGSC	39.09	41.30	45.72	52.75	64.48	48.67
IWAB	37.18	39.02	42.25	47.68	56.80	44.59

Table 4.2: *Word error rate (%) speech recognition results of the beamformers under study and the fourth unprocessed microphone array signal.*

Speech enhancement experiments with real microphone array data

Experiments with different acoustical and geometric conditions with real array data were carried out in order to further evaluate speech enhancement obtained by the various methods. For evaluation purposes the LAR distance have been computed as the difference between a close-talking recording ($s[n]$) used as the reference and the complete output of the processed signals ($z[n]$) arriving to the array. This difference with the previous evaluations is due to the fact that it is not possible to separately process the speech components. In this way, the results shown in this section represent not only the speech distortion introduced by beamformers, also the effect of noise and interference reduction is considered.

The first set of data was recorded in a noisy computer laboratory with an array of 3 cm of inter-sensor separation in one case and with 4 cm spacing in a second set of recordings. The target speaker was located in the broadside at 1 meter of the array. Figure 4.7 shows LAR distances obtained in this scenario for each beamformer and also for the fourth unprocessed microphone (MIC) of the array. It can be observed a clear superiority of the proposed beamformer in comparison to the other beamformers, and specially to the conventional post-filtering of the output of the GSC (PGSC). The lost of performance in the case of 4 cm separation is due to the larger divergence of the assumed plain propagation model and the real scenario, where the far-field condition is not accomplished for the high frequency components in the case of the target speaker at only 1 meter distance.

Figure 4.8 shows the results obtained with data collected in a conference room with the 4 cm inter-element spacing array for the case of a speaker located in the broadside at 1 meter and at 3 meters of the array. In these experiments the proposed beamformer is again the one that obtains the lower LAR distance results among the beamformers under study in both 1 m and 3 m distance scenarios. In the case of 3 meters distance, an expected lower performance is observed since noise and reverberation are supposed to be more harmful in this situation.

Finally, in the same scenario of the last experiment an interfering AM talk-radio has been

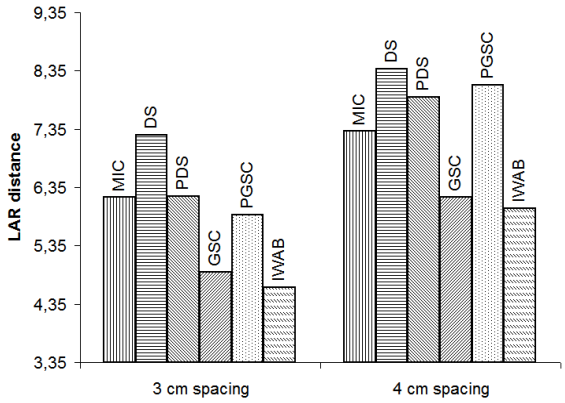

Figure 4.7: *LAR distance results for an array with element spacing of 3 and 4 cm of the beamformers under study and the fourth unprocessed microphone array signal.*

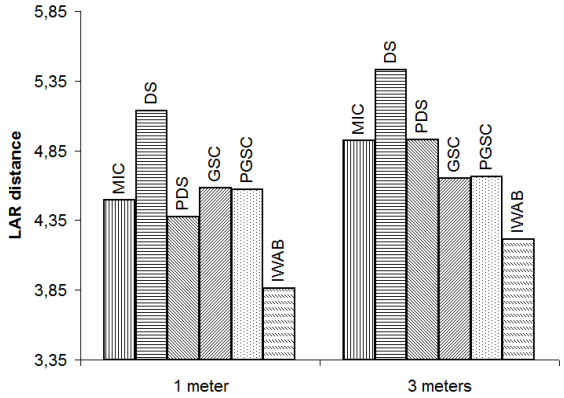

Figure 4.8: *LAR distance results for a speaker located at 1 and 3 meters of the beamformers under study and the fourth unprocessed microphone array signal.*

introduced at approximately 45 degrees off-axis from the center of the microphone array. The target speaker is in the broadside at 1 m of the array. Results shown in Figure 4.9 confirm a general better performance of the adaptive based techniques and particularly the IWAB beamformer is the one that obtains better noise and interference reduction, besides a low level of degradation considering these LAR distance measures.

4.1.4 Conclusions

A new robust beamformer with microphone arrays called integrated Wiener-filtering with adaptive beamformer has been presented. Novelty of the proposal is that a post-filtering technique is integrated in the fixed beamformer path of a GSC-like structure beamformer with adaptive

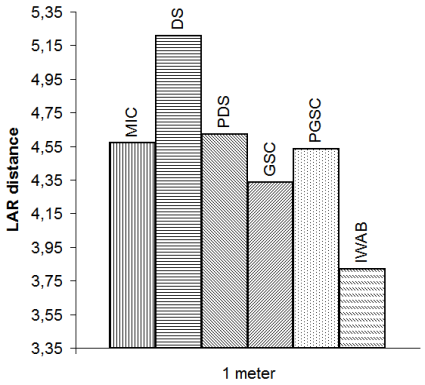

Figure 4.9: *LAR distance results for a speaker located at 1 meter and a radio jamming signal from 45 degrees competing with the target speaker of the beamformers under study and the fourth unprocessed microphone array signal.*

blocking matrix. As a consequence of this integration, a cleaner output of the fixed path is obtained, resulting in a robust to noise and to interference beamformer with a low level of speech degradation introduced by the Wiener filter.

According to experimental results, it can be stated that both conventional post-filtering of a robust beamformer (PGSC) and the proposed IWAB beamformer show similar good performances in noise and interference reduction. In low noise scenarios, the two techniques do not present a robust performance, but when noise increases their noise reduction capability turns out better. Furthermore, IWAB clearly outperforms PGSC in terms of speech distortion for various noise scenarios in the case of simulated data and obtains better speech recognition results. Attending to the results with real array data, IWAB beamformer has shown to obtain the best results among all the beamformers under study for all the scenarios.

Therefore, the proposed IWAB technique can be considered as a convenient beamformer for a wide range of conditions, particularly when noise is not very low, and in general it is preferable to conventional post-filtering of the output of an adaptive beamformer.

4.2 Development of an ASR system for a smart-room environment

There are many innovative services that can be offered in smart-rooms, for instance, lecture summarization, identification of people attending a conference, or composition of a draft about what was said in a meeting. Such highly sophisticated multimodal services are based on the information provided by many basic technological components of the various modalities. Particularly, one key functionality for most of the developed services is Automatic Speech Recognition

(ASR).

Nowadays ASR technologies perform reasonably well in controlled scenarios. However, in the context of the development of hands-free speech applications where close-talking microphones are not allowed, speech recognition becomes a difficult challenge. Audio signals recorded with microphones that can be located several meters away from the source of interest are severely degraded by noise, reverberation, and also position, orientation and dynamics of multiple concurrent speakers. As a consequence of these sources of degradation, speech recognition shows a dramatic lost of performance.

As commented in previous Chapter 3, a possible alternative to reduce problems introduced by distant microphone recordings is to use microphone arrays as a pre-processing stage in order to enhance the recorded signal that might be used for speech recognition.

In practice, although microphone arrays have shown to improve speech recognition, the gain in many cases is clearly insufficient. There are many reasons for low performance with microphone arrays. For instance, most of the conventional beamformers are not specifically designed for speech applications and particularly for speech recognition purposes (some recent alternatives where presented in Section 3.2.3). Furthermore, microphone arrays are very sensible to calibration upsets, such as different microphone gain, resulting in low performance improvement compared to single microphone approaches, which in many cases does not justify the increase of complexity due to the incorporation of sophisticated microphone array processing systems.

On the other hand, it is well-known that the larger amount of data is used for acoustic modelling the better the resulting system is expected to be. More concretely, using similar training or adaptation sets of data to testing speech will cause an improvement of the overall system performance.

In this section, first a speaker independent medium size vocabulary ASR system for distant-talking recognition in the UPC smart-room is presented. The system is derived from a free of noise dedicated baseline system for American accented speakers (De Boissieu, 2005), which is upgraded and improved by means of acoustic model re-training and adaptation methods. In fact, this resulting system is actually a fundamental technology component in the demonstrations that are usually carried out in the context of the EU funded CHIL project research activities.

Additionally, speech recognition experiments of the current on going research combining microphone array processing and speech recognition are presented. In this very recent work, the use of microphone array beamformers is proposed not only as part of the front-end of the ASR system. Rather, microphone arrays are also introduced in the process of acoustic model adaptation. In this way, one can expect to solve problems related to calibration and to signal processing mismatch that could occasion the use of microphone arrays. Similar approaches combining microphone arrays and model adaptation can be found in (Matassoni et al., 2000)

and (Chien et al., 2001) for connected digit recognition tasks.

4.2.1 Data resources

Three different sets of data have been used to build and test the ASR system: the WSJ0 database that is used for the construction of the basic models, the TED database that is used for speaker adaptation purposes and the UPC Microphone Array Database that is used for adaptation and evaluation. Additionally, the process followed to build new synthesized data is described next.

ARPA Continuous Speech Recognition Pilot Corpus (WSJ0)

The WSJ0 database (Paul & Baker, 1992) is an American speech database that contains high-fidelity speech recordings with excerpts from the Wall Street Journal. Throughout the design of the ASR system, one of the AURORA short development sets containing 7183 utterances of clean speech with 83 speakers and over 14 hours of data is used to design a medium sized vocabulary speaker independent recognizer for particular use of US English native speakers. In addition to the speech data, the database contains complete orthographic transcriptions and complete bigram language models for the WSJ text data from which the prompting was taken. For the test set, an AURORA short set was chosen, with 330 clean speech utterances with 10 speakers. Along the section the WSJ0 data sets will be referred to as *wsj_ctm*. The *ctm* suffix stands for the nature of clean speech material recorded with a close talking microphone.

Translanguage English Database (TED)

The Translanguage English Database (TED) (Lamel et al., 1994) is a corpus of recordings made of oral presentations given by non-native English speakers. The recordings provide a relatively large number of speakers speaking a variant of English over a relatively large amount of time (15 min each presentation) on a specific topic. Speakers with Italian, German or Japanese accents are included among others. The total number of speeches fully transcribed is thirty-nine and these are the ones that have been used in this work. In fact, the speeches were segmented into short utterances and some segments including undesired noisy events or unclear transcriptions were eliminated of the final selected data. The resulting processed corpus (*ted_ctm*) finally consists of 3488 close-talking speech utterances of about two hours and a half of total duration.

UPC Microphone Array Database

UPC Smart-Room The system developed is for use in room environments and particularly in the UPC smart-room. Thus, a database was collected in UPC smart-room shown in Figure 4.10. It is a meeting room equipped with several multimodal sensors such as microphone

arrays, table-top microphones, and fixed or pan-tilt-zoom video cameras. The room dimensions are 3966 x 5245 x 2800 mm, which correspond to x, y, z coordinates respectively, and its measured reverberation time is approximately 550 msec.

For the current work, the important sensors are the 64 element linear MARK III NIST Microphone Array and the reference close-talk microphone recordings (*upc_ctm*). The distance between the neighboring microphones of the array is 20 mm. In the experiments with a single distant microphone, channel 33 of the array has been used as the far-field microphone (*upc_ff*).

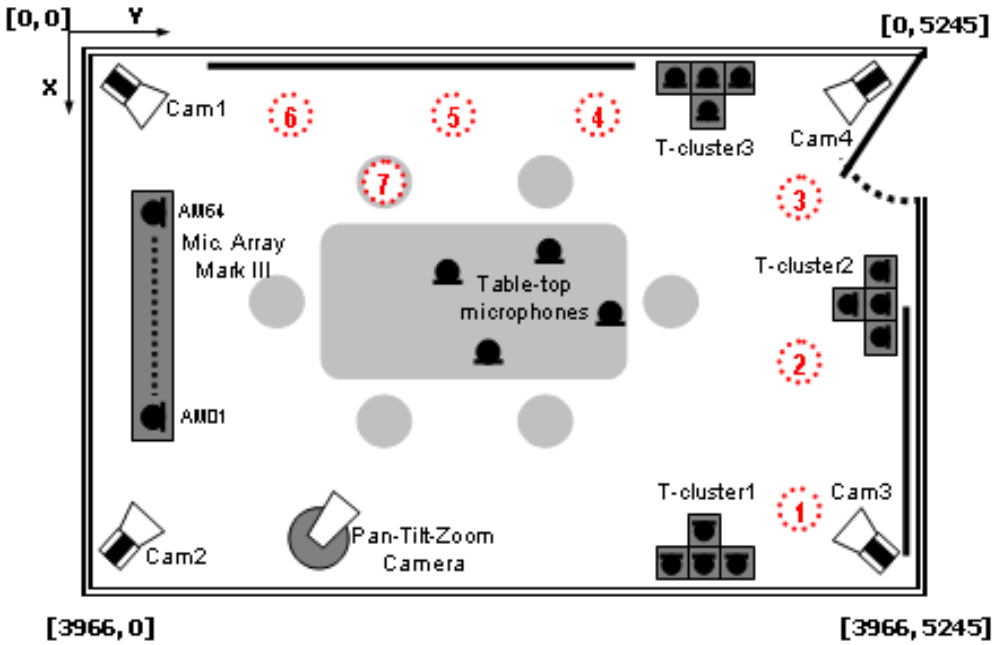

Figure 4.10: *Sensors set-up in the UPC smart-room and positions of the recorded speaker..*

Database recording UPC database was built on the WSJ0 model. The speech text was taken from WSJ0 prompts. 210 utterances were recorded for development and 277 were collected for testing. The collected data was split into three testing sessions and three training sessions to avoid environmental dependence. The recorded speaker was a male with accent very different from American, but with good level of English. The speaker read between 10 and 15 sentences in each of the seven pre-established points. In the case of points 1 to 6 the subject was stand up, while in the case of point 7 the subject was sat down in a chair in front of the table. All the signals were collected at 44.1 kHz sampling rate and 16 bits per sample.

Generation of new resources

It is well-known in automatic speech recognition that unmatched training and testing conditions yields to poor performance systems. A general problem of distant-talking recognition development is the lack of matched data to each specific application condition, which is particularly problematic when a microphone array based system is intended to be developed.

To address the problem of unmatched training and development data in some of the experiments described below, it was necessary to artificially generate synthesized data simulating the acoustic environmental conditions of the UPC smart-room. Concretely, the two main databases used for HMM construction, the WSJ0 and the TED database, were processed to generate new simulated microphone array databases.

The process followed to generate new data first consisted in the measurement of the source-to-microphone room impulse response of each microphone of the UPC smart-room, particularly of the MARK III array sensors. To achieve this, a chirp time-stretched pulse was played from a loudspeaker at several different positions in the room. The characteristic of the chirp signal is that when it is convolved with itself the result is a delta function, consequently the recorded chirp signals when are convolved with the original played signal provides a good measurement of the impulse response. In this way, multiple impulse responses from different positions to each sensor were obtained. Additionally, background noise was recorded in quiet conditions.

The various sets of simulated data were finally generated by convolving the original clean speech material with the measured impulse responses and adding background noise as follows:

$$x_q(t) = s(t) * h_{q,l}(t) + \alpha n_q(t) \tag{4.6}$$

where q is the microphone that is desired to be simulated, $s(t)$ is the clean speech material, $h_{q,l}(t)$ is the impulse response from randomly selected position l to microphone q, $n_q(t)$ is a random segment of appropriate length of noise recorded at microphone q and α is a random SNR factor constrained according to observations of SNR of real recorded data.

This process was followed to generate the distant recording WSJ0 database (*wsj_ff*) and the distant recording TED database (*ted_ff*).

4.2.2 Baseline Automatic Speech Recognition system

The baseline automatic speech recognition system is the one obtained simply on the basis of the models constructed with the WSJ0 train material. The main problems that must be addressed in the development of the distant-talking system for the robust speaker independent recognition of non-American accented speakers are derived from the results obtained with this very basic

system. Then, in subsequent sections, these problems are tackled in order to achieve a robust system for particular use in the UPC smart-room environment.

Throughout the entire work, the HTK Toolkit v3.2.1 (Young et al., 2002) has been the essential software for acoustic model construction and development of the different systems.

The front-end and back-end description provided next is common to all the evaluated systems.

Front-end

Several front-ends were tested in the various developed systems, some of them based on MFCCs (see Section 3.2.1) and others on alternative feature representations such as the Frequency Filtering (FF) parameters (Nadeu et al., 2001). Additionally, robust feature post-processing techniques like CMS and CMVN (see Section 3.2.1) were also evaluated. Although, the use of these techniques showed a positive impact on the baseline system, in practice their influence on the performance provided by the systems developed next, which incorporate acoustic model adaptation, is nearly irrelevant.

It is for that reason that in the experiments shown in this section, a simple conventional front-end is considered . Input signal was pre-emphasized using 0.97 pre-emphasis coefficient. Hamming windowed frames of 25 msec and 10 msec shift were used. Feature vectors of 12 MFCCs plus the energy of the frame, in addition to its velocity and acceleration formed the final 39 elements vector.

Back-end

Language model and vocabulary A quite general vocabulary speech recognizer with the AURORA ASR system files was built, giving a 5099 word bigrams language model. A general pronunciation dictionary was used, and non specific pronunciation adaptation was introduced in the dictionary for the case of developing and evaluating non-American accented systems.

Acoustic model training Acoustic model construction for the speaker independent system was done in various development steps. In a first step, 3 states and one mixture per state monophone Hidden Markov Models (HMMs) were trained. The second step was the enhancement of the system by converting monophones to context-dependent phones (triphones). A phonetic decision tree-based clustering method was applied to effectuate the state tying. Finally, mixture upgrading was iteratively realized until the final speaker independent tied-state context dependent triphones of 3 states and 32 mixtures per state were obtained.

Evaluation of the baseline system

Table 4.3 shows word error rate (WER, see Section 4.1.3) results of the baseline system with clean speech trained models – *wsj_ctm* (train) – and testing with both the WSJ0 selected test set – *wsj_ctm* (test) – and with the close talking microphone recording of the UPC database *upc_ctm*. Results with far-field recording are also shown to evidence the problem of the environmental mismatch. The *upc_ff* test set corresponds to recordings of one of the microphones of the array, while *wsj_ff* is the test set of the artificially synthesized WSJ0 database as it was described in Section 4.2.1

Test data	WER (%)
wsj_ctm	11.0 %
upc_ctm	51.2 %
wsj_ff	91.0 %
upc_ff	103.8%

Table 4.3: *Speech recognition word error rate with HMMs trained with WSJ0 clean data set.*

From these results, it becomes evident that two different and fundamental problems need to be addressed to develop a system that is able to perform reasonably well in our smart-room environment. On the one hand, a clear speaker pronunciation mismatch is observed comparing the results of the *wsj_ctm* and *upc_ctm* test sets. The American accented database used for model training is not in well-accordance to the accents of the speakers that are going to use more frequently the ASR system, which are mostly Spanish accented speakers. On the other hand, an even more important environmental mismatch problem is stated. When comparing results of the clean *wsj_ctm* test set and the artificially noised *wsj_ff* test set and 80 % error increase is observed. As a consequence of these two mismatch sources, the baseline system results completely useless when is used for automatic speech recognition of distant-talking speech in the UPC smart-room environment.

4.2.3 Speaker adaptation

The baseline recognition system was designed for particular class of speakers with good American accent. Due to the mismatch between the subjects present in the training data (US English) and the potential users of our application (most of them Catalan or Spanish natives), speaker adaptation (see Section 3.2.2) becomes necessary in order to get a reasonable level of performance, as it has been shown above.

Concretely, maximum likelihood linear regression (MLLR) technique is considered in this section for speaker adaptation purposes. In general, in the following speaker adaptation ex-

periments, the MLLR technique is applied off-line in a supervised and batch mode, obtaining modified HMMs for specific use. Since a considerable amount of enrollment data has been used in the two experiments described below a two pass process from more generic to more specific adaptation has been applied. First, global adaptation is generated. Then, a second pass that uses the global transformation to transform the model set, producing better frame/state alignments, is applied to estimate a set of more specific transforms using a regression class tree. The regression class tree used to group the Gaussians in the model set is up to 256 leaves. Additionally, full matrix transformations have been considered.

Speaker dependent system based on MLLR adaptation

A speaker dependent recognizer based on MLLR adaptation technique has been built in order to asses the usefulness of the MLLR approach in our system, and also in order to obtain an upper-bound performance limit in the case of perfect matched speaker conditions.

For that purpose, the train data set of the UPC Microphone Array database is used as the enrollment or adaptation data.

In Table 4.4 it can be seen how the speaker is almost perfectly adapted with *upc_ctm* adaptation data set and testing with the *upc_ctm* test set, thus obtaining almost as good performance as the reference results obtained when training and testing with the WSJ0 corpus. Almost 40% of word error reduction is obtained with new adapted speaker dependent models.

Test data	Adapt data	WER
wsj_ctm	—	11.0%
upc_ctm	—	51.2%
upc_ctm	upc_ctm	13.0%

Table 4.4: *Speech recognition word error rate comparing HMM non-adapted and HMM adapted performances.*

It must be kept in mind that the aim of the developed system is to obtain a speaker independent recognizer. The adapted models obtained by this experiment are only matched to the particular characteristics of the speaker in the UPC Microphone Array database. Consequently, a speaker adaptation processing such as the one applied in this experiment is not allowed for the purpose of the system.

Speaker independent system based on MLLR adaptation

An alternative proposal to speaker dependent adaptation is to adapt models with a whole database whose speakers should have similar characteristics to the real users of the developed

tool. Following this idea, the TED database described above has been used for supervised adaptation of the models of the baseline system. The adaptation process followed is exactly the same as the prior experiment, but in this case it is expected that the resulting system can be more adequate for the general case of non-native American speakers, rather than for a single speaker.

Table 4.5 shows comparative results testing with the clean reference of the recorded UPC database for the case of the baseline system, the baseline system adapted to the particular speaker characteristics of the UPC database and the baseline system adapted with the TED database. Although, this general speaker adaptation does not obtain such an impressive performance as specific adaptation, almost 20 % of word error reduction is achieved by using the general non-native English speaker models, which is in practice a quite significant improvement with respect to the use of the non-adapted models.

Test data	Adapt data	WER
upc_ctm	—	51.2%
upc_ctm	upc_ctm	13.0%
upc_ctm	ted_ctm	32.8%

Table 4.5: *Speech recognition word error rate with non-adapted, speaker adapted and general non-native English adaptation HMMs.*

It is likely that using a specific transcription dictionary for each speaker of the TED database taking into account the possible characteristic pronunciations would provide an enhanced model set for speaker independent recognition of non-native US speakers. Unfortunately, only word level transcriptions were provided with the database. Additionally, specific pronunciation dictionaries for the common characteristics of Spanish and Catalan native speakers based on phonetic expertise knowledge would also likely improve the performance of the system. However, for the case of the speaker of the UPC database, some specific pronunciation information was manually generated, and in preliminary experiments the use of an adapted pronunciation dictionary did not show a significant additional improvement. It was mainly for that reason, and because generating specific adapted pronunciation for each speaker of the TED database is a tough task, that phonetic pronunciation adaptation was avoided in the experiments shown in this work and a general purpose phonetic transcription dictionary was used as commented previously in Section 4.2.2.

4.2.4 Acoustic matched training and adaptation

Regarding the performance of the baseline system when distant-talking microphones are used, it becomes clear that HMMs trained with clean speech data are deeply mismatched with the real operation conditions for which the system is intended to be used. Consequently, a more robust

set of HMMs needs to be retrained or adapted to the actual testing environmental conditions.

Robust HMM construction

The first approach to increase robustness of the baseline system to acoustic environmental mismatch would consist in retraining a new complete set of HMMs with the artificially generated *wsj_ff* data set, that imitates the UPC room acoustic conditions. For that purpose, the same building process described for the baseline system has been applied to the synthesized data.

Table 4.6 shows the results obtained with the new robust models testing with the synthesized *wsj_ff* data and the distant talking data of the UPC database.

Test data	WER
wsj_ff	39.3%
upc_ff	79.4%

Table 4.6: *Speech recognition word error rate with robust HMM set, testing with the US native speaker data set and the UPC unmatched speaker data set.*

Comparing these results to the ones shown in Table 4.3 obtained with the baseline HMM model set, it can be stated that the generation of new models on the basis of artificially simulated data is a feasible method to better match to the particular characteristics of the UPC environment and consequently, to improve distant-talking recognition. The results obtained are still far from the clean matched conditions, however the matched distant-talking models provide acceptable results in the case of matched accent speaker characteristics, and it can be considered a good starting point to develop some additional solutions to further increase robustness to adverse acoustic conditions, such as microphone array processing.

