Universidad de Alcalá

DEPARTAMENTO DE TEORÍA DE
LA SEÑAL Y COMUNICACIONES
Escuela Politécnica Superior
Campus Universitario s/n
28805 Alcalá de Henares (Madrid)
Telf: +34 91 885 66 90
Fax: +34 91 885 66 99
dpto.teoriadelasenal@uah.es

D. ROBERTO GIL PITA Y D. MANUEL ROSA ZURERA, Profesores Titulares de Universidad del Área de Conocimiento de Teoría de la Señal y Comunicaciones de la Universidad de Alcalá,

CERTIFICAN

Que la tesis **"Multi-channel Speech Separation in Reverberant Environments"**, presentada por D. Cosme Llerena Aguilar, realizada en el Departamento de Teoría de la Señal y Comunicaciones bajo nuestra dirección, reúne méritos suficientes para optar al grado de Doctor, por lo que puede procederse a su depósito y defensa.

Alcalá de Henares, 27 de Enero de 2016.

Fdo. Dr. D. Roberto Gil Pita

Fdo. Dr. D. Manuel Rosa Zurera

Universidad de Alcalá

DEPARTAMENTO DE TEORÍA DE
LA SEÑAL Y COMUNICACIONES
Escuela Politécnica Superior
Campus Universitario s/n
28805 Alcalá de Henares (Madrid)
Telf: +34 91 885 66 90
Fax: +34 91 885 66 99
dpto.teoriadelasenal@uah.es

D. SANCHO SALCEDO SANZ, Coordinador del Programa de Doctorado en Tecnologías de la Información y las Comunicaciones,

CERTIFICA

Que D. Cosme Llerena Aguilar ha realizado en el Departamento de Teoría de la Señal y Comunicaciones y bajo la dirección de los Doctores D. Roberto Gil Pita y D. Manuel Rosa Zurera, la tesis doctoral titulada **"Multi-channel Speech Separation in Reverberant Environments"**, cumpliéndose todos los requisitos para la tramitación que conduce a su posterior defensa.

Alcalá de Henares, 27 de Enero de 2016.

Fdo. Dr. D. Sancho Salcedo Sanz

Escuela Politécnica Superior

Departamento de Teoría de la Señal y Comunicaciones

Ph. D. Thesis

Multi-channel Speech Separation in Reverberant Environments

Author:
Cosme Llerena Aguilar

Supervisors:
Roberto Gil Pita, Ph. D.
Manuel Rosa Zurera, Ph. D.

March, 2016

Abstract

Humans are capable of following a single speaker and understanding what that speaker is saying in rooms where many people are speaking. This ability is enormously complex since many processes are included. In the scientific community this problem is known as cocktail party problem and has been subject of study for decades. The development of technical solutions presents a challenge due to its complexity. Indeed, many of the processes performed by the human auditory system are still unknown. In general terms, there are two ways of understanding the cocktail party problem, as a human speech recognition problem or as a speech separation issue. In this thesis we will focus on the second class of problems.

Concerning separation problems, many methodologies can be found in the literature. The type of separation solution depends on the circumstances in which the separation problem take places, that is, on factors as for instance, the number of speakers and microphones, the level of noise and reverberation, among others. Perhaps one of the main problems for speech separation algorithms is reverberation, which causes that many of them do not perform correctly. And so, the necessity of robust separation algorithms against reverberation seems to be clear. In order to solve the separation problem in reverberant environments, three main groups of techniques can be distinguished. CASA techniques are the first group of solutions, but it must be said that this line of research is not the most successful one. Other types of proposals are beamforming techniques which present important problems in the presence of reverberation. Beamforming techniques are based on the principle of spatial filtering, that is, they are specially designed to enhance and eliminate signals coming from specific directions. Reverberation can be considered as interferences that can come from any direction and so, beamforming is not the most suitable solution for eliminating the effects of reverberation. Furthermore, broadly speaking, beamforming methods require sophisticated sensor arrays and entail important computational load. These drawbacks limit the use of beamforming in many applications. The third group of solutions are BSS techniques, which are based on statistical and other signal properties. In this PhD thesis we focus on the latter group of techniques.