Regarding the UPC data results, although the new artificially noised models generated to match with the environmental conditions permit a considerable improvement with respect to the baseline results, it is clear that the unmatched speaker problem still persists.

HMM environment and speaker adaptation

It has been shown above that both the speaker mismatch and the environmental mismatch problems can be independently addressed based on the generation of more robust models by means of acoustic adaptation or the use of artificially generated data re-training approaches.

Actually, the objective is to develop a system for distant-talking recognition of non-American accented speakers and in this sense the two problems must be faced simultaneously.

In previous Chapter 3, it was commented that acoustic model adaptation techniques, such

as the MLLR approach, can be applied to adapt to the acoustic environment, in addition to the speaker characteristics, if the enrollment data provided is representative of particular environment conditions. In this sense, the first experiment that has been carried out for comparison purposes is to develop a speaker dependent system for distant talking speech recognition. This can be naturally achieved if speaker independent acoustic models are adapted with the distant microphone recordings of the UPC database.

Table 4.7 shows the results of adapting with distant-talking UPC data compared to the speaker matched result obtained previously when robust models were built. The difference between the two speaker adaptation experiments is the starting models that were used in each case. Best results are obtained with the speaker dependent system based on robust models as it could be expected, since the robust models are more close to the environmental conditions of the testing data.

Test data	Train data	Adapt data	WER
wsj_ff	wsj_ff	—	39.3%
upc_ff	wsj_ff	upc_ff	39.0%
upc_ff	wsj_ctm	upc_ff	49.6%

Table 4.7: *Speech recognition word error rate comparing the US native matched noised HMMs system and specific MLLR adaptation of the unmatched speaker stating from baseline clean models and noised models.*

Attending to the results, it is shown that MLLR technique is able to perform reasonably well for adaptation of speaker and acoustic environment characteristics. In fact, the speaker dependent system that adapts the robust HMMs shows a slight better performance than the speaker matched case. With regard to the clean speech results of Table 4.4, MLLR was also able to greatly improve the system performance, but it was not able to obtain such a good performance as in the speaker matched case. This different behaviour can be due to the approximation carried out to synthesize the distant-talking data, which was experimentally tested to provide data slightly more noisy than the real recorded data.

With respect to the problem of generalizing the above system to the case of speaker independent recognition of non-American accented users, the solution adopted is equivalent to the one described previously for building an speaker independent system based on MLLR adaptation. In this case, artificially generated distant-talking data from the TED database is used for adaptation of the noise robust HMMs.

Table 4.8 shows the results of the distant-talking systems in US native matched speaker conditions, unmatched speaker conditions and when adapting with the synthetic *ted_ff* data set. Again a performance improvement is obtained due to adaptation with non-native English

speakers, although results are still poorer than the ones of the matched speaker case or the ones obtained with the speaker dependent adapted system.

Test data	Train data	Adapt data	WER
wsj_ff	wsj_ff	—	39.3%
upc_ff	wsj_ff	—	79.4%
upc_ff	wsj_ff	upc_ff	39.0%
upc_ff	wsj_ff	ted_ff	57.2%

Table 4.8: *Speech recognition word error rate of the artificially noised HMM system tested with the US matched speaker, the UPC unmatched speaker, adapting and testing with the UPC speaker, and adapting with the artificially noised TED database and testing with the UPC unmatched speaker.*

An example of a future smart-room service

One of the alternative technological components that has been successfully developed in the UPC smart-room is the speaker identification tool (Luque et al., 2007). In fact, the possibility of knowing the identity of the actual active speaker would permit improving speech recognition in various ways. For instance, it would be possible to apply unsupervised on-line MLLR adaptation with the speech data that is recognized to obtain an adapted set of models to the particular characteristics of each speaker. A general drawback of unsupervised model adaptation is that it is necessary to have a set of well-trained initial models in order to perform properly. To assess the plausibility of this approach, the developed robust distant-talking model set for non-native US English speakers have been used for simultaneous recognition and unsupervised adaptation. Thus, a 46.0% WER was obtained, which is more than an 11% error reduction with respect to not applying unsupervised adaptation. Then, it is shown that in a smart-room environment as the one at the UPC, the developed set of models can be efficiently used jointly with speaker identification to obtain more specific HMM sets by means of unsupervised adaptation.

4.2.5 Impact of beamforming on ASR

Throughout previous sections a baseline automatic recognition system, initially designed for a particular class of speakers with good American accent and for close talking microphone recognition, has been generalized to a speaker independent system for non-native American speakers and for distant talking recognition in a smart room environment. By means of acoustic model adaptation and with the aid of simulated data, the initial WER obtained with the baseline system (103.8 %) is reduced to 57.2% in a medium size vocabulary task.

The remarkable error reduction achieved with this robust set of HMMs can permit developing successful speech recognition low perplexity tasks (as many of the dialogue demonstrations that

are carried out in the UPC smart room). However, in more complex tasks, distant-talking recognition is still far from close talking speech recognition. In order to further increase robustness to noise and reverberation, which are the two main degrading effects of the far-field recordings, many different techniques have been proposed and some of them have been commented in previous Chapter 3.

In order to take advantage of the availability of multiple microphones in the UPC smart-room, and particularly to benefit from the presence of a 64 linear microphone array, the incorporation of microphone array beamforming techniques to automatic speech recognition is investigated in this section. Then, the objective is to provide an insight on the convenience of introducing beamforming into the recognition systems described above. In fact, the experiments presented in this section are the first results obtained in the context of the current research that is being carried out by the author in the field of robust speech recognition with microphone arrays. Microphone array beamforming algorithms and strategies are first evaluated on speaker dependent systems. Thus, the speaker mismatch and environmental mismatch problems are separated, which permits evaluating the benefits of array signal processing algorithms more accurately. Then, recent results incorporating beamforming to the speaker independent system are also provided.

Clean models vs robust models

The first question that arises when applying beamforming approaches to speech recognition as a pre-processing step is to know whether the degree of speech enhancement provided is adequate to use the testing beamformed data with models trained with free of noise data. Thus, Table 4.9 shows the comparative performance of the close-talking and distant-talking speaker dependent systems (adapted to the speaker of the UPC database) described above when testing with matched environment data (*upc_ctm* and *upc_ff*) and beamformed data (*upc_ds*) with a simple delay-and-sum of the 64 channels of the array.

Test data	Train data	Adapt data	WER
upc_ctm	wsj_ctm	upc_ctm	13.0%
upc_ds	wsj_ctm	upc_ctm	86.0%
upc_ff	wsj_ctm	upc_ff	49.6%
upc_ds	wsj_ctm	upc_ff	47.4%
upc_ff	wsj_ff	upc_ff	39.0%
upc_ds	wsj_ff	upc_ff	40.4%

Table 4.9: *Speech recognition word error rate of the speaker adapted close-talking and distant-talking systems when tested with matched environment data and delay-and-sum beamformed data.*

Regarding results obtained by the system adapted with the clean reference UPC microphone

data in Table 4.9, it becomes evident that speech enhancement achieved with microphone arrays is inadequate to test this concrete speech recognition system. Beamformed data, although showing a certain degree of noise and reverberation reduction, is not well-matched with the close-talking microphone environmental characteristics, and consequently, recognition performance is far from the results obtained in this environmental matched case.

On the other hand, with respect to the recognition systems adapted with far-field UPC data, it is shown that testing with delay-and-sum data only outperforms single microphone distant talking recognition when the recognizer used is the one originally trained with the *wsj_ctm* train corpus. When beamformed data is used to test the system originally trained with the artificially noised *wsj_ff* corpus, a performance drop is observed. The different impact of the use of beamforming on the performance of the two distant-talking adapted speech recognition systems is related to the degree of adaptation of the acoustic models to the single distant microphone environment characteristics. The system trained and adapted with distant-talking data is well-matched to the UPC far-field environment, thus, achieving a better performance in the matched environmental case, in exchange of a lost of robustness when it is tested with less noisy data. Consequently, microphone arrays fail to improve the speech recognition performance of this system because beamforming provides enhanced speech data for testing, which is less matched to the particular characteristics of the acoustic models. In contrast, acoustic models of the distant-talking adapted system originally trained with *wsj_ctm* data are not so closely adapted to the distant-talking environment, resulting in "cleaner" models that can benefit from testing with microphone array beamformed data.

The claimed unclear and difficult relationship between beamforming (or more generally speech enhancement) and speech recognition accuracy raises in these last experiments. Concretely, enhancing test data can cause a performance drop on systems specifically adapted to high noise and reverberant conditions.

Testing various beamformers

Various conventional beamforming approaches have been applied as a pre-processing stage prior to feature extraction to test the two speaker dependent systems adapted with distant talking real data of the the UPC database: the adapted speech recognizer originally trained with clean speech material (*wsj_ctm*), and the the adapted system that was initially trained with artificially noised speech data (*wsj_ff*).

Some of the beamformers tested have been the delay-and-sum (*upc_ds*), a weighted delay-and-sum beamformer depending on SNR estimations of each channel (*upc_wds*), an harmonically nested beamformer similar to the one described in Section 3.1.1 and the robust GSC described in Section 4.1.1. Indeed, the performance obtained by these two last beamformers was below the

expectations and they have not been included in this section.

In the case of the GSC beamformer, low performance was partially due to the use of a simple speech activity detector based on SNR estimations for alternatively updating the adaptive blocking matrix (when speech was detected) and the multiple input canceler block (when speech was not detected). This simple detection strategy resulted in adaptation problems, and consequently in speech output degradation, that affected the performance of the recognition system. Additionally, tuning the size of the filters, the adaptation parameters or the delays described in Section 4.1.3, according to the sampling rate (44.1 kHz) should also improve the performance.

In the case of the nested beamformer, the dimensions of the array and the high sampling frequency rate did not allow to design an appropriate harmonically nested array distributed in octaves as it would be desired. Additionally, the frequency band filtering operation was done in the DFT domain (different nested beamformers applied to each corresponding frequency bin), resulting in possible discontinuity problems that could affect the results.

Additionally, experiments with post-filtering stages similar to the beamformers described in previous Section 4.1 were also completed, but they have not been included either. In general, the harmful influence of speech degradation was checked to be more important than the noise reduction ability, resulting in low speech recognition performance when beamformers based on post-filtering were used as a pre-processing stage.

Attending to the results shown in Table 4.10 and also the experiments not reported here, it is seen again that beamforming only provides enhanced recognition in the case of the speech recognition system with acoustic models initially trained with close-talking microphone data, independently on the beamformer used for testing. Consequently, it is confirmed that the HMM set that was trained with the *wsj_ff* corpus and then adapted with the *upc_ff* data set is not well-matched with beamforming, due to specific adaptation of the recognizer to the distant-talking environment.

Regarding the results of the system initially trained with the *wsj_ctm* data set, only slight improvements with respect to the distant-talking recording recognition experiment are obtained thanks to beamforming. The delay-and-sum and weighted delay-and-sum show almost equivalent performance. As commented previously, they outperformed in general the other more sophisticated beamformers evaluated.

In conclusion, it can be stated that using beamforming as a pre-processing stage prior to automatic speech recognition permits obtaining enhanced versions of speech signals impinging a microphone array, but it can have a beneficial or damaging effect on ASR performance depending on the characteristics of the acoustic model set. On the one hand, when beamformers are used with recognition systems that are adapted to clean (or cleaner) conditions than the actual testing data, an enhanced speech recognition performance can be expected. On the other hand, when

Test data	Train data	Adapt data	WER
upc_ff	wsj_ctm	upc_ff	49.6%
upc_ds	wsj_ctm	upc_ff	47.4%
upc_wds	wsj_ctm	upc_ff	47.2%
upc_ff	wsj_ff	upc_ff	39.0%
upc_ds	wsj_ff	upc_ff	40.4%
upc_wds	wsj_ff	upc_ff	40.2%

Table 4.10: *Speech recognition word error rate of the two speaker dependent systems adapted to the distant-talking environment when they are tested with beamformed data (delay-and-sum and weighted delay-and-sum).*

acoustic models are adapted to the actual noise and reverberant environment, microphone array beamforming might result in a drop of recognition performance.

Adaptation with beamformed data

It has been shown that microphone arrays have not such a beneficial impact in speech recognition performance as one could probably expect, particularly if it is taken into account that a 64 microphone array is being used. For instance, in the case of delay-and-sum only a 2.2% word error rate reduction is obtained by one of the recognizers, while even a worse performance with respect to the single distant microphone experiment is obtained by the other system.

In order to successfully apply beamforming to speech recognition systems it becomes necessary to solve the mismatch between beamforming data and the acoustic models. Then, to increase the robustness of the acoustic models and the performance of the overall system one can extend the idea of using model adaptation for environmental compensation. The fact is that if model adaptation with far-field data is shown to be useful, then it seems also possible to apply beamforming to the data that is going to be used for adaptation. That is, a well-known principle in speech recognition is recalled: training – adapting in this case – with the most similar data to the one used in testing yields to improved results. A block diagram of the proposed HMM adaptation with beamformed data for robust construction of acoustic models is shown in Figure 4.11.

Table 4.11 shows the results of the proposed idea: HMM models are adapted with data resulting of delay-and-sum processing and then the beamformed data set (*upc_ds*) is used to evaluate these supposed better matched speech recognition systems. First, in contrast to the previous experiment, the proposed approach permits improving the performance of the two speaker dependent ASR systems considered. Using beamformed data for adaptation purposes clearly outperforms the conventional application of microphone arrays as a simple pre-processing

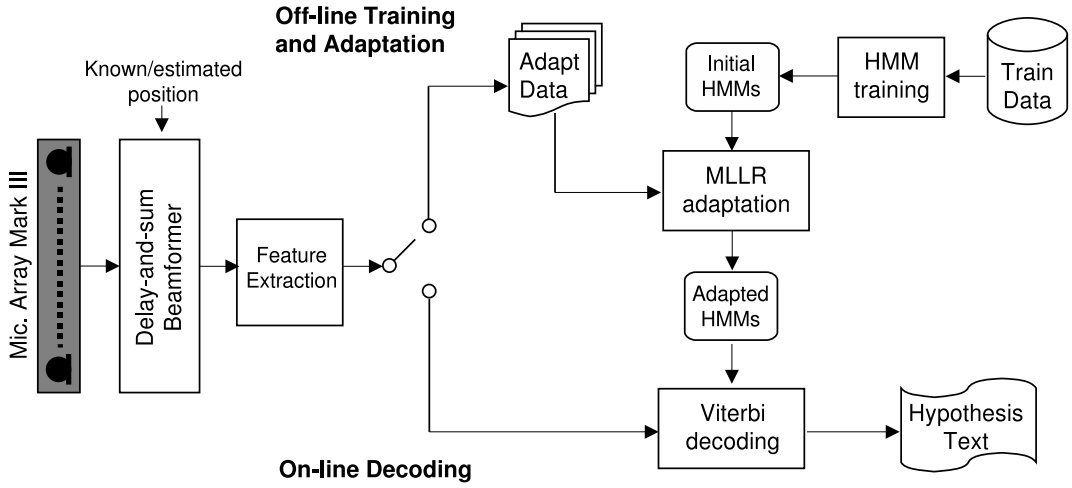

Figure 4.11: *Block diagram of the proposed scheme for adaptation with beamformed data.*

stage, independently on the ASR system. Comparing these results with the ones obtained with single distant microphone model adaptation and testing, 14.5% and 8.5% error rate reduction are achieved by the system originally trained with clean data and the system trained with noised data respectively.

Test data	Train data	Adapt data	WER
upc_ff	wsj_ctm	upc_ff	49.6%
upc_ds	wsj_ctm	upc_ff	47.4%
upc_ds	wsj_ctm	upc_ds	35.1%
upc_ff	wsj_ff	upc_ff	39.0%
upc_ds	wsj_ff	upc_ff	40.4%
upc_ds	wsj_ff	upc_ds	30.5%

Table 4.11: *Speech recognition word error rate comparing adaptation with single distant microphone data and testing with both single microphone and beamformed data, in contrast to model adaptation and testing with beamformed data.*

Hence, it is confirmed that the proposed approach combining acoustic model adaptation with microphone array processing is a convenient option to further enhance performance of automatic speech recognition systems in adverse multi-microphone scenarios, such as smart-room environments. Although, only results with delay-and-sum beamformer are shown, the adaptation scheme can be generalized to any other beamforming algorithm, resulting in general in more robust performance than conventional application of microphone arrays as part of the front-end of automatic speech recognizers.

However, two key questions still need to be clarified with respect to the proposed approach:

the influence of unknown source position and its possible generalization to the case of the previously described speaker independent speech recognizer.

Localization dependence

In all the previous experiments using microphone arrays, it was assumed that the exact position of the speaker was known, since it was manually labelled when the data was collected. However, it is evident that this is not the usual situation in real world applications and the source position generally needs to be estimated by means of some source localization algorithm. In next Chapter 5, several acoustic tracking methods are reviewed, and in Chapter 6 a robust speaker localization system is proposed. This proposed acoustic source localization system is applied in this experiment to obtain a reliable position estimation for each speech utterance of the UPC microphone array test set, and then, this estimated position is used for beamforming the testing data obtaining the new test data set referred to as *upc_dsloc*. Table 4.12 shows the recognition results of the experiment.

Test data	Train data	Adapt data	WER
upc_ds	wsj_ctm	upc_ds	35.1%
upc_dsloc	wsj_ctm	upc_ds	42.4%
upc_ds	wsj_ff	upc_ds	30.5%
upc_dsloc	wsj_ff	upc_ds	33.8%

Table 4.12: *Speech recognition word error rate of the proposed beamforming and adaptation approach tested with beamformed data based on actual labeled speaker position or estimated speaker position.*

In both distant-talking systems, the error rate increases as a consequence of targeting estimated speaker positions compared to ideal source position targeting. In the case of the system originally based on clean training data the error increase is of 7.2%, while the system based on noised training data shows an error increase of only 3.5 %. This difference of the influence of estimated speaker positions on speech recognition performances are likely related to the commented characteristics of the two model sets: the former is less robust to noise and reverberation, while the latter is more adapted to highly noise and reverberant conditions. As a consequence of wrong source targeting, the resulting beamformed speech signals are more noisy and reverberated, which affects more importantly the performance of the former recognition system.

In any case, when microphone array beamforming is considered, an error increase due to source position estimation should be expected. In fact, the performance of the proposed adaptation scheme when speaker position needs to be estimated is still better than using microphone arrays only as a pre-processing stage, even when actual position in assumed to be known beforehand. Thus, the proposed adaptation and beamforming scheme can be still considered the

most convenient option.

Generalizing adaptation with beamformed data to speaker independent ASR

The impact of microphone array beamforming approaches to the previously described robust speaker independent system for non-native US English speakers based on adaptation with the artificially noised TED database is preliminary assessed in this section.

In order to evaluate the proposed adaptation and beamforming scheme it has been necessary to generate a complete microphone array database based on the original utterances of the TED corpus simulating the 64 channels of the microphone array. Then, delay-and-sum beamforming has been applied to this artificially generated microphone array database, and the resulting data has been used for adaptation of the robust HMM set to both the characteristics of the non-native speakers and to microphone array processing simultaneously.

Table 4.13 shows first the speaker independent performance previously reported based on single distant microphone adaptation and testing, second the same system evaluated with beamformed data, and third the performance achieved when beamformed data is used for both adaptation and testing purposes, as proposed.

Test data	Train data	Adapt data	WER
upc_ff	wsj_ff	ted_ff	57.2%
upc_ds	wsj_ff	ted_ff	68.0%
upc_ds	wsj_ff	ted_ds	56.7%

Table 4.13: *Speech recognition word error rate of the proposed adaptation scheme compared to single distant-talking microphone adaptation and testing, and to conventional application of beamforming as a pre-processing step in the case of the speaker independent system for non-native US English speakers.*

Similarly to what has been observed in the speaker dependent system, the use of the delay-and-sum beamformer is not well-matched with the robust set of models. The models are particularly adapted to the noise and reverberation characteristics of the single distant microphone, and the noise and reverberant reduction achieved by the beamformer, results in a lost of performance of the recognizer. However, when beamformed data is used for adaptation, the performance of the resulting system is more robust and outperforms the single distant microphone recognition system. In this particular case, the error reduction achieved is not as significant as the ones reported with the speaker dependent systems. This lower improvement is very likely related to the use of simulated microphone array data for adaptation, in contrast to the experiments with the speaker dependent systems, where real microphone array data was used.

4.2.6 Conclusions

In this work, an automatic speech recognition system addressed to perform robustly in the UPC smart-room environment with distant talking non-native US English speakers has been developed based on both acoustic model re-training and adaptation with artificially synthesized data. The robust recognizer achieves a 57.2% word error rate, that compared to the performance of the initial baseline system, it is a considerable improvement. Additionally, if unsupervised MLLR adaptation is applied to the recognized speech utterances, even better performances can be obtained. However, unsupervised adaptation in real world applications is determined by the need of knowing who is speaking, which can be achieved in smart-room environments based on speaker identification technology.

Finally, some strategies based on microphone array beamforming have been investigated, and it has been shown that the combination of simple microphone array strategies together with HMM adaptation can improve the performance of speech recognition systems developed in smart-room environments. Using microphone array processed data for both adaptation and testing is more adequate than conventional use of microphone arrays as a pre-processing step of an automatic speech recognizer. Experiments on speaker dependent systems show a remarkable error reduction of 12.3 % and 9.9 % with respect to conventional microphone array application depending on the initial model set used prior to adaptation. Furthermore, it has been shown that the performance drop due to targeting speaker position estimation is not dramatical. Additionally, recent research confirms the potential utility of the approach when using microphone array data from multiple speakers to construct a non-native English speaker recognizer, even when this microphone array data is simulated.

Chapter 5

Audio Source Tracking and Head Orientation Estimation

The main state of the art contributions to the audio source localization and tracking task are reviewed in this chapter. Additionally, some recent works in the closely related head orientation estimation task are commented.

The application of microphone array beamforming might result useful in smart-room environments with multiple far-field microphones for both speech enhancement and speech recognition tasks as shown in the previous Chapter 3 and Chapter 4. Unfortunately, most of these techniques based on multi-microphone processing rely on one fundamental cue that is mostly unknown: the source position. The need for a reliable target position estimation in the beamforming applications is one of the reasons for the increasing interest in the acoustic source localization and tracking topic. Furthermore, accurate knowledge of the position of the events or the speakers present in a room is also useful for other multimodal services like analyzing group dynamics or behaviors, deciding which is the active speaker among all the presents, or providing information for an automatic steering camera system (Wang & Chu, 1997).

The problem of acoustic source localization and tracking on the basis of signals recorded by one or more microphone arrays can be split into three basic stages. In the first stage, estimations of such information as Time Difference of Arrival (TDOA) or Direction of Arrival (DOA) is usually obtained from the combination of the different microphones. In general, in a second step the set of relative delays or directions of arrival estimations are used to derive the 3-D source position that is in the best accordance with them and with the given geometry. In a third optional stage, a tracking of the possible movements of the sources according to a motion model can be employed.

The decision about using a concrete technique depends on many factors such as the level of noise and reverberation, the presence of single or multiple sound sources, their possible dynamics and orientations, the geometric distribution of the microphones available or the concrete purpose of application. In this sense, some applications might be interested in detecting multiple sources simultaneously or even detecting acoustic events such as door-slams that highly differ from speech characteristics.

In the particular case of person tracking applications, the estimation of the head orientation has recently emerged as an interesting field of research. Both knowing the position and the orientation of the persons located in a room can be exploited by scene analysis modules to better understand what is happening in the room. Thus, enhanced versions of the speech of each speaker can be obtained thanks to an adequate selection of microphones involving both position and orientation cues.

In this chapter, it is intended to give a representative overview of the main contributions in the field of audio source localization and tracking organized according to the three basic problems commented. Additionally, the closely related head orientation estimation task will be also reviewed.

5.1 TDOA and DOA estimation approaches

There are basically two observable characteristics of the signals arriving to microphones that can be directly estimated to infer information from the acoustic source localization: the Direction of Arrival (DOA) to a microphone array and the Time Difference of Arrival (TDOA) between a pair of microphones.

On the one hand, DOA estimation is solved based on the array signal processing theory introduced in Chapter 2 that mostly has been developed for narrow band signals located in the far-field. On the other hand, TDOA estimation is generally solved based on Time Delay Estimation (TDE) techniques which are usually considered to be the most convenient solution for the particular speech characteristics.

In general, DOA techniques are not as popular as TDOA techniques for speech applications mainly because they are generally more complex, present more geometric restrictions and specially because it is difficult to adapt the particular characteristics of the speech signal to the theoretical array signal model that is assumed. However, the great impact of array processing in other fields like sonar or radio-communication (Krim & Viberg, 1996), the possibility of detecting multiple sources or in some cases the theoretical superiority in terms of resolution and precision are some of the motivations that makes of these techniques an interesting trend of research. In fact, some recent appealing works have combined efficiently both approaches to derive high

performance audio localization systems.

5.1.1 Direction of Arrival estimation

In general, all the microphone array signal processing techniques for DOA estimation are based in the two fundamental assumptions already commented: the far-field and the narrow band condition.

Most popular DOA estimation techniques found in the speech related literature belong to one of the two already commented main categories: Beamforming based methods, also known as steered response power techniques (SRP), and the subspace based methods, that could be named High-Resolution Spectral Estimation (HRSE) techniques.

In the first case it is possible to apply narrow band beamformers to the broad band problem with the consequential known problems of variable spatial resolution and possible spatial aliasing at the high frequency range. Alternatively, to solve the broad band problem the speech signal can also be split in sub-bands as explained in Chapter 2. In this way, DOA techniques can be applied independently to each sub-band as if they were narrow band signals. As a consequence of this frequency division an independent DOA estimation for each sub-band can be obtained, but commonly it is preferable to obtain a unique coherent estimation applying focussing transformations (Hung & Kaveh, 1988; Krolik & Swingler, 1990). Alternatively to the focussing transformations it is also possible to apply sub-optimal incoherent solutions like averaging the covariance matrices or even averaging the resulting spatial spectrum of each frequency bin.

Steered-beamformer techniques

SRP techniques are based in the exploration of all the space positions with a steered beamformer in order to estimate the direction with more energy. For consistency with traditional literature, these techniques have been presented as DOA estimation techniques in Chapter 2 and in this section. However, it must be understood that the idea of space exploration can be further extended to time delay estimation or even position estimation, rather than only direction of arrival estimation. This can be done by simply adjusting the weights of the beamformer used depending on the time delays of the potential positions of the source explored.

To better understand this idea of exploration one can imagine the simplest case of a delay-and-sum beamformer of two microphones used for localization. In this case, the two captured signals are delayed and summed to compute the energy at different potential time delays of arrival resulting in a functional of the time delay. However, the time delays can be mapped to potential directions of arrival assuming far-field propagation, thus resulting in a functional of the

DOA as in Equation 2.27. Alternatively, the location of the sound source can be constrained to the set of all points in space corresponding to the given TDOA, that in this case is an hyperbola in two-dimensional space.

Consequently, SRP approaches can be viewed as a problem of finding the steering position that maximizes any Spatial Likelihood Function (SLF) $F(\mathbf{x})$, which must be a map of likelihoods for each position $\mathbf{x}$ in space, for instance the energy received by a delay-and-sum beamformer. It can be expressed as:

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} F(\mathbf{x}) \tag{5.1}$$

This interpretation of the SRP techniques permits defining flexible Spatial Likelihood Functionals combining several microphone arrays and will be further commented below because it is the basis of some of the most successful current localization systems.

Some recent works in the field of speech applications can be found in the literature for speaker localization based on beamforming techniques. As part of a complete system for localization and tracking non-adaptive delay-and-sum is proposed in (Ward & Williamson, 2002) for DOA estimation previously to the tracking stage, while in a similar work (Warsitz et al., 2004) a filter-and-sum adaptive beamformer is used to derive speaker localization information from the filter coefficients. Also speaker localization using a steered filter-and-sum beamformer is proposed in (Strobel et al., 1999) that implements a ML steered beamformer such its output energy is obtained by sampling cross-correlation functions and summing the results.

HRSE based techniques

As shown in Section 2.2.2, HRSE techniques are based on the decomposition and exploitation of the properties of the covariance matrix and the main representing technique of this group is the MUSIC algorithm (sch,).

In the speech related literature one can find some recent works based on HRSE techniques for localization. In (Choi et al., 2004) incoherent wide band MUSIC is used as part of an audio-visual system for separation of multiple concurrent speakers. The same approach is considered in (Potamitis et al., 2003a) in the development of a multi-speaker tracking system that considers different kinds of motion and sudden change in their course. Multi-source localization in reverberant environments is also addressed in (Di Claudio et al., 2000) based on DOA estimation derived from the application of the ROOT-MUSIC algorithm (Rao & Hari, 1989) as a previous step to the clustering of the raw DOA estimates to obtain candidate source positions. To finish with this review of the works related to the subspace based methods an interesting contribution for DOA estimation of speech with two channel arrays can be found in (Hioka et al., 2002). In

this work it is proposed to generate virtual multi-dimensional array data extracting the harmonic components of the estimated fundamental frequency and forming a new data vector that can be directly applied to narrow band MUSIC after some transformations of the covariance matrix.

5.1.2 Time Difference of Arrival estimation

There are several techniques that have been used in the last two decades to solve the problem of Time Delay Estimation and more concretely to estimate the time difference between a pair of signals. Earlier approaches in the field of sonar were based on the usage of LMS adaptive filters to estimate delay between a pair of signals (Reed et al., 1981; Youn et al., 1982).

The approaches that have become most common and popular in speech applications are those based on the computation of cross-correlation, which deserve to be further explained later due to their relevance in current state of the art localization systems.

Recently, a promising multi-channel cross-correlation method that generalizes the idea of cross-correlation coefficient between two signals to the multi-channel case is proposed in (Chen et al., 2003). By using the notion of spatial prediction and interpolation a multi-channel spatial correlation matrix is deduced, and then used for time delay estimation. Thus, the estimator can take advantage of the redundancy when more than two microphones are available to better cope with noise and reverberation.

Some alternative approaches not related to conventional cross-correlation for TDE have been also recently proposed. In (Benesty, 2000) it is intended to blindly estimate the propagation channel in order to properly model reverberation and, in this way, to obtain TDOA from the differences of the main peaks of the estimated impulse responses. In (Chen et al., 2004) the same blind technique is generalized to the multi-channel case. Apparently, these algorithms can perform robustly in high reverberant environments, since they do not assume simple delaying propagation channel taking fully into account the reverberation .

A nice review and comparison of the most successful TDOA techniques for audio applications can be found in (Chen et al., 2006).

Correlation based techniques

Correlation based techniques are doubtless the most popular and widely used approaches to TDOA estimation in speech applications. These efficient techniques are based in a simple reasoning as follows. Given the auto-correlation $R_{x_1x_1}(\tau)$ of a signal $x_1(t)$ with Fourier transform $X_1(f)$ it can be expressed in the time and frequency domain as

$$R_{x_1x_1}(\tau) = \int_{-\infty}^{+\infty} x_1(t)x_1(t-\tau)dt = \int_{-\infty}^{+\infty} X_1(f)X_1^*(f)e^{-j2\pi f\tau}df \tag{5.2}$$

It becomes evident from this expression that the peak of the auto-correlation function is obtained when time lag τ is equal to 0 and corresponds to the energy of $x_1(t)$. As a consequence, if we define the cross-correlation between two signals $x_1(t)$ and $x_2(t)$ as

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{+\infty} X_1(f)X_2^*(f)e^{-j2\pi f\tau}df \tag{5.3}$$

and we assume that $x_2(t)$ corresponds to a delayed version of $x_1(t)$, the cross-correlation should ideally generate a peak in the correlation function corresponding to the delay between the pair of signals:

$$\tau_{12} = \arg\max_{\tau} \int_{-\infty}^{+\infty} X_1(f)X_2^*(f)e^{-j2\pi f\tau}df \tag{5.4}$$

In the case of the signals captured by a pair of microphones in a real room environment, the presence of disturbing factors such as noise, path differences from the audio source to the microphones and the effect of reverberation can considerably mask this peak (Champagne et al., 1996). In order to increase robustness against these factors, the cross-correlation function is usually weighted with a frequency dependent function attending to different optimality criteria, in what is named generalized cross-correlation (GCC) (Knapp & Carter, 1976):

$$\tau_{12} = \arg\max_{\tau} \int_{-\infty}^{+\infty} \psi(f)X_1(f)X_2^*(f)e^{-j2\pi f\tau}df \tag{5.5}$$

It is worth noting at this point that maximizing the cross correlation in 5.4 is equivalent to time-shifting $x_2(t)$ by τ and adding it to $x_1(t)$, and maximizing the energy in the resulting signal. In other words, it is equivalent to the example of SRP with a delay-and-sum beamformer of only two microphones and for this reason, the technique in 5.4 is said to maximize the delay-and-sum energy. In the same way, the GCC technique in 5.5 can be interpreted as the maximization of the energy of a filter-and-sum beamformer. In this respect, the weight function $\psi(f)$ can be viewed as a pre-filter that is applied to the two observed spectra, before carrying out the standard delay-and-sum energy maximization.

Besides the cross-correlation in 5.4, that can be considered as unweighted cross-correlation (UCC) with weight $\psi_{UCC}(f) = 1$ for every frequency, the two functions that have been more extensively used in audio processing are the Maximum Likelihood (ML) and the Phase Transform (PHAT) weighting functions.

In low noise and reverberant environments, the use of the UCC is justified since the importance of each frequency depends only on its energy. However, in environments characterized by the presence of high additive noise, the ML weighting function is more convenient if reliable estimations of the additive noise spectrum corrupting each microphone are available. The ML weights are defined as:

$$\psi_{ML}(f) = \frac{|X_1(f)||X_2(f)|}{|X_1(f)|^2|N_2(f)|^2 + |X_2(f)|^2|N_1(f)|^2} \tag{5.6}$$

where $N_1(f)$ and $N_2(f)$ are the noise spectrums affecting $X_1(f)$ and $X_2(f)$ respectively. The main drawback of this technique is that in many cases it is not possible to obtain reliable prior knowledge of noise spectra. Furthermore, although noise is a very important source of disturbance, in real room applications the most important harming effect is usually the reverberation and the ML weight is not particularly efficient in highly reverberant scenarios.

Actually, the most commonly used weighting function that has become almost standard in acoustic event localization is the Phase Transform, also known in the literature as Crosspower-Spectrum Phase (CSP) (Omologo & Svaizer, 1994; Omologo et al., 1997) technique:

$$\psi_{PHAT}(f) = \frac{1}{|X_1(f)||X_2^*(f)|} \tag{5.7}$$

The PHAT weighting function has several advantageous properties that makes it a very appropriate choice for TDE. First, its simplicity: the function does not depend in noise spectra. In fact, this simple weight implies a normalization of the speech spectra making the estimation of the delay only dependent on the phase components of the spectrum and not on its magnitude, which is convenient following the reasoning that the delay information is retained in the phase. Moreover, speech (which is the usual input in audio localization) can show some periodicity (voiced sounds) causing unwanted periodic maximums in the cross-correlation and the PHAT function whitens the spectra avoiding this problem. Finally and most importantly in reverberant environments, the equal importance of each frequency makes PHAT a technique robust to reverberation under the assumption that the signal to reverberant ratio (SRR) is constant for all the frequencies. This assumption is very reasonable, since the amount of reverberation at each frequency is usually considered to be directly proportional to the signal energy at that frequency.

In fact, most of the more recent contributions are modifications or enhancements of the GCC technique, and specifically of the PHAT weighting technique. In (Yan et al., 2003) the GCC-PHAT computation is divided into two different processing for the high and the low band frequencies, in (Kwon et al., 2004) GCC-PHAT function is weighted according to the phase linearity in each band, and a modified version that takes into account SNR variations for various

frequencies is proposed in (Rabinkin, 1998). Also inspired in the GCC, in (Brandstein, 1999) it is attempted to exploit the estimated periodicity of harmonic spectral intervals to design a GCC filter appropriate for the combination of noise and multi-path signal distortions. In (Raykar et al., 2003) it is suggested to compute cross-correlation between features corresponding to the excitation source information as it is expected to be less affected by degradation, instead to the spectral content of the speech signal.

5.2 Position estimation approaches

The availability of multiple TDOA (or DOA) estimations lead to a minimization of an over-determined and non linear error function to obtain a unique estimation of the position. The objective of this error function is to find, by means of minimization, the position from which the theoretical delays or angles are in more accordance with the ones estimated. For instance, in the case of obtaining P TDOA estimations from P different microphone pairs the ML error function (Brandstein, 1995) can be expressed as:

$$E(\mathbf{x}) = \sum_{p=1}^{P} \frac{1}{\sigma_{\hat{\tau}_p}^2} |\hat{\tau}_p - \tau_p(\mathbf{x})|^2 \tag{5.8}$$

where $\hat{\tau}_p$ is the estimated TDOA of the p-th microphone pair, $\tau_p(\mathbf{x})$ is the theoretical TDOA from a potential source at position $\mathbf{x}$ to that microphone pair p and $\sigma_{\hat{\tau}_p}^2$ is the variance of the estimated TDOA. The theoretical delay $\tau_p(\mathbf{x})$ depends on the assumed position of the source $\mathbf{x}$, the known positions $\mathbf{x}_{p_1}$ and $\mathbf{x}_{p_2}$ of the two microphones of the p-th pair and the speed of sound c as follows

$$\tau_p(\mathbf{x}) = \frac{|\mathbf{x} - \mathbf{x}_{p_1}| - |\mathbf{x} - \mathbf{x}_{p_2}|}{c} \tag{5.9}$$

Finding the most likely position given a set of TDOAs, consists in minimizing 5.8,

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} E(\mathbf{x}) \tag{5.10}$$

One can find in the literature several approaches to solve Equation 5.10. Some of them will be commented below classified into direct approaches or closed-form approximations classes. Alternatively, in Section 5.1.1 was already commented that it is possible to find the position of a sound source based on the maximization of a Spatial Likelihood Function. These techniques will be referred as SRP-based approaches and will be also reviewed.

5.2.1 Direct approaches

Possible direct approaches to solve the minimization problem of Equation 5.10 are based on efficient iterative search algorithms. For instance, the gradient-descent (Widrow & Stearns, 1985) is a well-known iterative algorithm for minimization that can be applied to this problem:

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mu \nabla E(\mathbf{x}) \tag{5.11}$$

where $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ are the successive estimates of the source position at time instant t and $t-1$, μ is the actualization step and $\nabla E(\mathbf{x})$ is the gradient of the error function $E(\mathbf{x})$.

In (Yu & Silverman, 2004) the simplex method is also applied to solve Equation 5.10. Unfortunately, these direct methods in some cases present drawbacks as the need of an initial estimate close to the real position to avoid problems of instability, getting to local minimums or slow convergence, that means that it would be necessary a lot of iterations and as a consequence a lot of TDOA estimations to obtain a reliable position estimate.

5.2.2 Closed-form approximations

Alternatively to the iterative methods, some closed-form estimators that approximate to a sub-optimal solution at an inexpensive cost have been developed to solve the minimization problem. Many different approaches can be found in the literature and they basically differ in the degree in which they approximate to the ML solution. For example, in (Smith & Abel, 1987) it is presented the spherical-interpolation method that estimates position through a linear least-squares matrix solution derived from the range-difference measurements that can be directly obtained from TDOA estimations. In that work the proposed method is compared showing a significant better performance to a similar approach named spherical-intersection (Schau & Robinson, 1987) and to the plane-intersection (Schmidt, 1972) method in which the source location is found as a focus of conic specified by the sensor locations and the range-difference measurements. More recent contributions are proposed in (Chan & Ho, 1994) in which source localization is based on intersections of hyperbolic curves defined by the TDOAs estimated and showing a better performance than the spherical-interpolation method. In (Brandstein et al., 1995) a simple method for 2-D position estimation is proposed together with a frequency domain time-delay estimator. It basically consists on the average of the best candidate crossing points of the various bearing lines, derived from the angle estimation of each microphone pair, according to an error estimation of each potential crossing point. In (Brandstein et al., 1997) a method called linear-intersection approximates closed-form ML solution for 3D localization in rooms with square shaped microphone clusters showing also better performance than the spherical-intersection method.

In general, most of these source-location procedures are very sensitive to errors in the delay estimate. In real reverberant room applications with distant sources their performance is not satisfactory enough and the SRP-based approaches commented next are usually preferable.

A simple crossing bearing line example

A more detailed description of the technique proposed in (Brandstein et al., 1995) for source localization is provided here as an example of these techniques and because it will be considered in some of the experiments of the next chapter.

The algorithm is originally based on the TDOA estimations provided by a new time delay estimator proposed in (Brandstein et al., 1995). However, the TDOA estimations can be obtained by means of any other technique, for instance the GCC-PHAT described above.

First, the delay estimate $\hat{\tau_p}$ and the corresponding source bearing angle $\hat{\varphi}_p$ of each adjacent microphone pair p are computed. A bearing line for each microphone pair is obtained with angle $\hat{\varphi}_p$ that crosses the middle point of the corresponding microphone pair. Then, a set of candidate points is obtained as the intersection point of all pairwise bearing lines. In fact, it is better only to cross the bearings of microphone pairs placed in different surfaces or walls, for instance if P microphone pairs are distributed in two walls ($P/2$ each one), the total number of candidate points would be $P^2/4$. To keep generality we will assume that the number of candidate points is K. For each k candidate point (x_k, y_k) a weighted error estimate is evaluated as follows

$$E(x_k, y_k) = \frac{\sum_{p=1}^{P} w_p(x_k, y_k)\Delta_p(x_k, y_k)^2}{\sum_{p=1}^{P} w_p(x_k, y_k)} \tag{5.12}$$

where $\Delta_p(x_k, y_k)$ is the orthogonal distance from bearing line p to candidate point k. Using trigonometric relations it can be computed as $\Delta_p(x_k, y_k) = R_p(x_k, y_k)sin(|\hat{\varphi_p} - \phi_p(x_k, y_k)|)$, where $R_p(x_k, y_k)$ is the distance to the candidate position from the microphone pair midpoint and $\phi_p(x_k, y_k)$ is the bearing angle of a candidate point relative to the microphone pair p. The weighting coefficient for the microphone pair p, $w_p(x_k, y_k)$, is the reciprocal of the estimated spatial variance for that microphone pair

$$w_p(x_k, y_k) = \frac{1}{var\{\Delta_p(x_k, y_k)\}} \approxeq (\frac{R_p(x_k, y_k)(c/f_s)cos(\hat{\varphi_p} - \phi_p(x_k, y_k))}{d_p sin\hat{\varphi}_p})^2 var\{\hat{\tau}\} \tag{5.13}$$

where d_p, c, f_s and $var\{\hat{\tau}\}$, are the inter-microphone distance of pair p, the velocity of sound, the sampling frequency and the variance of the time estimate respectively.

The final estimate position $(\hat{x}, \hat{y})$ is obtained as the average of the $K/2$ candidate points

with the lowest error and a a location error estimate is also taken by means of the average of the $K/2$ corresponding error estimates. A summary of the algorithm is given in Table 5.1.

Crossing lines algorithm
1. Time delay and angle estimation of each selected microphone pair for each analysis frame. 2. Compute the lines with the angle estimations that crosses the midpoints of the microphone pairs. 3. Compute candidate points from the crossing lines and compute the error associated o each candidate. 4. Final estimation (and error) obtained with the average of the half of the points with less error.

Table 5.1: *Summary of the basic steps of the localization algorithm described in Brandstein et al. (1995).*

5.2.3 SRP-based approaches

In previous section was shown that steered-beamformer approaches for DOA estimation can be generalized and interpreted as a problem of maximizing a likelihood map of every possible exploration position (Mungamuru, 2003). This interpretation gives us a flexible manner of defining these likelihood functions, that are not only necessary to be constrained to the energy received by a beamformer at different spatial positions. It was also commented that finding the delay for which the cross-correlation is maximized for a pair of microphones is equivalent to explore with a delay-and-sum beamformer with two microphones. Thus, if we have several pairs of microphones a possible likelihood map can be defined as the sum of the contributions of the cross-correlation of every microphone pair to each concrete spatial position. The position that maximizes this function is the most likely position of the source. Hence, known that the theoretical delays between a pair of microphones is given by 5.9, Equations 5.1 and 5.4 can be joined for an arbitrary number P of microphone pairs, resulting in the following Spatial Likelihood Function (SLF) and maximum search:

$$F(\mathbf{x}) = F(\mathbf{T}(\mathbf{x})) = \sum_{p=1}^{P} \int_{-\infty}^{+\infty} X_{p_1}(f) X_{p_2}^*(f) e^{-j2\pi f \tau_p} df$$

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \sum_{p=1}^{P} \int_{-\infty}^{+\infty} X_{p_1}(f) X_{p_2}^*(f) e^{-j2\pi f \tau_p} df \tag{5.14}$$

That is, for each spatial position $\mathbf{x}$, the time delay vector $\mathbf{T}(\mathbf{x}) = \begin{bmatrix} \tau_1(\mathbf{x}) & \tau_2(\mathbf{x}) & \dots & \tau_P(\mathbf{x}) \end{bmatrix}$ formed by theoretical delays to each microphone pair is found and the pairwise cross-correlations are computed. The likelihood assigned to each position is equal to the sum over all pairwise cross-correlations, and the position with a maximum likelihood is the estimated source position.

The SRP-PHAT algorithm

Rather than using the cross-correlation for obtaining the SLF, any generalized cross-correlation function can be considered. Particularly, when the PHAT weighting function is used the resulting likelihood map is:

$$F(\mathbf{x}) = F(\mathbf{T}(\mathbf{x})) = \sum_{p=1}^{P} \int_{-\infty}^{+\infty} \frac{X_{p_1}(f) X_{p_2}^*(f)}{|X_{p_1}(f)||X_{p_2}^*(f)|} e^{-j2\pi f \tau_p} df \tag{5.15}$$

The technique in 5.15 is known as SRP-PHAT algorithm (DiBiase et al., 2001). The elegant and robust way in which the TDOA estimation problem based on GCC with PHAT weighting and the position estimation problem based on SRP search is jointly solved by this technique, has actually turned it out in the most successful state of the art approach to microphone array sound localization. A summary of the algorithm is given in Table 5.2.

SRP-PHAT algorithm
1. Pre-compute theoretical delays from each possible exploration position to each microphone pair. 2. For each analysis frame compute the cross-correlations of each microphone pair. 3. For each position accumulate the contribution of cross-correlations (using delays pre-computed in 1). 4. Select the position with the maximum score.

Table 5.2: *Summary of the basic steps of the SRP-PHAT algorithm.*

Many of the most outstanding recent proposals for robust sound source localization are based on the SRP-PHAT algorithm. In (Aarabi, 2003), SLFs from several microphone arrays are integrated taking into account the different level of access that a microphone array has to different positions. That is, the SLFs obtained from each microphone array are weighted depending on the distance to the position explored. In (Mungamuru & Aarabi, 2004) an enhanced approach is introduced, that takes into account other factors for the microphone level of access such as the source directivity and microphone directivity, rather than only the source to microphone distances.

Some other works are addressed to reduce the known computational expense of the SRP based methods. In (Zotkin & Duraiswami, 2004), an efficient search algorithm is developed significantly reducing complexity by using coarse-to-fine search strategies in both space and frequency domain. In (Peterson & Kyriakakis, 2005), the computational expense reduction is achieved by reducing the number of potential source points to a small set of possible positions using in a first step a fast closed-form localization algorithm.

5.3 Source Tracking approaches

In the case of multiple or even in a single moving speaker scenario it is well-known that successive raw estimation of the position does not yield in tracking of the target (Potamitis et al., 2003b; Sturim et al., 1997). Some of the reasons for this fact are the speech spectrum varying characteristics, or the noise and reverberation influence that cause the appearance of inconsistent spurious measurements also known as clutter. As a consequence, in order to obtain a reliable track of the sources present in a room it becomes necessary the use of a spatial filter according to a motion model capable of smoothing the noise-corrupted location estimates.

Traditionally the spatial filtering operation has been carried out by an appropriate Kalman filter based on a simple Newtonian model. For instance, in (Sturim et al., 1997) and (Potamitis et al., 2003a) Kalman filter is used to smooth the results obtained from a TDOA based localization system and a wide band MUSIC based localization system respectively. In these two works, the Interaction Motion Model algorithm (Mazor et al., 1998) that operates with multiple Kalman filters running in parallel is applied to let different motion models to be chosen from the observed data in order to properly model various activity states of a speaker.

Particularly, in the case of multiple sources to be tracked appears a new problem and it is that of associating each raw localization to a concrete source prior to spatial filtering. This is named data association (Bar-Shalom & Fortmann, 1988) and it can be accomplished through the use of acceptance regions and algorithms that consistently assigns measures to a concrete track.

Recently a promising alternative to the Kalman filtering approach for the target tracking task named Particle filtering (Arulampalam et al., 2002; Vermaak & Blake, 2001) has been intensively proposed. The Kalman Filter is an optimal solution to the tracking problem when the state and the measurement equations are linear and if the system and measurement noise are Gaussian, however the resulting state-space model of the speaker motion and likelihood of the speaker position based on the TDOA estimations is non-linear and non-Gaussian (see below for details of the Kalman filter). The Particle filtering technique, also known as Sequential Monte Carlo methods, allows an efficient and elegant integration of the non-linear relationship between the TDOA values and source position, providing a robust and accurate tracking operation in adverse environments. In some of the recent works Particle filtering is studied being provided by GCC estimations in comparison to delay-and-sum estimations (Ward et al., 2003), by filter-and-sum estimations (Warsitz et al., 2004) or in a multiple speaker scenario (Vo et al., 2004).

Finally, it must be highlighted that many of the recent literature about speaker tracking is based on both audio and video streams simultaneously (Strobel et al., 2001; Vermaak et al., 2001; Gatica-Perez et al., 2003). This is basically because both Kalman or Particle filtering

approaches form an easily extendable framework that makes possible the fusion of multimodal cues in a very efficient way.

Kalman filtering basics

The Kalman filter (Kalman, 1960; Welch & Bishop, 1995) is a set of mathematical equations that provides an efficient computational (recursive) means to estimate the state of a process, in a way that minimizes the mean of the squared error. Tracking moving objects in any field of digital signal processing is probably the most outstanding application of the Kalman filter. Since it can provide accurate continuously-updated information about the position and velocity of a source, given only a sequence of raw observations about its position, the Kalman filter has also successfully been used in audio source localization applications to track and smooth the position of moving speakers.

In order to use the Kalman filter to estimate the internal state of a process, that is the "true" position and velocity of a moving speaker, given only a sequence of noisy position observations, one must model the process in accordance with the framework of the Kalman filter. The Kalman filter model assumes the true state at time k is governed by the linear stochastic difference equation

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{w}_k \tag{5.16}$$

where $\mathbf{A}$ is the state transition model which is applied to the previous state $\mathbf{x_{k-1}}$ and $\mathbf{w}_k$ is the process noise. The equation 5.16 can additionally incorporate an optional control term that has been omitted for convenience and simplicity reasons (Welch & Bishop, 1995).

The observation or measurement vector $\mathbf{z}_k$ is related with the state vector as

$$\mathbf{z}_k = \mathbf{H}\mathbf{x}_k + \mathbf{v}_k \tag{5.17}$$

where $\mathbf{H}$ is the observation model which maps the true state space into the observed space and $\mathbf{v}_k$ is the measurement or observation noise vector. Both the process ($\mathbf{w}_k$) and measurement ($\mathbf{v}_k$) noises are independent of each other and are assumed to be zero mean Gaussian white noise with covariance matrices $\mathbf{Q}$ and $\mathbf{R}$ respectively. Here, for simplicity, it has been assumed that the state transition matrix, the control matrix, the observation and the statistics of the noises do not change over time.

The Kalman filter state is represented by the estimate of the state at time k $\hat{\mathbf{x}}_{k|k}$ and the error covariance matrix $\mathbf{P}_{k|k}$, which is a measure of the estimated accuracy of the state estimate. These variables are estimated in a two step recursive procedure. In a first step, the time update

equations are applied to obtain the a priori estimates for the next time step, thus it can be interpreted as a prediction step:

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{A}\hat{\mathbf{x}}_{k-1|k-1}$$
$$\mathbf{P}_{k|k-1} = \mathbf{A}\mathbf{P}_{k-1|k-1}\mathbf{A}^T + \mathbf{Q} \tag{5.18}$$

The measurement update equations are responsible for incorporating the measurement into the a priori estimate to obtain an improved posterior estimate. First, the Kalman gain $\mathbf{M}_k$ is computed with Equation 5.19. Then, the Kalman gain and the observation measurement $\mathbf{z}_k$ are used to generate the a posteriori state estimate and error covariance as described by next Equations 5.20 and 5.21. It is commonly said that the measurement equations act as a correction update step.

$$\mathbf{M}_k = \mathbf{P}_{k|k-1}\mathbf{H}^T(\mathbf{H}\mathbf{P}_{k|k-1}\mathbf{H}^T + \mathbf{R}) \tag{5.19}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{M}_k(\mathbf{z}_k - \mathbf{H}\hat{\mathbf{x}}_{k|k-1}) \tag{5.20}$$

$$\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{M}_k\mathbf{H})\mathbf{P}_{k|k-1} \tag{5.21}$$

The Kalman gain $\mathbf{M}_k$ can be interpreted as a weighting that gives more or less importance to the measured observation $\mathbf{z_k}$ or to the predicted measurement $\mathbf{H}\hat{\mathbf{x}}_{k|k-1}$, depending on the measurement error covariance $\mathbf{R}$ and the a priori estimate error covariance $\mathbf{P}_{k|k-1}$. On the one hand, when $\mathbf{R}$ approaches zero, the actual measurement is trusted more, while the predicted measurement is trusted less. On the other hand, as $\mathbf{P}_{k|k-1}$ approaches zero the actual measurement is trusted less, while the predicted measurement is trusted more.

This Kalman gain dependence gives rise to an important practical issue of Kalman filter applications: the correct determination of the measurement noise and process noise covariance matrices $\mathbf{R}$ and $\mathbf{Q}$.

In practice, it is more or less feasible to estimate $\mathbf{R}$, however, the determination of the process noise covariance $\mathbf{Q}$ is quite more difficult. Consequently, the filter performance is generally adjusted by off-line tuning of the parameters $\mathbf{Q}$ and $\mathbf{R}$. The problem of off-line tuning is that in many applications these noises can importantly change over time: the values that are found to provide enhanced performance for a concrete noisy situation may not be addequate when the scenario changes. For instance, in speaker tracking applications, the measurement noise can vary with the source position (depending on the microphone distribution) or the process noise can be very different if the speaker suddenly changes from being static to move around the room. Hence, determining these filter parameters is a tough and critical aspect for the correct performance of the Kalman filter.

To end this brief review of Kalman filter basics, it is worth noting that many real dynamical systems do not exactly fit this linear model. Because the Kalman filter is designed to operate in the presence of noise, an approximate fit is often good enough for the filter to be useful. However, in some cases variations on the Kalman filter that allow richer and more sophisticated models are necessary. For instance, this is the case of the extended Kalman filter (EKF) (Welch & Bishop, 1995), which generalizes to the case of non-linear state transition and observation models.

5.4 Head orientation estimation

Orientation estimation refers to the process of determining the direction that a sound source is facing. Although orientation estimation can be applied to any kind of directive sound source, the most typical and useful application, is that of trying to estimate the orientation of a human speaker, which one can refer to as head orientation estimation.

As already commented, knowledge about both the position and the orientation of human speakers in applications that require human-computer interaction, allows a better understanding of what users do or what they refer to. Thus, this information can be exploited by several applications, such as automatic camera steering in video conferences or enhanced microphone network management for Automatic Speech Recognition or any other speech related applications. In fact, the recent significant research efforts devoted to the development of human-computer interfaces in intelligent environments such these, has been the main reason for the recent increasing interest in the orientation estimation problem.

Actually, the problem of head orientation estimation has been mostly tackled based only on visual cues. The interest in this problem based on multi-channel speech observations is so recent and challenging that very few works can be found in the speech related literature. However, since humans rely on both acoustic and visual information to perform orientation estimation, one can think that speech should also be used to infer some information about source orientation. Thus, approaches to head orientation estimation can be coarsely classified in terms of the information they rely on, that is only visual, only audio or audio-visual approaches.

Regarding the audio based approaches, rather than stand-alone orientation estimation algorithms, most of the methods have been proposed in relation to robust sound localization systems. The main motivation is that head orientation is a factor that can degrade performance of localization algorithms. A detailed description of the effects of head orientation in two typical speaker localization algorithms will be shown in next Chapter 6.

Then, to overcome the possible degrading effect of head orientation, some robust localization proposals try to incorporate head orientation as a new search parameter, to first obtain a more

reliable source position estimation, and second to obtain also an estimation of its orientation. This is the case of (Mungamuru & Aarabi, 2004), that based on the SRP-PHAT algorithm, extends the search with the orientation variable by weighting the contribution of each microphone pair for different possible orientations. A similar approach also based on the SRP-PHAT algorithm can be found in (Brutti et al., 2005), named the Oriented Global Coherence Field (OGCF) method. An earlier work on speaker orientation estimation not related to the the extension of a robust localization system was proposed in (Sachar & Silverman, 2004), which is based on acoustic energy measurements with a large microphone array.

Chapter 6

Contributions to Speaker Tracking and Head Pose Estimation

The main proposals and the related experiments developed by the author in the context of the research activity in the fields of acoustic source localization and tracking and head orientation estimation are provided. First, the effect of head orientation is assessed in different speaker localization algorithms. Then, a complete robust acoustic source localization and tacking system is proposed. Finally, some recent approaches to the problem of head orientation estimation are described and tested.

Speaker localization is a basic functionality for computational perception of human activities in a smart-room environment. In previous chapters, it has been shown that a reliable measure of the talker position is needed for technologies that are often deployed in that environment, like for instance microphone array beamforming.

The degree of reliable information provided by speaker localization systems on the basis of the audio signals collected in a smart-room environment with a distributed microphone network is known to be dependent on a great number of factors such as environmental noise, room reverberation, talker movement and head orientation.

Concretely, the effect of the head orientation on the speaker localization performance has received few research attention. In (Abad et al., 2005), the influence of the head orientation in smart-room environment was studied by the author and it is now reviewed in next section. This work constitutes the first early approximation of the author to the problem of source localization. Two representative systems are built to investigate the effect of talkers head orientation on them. It is shown that techniques that join the estimated cross-correlations in a collaborative way, such as SRP-PHAT, are in general a convenient choice for developing a system that, in addition to being robust to noise and reverberation, is almost independent on talker head orientation, if the microphone network is distributed appropriately in the room.

In this chapter, the development of a robust audio person tracking system for smart-room environments in the context of the EU funded CHIL project research activities carried out by the author is also described. The proposed system is based on the SRP-PHAT algorithm due to its commented robust performance in real conditions. Some novelties (Abad et al., 2006), aimed to enhance the accuracy of the system independently on the application scenario and to reduce the computational complexity, are proposed and separately evaluated. Additionally, the proposed system was evaluated in the CLEAR 2006 workshop obtaining remarkable results, which are also provided (Abad et al., 2007).

Finally, last section describes the recent work carried out on head orientation estimation. Two completely different approaches –one based on source localization exploration techniques and an inexpensive method based on talker directivity properties– are presented and compared. In (Segura et al., 2007), one can find a multimodal head pose orientation system which partially relies on the simple method based on speaker directivity characteristics.

6.1 Study of head orientation influence in a smart-room environment

Head orientation is one of the major factors that influence the performance of speaker localization systems. However, the study of the effects of head orientation and its estimation is a matter of very recent interest and few works can be found addressing these problems (see Section 5.4).

In this section, we try to get an insight into the effect of talker's head orientation. The relationship of orientation with speaker directivity pattern and reverberation is first described. Then, the effect of head orientation on the performance of two selected acoustic source localization techniques in the UPC smart room environment is investigated: a steered response power based technique and a closed form localization method (see Section 5.2). The former is evaluated with different microphone arrangements to show the dependence of the localization error on the head orientation. However, the steered response technique, due to the way in which the contributions from the various microphones are combined, is shown to be more robust to head orientation changes if microphones are correctly distributed. In contrast, the closed form method is too much dependant on head orientation and on the selection of the most adequate microphones.

6.1.1 Talker directivity and reverberation: The effect of orientation

The extent of head orientation and its influence on the performance of source localization algorithms in smart room environments is related with the directivity of human speakers and the room reverberation characteristics.

On the one hand, the measurements reported in (Chu & Warnock, 2002) show that human talkers do not radiate voice sound uniformly in all directions; more energy is radiated in talker's forward direction than towards the side or the rear direction. Figure 6.1 shows the A-weighted radiation pattern of human talker in horizontal and vertical plane passing through the talker's mouth. With regard to the horizontal radiation pattern, it shows about -2dB attenuation on the side of the talker (90° or 270°) and the attenuation behind the talker (180°) is stronger than -6dB. Similarly, the vertical radiation pattern of human talker is not uniform; e.g. there is about -3dB attenuation above the talker's head. In addition, the radiation pattern is frequency dependent (Chu & Warnock, 2002); behind the talker, the low speech frequencies are attenuated less than the high speech frequencies. In other words, speech high frequency components are more directive.

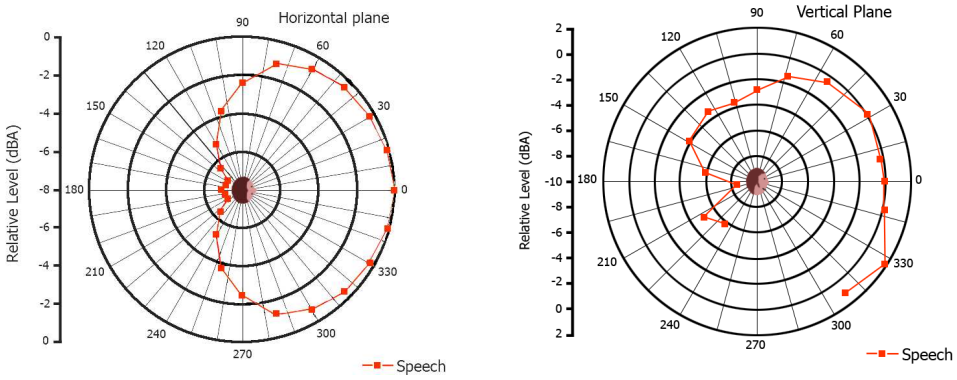

Figure 6.1: *Talker diagram in the horizontal and vertical plane (after Chu and Warnock (2002)).*

On the other hand, in Section 2.1.2 was described the effect of room reverberation. Two kinds of reflections were distinguished: early reflections and late reflections. Although, it is mostly the late reflections that smear the speech spectra and reduce the intelligibility and quality of speech signals, in speaker localization, we are not concerned about the speech intelligibility; actually, the early reflections are usually of high energy and they can arrive to the localization sensor from positions that are very different from the talker position. Figure 6.2 illustrates a typical situation where the various microphones collect not only the sound wave directly coming from the sound source (labeled by 0 in Figure 6.2) but also indirect sound waves – the ones originated at the same source and reflected from the walls or other surfaces in the room (waves 1-3 in Figure 6.2) – in addition to other possible interference signals and background noise.

In that talker-localization sensor configuration (talker is facing the closest wall), the energy of the direct wave is attenuated due to the talker's radiation pattern. But the wave number 2 before it is reflected is stronger than the direct wave number 0 and it can reach the localization sensor with a higher energy than the direct wave, depending on the absorption coefficient of the wall.

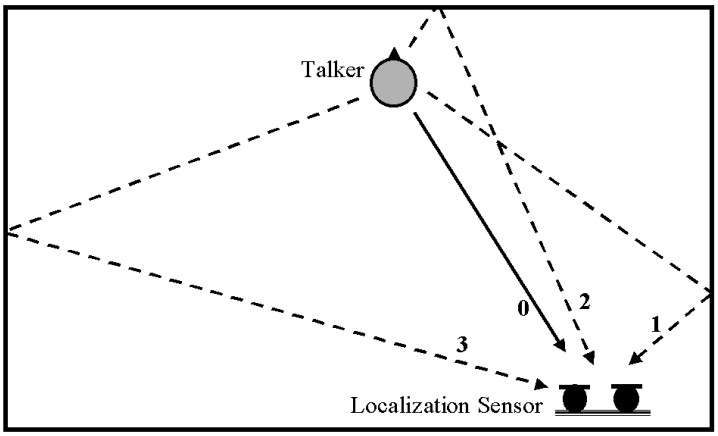

Figure 6.2: *Multi-path example where direct sound wave is 0 and reflective waves are 1, 2 and 3. In this situation, indirect sound waves could reach sensors with higher energy.*

This example illustrates that given the room properties and the positioning of the localization sensor(s), the reflected speech sound waves can represent a serious competition to the direct speech sound wave within a talker localization algorithm and they can degrade its performance. It is obvious that the the accuracy of the localization measure depends on the ratios between the energy of the direct sound wave and the energies of the indirect waves. Those ratios are affected by the position and orientation of talker's head with respect to the sensors and the reverberation properties of the room.

6.1.2 Effect of head orientation on the speaker localization performance

A small database collected in the UPC's smart room of a male subject at several positions and with varying head orientations, is used to assess the effect of head orientation on two representative speaker localization techniques. First, the SRP-PHAT technique described in Section 5.2.3 and summarized in Table 5.2, is selected to show the dependence of localization performance on the microphone network set-up and the head pose orientation with respect to the microphone distribution. Second, the SRP-PHAT technique –representative of steered response power based techniques– and the simple crossing lines (CL-PHAT) technique described in Section 5.2.2 and summarized in Table 5.1 –representative of closed form localization approaches– are tested to find out the best relative improvement that can be obtained in the case of selecting the microphones that are more adequate for source localization in our smart-room environment, assuming that the actual position and head orientation is known. Thus, the degree of dependence of this two representative techniques on talker's head orientation is assessed.

In the SRP-PHAT experiments, the size of the exploration step is 4 cm and only cross-correlations of the microphones belonging to the same T-array are used. In the case of CL-PHAT

3 bearing lines are estimated for each T-array. In both cases, the delay estimation is achieved by means of the GCC-PHAT of Hanning windowed frames of 4096 samples with a 50% overlap. Furthermore, cross-power spectral density estimations are smoothed over time with a forgetting factor equal to 0.3 (see below Section 6.2.2).

Database for talker orientation experiments

UPC Smart-Room The testing database was collected in the UPC smart-room described in Section 4.2.1.

For speaker localization, the important sensors are the three T-shaped microphone arrays and the 64 element linear microphone array (from NIST). In this study only the T-shaped microphone arrays were used (from now on referred as T-arrays). The T-arrays are placed on three different walls of the room at the height of 230 cm (see Figure 6.3). Each T-array has 4 microphones (Shure Microflex), three of them form a horizontal line with an inter-sensor separation of 20 cm and the last one is placed vertically 30 cm above the central microphone of the horizontal line.

Database recording To collect the database, a male subject was moving through a predefined set of six points depicted at Figure 6.3 (from P1 to P6). At each point, the subject stopped and uttered a few Spanish sentences facing four different orientations in such a way that he headed towards each of the four different walls of the room for 10-15 sec. The orientations are denoted as North (N), West (W), South (S) and East (E).

For the recording, the three T-arrays and a close-talk microphone were used. The signals coming from microphones were acquired by Hammerfall audio card at 44.1 kHz sampling rate. The position of subject was marked manually by listening to the audio signal and using the coordinates of the predefined six points (the precision of a reference obtained by using a recording from a calibrated Zenithal camera was not satisfactory). The z coordinate was added based on subject's height.

Experimental results and discussion

In a first set of experiments the SRP-PHAT technique is used to show the importance of orientation on source localization performance.

Figure 6.4 shows the distance error of the SRP-PHAT position estimation at the point P4 for the all four considered orientations (N, W, S, E). The three different error curves correspond to using a) only the T-array from the S wall (upper curve), b) the two T-arrays from the S and E walls (middle), and c) the all three T-arrays (bottom). The dashed line in each graph denotes the threshold when the localization is considered as correct (30 cm).

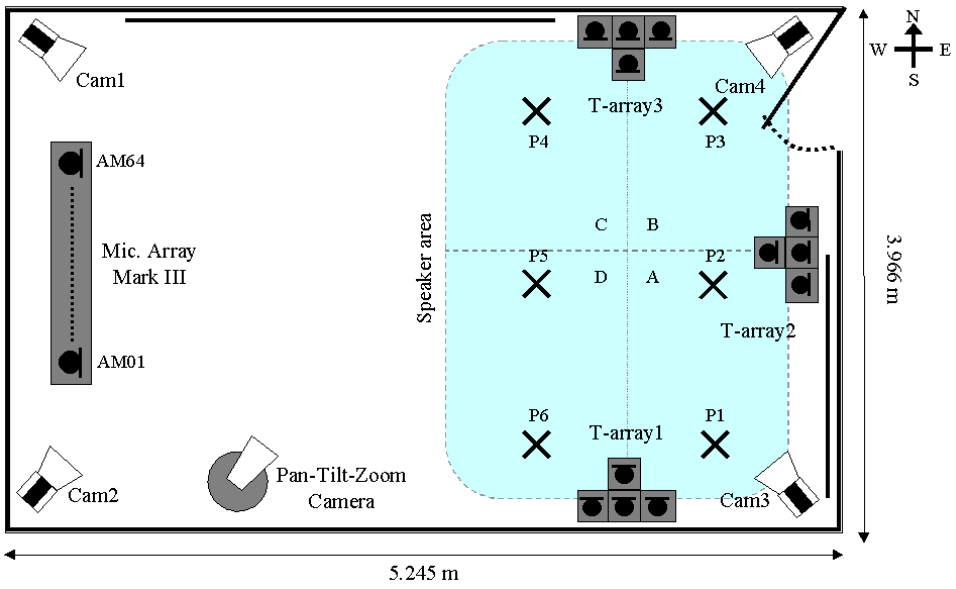

Figure 6.3: *Sensors set-up in the UPC smart-room.*

A high dependence of the localization error on the talker orientation is obvious from the results when using only one T-array. In this case, when the subject faces S, the direct sound wave is stronger than the reflected waves and the obtained localization errors are near the correct localization threshold. For the other orientations (N, W, E), the effect of reverberation together with the loss of directivity causes the reflected sound waves are stronger than the direct one and the localization error increases. Particularly, the combined effect of the talker directivity and the room reverberation can be observed when the subject faces N and W. Attending only to the talker directivity pattern, we would expected a better performance when the subject faces W than when he faces N. However, better average results are obtained when facing N – Root Mean Square Error (RMSE) of 803mm vs. 1205mm for W – basically due to the fact that in this case the main contribution of the indirect reflections comes from a position closer to the actual subject position.

When using the two T-arrays from the S and E walls, there is a clear improvement when the subject faces E in comparison to using only the S wall T-array. A slight improvement can also be observed when he faces S, but the performance for the other two orientations remains poor. As expected, the performance improves further when using the all three T-arrays; the errors are below the correct localization threshold for the cases when the subject faces N, S, and E. When he faces W (the wall without T-array), the performance improved significantly compared to the one or two T-array case, but it still remains poor.

From these observations some remarkable properties of the SRP-PHAT technique in relation to the orientation effect can be extracted. This technique combines the contributions of the dif-

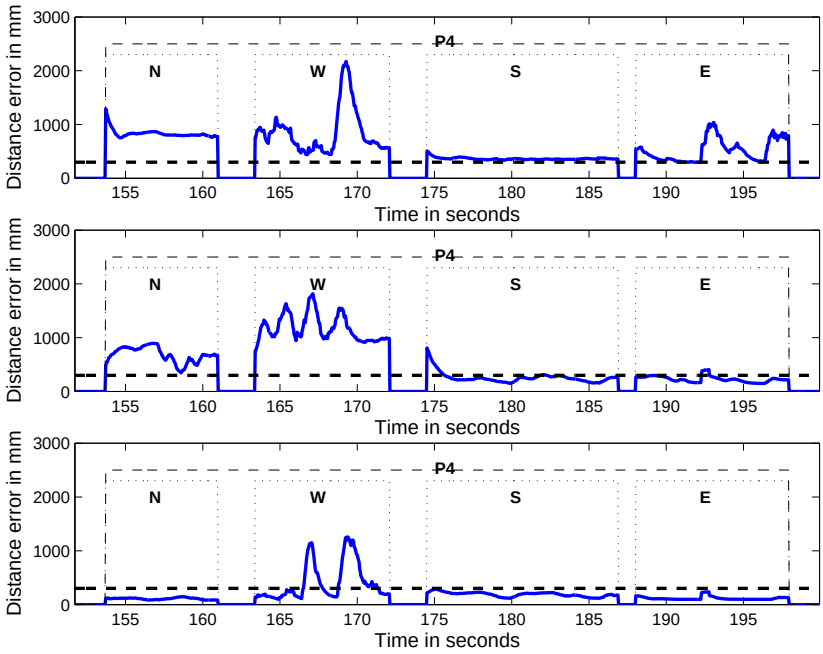

Figure 6.4: *Error in mm for various orientations in point P4 using T-array 1 (up), T-arrays 1&2 (middle) and the 3 T-arrays (bottom).*

ferent GCC estimations in an additive and collaborative way, which means that the contribution of worse microphone pairs (in the sense of their relative position and orientation with respect to the talker) does not seriously affect the performance of the technique if the best microphones are included in the selected set. In other words, with an appropriate distribution of microphones in the room the effect of talker orientation on the localization performance can be importantly diminished using this technique.

In a second set of experiments to measure the effect of head orientation on alternative localization algorithms, the SRP-PHAT and CL-PHAT are evaluated for the two following cases: a) using the all three T-arrays and b) using the best possible selection of T-arrays according to experimental results with all the combinations assuming known position and orientation, that is, given a concrete position and a concrete orientation the selection of T-arrays that provides the best localization result. Notice that the case b) can be considered as an upper-bound in a hypothetical case of an ideal selection of T-arrays.

Table 6.1 shows the results of this experiment. It must be taken into account that the object of this experiment is not to compare in terms of absolute localization performance the two techniques, since it is clear that SRP-PHAT provides better performances. The experiment is aimed to compare the different impact of the orientation. Thus, the relative error reduction in the case b) with respect to case a) for the two techniques is the important result.

By using the ideal selection of T-arrays, a much higher error reduction in relative terms is

Set-up	SRP-PHAT	CL-PHAT
3 T-arrays	228 mm	533 mm
Best T-array selection	184 mm	335 mm
Relative difference	19.37%	37.15%

Table 6.1: *RMSE in mm of SRP and CL techniques (mean of all the points and orientations).*

achieved for CL-PHAT (37.13%) than for SRP-PHAT (19.37%). This result indicates that the influence of orientation, and as a consequence the influence of the correct selection of microphones, on the localization performance is stronger in CL-PHAT than in SRP-PHAT.

This difference between SRP-PHAT and CL-PHAT in the orientation effect is mainly due to the multiplicative way in which the microphone pairs are combined in the CL-PHAT technique, i.e., the candidate points are obtained from the crossings of bearing lines and furthermore these lines are estimated based on the maximum of the cross-correlation, which is a hard decision. Hence, it can be deduced that in contrast to SRP-PHAT, the contributions of different T-arrays do not combine in a collaborative way in the CL-PHAT technique. As a result, using extra microphones in addition to the "right" ones not only does not help, but it can harm the final performance, unless a reliable best array indication is available.

Finally, some preliminary experiments were carried out to investigate simple strategies to alleviate the effect of the talker orientation on the localization performance in a more realistic scenario where it is assumed that the absolute orientation is known. For this purpose, in the case of the SRP technique we tried to use a different combination of T-arrays for the computation of the score at each position depending on the zone being explored. We divided the talker area in 4 sub-areas (see A, B, C and D in Figure 6.3; the boundaries of the sub-areas were determined by the T-arrays positioning) where a different set of T-arrays is used depending on the orientation. In these experiments, only a 3% relative improvement was obtained compared to the usage of the all three T-arrays, in part due to the already mentioned robustness of the technique to the effect of orientation. In Section 6.3 more elaborated strategies are commented, that consist in including orientation as a new parameter search in the maximization of Spatial Likelihood function computed by the SRP-PHAT algorithm

In the case of CL-PHAT, as a first attempt, we computed simultaneously the solution for the given orientation and for the four speaker areas A, B, C and D with different T-clusters set, in order to select the position with the lowest estimated error given by the error function employed in CL-PHAT. The results obtained in this case were not satisfactory. Then we used a table indicating the best selection of T-arrays depending on the orientation and independent on the zone. In this case, a 7% relative improvement was achieved in comparison to the usage of the all three T-arrays.

6.1.3 Conclusions

It has been shown that talker orientation strongly affects the performance of acoustic localization in smart-rooms due to the combinative effects of talker directivity pattern and room reverberation. However, techniques that join the estimated cross-correlations in a collaborative way, such as SRP-PHAT, have shown to be able to perform nearly independently on the talker orientation if the microphones are distributed appropriately in the room. In the case of closed form approximation techniques based on crossing lines, it has been shown that the multiplicative nature of the fusion of angle estimations yields in a highly dependent orientation method that needs a reliable best array indication to perform robustly. In addition, some preliminary basic ideas for introducing the orientation cue into the selected talker localization techniques have been assessed obtaining slight improvements.

6.2 Person tracking system for smart-room environments

In audio person tracking in smart-room environments, noise, reverberation, multiple moving speakers and head pose orientation are factors that demand an effort on the development of new robust methods capable of dealing with independence on the environmental conditions.

In this section a complete audio person tracking system is proposed. The overall system design is fundamentally aimed to develop a robust system with independence on the acoustic and room conditions, such as the number of sources, their maneuvering modes or the number of microphones.

The proposed system is based on the SRP-PHAT algorithm with some additional robust modifications. First, an adaptive smoothing factor for cross-power spectrum estimation derived from the estimated velocities of sources is proposed. Furthermore, motivated by the need for reducing the computational complexity and by experimental observations of the frequency properties of the cross-correlation in SRP algorithms, a two-pass search procedure that enhances the accuracy results and reduces the computational cost is proposed. Additionally, other development aspects are discussed.

6.2.1 Audio person tracking system baseline

In previous Chapter 5 many different approaches to tackle acoustic source localization in a smart room environment have been presented. Most remarkable differences between them are related with the way in which two basic problems are faced: a) how to infer information from the position of the sources on the basis of the microphone captures and b) how to use these information to obtain a reliable 3D position in the room space.

On the one hand, Direction of Arrival (DOA) to a microphone array or Time Difference of Arrival (TDOA) between a pair of microphones, can be obtained on the basis of High Resolution Spectral Estimation techniques (Section 5.1.1), or cross-correlation techniques (Section 5.1.2), among others. On the other hand, the availability of multiple TDOA/DOA estimations has been shown to lead to a minimization of an over-determined and non linear error function to obtain a unique estimation of the position that can be faced in many different ways, for instance, by means of iterative search algorithms (Section 5.2.1), closed-form estimators that approximate to a sub-optimal solution (Section 5.2.2) or space exploration based techniques, also known as Steered Response Power (SRP) techniques (Section 5.2.3).

The SRP-PHAT algorithm described in Section 5.2.3 and summarized in Table 5.2 tackles and integrates these two basic problems of localization in a robust and smart way. The main strength of this technique consists on the combination of the simplicity of the steered beamformer approach with the robustness offered by the generalized cross-correlation with PHAT weighting (GCC-PHAT). In practice, it has become the most successful state of the art source localization system. In addition to its robustness mainly to reverberation, in previous section it has been shown its superiority compared to a simple closed form localization approach and its relative independence on head orientation. In other preliminary experiments not shown in this dissertation, other closed-form approximations were also tested an resulted less robust than the mentioned SRP-PHAT algorithm. Consequently, the proposed system in this section is based on it. Figure 6.5 shows an example of the Spatial Likelihood Function obtained with the SRP-PATH algorithm.

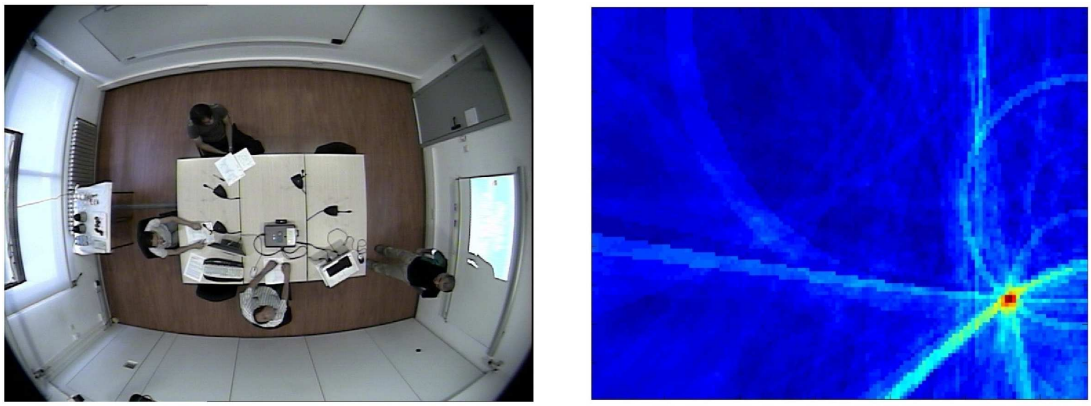

Figure 6.5: *On the left, zenithal camera snapshot. On the right, example of the Spatial Likelihood Function obtained with the SRP-PHAT process. A maximum peak can be clearly observed close to the position of the main speaker.*

6.2.2 Adaptive smoothing factor for Cross-Power Spectrum (CPS) estimations

Smoothing across time of the GCC-PHAT estimations is a simple and efficient way of adding robustness to the system. In fact, the temporal smoothing can be done in the time domain (GCC-PHAT) or in the frequency domain (CPS). Considering the smoothed cross-power spectrum $\widehat{G}_{x_1x_2}(t,f)$ in time instant t and the instantaneous estimation $G_{x_1x_2}(t,f) = X_1(t,f)X_2^*(t,f)$, our system performs the smoothing along time as follows,

$$\widehat{G}_{x_1x_2}(t,f) = \beta\widehat{G}_{x_1x_2}(t-1,f) + (1-\beta)G_{x_1x_2}(t,f) \tag{6.1}$$

The influence of the past $t-k$ instantaneous cross-power spectrum ($G_{x_1x_2}(t-k,f)$) in the smoothed cross-power spectrum estimation $\widehat{G}_{x_1x_2}(t,f)$ depends exponentially on this β factor as $\beta^k(1-\beta)$. Consequently, applying Equation 6.1 for the estimation of the cross-correlations in a source localization system like the one proposed, means that the cross-correlations are affected not only by the current position of an hypothetical acoustic source, they also depend on its past positions.

Table 6.2 shows the localization accuracy of the SRP-PHAT algorithm with different smoothing factors for the CPS estimation in 8 seminars randomly extracted from the evaluation data. More details on the evaluation data and metrics are provided in next Section 6.2.4. From these experimental observations it can be generally seen that a low forgetting factor have a beneficial influence in the overall source localization system. However, the most adequate value is highly dependent on the concrete data. On the one hand, increasing this factor provides enhanced cross-correlation estimations and consequently, better performance of the source localization system in the particular case of slow moving or almost static sources. On the other hand, an high smoothing value can be dramatically inconvenient in a scenario with many fast moving speakers.

β	Set1	Set2	Set3	Set4	Set5	Set6	Set7	Set8
0	51.2%	66.5%	50.9%	35.3%	79.3%	40.9%	54.8%	82.4%
0.25	61.5%	73.8%	63.8%	55.3%	81.3%	47.5%	48.9%	83.5%
0.5	62.4%	77.1%	67.7%	59.8%	82.5%	47.5%	53.8%	84.0%
0.75	67.3%	81.5%	68.1%	59.0%	80.1%	42.9%	53.8%	83.0%
0.9	70.2%	77.5%	62.7%	59.8%	75.6%	36.9%	54.3%	88.3%

Table 6.2: *A-MOTA results of 8 randomly selected seminars with different forgetting factors for the computation of the cross-power spectrums. Depending on the specific data, high forgetting factor values show a very different behaviour.*

Since the object of the proposed system is to perform robustly independently on the scenario and the right selection of this β factor is clearly a crucial aspect, an adaptive smoothing factor selection strategy has been developed. The strategy consists on selecting this factor depending on the estimated velocity of the audio sources present. Figure 6.6 shows a β selection function depending on the velocity that has been experimentally found to provide remarkable results

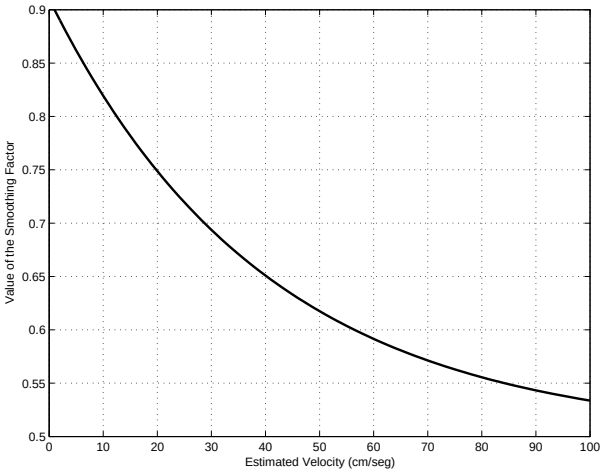

Figure 6.6: *Value of the adaptive smoothing factor depending on the estimated velocity.*

The problem of estimating the velocity of the sources is solved on the basis of the state estimations provided by the Kalman based tracker described next.

Kalman based tracker

The tracker developed, based in the one proposed in (Segura, 2004)), consists on the application of simple rules for the detection of simultaneous active tracks and for the association of the new observations (position estimations provided by the source localization system) to the current active tracks. These observations associated to each active track are filtered with a Kalman filter. Figure 6.7 shows an schema of the basic functionality of the tracker.

Each new observation received is assigned to the nearest active track (if any) whenever its distance is below a fixed threshold. In the case that the observation can not be associated to any active track the position given by the tracker is the one observed and the velocity is fixed to an high value to assure the lowest possible smoothing factor. To create new active tracks the system keeps a history window of the past W observations. When C observations are close enough between them (for instance, $C > W/2$ with distances below a constant threshold), a new track is activated at the average position of the C close observations if there is not already an active track at that position. When no one observation is assigned to a concrete track in the past K observations, that track is eliminated.

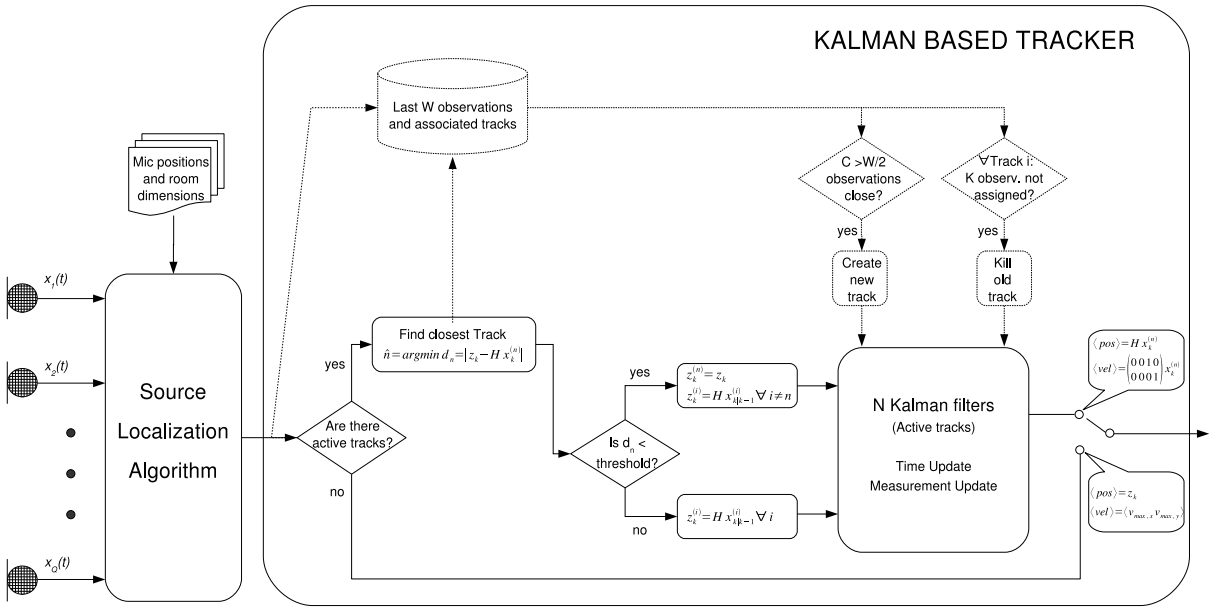

Figure 6.7: *Blocks diagram of the Kalman based tracker used for velocity estimation.*

Each active track is in fact a Kalman filter of four state components, that is the position and the velocity in 2D. When an observation is associated to a concrete track, then the prediction and measurement equations of 5.18 and 5.21 are applied to that concrete Kalman filter to obtain a new state (position and velocity) estimation. In this case, these new estimated position and velocity are the output of the tracker. Regarding the tracks (Kalman filters) to which the current observation has not been assigned, an observation equal to the projection of their last state estimation, that is their a priori state estimation (see Section 5.3), is assigned and prediction and measurement equations are also applied to obtain a new state estimation for each Kalman filter.

The state transition matrix $\mathbf{A}$, the observation matrix $\mathbf{H}$ and the process noise matrix $\mathbf{Q}$ are the same for each Kalman filter and are:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & T & 0 \\ 0 & 1 & 0 & T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} T^2/2 & 0 \\ 0 & T^2/2 \\ T & 0 \\ 0 & T \end{bmatrix} Q \begin{bmatrix} T^2/2 & 0 & T & 0 \\ 0 & T^2/2 & 0 & T \end{bmatrix}, \mathbf{R} = \begin{bmatrix} R & 0 \\ 0 & R \end{bmatrix} \tag{6.2}$$

where T is the time rate of new observations and Q and R are the parameters that control the magnitude of the process and measurement noise covariance matrices respectively. The

importance of the right selection of this Q and R parameters was already discussed in Section 5.3. In general, $\mathbf{Q}$ must be experimentally determined based on off-line tuning for the concrete scenario of application, while the measurement noise matrix $\mathbf{R}$ can be usually estimated in some way.

In the tracker presented here, the $\mathbf{Q}$ matrix is determined based on experimental observation, while $\mathbf{R}$ is estimated based on the errors associated to each cross-correlation microphone pair estimation and their relative positions to the observed position (Segura, 2004).

Unfortunately, even in the case of estimating the measurement noise as described, the performance of the tracker for position tracking was experimentally verified to be too much dependant on the different evaluation data and significant improvements were not observed for a fixed Q value. This is basically due to large differences on the evaluation data and concretely on the number of sources and their dynamics. Consequently, in the proposed source localization system and in the experiments shown in next sections 6.2.4 and 6.2.5, the tracker has only been used for determining the adaptive CPS smoothing factor based on the estimated velocity.

6.2.3 The two-pass search algorithm

Motivated by different observations on the influence of the high and low frequency range of the cross-correlations in the source localization performance and by the need for reducing the computational expense of the proposed system an efficient two step SRP search algorithm is developed.

The GCC-PHAT described in Section 5.1.2 was claimed to be a convenient choice for time delay computation in reverberant environments. This is justified assuming that the reverberation at each single frequency is proportional to the energy of the direct speech wave at that frequency, and in this way the signal to reverberant ratio (SRR) is constant for all the frequency range. Consequently, each frequency should be equally important in the computation of the generalized cross-correlation when the reverberation is the major degrading factor affecting the microphone signals.

However, one can also interpret the importance of each frequency band in relation with the reverberation in an alternative way attending to radiation and room acoustics considerations. In Section 6.1.1, it was commented that in (Chu & Warnock, 2002) it is shown that talker's radiation pattern is more directive for high frequencies. Furthermore, most of the construction materials of typical rooms show higher absorption of sound for higher frequencies. Consequently, the speech high frequency range can be considered to be less reverberated. Thus, if one can consider a source localization system relatively independent on orientation (like SRP-PHAT based systems as shown in previous Section 6.1), enhancing high frequencies will theoretically improve the GCC-PHAT estimation and the performance of the overall system.

Hence, in order to investigate the role of each frequency in the cross-correlation computation between microphone signals captured in real smart-room environments, several preliminary audio source localization experiments were conducted. Some of them taking into account only the cross-correlation information provided by individual frequency bands, for instance of a single octave or a mel-scaled band. Also, different experimental frequency masks were designed and tested for the computation of the GCCs according to the idea of giving more importance to the less reverberated frequencies, that is the highest ones.

Although a strong conclusion about the relationship between reverberation and frequency importance was not achieved as a result of these experiments, a remarkable property of the SRP-PHAT technique could be generally observed. Most of the information for a rough localization based on an SRP-PHAT like algorithm is concentrated in the low-frequency bins of the GCC-PHAT, while high frequency bins are needed for providing finest position estimations.

On first thought, it was considered that this observation was somehow related to the less reverberated assumption of the high frequency band discussed here. However, the actual reason is related to the frequency dependent beam-width of beamformers and consequently of the SRP-PHAT algorithm, since in last chapter was shown that it can be interpreted as an estimator of the energy received by a filter-and-sum beamformer.

Taking into account this interpretation of the SRP-PHAT technique, the observation about the influence of each frequency can be understood as follows. The low frequency beamformer response is wider than the high frequency beamformer response for a fixed geometry. Consequently, if the spatial exploration is done based on large search cells (coarse search), only lower frequency components must be considered in order to beam the energy emitted from the whole cell. Otherwise, if high frequency components are considered in a coarse exploration, only the high frequency energy of a small region around the concrete position explored (smaller that the search cell) will be taken into account for the computation of the likelihood score. That is, high frequency components permit higher resolution explorations, but they should be only taken into account when accurate position estimations are desired. This idea was exploited in (Zotkin & Duraiswami, 2004) to efficiently approximate the SRP-PHAT exhaustive search by a coarse-to-fine hierarchical search strategy. In practice, that work can be considered a generalization of the proposed two-pass search algorithm described below.

Two-pass SRP Search

The two-pass search procedure proposed in this section permits reducing the computational complexity of the exploration, which is known to be one of the main drawbacks of the SRP based techniques, particularly if an high space resolution is desired. Additionally, it is shown by experimental evaluation that better localization performances than the exhaustive search of

conventional SRP-PHAT are obtained as a consequence of this two step search strategy. The proposed two-pass SRP search consists on the following two stages:

Coarse Search This search procedure is performed only in the *x-y* axis (z is assumed to be 1.5 m), with a searching cell dimension of 20 cm and only using the low frequency information of the cross-correlations ($f < 9kHz$). A first coarse estimation is obtained from this search, say $(x_1, y_1, 150)$ cm.

Fine Search A new limited search area around the obtained *coarse* estimation is defined $(x_1 - 25 : x_1 + 25, y_1 - 25 : y_1 + 25, 110 : 190)$ cm. In this new fine search, dimension of the cell search is fixed to 4 cm for the *x-y* axis and to 8 cm for the z-axis. In the *fine search* all the frequency information of the cross-correlations is used and a more accurate estimation is obtained.

Regarding the computational complexity reduction, assume for instance that acoustic source localization is performed in a typical room of 400x500x300 cm (quite similar to the UPC smart-room). In the particular case of an exhaustive search with a resolution of 4cm in 3D space, as achieved by the system proposed, the number of points in which the spatial likelihood should be computed for each analysis frame is 100x125x75. That is, almost 10^6 evaluation points, which is completely unpractical. The proposed two-pass search strategy in the same room would require 20x25 likelihood score computations in the coarse search step, and additionally 12x12x10 score computations in the fine search step. That is, a total of approximately 2000 spatial likelihood points. This huge difference of almost three orders of magnitude constitutes by itself a considerable motivation for the use of the two-pass search strategy.

Additionally, the experiments of next section 6.2.4 show that the proposed two-pass search strategy provides enhanced source localization performance, which was also reported in the similar approach in (Zotkin & Duraiswami, 2004). The most likely reason hold by the author to explain the improvement is that it is related with spatial aliasing problems. Spatial aliasing for a fixed array geometry usually appears in the high frequency range as explained in Chapter 2. To better understand this problem, consider again the example of a two microphone delay-and-sum beamformer computing the energy arriving from the different positions in the space. On the one hand, if a fine position resolution is desired, then the higher frequency components of the speech signal must be considered – since the beamformer is narrower in the high band –, but at the same time spatial aliasing is more likely. For instance, when steering any point in the space, it is possible that at the same time the beamformer is steering the actual source position in the high frequency due to spatial aliasing, and consequently this point can become a false competitor. On the other hand, if a coarse position resolution is needed, then only low frequency range of the speech signal must be used to account for the energy received from a large spatial region. Consequently, the spatial aliasing problem is reduced. Generalizing these considerations to the

case of more than one filter-and-sum beamformer, then it can be stated that the SRP-PHAT with two-pass search strategy enhances the source localization performance because the first coarse estimation is partially free of the spatial aliasing problems, and the fine search –although not being free of the spatial aliasing problem– is carried out in a reduced space region around the first coarse estimation which is very unlikely to contain competitor aliases.

Table 6.3 summarizes the basic steps of the proposed localization algorithm with both the adaptive forgetting factor for cross-power spectrum estimations and the SRP-PHAT two-pass search strategy.

The proposed localization algorithm
1. Pre-compute theoretical delays from each possible coarse exploration position to each microphone pair.
2. Pre-compute theoretical delays from each possible fine exploration position to each microphone pair.
3. For each analysis frame compute the cross-correlations of each microphone pair (applying a first order smoothing filter).
4. For each position accumulate the contribution of the low frequency components of the cross-correlations (using delays pre-computed in 1).
5. Select the position with the maximum score as the coarse position estimation.
6. For each position of a small region around the coarse estimation accumulate the contribution of cross-correlations (using delays pre-computed in 2).
7. Select the position with the maximum score as the fine position estimation.
8. Estimate velocity and the adaptive smoothing factor for the next iteration.

Table 6.3: *Summary of the basic steps of the proposed localization system based on SRP-PHAT algorithm with a two-pass search strategy and adaptive forgetting factor for the estimation of the cross-power spectrums.*

6.2.4 Comparative evaluation

Evaluated systems

Due to excess of computational load, first the double search strategy is applied without frequency masking – that is, all the frequency band is considered in the coarse search step – in the system that in this work is considered the baseline SRP-PHAT (*Baseline*). The same system with adaptive smoothing factor for the CPS estimation (*Adaptive*) is also assessed. Then, SRP-PHAT algorithm with the two-pass proposed search strategy with frequency masking (*Two-pass*) and the complete proposed system including adaptive smoothing factor and two-pass algorithm (*Proposed*) are tested. Finally, results of the exhaustive SRP-PHAT algorithm with and without adaptive smoothing factor (*Exhaustive-A* and *Exhaustive* respectively) are provided to compare the two-pass search approach to the exhaustive search. The non-adaptive β factor of the *Baseline*, the *Two-pass* and the *Exhaustive* systems is fixed to 0.7.

Regarding implementation details of the algorithms, the analysis frame consists of Hanning

windowed blocks of 4096 samples, 50% overlapped, obtained at a sampling rate of 44.1 kHz. The FFT computation dimension is fixed to 4096 samples. The units used in all the internal computations are the centimeters for distances and centimeters per second for velocity. With regard to the Kalman based tracker, two maximum active tracks are allowed, the size of the observation window W is 20, the process noise parameter Q is fixed to 15 and the matrix **R** is re-estimated at each new analysis frame.

Finally, a confidence threshold is used to select or reject the position with the maximum value obtained from the accumulated contributions of all the correlations (see *Step 4* of Table 6.3), since this value represents the likelihood of the estimation given. The confidence threshold has been experimentally fixed to 0.5 for each 6 microphone pairs.

Data description

Person Tracking evaluation is run on an extract of the data collected by the CHIL consortium for the CLEAR 2006 evaluation (see next Section 6.2.5).

The data are audiovisual recordings of seminars given at each partner site. Two types of seminars were recorded: Non-interactive seminars, where mostly one presenter is speaking in front of a larger audience, and highly interactive seminars, where a smaller group of attendees listen to a presentation, ask questions, maybe take turns, etc.

The intention of the experiments is to evaluate the two proposals in two different but comparable environments. For this purpose, only data collected by IBM (interactive) and University of Karlsruhe, UKA (non-interactive) is considered. Each room is equipped with 4 T-shaped 4-channel microphone clusters appropriately distributed. In general, only microphone pairs of the same *T-cluster* array are used by the algorithms.

Evaluation metrics

Metrics and scoring of the systems has been done following the common agreement of the CHIL consortium for audio person tracking evaluation. The evaluation is run comparing the systems to 3D references at a rate of 1 label per second. Only time periods with one active speaker are considered. Two basic metrics are defined:

Multiple Object Tracking Precision (MOTP) [mm] This is the precision of the tracker when it comes to determining the exact position of a tracked person in the room. It is the total Euclidean distance error for matched *ground truth–hypothesis* pairs (i.e. Euclidean distance less than 500 mm) over all frames, averaged by the total number of matches made.

Acoustic Multiple Object Tracking Accuracy (A-MOTA) [%] This is the accuracy of the tracker

evaluated only for the active speaker at each time instant. It is calculated as one minus the ratio of the sum of errors over all frames and the total number of frames; the errors can be misses (i.e. Euclidean distance higher than 500mm or not hypothesis for a given ground truth label) or false positives.

Experimental results

Table 6.4 shows comparative results of the four accelerated systems in interactive environments. Comparing *Baseline* with *Two-pass* and *Adaptive* with *Proposed* results, a clear improvement thanks to the frequency masking of high frequency components in both precision and accuracy scores can be confirmed. The adaptive smoothing factor technique (*Baseline* vs. *Adaptive* and *Two-pass* vs. *Proposed*) does not show such an important influence in the results, with a slight improvement in MOTP and a slight decrease in A-MOTA score.

System	MOTP	Misses	False Positives	A-MOTA
Baseline	189,5mm	31,08%	21,90%	47,05%
Adaptive	185,8mm	31,48%	22,78%	45,78%
Two-pass	184,5mm	19,50%	11,10%	69,40%
Proposed	180,3mm	19,08%	12,13%	68,80%

Table 6.4: *Audio person tracking results of interactive (IBM) seminars.*

Person tracking results in non-interactive environments are presented in Table 6.5. Similar improvements are obtained with the proposed two-pass algorithm with frequency masking to those observed in the interactive environment, while in this case a generalized improvement is observed also with the adaptive smoothing technique.

System	MOTP	Misses	False Positives	A-MOTA
Baseline	161,2mm	28,25%	20,95%	50,81%
Adaptive	155,0mm	26,79%	19,29%	53,93%
Two-pass	147,0mm	19,32%	12,64%	68,04%
Proposed	142,4mm	17,98%	11,04%	70,97%

Table 6.5: *Audio person tracking results of non-interactive (UKA) seminars.*

Attending to results in both scenarios, the proposed adaptive smoothing factor technique compared to the application of a fixed value shows equivalent results in the interactive scenario, while a slight improvement in the non-interactive scenario is observed. Indeed, this difference justifies the need for the adaptive technique. For instance, we could likely find (experimentally)

a fixed smoothing value to improve the interactive results in exchange of probably degrading the performance in other scenarios. However, with the adaptive factor, we can develop a system working reasonably well independently on the number of speakers and their dynamics. In other words, the objective is not to improve the best possible results, the objective is to obtain good results in most environments. Anyway, the adaptive smoothing factor for the CPS estimations in the SRP-PHAT algorithm can be further investigated, for example, the relation between the velocity and the assigned value (see Figure 6.6) or the possible impact of less accurate Kalman estimations due to data association rules applied to assign estimated positions to tracks in multiple-speaker environments.

The proposed two-pass SRP algorithm based on frequency masking shows a great performance compared to the use of the complete frequency band of the cross-correlations in the coarse search step. Masking high frequency components is clearly necessary if coarse-to-fine search strategies are used to speed-up the SRP-PHAT algorithm for tracking applications, independently on the environment.

Finally, Tables 6.6 and 6.7 show again the results of the proposed two-pass proposed search based systems in comparison to the SRP-PHAT exhaustive search based systems for the case of interactive and non-interactive environments respectively. It can be seen that, in addition to the great computational reduction described in Section 6.2.3, a slight general source localization enhancement in both accuracy and precision is also achieved, that, as it was already commented, is probably caused by the reduction of the effect of spatial aliasing.

System	MOTP	Misses	False Positives	A-MOTA
Two-pass	184,5mm	19,50%	11,10%	69,40%
Proposed	180,3mm	19,08%	12,13%	68,80%
Exhaustive	205mm	17.66%	13.46%	68.88%
Exhaustive-A	202mm	17.07%	14.54%	68.39%

Table 6.6: *Audio person tracking results of interactive (IBM) seminars comparing the proposed two-pass search strategy and the computational expensive exhaustive SRP search.*

6.2.5 The CLEAR 2006 evaluation campaign

The 2006 CLEAR evaluation and workshop was an international effort to evaluate systems designed to recognize events, activities, and their relationships in interaction scenarios. It was meant to bring together projects and researchers working on related technologies in order to establish a common international evaluation in this field. Person Tracking based on audio, video and audio-video modalities, was one of the technologies considered in the CLEAR 2006 evalua-

System	MOTP	Misses	False Positives	A-MOTA
Two-pass	147,0mm	19,32%	12,64%	68,04%
Proposed	142,4mm	17,98%	11,04%	70,97%
Exhaustive	151mm	18.95%	15.59%	65.46%
Exhaustive-A	147mm	17.20%	13.51%	69.29%

Table 6.7: *Audio person tracking results of non-interactive (UKA) seminars comparing the proposed two-pass search strategy and the computational expensive exhaustive SRP search.*

tion campaign.

The proposed robust audio source localization system described above was developed and presented in the context of the Person Tracking evaluation of the CLEAR 2006 campaign. It was the UPC proposal submitted for both audio and audio-video evaluations. The results obtained in the audio modality were very satisfactory and the proposed system resulted in one of the most robust and outstanding among the different presented systems. Brief description of the audio person tracking evaluation and the performance of the proposed system is reproduced in this section. For details about the video and multimodal Person Tracking systems developed at UPC and their performances in the evaluation, one can refer to (Abad et al., 2007).

Brief evaluation description

The evaluation tasks considered were the same of the experiments of previous Section: single and multi-person tracking, based on non-interactive seminar (collected by ITC and UKA) and highly interactive seminar (collected by IBM, RESIT and UPC) recordings, respectively.

Room set-ups of the contributing sites present two basic common groups of devices: the *audio* and the *video* sensors. Audio sensors set-up is composed by 1 (or more) NIST Mark III 64-channel microphone array, 3 (or more) T-shaped 4-channel microphone cluster and various table-top (T-T) and close-talk microphones (CTM). Video sensors set-up is basically composed by 4 (or more) fixed cameras (FixCam). In addition to the fixed cameras, some sites are equipped with 1 (or more) pan-tilt zoom (PTZ) camera. Table 6.2.5 summarizes the sensors set-up and the size of the collected data by each site.

The evaluation metrics used are also the ones described in 6.2.4. More complete description of the data and the evaluation can be found in (Stiefelhagen & Garofolo, 2007).

Site	Type	Minutes	MarkIII	T-cluster	T-T	CTM	FixCam	PTZ
IBM	Interactive	15	2	4	3	5	5	3
ITC	Non-interactive	15	1	7	4	3-4	4	0
RESIT	Interactive	14	1	3	4	4	5	0
UKA	Non-interactive	180	1	4	4	3-4	4	2
UPC	Interactive	13	1	3	7	5	5	1

Table 6.8: *Summary of the recorded data by IBM, Istituto Trentino di Cultura (ITC), Research and Education Society in Information Technology (RESIT), University of Karlsruhe (UKA) and Universitat Politènica de Catalunya (UPC).*

Audio person tracking results

The proposed SRP-PHAT with adaptive smoothing factor and two-pass search strategy algorithm previously described was submitted to the evaluation. In practice, some slight differences in the size of the cell search in the coarse and fine search steps and some other implementation slight refinements were made. This is the reason for small differences with respect to the results of Section 6.2.4.

Regarding the evaluation data and the sensors set-up, it was decided to use all the *T-clusters* available in the different seminars and only to use the *MarkIII* data of those sites where the *MarkIII* is located in a wall without a *T-cluster* (IBM, RESIT and UPC). In general, only microphone pairs of the same *T-cluster* or *MarkIII* array were considered by the algorithm.

In the experiments where the *MarkIII* is used, 6 microphone pairs are selected for GCC-PHAT computation (the same number of microphone pairs obtained from a *T-cluster*, thus, resulting equivalent for the selection of the confidence threshold). The pairs selected out of the 64 microphones of *MarkIII* are *1-11*, *11-21*, *21-31*, *31-41*, *41-51* and *51-61*. Hence, an inter-microphone separation of 20 cm for each microphone-pair is considered.

In Table 6.9 individual results of the proposed system for each data set and average results for both tasks are shown. Notice that task results are not directly the mean of the individual results, since the scores are recomputed jointly. The evaluating system in both tasks is the one described. The multi-person task is only evaluated when only one speaker is active, thus, mean performances obtained, as it could be expected, are quite similar. In fact, there is a certain drop in the performance of the multi-person task, but it is more related with the particular characteristics of each data set, that with the task indeed. For instance, UPC data is particularly noisy and present some challenging situations such as *coffee breaks*. Attending to these results, it can be stated that the system presented for acoustic tracking performs reasonably well in controlled scenarios with one or few alternative and non-overlapping speakers, while it shows a considerable decrease in difficult noisy scenarios with many moving and overlapping speakers.

Task	MOTP	Misses	False Positives	A-MOTA
ITC data	108mm	8.56%	1.46%	89.98%
UKA data	148mm	15.09%	10.19%	74.72%
Single Person	145mm	14.53%	9.43%	76.04%
IBM data	180mm	17.85%	10.54%	71.61%
RESIT data	150mm	12.96%	6.23%	80.80%
UPC data	139mm	32.34%	28.76%	38.89%
Multi Person	157mm	20.95%	15.05%	64.00%

Table 6.9: *Audio results of the UPC system in the CLEAR 2006 evaluation for both single and multi-person tracking.*

Next tables 6.10 and 6.11 show the acoustic person tracking performance results of the systems presented in the evaluation, in both single person and multi-person task respectively. The participant sites, in addition to the UPC, are AIT (Athens Information Technology), ITC (Istituto Trentino di Cultura), TUT (Tampere University of Technology) and UKA (University of Karlsruhe). The details of each system and the reported results can be found in (Stiefelhagen & Garofolo, 2007).

Site/System	MOTP	Misses	False Positives	A-MOTA
AIT	226mm	51.16%	51.16%	-2.32%
ITC	144mm	46.61%	5.17%	48.22%
TUT I	245mm	27.87%	27.84%	44.29%
TUT II	245mm	27.93%	27.86%	44.21%
UKA I	137mm	10.28%	10.28%	79.43%
UKA II	138mm	16.88%	16.88%	66.23%
UKA III	186mm	22.58%	22.58%	54.84%
UPC	145mm	14.53%	9.43%	76.04%

Table 6.10: *Results for acoustic single person tracking task in the CLEAR 2006 evaluation.*

Regarding the single person tracking results, high localization accuracies of up 14 cm with 79.43% correct estimations was obtained by the leading system presented by UKA. The second ranking system was the one proposed in this thesis, showing a similar precision performance with a small decrease in the A-MOTA score. The accuracy of the other trackers presented was quite below of these two systems. With respect to the multiple person scenario, the scores were quite low, compared to the single person case. Except for the UPC system, that was clearly the leading one, reaching 64 % accuracy an 16 cm precision. All the other systems were well-below the expectations, due the large number of misses or large inaccuracies in localization itself.

Site/System	MOTP	Misses	False Positives	A-MOTA
AIT	230mm	56.19%	56.19%	-12.38%
ITC	218mm	65.03%	19.32%	15.65%
TUT II	334mm	83.32%	83.22%	-66.53%
UKA I	240mm	52.45%	52.45%	-4.90%
UKA II	247mm	54.39%	54.39%	-8.78%
UPC	157mm	20.95%	15.05%	64.00%

Table 6.11: *Results for acoustic multi-person tracking task in the CLEAR 2006 evaluation.*

Attending to the both scenarios, the UPC system presented in this thesis was the only one able to show a reasonable precision and accuracy performance. Thus, it can be stated that the proposed system was in average the most successful acoustic tracker presented.

6.2.6 Conclusions

A robust Audio Person Tracking system for smart-room environments based on the well-known SRP-PHAT algorithm has been developed and described. Two simple and efficient novelties have been described and evaluated in interactive and non-interactive seminars collected by the CHIL consortium. Firstly, using an adaptive smoothing factor for the cross-power spectral estimations has shown to be convenient independently on the dynamics of the speakers. Secondly, the Two-Pass Search algorithm based on frequency masking, significantly improves the precision and accuracy of the tracker with respect to not masking the speech signal. Additionally, a slight improvement by the two-pass strategy compared to exhaustive SRP search is achieved , besides reducing the computational load of the search procedure in a very significant manner. Finally, the proposed system was evaluated in the CLEAR 2006 campaign obtaining remarkable results in the two tasks considered. The system was able to efficiently solve the problem of source localization in situations with few non-overlapping speakers, while it shows a considerable loss of performance in some challenging and noisy situations that must be addressed.

6.3 Head Orientation estimation

It has been shown in previous Section 6.1 that head orientation has an important influence on the correct performance of localization systems developed in smart-room environments. The impact varies in each case depending on the concrete localization technique considered and on the distribution of the microphone network available. Consequently, the need for head orientation is justified by the potential positive impact on audio source tracking systems. Additionally, it is a complementary cue that might be used in the development of various multimodal services

deployed in smart-rooms.

As commented in Section 5.4, in some approaches the problem of head orientation is very closely related to the audio source tracking problem, resulting the orientation an additional searching parameter in SRP based techniques, and consequently converting source localization and orientation estimation in a coupled problem. However, the use of these type of approaches might result too sophisticated if the object is to simply estimate the head orientation. Then, the development of alternative simple methods is an appealing field to explore.

In this section, two alternative head orientation methods are proposed and evaluated: an estimator based on the SRP-PHAT algorithm (similar to the ones of (Mungamuru & Aarabi, 2004) and (Brutti et al., 2005)) and a simple estimator based on talker directivity considerations described in Section 6.1.1.

6.3.1 The SRP-PHAT based head orientation estimator

Head orientation estimation can be tackled based on the joint maximization of a Spatial Likelihood Function depending simultaneously on the potential source positions and orientations. For instance, the SRP-PHAT likelihood function of Equation 5.15 can be extended to incorporate orientation as follows:

$$F(\mathbf{x};o) = F(\mathbf{T}(\mathbf{x}),\mathbf{O}(\mathbf{x};o)) = \sum_{p=1}^{P} \Omega_p \int_{-\infty}^{+\infty} \frac{X_{p_1}(f)X_{p_2}^*(f)}{|X_{p_1}(f)||X_{p_2}^*(f)|} e^{-j2\pi f \tau_p} df \tag{6.3}$$

For each spatial position $\mathbf{x}$ and orientation o, in addition to the time delay vector $\mathbf{T}(\mathbf{x}) = \begin{bmatrix}\tau_1(\mathbf{x}) & \tau_2(\mathbf{x}) & \dots & \tau_P(\mathbf{x})\end{bmatrix}$ formed by theoretical delays to each microphone pair, a weight vector $\mathbf{O}(\mathbf{x};o) = \begin{bmatrix}\Omega_1(\mathbf{x},o) & \Omega_2(\mathbf{x},o) & \dots & \Omega_P(\mathbf{x},o)\end{bmatrix}$ formed by an appropriate weight representing the influence of each cross-correlation -in terms of the relative orientation - must be computed.

With respect to the values that $\Omega_p(\mathbf{x},o)$ must take, it seems a reasonable option to weight the p-th microphone pair depending on the energy that would be received from an hypothetical source at the concrete position $\mathbf{x}$ and orientation o. Concretely, the solution adopted in this work is quite simple and consists in first computing the angle difference between the orientation explored and the line that crosses the middle point of the microphone pair and the position explored ($\theta_p(\mathbf{x})$). Then, the weight $\Omega_p(\mathbf{x},o)$ is computed as the value given by a function depending on the angle differences that approximates the normalized talker directivity pattern. Figure 6.8 shows the function considered to approximate the source directivity pattern and next equations mathematically define the computation of $\Omega_p(\mathbf{x},o)$.

$$\Delta\theta_p(\mathbf{x};o) = \theta_p(\mathbf{x}) - o$$

$$\Omega_p(\mathbf{x}, o) = F(\Delta\theta_p(\mathbf{x}; o)) = \frac{1}{1 + 0.6sin^2(\Delta\theta_p(\mathbf{x}; o)/2)} \tag{6.4}$$

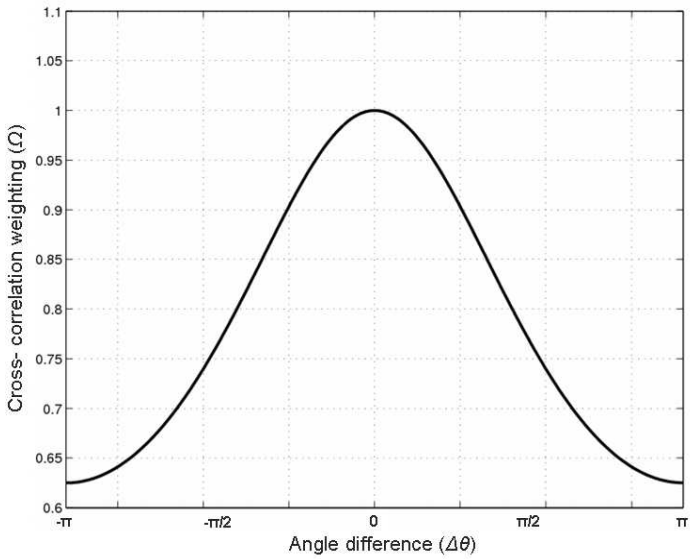

Figure 6.8: *Normalized directivity function depending on the angle differences for the computation of microphone pairs weighting.*

Similarly to conventional SRP-PHAT, the likelihood assigned to each position and orientation is equal to the sum over all the pairwise cross-correlations weighted according to equation 6.4. The joint estimated position and orientation are the ones that maximize the likelihood function:

$$\{\hat{\mathbf{x}}, \hat{o}\} = \arg\max_{\mathbf{x},o} F(\mathbf{T}(\mathbf{x}), \mathbf{O}(\mathbf{x}; o)) \tag{6.5}$$

By means of the method defined by equations 6.3 and 6.5, it is possible to jointly solve the problem of audio source localization and orientation estimation in a robust and smart way. In fact, coupling the orientation parameter in the search procedure might presumably provide more robust estimations of the source position, since the contribution of each cross-correlation is weighted according to the degree of reliable information that provides a microphone pair depending on its relative position with respect to the source. A summary of the proposed algorithm for joint estimation of source position and orientation is shown in Table 6.12.

One major problem arises with the proposed algorithm. In practice, the above formulation is equivalent to compute a different SLF for each possible orientation, consequently, the problems of computational load already noticed in the SRP-PHAT algorithm is of major relevance in this case due to the growth of operations and memory requirements. As a result, the exhaustive search of Equation 6.5 is unfeasible in real time applications. One possibility to partially solve

Joint position and orientation estimation method

1. Pre-compute theoretical delays from each possible exploration position to each microphone pair.
2. Pre-compute the cross-correlation weights from each possible exploration position to each microphone pair for every possible orientation.
3. For each analysis frame compute the cross-correlations of each microphone pair.
4. For each position and orientation accumulate the weighted contribution of cross-correlations (using delays pre-computed in 1 and the weights pre-computed in 2).
5. Select the pair position and orientation with the maximum score.

Table 6.12: *Summary of the basic steps of the joint source position and orientation estimation method based on SRP-PHAT algorithm (SRPPHAT-J).*

this problem would consist on applying efficient searching strategies, like the proposed two-pass search algorithm described in Section 6.2.3.

Alternatively, to tackle the problem of excessive number of operations, it is possible to decouple the problem into two different stages. In the first stage, the source position is estimated by means of conventional SRP-PHAT algorithm (or applying faster search strategies). In a second stage, the likelihood of the various orientations is computed only at the estimated source position. Then, the orientation that maximizes the likelihood at this position, is the estimated head orientation. Although the benefits of introducing orientation information in the source localization problem are lost in this way, it is a practical and reliable way of estimating the head orientation. Additionally, if the microphone network is well distributed, the reliability of the source position and orientation estimations is not expected to be considerably diminished. Table 6.13 summarizes the fast head orientation estimation based on SRP-PHAT algorithm.

Fast orientation estimation method

1. Pre-compute theoretical delays from each possible exploration position to each microphone pair.
2. Pre-compute the cross-correlation weights from each possible exploration position to each microphone pair for every possible orientation.
3. For each analysis frame compute the cross-correlations of each microphone pair.
4. For each position accumulate the contribution of cross-correlations (using delays pre-computed in 1).
5. Select the position with the maximum score.
6. Compute the score at the selected position for every possible orientation (using delays pre-computed in 1 and the weights pre-computed in 2).
7. Select the orientation with the maximum score.

Table 6.13: *Summary of the basic steps of the fast orientation estimation method based on SRP-PHAT algorithm (SRPPHAT-F).*

6.3.2 The HLBR head orientation estimator

In this section a new approach for estimating the head orientation from acoustic signals is presented (Segura et al., 2007). The proposed method is very efficient in terms of computational load due to its simplicity.

Assuming that a well-distributed network of microphones is available, the knowledge of the human radiation pattern can be used to estimate the head orientation of an active speaker by simply computing the energy received at each microphone and searching the angle that best fits the radiation pattern with the energy measures. This is done in (Sachar & Silverman, 2004) on the basis of a large aperture linear microphone array.

However, this simple approach has several problems since the microphones should be perfectly calibrated and different attenuation at each microphone due to propagation must be accounted for, thus requiring the use of sound propagation models. In the proposed approach, the computational simplicity is kept by using acoustic energy normalization to solve the aforementioned problems. Additionally, there is not need for a large array of microphones.

In Section 6.1.1, where speaker's diagram pattern is discussed, it was commented that the energy at the low frequency band radiated by an active speaker is low directional, while, at the high frequency range the radiation pattern is highly directive (Chu & Warnock, 2002). One can make use of this fact to define the High/Low Band Ratio (HLBR) of a radiation pattern. The HLBR of a radiation pattern is defined as the ratio between high and low bands of frequencies of the radiation pattern. According to experimental observations, the low band considered is the range of frequencies from 200 Hz to 400 Hz, while the high band is the range from 3500 Hz to 4500 Hz.

Hence, instead of computing the absolute energy received at each microphone, it is proposed to estimate the HLBR of the acoustic energy for each individual sensor in order to find the orientation that fits better these measures, according to a model of the HLBR radiation pattern depending on the head pose. The advantage of this measure is that its value is directly comparable across all microphones since, after the normalization, the effects of bad calibration and propagation losses are cancelled.

More concretely, the proposed technique first needs the position of the source $\mathbf{x}$ to be known beforehand or estimated by means of any source localization method. Then, the vectors $\mathbf{v}_q$ from the speaker to each microphone $\mathbf{m}_q$ with module $|\mathbf{v}_q|$ equal to the HLBR measure of the microphone are computed. The angle of $\mathbf{v}_q$ is θ_q and $\mathbf{V}(\theta)$ is the function that relates the HLBR of acoustic energy at each microphone with each angle as follows:

$$\mathbf{V}(\theta) = \sum_{q=1}^{Q} |\mathbf{v}_q| \delta(\theta - \theta_q) \tag{6.6}$$

The estimated speaker orientation can be computed by searching the angle that maximizes the correlation between a mathematical model of the HLBR of a radiated pattern $\mathbf{G}(\theta)$ and the HLBR of the acoustic energy measured at each microphone:

$$\hat{o} = \arg\max_{\theta} \mathbf{G}(\theta) * \mathbf{V}(\theta) = \sum_{q=1}^{Q} |\mathbf{v}_q| \mathbf{G}(\theta - \theta_q) \tag{6.7}$$

In order to model $\mathbf{G}(\theta)$, an appropriate Gaussian function or the function of Figure 6.8 have been experimentally tested to provide good results. In the following experiments, the function of Figure 6.8 is the one considered.

The technique described by Equations 6.6 and 6.7 is referred to as the basic HLBR head orientation estimation method (HLBR-B).

Additionally, one can find alternative solutions to the head orientation estimation problem also based on HLBR measures. A simple method that would not depend on a model of the HLBR radiated pattern, would consist on computing the sum vector of all the HLBR vectors of each microphone. Then, the angle of the sum vector would be an estimation of the head orientation:

$$\mathbf{v}_{sum} = \sum_{q=1}^{Q} \mathbf{v}_q \qquad\qquad \hat{o} = \angle \mathbf{v}_{sum} \tag{6.8}$$

The alternative algorithm defined by Equation 6.8 will be referred to as the vectorial HLBR head orientation estimation method (HLBR-V).

One of the advantages of the HLBR orientation estimation methods, in addition to their simplicity, is that they can provide an accurate resolution without a significant increase in the computational expense. In the case of the the HLBR-B method, Equation 6.7 can be evaluated for an high number of different orientations, since it is almost an inexpensive operation. In the case of the HLBR-V method, resolution is naturally achieved since the orientation estimated is provided by the angle of the vector formed by the sum of the HLBR vectors of each microphone, and consequently, it is not restricted to a fixed set of candidate values.

Finally, it must be mentioned that a motion model for head pose tracking – for instance, a Kalman filter – can be used to obtain better estimations in both the HLBR based techniques and the SRP-PHAT based approaches described in previous section. In fact, it has been experimentally tested to provide enhanced orientation estimations, however it has not been introduced

in the following experiments in order to better assess and compare the abilities of each proposed method by theirselves.

6.3.3 Experimental evaluation

The two proposed head orientation estimation methods described above are tested and compared in this section: the SRP-PHAT based method in its two versions – the joint source localization and head orientation estimation method (SRPPHAT-J) and the fast head orientation estimation method (SRPPHAT-F) – and the simple HLBR method in its two versions also – the basic HLBR head orientation estimation method (HLBR-B) and the vectorial HLBR head orientation estimation method (HLBR-V)–.

Regarding implementation details of the SRP-PHAT based methods, in the case of the SRPPHAT-F algorithm, the source position estimation is obtained with the system proposed in Section 6.2, with both the adaptive forgetting factor and the two-pass search strategy. In the case of the SRPPHAT-J algorithm, the two-pass search strategy is slightly modified: first, the coarse search step is done as usual (without incorporating orientation information); then, the orientation parameter is incorporated into the fine search step. In both cases, the number of possible orientations is fixed to 8.

With respect to the HLBR based techniques, the source position is previously estimated also with the system described in Section 6.2. Additionally, a first order filter is used to smooth the estimations of the high and low frequency band with a forgetting factor equal to 0.7.

Data description

Head orientation estimation is evaluated with the CLEAR head pose database (Stiefelhagen & Garofolo, 2007). It consists on an extract of 3 seminars from the data collected by the CHIL consortium for the CLEAR 2006 evaluation (see above Section 6.2.5) that was labelled for particular head pose evaluation purposes. The seminars are from the UKA site, that is, a non-interactive indoor scenario where a person is giving a talk, for a total of approximately 15 min.

All results described in this work were obtained using a set of four T-shaped 4-channel microphone clusters.

Evaluation metrics

Metrics and scoring of the systems has been done following the common agreement of the CHIL consortium for head pose evaluation. Three basic metrics are defined:

Pan Mean Average Error (PMAE) [degrees] This is the precision of the head orientation angle estimation in terms of degrees.

Pan Correct Classification (PCC) [%] This is the ability of the system to correctly classify the head position within 8 classes spanning 45° each.

Pan Correct Classification within a Range (PCCR) [%] This is the ability of the system to correctly classify the head position within 8 classes spanning 45° each, allowing a classification error of ±1 adjacent class.

Experimental results

Table 6.14 summarizes the results obtained by the methods under study. Regarding these results, it is clear that methods based on SRP-PHAT show a generalized better performance than methods based on HLBR. On the one hand, the SRPPHAT-J and SRPPHAT-F show in practice almost identical head orientation estimation performances. Since SRPPHAT-F is much less computationally demanding, it can be considered the most convenient option for head pose orientation estimation. On the other hand, it seems that the HLBR-V method is slightly superior to the HLBR-B, consequently it can be considered the most convenient option to estimate head orientation if the source position is known beforehand or obtained with a different source localization estimator than the SRP-PHAT algorithm. Notice that applying sequentially the SRP-PHAT algorithm and the HLBR-V to estimate respectively the source position and the head orientation is almost equivalent, in terms of number of operations, to the application of the SRPPHAT-F method. However, the memory requirements of the SRPPHAT-F for storing the weights associated to each microphone pair, position and orientation can be a problem in some cases, and alternative solutions like the proposed HLBR-V method might result more convenient.

Method	PMAE	PCC	PCCR
SRPPHAT-J	44.68°	37.32%	73.38%
SRPPHAT-F	44.23°	37.71%	73.89%
HLBR-B	52.92°	29.85%	67.99%
HLBR-V	50.98°	32.61%	68.94%

Table 6.14: *Head pose orientation results of the four methods evaluated.*

Hence, it has been shown that SRPPHAT-J and SRPPHAT-F algorithms obtain similar results in the head pose orientation estimation task, but it must be also confirmed that source localization performance provided by the SRPPHAT-F method is equivalent to the SRPPHAT-J technique, despite it does not incorporate the orientation parameter in the search process. Table 6.15 shows the source tracking results of the two algorithms with the seminars of the CLEAR head pose database.

System	MOTP	Misses	False Positives	A-MOTA
SRPPHAT-J	157mm	18.28%	10.18%	71.53%
SRPPHAT-F	160mm	17.42%	11.90%	70.67%

Table 6.15: *Audio person tracking results of the SRPPHAT-J and SRPPHAT-F algorithms evaluated with the CLEAR head pose database.*

From these results it can be stated that incorporating orientation information in the search of the SRP-PHAT algorithm provides an enhanced source localization performance as it was expected. However, the improvement is achieved in exchange of an high increase in the computational load of the algorithm. Since the enhancement obtained is not very remarkable and the orientation estimation performance of both approaches is equivalent, the SRPPHAT-F method can be still considered as the most convenient approach to both speaker localization and head orientation estimation.

6.3.4 Conclusions

It has been shown that head orientation estimation can be achieved by means of both SRP-PHAT based approaches and more simple techniques based on speech radiation considerations. On the one hand, coupling the source localization and orientation estimation problem in a joint search permits obtaining enhanced source localization estimations and reasonably good results in head pose estimation, but in exchange of high memory and computational expense requirements, which are key issues in this kind of exploration approaches. To partially solve these problems, a faster head orientation estimation consisting in the evaluation of the likelihoods of the various possible orientations at the estimated source position can be applied, without a remarkable drop of performance in both source localization and head pose estimation tasks. On the other hand, inexpensive head orientation estimation methods based on the high/low band ratio (HLBR) measure have been also presented. Although these approaches show in general a worse performance compared to the more sophisticated SRP based algorithms, they can become an interesting alternative for instance when source position is known beforehand, when the source localization algorithm is not based on SRP-PHAT, or when memory and computational cost are very limited.

Chapter 7

Conclusions and future work

This final chapter is aimed to summarize the contributions and major results of this thesis, in addition to highlight some directions for future research.

7.1 Summary and conclusions

The study of multi-microphone approaches to different challenging speech processing tasks of current interest, and particularly, the development of new proposals in the context of smart-room applications has been sought in this thesis. More concretely, two different but related research lines have been followed: speech enhancement and recognition with microphone arrays, and multi-microphone speaker tracking and head orientation estimation.

On the one hand, microphone array processing is presented as a possible solution to the problems that appear in distant-talking speech applications, which are mainly the noise and the reverberation. By means of spatial filtering or beamforming, one can select or reject concrete spatial directions, thus, obtaining enhanced versions of the signals captured by an array of sensors. However, the application of conventional beamforming algorithms to the speech case involves several difficulties that makes microphone array processing a very specific signal processing task. Consequently, an extensive and varied literature exists related to speech enhancement with microphone arrays. The most remarkable state of the art contributions have been summarized in this thesis; for instance, speech enhancement with microphone arrays is usually tackled by means of fixed beamformers such as the delay-and-sum, adaptive beamformers such as the Generalized Sidelobe Canceller or with the aid of additional multi-channel post-filtering stages.

Regarding the early work carried out by the author in speech acquisition with microphone arrays, a new beamformer called Integrated Wiener-filtering with Adaptive Beamformer (IWAB) has been discussed. The main novelty of the proposal is that a post-filtering technique is integrated on the structure of an adaptive beamformer. As a result of this integration, the proposed

IWAB beamformer, tested in several speech enhancement experiments, showed an equivalent noise and interference reduction with lower level of speech degradation, when it was compared to conventional two stage application of adaptive beamformers and post-filtering. However, attending to speech degradation measurements and some additional speech recognition experiments, the proposed beamformer, and in general those based on post-filtering techniques, were found to be more convenient in highly noise scenarios, while in low noise conditions it was seen that the effect of the remaining speech degradation introduced is more important than the noise reduction achieved. Nevertheless, the proposed IWAB technique can be considered an appropriate solution for high noise environments and, in general, it is preferable to conventional post-filtering of the output of an adaptive beamformer.

With respect to the use of microphone arrays for speech recognition, most of the works that can be found in the literature, and that have been discussed in this thesis, consist on the use of microphone array speech enhancement approaches as a pre-processing step prior to a conventional automatic speech recognizer. In general, this type of approaches do not provide a considerable speech recognition improvement, in part because of conventional beamformers are designed according to criteria that does not necessary match the recognition objective. Additionally, some promising alternatives introducing beamforming in the acoustic model training process, or designing the beamformer coefficients according to some functional of the speech recognition performance were also reviewed.

Most recent work by the author in the context of microphone array beamforming is related to the development of a distant-talking speech recognition system particularly adapted to the UPC smart-room environment and to the characteristics of non-native English speakers. The system was initially based on both acoustic model re-training and adaptation with artificially synthesized data, thus, achieving a remarkable word error rate reduction compared to the baseline system. Additionally, some strategies based on microphone array beamforming with a relatively large array were investigated in order to further improve the recognizer performance. First, it was confirmed that the use of conventional microphone array pre-processing is not well-matched with the speech recognition goal. On the one hand, although showing a certain degree of noise and reverberation reduction, the resulting beamformed data is far from being free of these harmful effects. Consequently, when beamformed data is used to test a recognizer trained with close-talking microphone data, some recognition improvement is obtained compared to testing with a single far-field microphone, but the performance achieved is still far from the close talking matched case. On the other hand, when microphone arrays are used as a pre-processing step in robust systems adapted to the actual acoustic environmental conditions, the enhanced signals provided by the beamformers can make drop the recognition performance, depending on the degree of adaptation of the acoustic models to the actual environment. In the thesis, the combination of microphone arrays together with HMM adaptation in order to improve the

performance of speech recognition systems developed in smart-room environments is suggested. Indeed, using microphone array processed data for both adaptation and testing is shown to be more adequate than conventional use of microphone arrays as a pre-processing step of an automatic speech recognizer. Experiments on speaker dependent systems show a remarkable error reduction with respect to conventional microphone array application, while recent research with speaker independent systems confirms the potential utility of the approach, even when the microphone array data used for adaptation is simulated. Furthermore, it is worth reminding that an on-line implementation of the recognizer derived from this work is currently used for demonstration purposes in the context of the EU funded CHIL project research activities carried out at the UPC, which was one of the initial goals of the thesis.

As a final thought on speech enhancement and recognition, it must be said that a great research effort still needs to be devoted to the problems that cause the use of distant-talking microphones in speech processing applications. Specifically in this thesis, these problems have been addressed on the basis of microphone array processing theory. In fact, attending to the experience obtained in the development of this work, the results provided in the thesis and also in accordance with the related literature, the author can stand that the use of microphone arrays for speech enhancement improves the quality of speech signals captured by multiple microphones, but it is still far of completely solving the problems of distant recordings. Particularly, the use of conventional beamformers for automatic speech recognition is not in well-accordance with the recognition paradigm, and in general, it results in poorer recognition performances than expected, when using for instance a 64 channel microphone array like in this thesis. In practice, some factors such as the increase on investment and sophistication that involves the use of microphone arrays in the development of real world applications must be taken into account and evaluated in comparison to the improvements that currently provide. Nevertheless, it is obvious that microphone array processing has made step forwards in the last times, and that research and new contributions in this field will help to bridge the current gap between close-talking and distant talking speech applications. In the next section, some general thoughts for future research, and some more particular future ideas related to the techniques proposed will be provided.

On the other hand, speaker position is a fundamental cue needed by microphone array processing to correctly steer beamformers towards the source of interest. Additionally, knowledge about the position of the acoustic sources is useful for other technologies that are usually deployed in smart-rooms. Consequently, some of the most remarkable state of the art approaches to acoustic source localization and tracking were pointed out in the thesis. The task has been split into three basic stages, which are aimed first to obtain some kind of information related with the sources position, such as time delays of arrival on the basis of cross-correlation between pairs of microphones; second, to estimate a spatial position according to the estimated informations

and the actual known geometry of the room and the microphones; and third, to track multiple sources by means of association rules and dynamic models, such as the Kalman filter.

The early work carried out by the author on source localization was initially devoted to compare several of the source localization approaches discussed. From these first experiments and comparisons, a study of the effect of head orientation on source localization performance in two completely different selected localization algorithms was highlighted. It was first demonstrated that talker orientation strongly affects the performance of acoustic localization in smart-rooms due to the combinative effects of talker directivity pattern and room reverberation. However, it was also shown that all the techniques are not affected in the same degree by head orientation. In fact, techniques based on space exploration, such as SRP-PHAT, join the estimated cross-correlations in a collaborative way and, consequently, they are able to perform nearly independently on the talker orientation if the microphone network is appropriately distributed in the room. Actually, attending to the related literature, and the early comparisons carried out by the author, the SRP-PHAT technique can be raised as the most convenient state of the art algorithm to tackle the problem of speaker localization on the basis of multi-microphone recordings. In general, it shows a robust performance to main degrading factors in real room environments, like noise, reverberation or speaker orientation.

Hence, since one of the fundamental objectives of the thesis is the development of a robust audio person tracking system for smart-room environments, an algorithm based on the SRP-PHAT technique has been designed and described. Two simple and efficient novelties have been introduced: first, the application of an adaptive smoothing factor for the cross-power spectral computations based on velocity estimations provided by a Kalman based tracker; and second, a two-pass search algorithm based on frequency masking. It has been shown that the proposed system is able to achieve significant improvements in both precision and accuracy speaker localization, independently on the application scenario. Indeed, it was demonstrated the convenience of the use of coarse-to-fine search strategies to reduce the computational load of the search process, and also to increase the robustness of the system. However, the use of these accelerated strategies was proved to be correct, only if narrow-to-broad band frequency cross-correlation filtering is also applied.

Additionally, the proposed person tracking system took part in the CLEAR 2006 international competitive evaluation campaign. The successful results obtained is one of the most remarkable achievements of the thesis. It was shown that the proposed system was one of the most outstanding participant systems performing quite robustly in the two proposed tasks: interactive and non-interactive scenarios. Concretely, the proposed system was the leading algorithm in the interactive task and the second ranked system, close to the leading one, in the non-interactive scenario. Thus, another fundamental goal of the present thesis, that is meeting the leading research laboratories in the field of multi-microphone processing, can be confirmed

to be accomplished according to the results obtained in that evaluation.

As a final remark related to the research and findings on the speaker tracking topic, it can be said that the performance of acoustic person tracking is quite satisfying in relatively controlled smart-room scenarios with few non-overlapping speakers, although there is obviously still room for future improvements. In contrast, more challenging situations with multiple moving and overlapping speakers or acoustic events pose many problems to the current available systems, which demand for new and more robust solutions.

Finally, a multi-microphone speech processing application aimed to estimate the head pose of the speakers present in smart-room environments has been stated as an appealing topic of research. The interest from a practical point of view states in its utility as an additional informative cue that can be exploited by several services deployed in intelligent applications, for instance, a microphone network management service aimed to automatically select the most adequate microphone for speech applications according to both speaker position and orientation information. In the past, head orientation estimation has been mainly addressed by means of video processing approaches. Consequently, the few audio works related to this problem that can be found in the literature have been briefly described in the thesis. Indeed, most of the approaches found are related to the development of robust source localization systems that introduce the head orientation as a new search parameter.

In the thesis, head pose estimation is investigated, and it is shown that it can be achieved by means of either joint source position and orientation exploration based approaches, similarly to some previous techniques, or more simple techniques based on speech radiation considerations. On the one hand, coupling the source localization and orientation estimation problem permits both obtaining enhanced source localization estimations and reasonably good results in head pose estimation, but in exchange of high memory and computational expense requirements, that can result in an unpractical solution in most cases. In order to partially solve these problems, the joint source and head orientation estimation method is simplified evaluating the various potential orientations only in the most likely estimated speaker position. This approximation permits achieving equivalent performances in both source localization and head pose estimation tasks, besides reducing the computational complexity of the algorithm. On the other hand, head orientation estimation methods based on speaker radiation considerations were also presented. In general, they show a reduced orientation estimation performance compared to the previous sophisticated methods, but they are almost computationally inexpensive. Thus, these simple methods can result in a practical alternative, for instance, when the source position is known beforehand.

In final conclusion, multi-microphone approaches to speech processing proposed in this thesis have shown to be a valid alternative to face the problems derived from distant-talking recordings that are usually met in a smart-room environment. Although some of the technologies are

shown to be below the expectations, such as in the case of speech recognition with microphone arrays, in general all the techniques developed achieved a relatively high performance thanks to the novelties introduced in multi-microphone processing. Particularly, remarkable results were obtained in the case of speaker tracking, and additionally, a cutting-edge research field like head orientation estimation, has been successfully initiated. Furthermore, as a consequence of the work carried out, some real running systems have been developed and are currently deployed in the UPC smart-room. Finally, a consistent basis for future research on multi-microphone topics at the UPC Speech Processing Group has been established with this thesis.

7.2 Directions for future work

While quite successful multi-microphone based approaches have been developed in this thesis for speech enhancement, speech recognition, speaker localization and head orientation estimation in smart-rooms, there is still a large room for future improvement in these research areas and particularly in the algorithms presented.

In fact, it has not been until recently that distant-talking automatic speech recognition has begun to appear feasible for medium to large vocabulary recognition tasks in meeting environments. In order to achieve a considerable advance in distant recognition in the basis of multi-microphone recordings, there is a need for an international effort to collect data and to establish a common framework for evaluation and reliable comparison. In this way, some initiatives like recent Rich Transcription Meeting Recognition Evaluations should provide this necessary basis.

From the author point of view, the future of speech recognition with microphone arrays consists in facing recognition with beamformers as a joint problem. On the one hand, taking into account beamformed data for the construction of speech recognizers, or on the other hand, taking into account speech recognition performance to design beamformers, are possible ways to tackle the problem in order to achieve enhanced speech recognition performance with microphone arrays.

With respect to the integrated Wiener filter and adaptive beamformer presented, and more generally, in relation to the use of post-filtering stages with beamformers, a deepest study of post-filters application is needed in order to reduce the speech degradation introduced and to make them useful for speech recognition applications. For instance, restricting maximum and minimum values of the filter coefficients attending to objective and subjective observations is a possibility to investigate.

The distant-talking speech recognition system presented was strongly conditioned by the development data available. For this concrete application, first there is a need for multi-accented non-native English speakers corpora that can provide improved matched speaker characteris-

tics. Second, real multi-microphone and microphone array data in smart-room environments is necessary to introduce beamforming in the development of enhanced acoustic models for speech recognition with microphone arrays. Additionally, the use of sophisticated beamformers, like adaptive beamformers or nested sub-array structures, was shown to not provide enhanced results when compared to simple delay-and-sum. This was partially due to misadjustment of some of the implementation parameters. Thus, more effort must be addressed in order to successfully apply these beamformers in speech recognition. For instance, in the particular case of adaptive beamformers a robust distant-talking speech detector would be necessary to avoid adaptation problems.

Regarding the speaker tracking topic future research must be oriented to more complex situations with multiple simultaneous and non-static speakers or acoustic events. One of the problems of multiple source localization is related with the need for an estimation of the number of sources. Multiple sources can be located by means of successive maxima search of an spatial likelihood map. However, it is usually necessary to know how many maximums must be searched. For instance, the number of sources to search can be based on the combination of the likelihood measure compared to a certain threshold and also on data association and tracking strategies. Additionally, collaboration from other audio based technologies, like acoustic event classification, can be useful to distinguish different types of audio sources of interest (speakers) or to reject others (noises).

In addition to the adaptation of the proposed speaker tracking system to handle these more complex scenarios, there is still a large room for future improvements in controlled scenarios, although it was shown that its performance is quite satisfactory. The Kalman based tracker and the simple association rules applied must be further investigated to improve the estimations provided by Kalman filters. Alternative appealing approaches like particle filtering must be also considered in the future. Additionally, more accurate functions relating estimated velocities and smoothing factors can be obtained. The two-pass search algorithm was mostly developed based on experimental observations, hence, it is needed a more consistent theoretical background, for instance relating appropriate frequencies for certain search dimensions given the room geometry properties.

Acoustic head orientation estimation is a very recent and immature investigation field. Consequently, there is relatively few research on this topic and large possibilities for future advances exist. For instance, one immediate approximation would consist on combining the two different approaches proposed, that is, combining in some way exploration based orientation estimation approaches and the High/Low Band Ratio (HLBR) measure. Additionally, with respect to the HLBR measure, alternative ways of measuring it or the most convenient frequency bands considered for its computation must be further investigated.

In general, related to the development of acoustic based applications and services in smart-

room environments, the future must be focused in the interaction between multiple technologies. For instance, many technologies can benefit from acoustic event classification or speaker identification. Furthermore, the future of the technologies deployed in a smart-room environment is also related to multi-modal integration. For instance, person localization, person identification or head pose estimation based on both audio and video modalities are likely the most successful way of tackling these problems. Currently, most of the multimodal approaches combine the results given by each individual modality by means of some kind of scoring, filtering or smoothing fusion approach, thus, obtaining in general an improved performance. However, it would be convenient to investigate other ways of multimodal integration. In this way, it is necessary to study the strengths and weaknesses of each individual monomodal technology in order to fully exploit multimodal cues of information and to find out the way in which each technology can benefit from each other.

Bibliography

Aarabi, P. (2003). The fusion of distributed microphone arrays for sound localization. *EURASIP Journal of Applied Signal Processing*, *4*, 338–347.

Abad, A., Cantón, C., Segura, C., Landabaso, J. L., Macho, D., Casas, J. R., Hernando, J., Pardàs, M., & Nadeu, C. (2007). UPC audio, video and multimodal person tracking systems in the CLEAR evaluation campaign. *Lecture Notes in Computer Science, Springer-Verlag*, *4122*, 93–104.

Abad, A., & Hernando, J. (2004a). Integrated adaptive beamforming and Wiener filtering for a robust microphone array. In *Workshop on Sensor Array and Multi-channel Processing*, pp. 367–371.

Abad, A., & Hernando, J. (2004b). Speech enhancement and recognition by integrating adaptive beamforming and Wiener filtering. In *International Conference on Spoken Language Processing*, pp. 2657–2660.

Abad, A., Macho, D., Segura, C., Hernando, J., & Nadeu, C. (2005). Effect of Head Orientation on the Speaker Localization Performance in Smart-room Environment. In *European Conference on Speech Communication and Technology*, pp. 145–148.

Abad, A., Nadeu, C., Hernando, J., & Padrell, J. (2003). Jacobian Adaptation based on the Frequency Filtered Spectral Energies. In *European Conference on Speech Communication and Technology*, pp. 1621–1624.

Abad, A., Segura, C., Macho, D., Hernando, J., & Nadeu, C. (2006). Audio Person Tracking in a Smart-Room Environment. In *International Conference on Spoken Language Processing*, pp. 2590–2593.

Abhayapala, T. D., Kennedy, R. A., & Williamson, R. C. (2000). Nearfield broadband array design using a radially invariant modal expansion. *The Journal of the Acoustical Society of America*, *107*, 392–403.

Allen, J. B., & Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, *65*(4), 943–950.

Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *50*(2), 174–188.

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America*, *55*(6), 1304–1312.

Bar-Shalom, Y., & Fortmann, T. E. (1988). *Tracking and Data Association*. Academic Press.

Benesty, J. (2000). Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. *The Journal of the Acoustical Society of America*, *107*(1), 384–391.

Bitzer, J., & Simmer, K. U. (2001). *Microphone Arrays: Signal Processing Techniques and Applications*, chap. 2: Superdirective Microphone Arrays, pp. 19–38. Springer-Verlag.

Bitzer, J., Simmer, K. U., & Kammeyer, K. D. (1999). Multi-Microphone noise reduction by post-filter and super-directive beamformer. In *International Workshop on Acoustic Echo and Noise Control*, pp. 100–103.

Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *27*(2), 113–120.

Brandstein, M. S. (1999). Time-delay estimation of reverberated speech exploiting harmonic structure. *The Journal of the Acoustical Society of America*, *105*(5), 2914–2919.

Brandstein, M. S., Adcock, J. E., & Silverman, H. F. (1995). A Practical Time-Delay Estimator for Localizing Speech Sources with a Microphone Array. *Computer Speech and Language*, *9*, 153–169.

Brandstein, M. S., Adcock, J. E., & Silverman, H. F. (1997). A closed-form location estimator for use with room environment microphone arrays. *IEEE Transactions on Speech and Audio Processing*, *5*(1), 45–50.

Brandstein, M. S., & Ward, D. B. (Eds.). (2001). *Microphone Arrays: Signal Processing Techniques and Applications*. Springer-Verlag.

Brandstein, M. S. (1995). *A Framework for Speech Source Localization Using Sensor Arrays*. Ph.D. thesis, Brown University.

Brayda, L., Wellekens, C., & Omologo, M. (2006). N-Best parallel Maximum Likelihood Beamformers for Robust Speech Recognition. In *European Signal Processing Conference*.

Brutti, A., Omologo, M., & Svaizer, P. (2005). Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays. In *European Conference on Speech Communication and Technology*, pp. 2337–2340.

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, *57*, 1408–1418.

Cardoso, J.-F. (1998). Blind signal separation: statistical principles. *Proceedings of the IEEE*, *86*(10), 2009–2025.

Champagne, B., Bedard, S., & Stephenne, A. (1996). Performance of time-delay estimation in the presence of room reverberation. *IEEE Transactions on Speech and Audio Processing*, *4*(2), 148–152.

Chan, Y. T., & Ho, K. C. (1994). A simple and efficient estimator for hyperbolic location. *IEEE Transactions on Signal Processing*, *42*(8), 1905–1915.

Chen, J., Benesty, J., & Huang, Y. (2006). Time Delay Estimation in Room Acoustic Environments: An Overview. *EURASIP Journal of Applied Signal Processing*, *2006*.

Chen, J., Benesty, J., & Huang, Y. A. (2003). Robust time delay estimation exploiting redundancy among multiple microphones. *IEEE Transactions on Speech and Audio Processing*, *11*(6), 549–557.

Chen, J., Huang, Y. A., & Benesty, J. (2004). An adaptive blind SIMO identification approach to joint multichannel time delay estimation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4, pp. 53–56.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979.

Chien, J.-T., Lai, J.-R., & Lai, P.-Y. (2001). Microphone array signal processing for far-talking speech recognition. In *International Workshop on Signal Processing Advances for Wireless Communications*, pp. 322–325.

Choi, C., Kong, D., Lee, H.-K., & Yoon, S. M. (2004). Separation of Multiple Concurrent Speeches using Audio-visual Speaker Localization and Minimum Variance Beam-forming. In *International Conference on Spoken Language Processing*, pp. 2301–2304.

Chu, W. T., & Warnock, A. C. C. (2002). Detailed directivity of sound fields around human talkers. Tech. rep. IRC-RR-104, National Research Council Canada.

Claesson, I., & Nordholm, S. (1992). A spatial filtering approach to robust adaptive beaming. *IEEE Transactions on Antennas and Propagation*, *40*(9), 1093–1096.

Cox, H., Zeskind, R. M., & Kooij, T. (1986). Practical supergain. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *34*(3), 393–398.

Cox, H., Zeskind, R. M., & Owen, M. M. (1987). Robust adaptive beamforming. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *35*(10), 1365–1376.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representation for monosyllablic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *28*(4), 357–366.

De Boissieu, F. (2005). Automatic Speech Recognition System in UPC Smart-Room. Master's thesis, Universitat Politècnica de Catalunya.

Devijver, P. A. (1985). Baum's forward-backward algorithm revisited. *Pattern Recognition Letters*, *3*(6), 369–373.

Di Claudio, E. D., Parisi, R., & Orlandi, G. (2000). Multi-source localization in reverberant environments by ROOT-MUSIC and clustering. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 5–9.

DiBiase, J., Silverman, H., & Brandstein, M. (2001). *Microphone Arrays: Signal Processing Techniques and Applications*, chap. 8: Robust Localization in Reverberant Rooms, pp. 157–180. Springer-Verlag.

Elko, G. (2000). *Acoustic Signal Processing for Telecommunication*, chap. 10: Superdirectional Microphone Arrays, pp. 181–237. Kluwer Academic Publishers.

Fischer, S., & Simmer, K. U. (1995). An adaptive microphone array for hands-free communication. In *International Workshop on Acoustic Echo and Noise Control*, pp. 44–47.

Flanagan, J., Berkeley, D., Elko, G., West, J., & Sondhi, M. (1991). Autodirective microphone systems. *Acustica*, *73*, 58–71.

Flanagan, J. L., Surendran, A. C., & Jan, E. E. (1993). Spatially selective sound capture for speech and audio processing. *Speech Communication*, *13*, 207–222.

Forney Jr, G. D. (1973). The Viterbi Algorithm. *Proceedings of the IEEE*, *61*(3), 268–278.

Frost, O. (1972). An Algorithm for Linearly Constrained Adaptive Array Processing. *Proceedings of the IEEE*, *60*(8), 926–935.

Gales, M. J. F., & Woodland, P. C. (1996). Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, *10*(4), 249–264.

Gales, M. J. F., & Young, S. J. (1993). Cepstral parameter compensation for HMM recognition in noise. *Speech Communication*, *12*(3), 231–239.

Gannot, S., Burshtein, D., & Weinstein, E. (2001). Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, *49*(8), 1614–1626.

Gatica-Perez, D., Lathoud, G., McCowan, I. A., Odobez, J.-M., & Moore, D. (2003). Audio-visual speaker tracking with importance particle filters. In *International Conference on Image Processing*, Vol. 3, pp. 25–28.

Gauvain, J. L., & Lee, C. H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, *2*(2), 291–298.

Gillespie, B. W., Malvar, H. S., & Florencio, D. A. F. (2001). Speech dereverberation via maximum kurtosis subband adaptive filtering. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 6, pp. 3701–3704.

Giuliani, D., Matassoni, M., Omologo, M., & Svaizer, P. (1995). Hands Free Continuous Speech Recognition in Noisy Environment using a Four Microphone Array. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 860–863.

Giuliani, D., Omologo, M., & Svaizer, P. (1996). Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation. In *International Conference on Spoken Language Processing*, Vol. 3, pp. 1329–1332.

Gong, Y. (1995). Speech recognition in noisy environments: A survey. *Speech Communication*, *16*(33), 261–291.

Gonzalez-Rodriguez, J., Sanchez-Bote, J., & Ortega-Garcia, J. (2000). Speech dereverberation and noise reduction with a combined microphone array approach. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 1037–1040.

Grenier, Y. (1992). A microphone array for car environments. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 305–308.

Griffiths, L., & Jim, C. (1982). An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, *30*(1), 27–34.

Haykin, S. (2001). *Adaptive Filter Theory* (4th edition). Prentice Hall.

Hioka, Y., Koizumi, Y., & Hamada, N. (2002). Improvement of DOA Estimation Method Using Virtually Generated Multichannel Data from Two-channel Microphone Array. In *International Symposium on Information Theory and Its Applications*, pp. 735–738.

Hoshuyama, O., & Sugiyama, A. (2001). *Microphone Arrays: Signal Processing Techniques and Applications*, chap. 5: Robust Adaptive Beamforming, pp. 87–109. Springer-Verlag.

Hoshuyama, O., Sugiyama, A., & Hirano, A. (1999). A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. *IEEE Transactions on Signal Processing*, *47*(10), 2677–2684.

Huang, X., Acero, A., & Hon, H. (2001). *Spoken language processing. A guide to theory.* Prentice Hall.

Hung, H., & Kaveh, M. (1988). Focussing matrices for coherent signal-subspace processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *36*(8), 1272–1281.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent Component Analysis.* John Wiley & Sons.

Hyvärinen, A., & Oja, E. (2000). Independent Component Analysis: Algorithms and Applications. *Neural Networks*, *13*(4-5), 411–430.

Iskra, D. J., Grosskopf, B., Marasek, K., van den Heuvel, H., Diehl, F., & Kiessling, A. (2002). SPEECON - Speech Databases for Consumer Devices: Database Specification and Validation. In *International Conference on Language Resources and Evaluation*, pp. 320–333.

Johnson, D. H., & Dudgeon, D. E. (1993). *Array signal processing.* Prentice Hall.

Junqua, J.-C., & Haton, J.-P. (1996). *Robustness in Automatic Speech Recognition: Fundamentals and Applications.* Kluwer Academic Publishers.

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing: An introduction to Natural Language Processing, Computational Linguistics, and Speech Processing.* Prentice-Hall.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME, Journal of Basic Engineering*, 35–45.

Kates, J. M. (1993). Superdirective arrays for hearing aids. *The Journal of the Acoustical Society of America*, *94*(4), 1930–1933.

Kellermann, W. (1991). A self-steering digital microphone array. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 5, pp. 3581–3584.

Kleban, J., & Gong, Y. (2000). HMM adaptation and microphone array processing for distant speechrecognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 1411–1414.

Knapp, C. H., & Carter, G. C. (1976). The Generalized Correlation Method for Estimation of Time Delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *24*(4), 320–327.

Krim, H., & Viberg, M. (1996). Two decades of array signal processing research: The parametric approach. *IEEE Signal Processing Magazine*, *13*(4), 67–94.

Krolik, J., & Swingler, D. N. (1990). Focussed wideband array processing via spatial resampling. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *38*(2), 356–360.

Kwon, H., Kim, S., & Bae, K. (2004). Time Delay Estimation Using Weighted CPSP Function. In *International Conference on Spoken Language Processing*, pp. 2853–2856.

Lamel, L., Schiel, F., Fourcin, A., Mariani, J., & Tillman, H. (1994). The Translanguage English Database (TED). In *International Conference on Spoken Language Processing*, pp. 1795–1798.

Leggetter, C. J., & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, *9*(2), 171–185.

Luque, J., Morros, R., Garde, A., Anguita, J., Farrús, M., Macho, D., F.Marqués, Martínez, C., Vilaplana, V., & Hernando, J. (2007). Audio, video and multimodal person identification in a smart room. *Lecture Notes in Computer Science, Springer-Verlag*, *4122*, 258–269.

Marro, C., Mahieux, Y., & Simmer, K. (1998). Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *SAP*, *6*(3), 240–259.

Masgrau, E., Aguilar, L., & Lleida, E. (1999). Performance comparison of several adaptive schemes for microphone array beamforming. In *European Conference on Speech Communication and Technology*, Vol. 6, pp. 2615–2618.

Matassoni, M., Omologo, M., & Giuliani, D. (2000). Hands-Free Speech Recognition Using Filtered Clean Corpus and Incremental HMM Adaptation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 1407–1410.

Mazor, E., Averbuch, A., Bar-Shalom, Y., & Dayan, J. (1998). Interacting multiple model methods in target tracking: A survey. *IEEE Transactions on Aerospace and Electronics Systems*, *34*(1), 103–123.

McCowan, I. A., Moore, D. C., & Sridharan, S. (2000). Speech enhancement using near-field superdirectivity with an adaptive sidelobe canceller and post-filter. In *Australian International Conference on Speech Science and Technology*, pp. 268–273.

McCowan, I. A., Moore, D. C., & Sridharan, S. (2002). Near-field Adaptive Beamformer for Robust Speech Recognition. *Digital Signal Processing: A Review*, *12*(1), 87–106.

McCowan, I. A. (2001). *Robust Speech Recognition Using Microphone Arrays*. Ph.D. thesis, Queensland University of Technology.

Miyoshi, M., & Kaneda, Y. (1988). Inverse filtering of room acoustics. *IEEE Transactions on Speech and Audio Processing*, *36*(2), 145–152.

Mungamuru, B., & Aarabi, P. (2004). Enhanced Sound Localization. *IEEE Transactions on Systems, Man and Cybernetics*, *34*(3), 1526–1540.

Mungamuru, B. (2003). *Enhanced Sound Localization*. Ph.D. thesis, University of Toronto.

Nadeu, C., Macho, D., & Hernando, J. (2001). Time and frequency filtering of filter-bank energies for robust HMM speech recognition. *Speech Communication*, *34*, 93–114.

Omologo, M., & Svaizer, P. (1994). Acoustic event localization using a crosspower-spectrum phase based technique. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 273–276.

Omologo, M., Matassoni, M., Svaizer, P., & Giuliani, D. (1997). Microphone Array Based Speech Recognition with Different Talker-Array Positions. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 227–230.

Parra, L. C. (2006). Steerable Frequency-Invariant Beamforming for Arbitrary Arrays. *The Journal of the Acoustical Society of America*, *119*(6), 3839–3847.

Paul, D., & Baker, J. M. (1992). The Design for the Wall Street Journal-based CSR Corpus. In *DARPA Speech and Natural Language Workshop*.

Pearce, D. (1998). Experimental Framework for the Performance Evaluation of Distributed Speech Recognition Front-ends. Tech. rep..

Peterson, J. M., & Kyriakakis, C. (2005). Hybrid algorithm for robust, real-time source localization in reverberant environments. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4, pp. 1053–1056.

Potamitis, I., Tremoulis, G., & Fakotakis, N. (2003a). Multi-Speaker DOA Tracking Using Interactive Multiple Models and Probabilistic Data Association. In *European Conference on Speech Communication and Technology*, pp. 517–520.

Potamitis, I., Tremoulis, G., Fakotakis, N., & Kokkinakis, G. (2003b). Multi-array fusion for beamforming and localization of moving speakers. In *European Conference on Speech Communication and Technology*, Vol. 3, pp. 1721–1724.

Rabiner, L., & Juang, B. H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–285.

Rabinkin, D. V. (1998). *Optimum sensor placement for a Beamforming Microphone Array.* Ph.D. thesis, Rutgers University.

Rao, B. D., & Hari, K. V. S. (1989). Performance Analysis of Root-Music. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *37*(12), 1939–1949.

Raub, D., McDonough, J., & Wolfel, M. (2004). A cepstral domain maximum likelihood beamformer for speech recognition. In *International Conference on Spoken Language Processing*, pp. 817–820.

Raykar, V. C., Duraiswami, R., Yegnanarayana, B., & Mahadeva Prasanna, S. R. (2003). Tracking A Moving Speaker using Excitation Source information. In *European Conference on Speech Communication and Technology*, pp. 69–72.

Reed, F., Feintuch, P., & Bershad, N. (1981). Time delay estimation using the LMS adaptive filter - Static behavior. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *29*(3), 561–571.

Sachar, J. M., & Silverman, H. F. (2004). A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4, pp. 65–68.

Sagayama, S., Yamaguchi, Y., Takahashi, S., & Takahashi, J. (1997). Jacobian approach to fast acoustic model adaptation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 835–838.

Sanchez-Bote, J. (2004). *Mejora de señal de voz en condiciones acústicas adversas mediante arrays de micrófonos (in Spanish).* Ph.D. thesis, Universidad Politécnica de Madrid.

Sanchez-Bote, J., Gonzalez-Rodriguez, J., & Ortega-Garcia, J. (2003). A real-time auditory-based microphone array assessed with E-RASTI evaluation proposal. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 5, pp. 477–480.

Satoshi, N., Kazuo, H., Futoshi, A., Takanobu, N., & Takeshi, Y. (2000). Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition. In *International Conference on Language Resources and Evaluation*, pp. 965–968.

Schau, H., & Robinson, A. (1987). Passive source localization employing intersecting spherical surfaces from time-of-arrival differences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *35*(8), 1223–1225.

Schmidt, R. O. (1972). A new approach to geometry of range difference location. *IEEE Transactions on Aerospace and Electronics Systems*, *8*, 821–835.

Segura, C., Cantón, C., Abad, A., Casas, J. R., & Hernando, J. (2007). Multimodal head orientation towards attention tracking in smart rooms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Segura, C. (2004). Untersuchung und Implementierung von Verfahren zum Nachführen akustischer Quellen (in German). Master's thesis, Technische Universität Berlin.

Seltzer, M. L., Raj, B., & Stern, R. M. (2004). Likelihood-maximizing beamforming for robust hands-free speech recognition. *IEEE Transactions on Speech and Audio Processing*, *12*(5), 489–498.

Seltzer, M. L. (2003). *Microphone Array Processing for Robust Speech Recognition*. Ph.D. thesis, Carnegie Mellon University.

Smith, J. O., & Abel, J. S. (1987). Closed-form least-squares source location estimation from range-difference measurements. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *35*(12), 1661–1669.

Stiefelhagen, R., & Garofolo, J. (Eds.). (2007). *Multimodal Technologies for Perception of Humans. First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006*, Vol. 4122. Lecture Notes in Computer Science, Springer-Verlag.

Strobel, N., Meier, T., & Rabenstein, R. (1999). Speaker Localization Using a Steered Filter-And-Sum Beamformer. In *Workshop on Vision, Modeling, and Visualization* Erlangen.

Strobel, N., Spors, S., & Rabenstein, R. (2001). Joint audio-video object localization and tracking. *IEEE Signal Processing Magazine*, *18*(1), 22–31.

Sturim, D. E., Brandstein, M. S., & Silverman, H. F. (1997). Tracking multiple talkers using microphone-array measurements. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 371–374.

Sullivan, T. M. (1996). *Multi-Microphone Correlation-Based Processing for Robust Automatic Speech Recognition*. Ph.D. thesis, Carnegie Mellon University.

Täger, W. (1998). Near field superdirectivity (NFSD). In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4, pp. 2045–2048.

Tebelskis, J. (1995). *Speech Recognition Using Neural Networks*. Ph.D. thesis, Carnegie Mellon University.

Van Compernolle, D. (1990). Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 833–836.

Van Compernolle, D., & Van Gerven, S. (1995). *COST 229: Applications of Digital Signal Processing to Telecommunications*, chap. Beamforming with Microphone Arrays, pp. 107–131.

Van Veen, B. D., & Buckley, K. M. (1988). Beamforming: A versatile approach to spatial filtering. *IEEE Acoustics, Speech and Signal Processing Magazine*, *5*(2), 4–24.

Vermaak, J., & Blake, A. (2001). Nonlinear filtering for speaker tracking in noisy and reverberant environments. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 3021–3024.

Vermaak, J., Gangnet, M., Blake, A., & Perez, P. (2001). Sequential Monte-Carlo fusion of sound and vision for speaker tracking. In *International Conference on Computer Vision*, Vol. 1, pp. 741–746.

Viikki, O., & Laurila, L. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, *25*, 133–147.

Viste, H. (2004). *Binaural localization and separation techniques*. Ph.D. thesis, École Polytechnique Fédérale de Lausanne.

Vo, B.-N., Singh, S., & Ma, W. K. (2004). Tracking multiple speakers using random sets. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 357–360.

Wang, H., & Chu, P. (1997). Voice source localization for automatic camera pointing system in videoconferencing. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 187–190.

Ward, D. B., Kennedy, R. A., & Williamson, R. C. (1995). Theory and design of broadband sensor arrays with frequency invariant far-field beam patterns. *The Journal of the Acoustical Society of America*, *97*, 1023–1034.

Ward, D. B., Lehmann, E. A., & Williamson, R. C. (2003). Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Transactions on Speech and Audio Processing*, *11*(6), 826–836.

Ward, D. B., & Williamson, R. C. (2002). Particle filter beamforming for acoustic source localization in a reverberant environment. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 1777–1780.

Warsitz, E., Häb-Umbach, R., & Peschke, S. (2004). Adaptive Beamforming Combined with Particle Filtering for Acoustic Source Localization. In *International Conference on Spoken Language Processing*, Vol. 5, pp. 2849–2852.

Welch, G., & Bishop, G. (1995). An introduction to the kalman filter. Tech. rep. TR 95-041, University of North Carolina, Department of Computer Science.

Widrow, B., & Stearns, S. D. (1985). *Adaptive signal processing.* Prentice-Hall.

Yamada, T., Nakamura, S., & Shikano, K. (1996). Robust speech recognition with speaker localization by a microphone array. In *International Conference on Spoken Language Processing*, pp. 1317–1320.

Yan, Z., Du, L., Wei, J., & Zeng, H. (2003). Time Delay Estimation Based On Hearing Characteristic. In *European Conference on Speech Communication and Technology*, pp. 557–560.

Youn, D., Ahmed, N., & Carter, G. (1982). On using the LMS algorithm for time delay estimation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *30*(5), 798–801.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., & Woodland, P. (2002). *The HTK Book* (3.2.1 edition). Cambridge University Engineering Department.

Yu, Y., & Silverman, H. F. (2004). An improved TDOA-based location estimation algorithm for large aperture microphone arrays. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4, pp. 77–80.

Zelinski, R. (1988). A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 5, pp. 2578–2581.

Zotkin, D. N., & Duraiswami, R. (2004). Accelerated speech source localization via a hierarchical search of steered response power. *IEEE Transactions on Speech and Audio Processing*, *12*(5), 499–508.