Many of these techniques are jointly used with different tools to develop signal separation. Within these tools, sensor networks play a key role. Nowadays, the use of wireless sensor networks is becoming very popular since they entail many advantages. However, these networks have some particularities that give rise to some problems for classical separation algorithms engineered to wired networks.

With this in mind, this thesis can be divided into two major parts. The first one is related to the design of new robust speech separation techniques against reverberation. Additionally, it will be very important to obtain computationally efficient separation methods, without requiring complex microphone networks. The second

part deals with, perhaps, the most important problem for classical speech separation algorithms when WASNs are used, the synchronization problem.

With respect to the first part, a new separation procedure has been introduced. The main requirements of this solution are: it must overcome classical BSS algorithms in reverberant environments, the simplest (two-) microphone array is used, and it must be computationally efficient. Aiming at comparing it with classical BSS methods, comparable algorithms should be chosen. One of the most successful types of separation methods are those based on sparsity. Within these methods, the well-known DUET algorithm has been selected since it works with only two microphones in echoic problems, achieving acceptable outcomes. To make a valid comparison, a study to determine the best microphone array configurations and frame lengths for DUET in our separation problems has been carried out. Using these configurations, the DUET algorithm is ready to be compared with our proposal. Moreover, the separation stage of DUET based on time-frequency binary masking has also be compared with another very popular masking technique that relies on l_1-norm minimization. From this study, we have observed that binary masking outperforms the other one.

Our separation procedure performs the estimation of the mixing matrix based on a geometric analysis of the separation scenario. Knowing available information, such as the microphone separation, the mutual angle or the type of microphones, the mixing matrix is estimated. Mixing parameters have two components, time and level differences. Using our mixing matrix estimation method, only time differences are calculated since a relationship between both differences has been established. It involves that our separation method is computationally less expensive. Furthermore, to avoid the estimation of level differences is important since it is a very difficult task in the presence of reverberation. To estimate time differences, TDE methods are used. In this sense, one of the most robust TDE method against reverberation has been used, the GCC-PHAT algorithm. A study has demonstrated its suitability in all our separation problems except when microphone separations are small. Considering it, a new TDE method has been developed for small arrays, obtaining very good results. Finally, comparing our separation solution with DUET, it has been demonstrated that our proposal outperforms DUET in all the separation scenarios (level of reverberation, number of talkers, etc.). It must also be mentioned that both DUET and our proposal have a separation stage based on binary masking which introduces an important problem in acoustic applications, the musical noise. To minimize this problem, a musical noise reduction algorithm has been proposed, demonstrating good outcomes.

In the second part of the thesis, we have tackled the introduction of classical BSS algorithms that use short-time analysis tools in WASNs. Perhaps the main problem of those BSS algorithms in WASNs is the desynchronization of the signals received at the different nodes. In this sense, a novel synchronization methodology based on signal processing has been proposed. The first aspect to be mentioned is the novelty of considering the differences in propagation delays, while traditional synchronization solutions only deal with the clock problem. A theoretical analysis has been developed to establish the theoretical delay between speech mixtures. Moreover, two new TDE methods according to our theoretical delay have been implemented. These methods have the additional bonus of using a reduce amount of information for transmission and they do not require many computational resources. A study reveals that with our synchronization solution, classical BSS algorithms can be introduced in WASNs.

Resumen

El ser humano es capaz de seguir y entender lo que un individuo está diciendo en salas en las que hay más personas hablando. Esta habilidad es enormemente compleja pues abarca numerosos procesos. En la comunidad científica este problema se conoce como atención selectiva (aunque es más conocido por su nombre en inglés, *cocktail party problem*) y ha sido objeto de estudio durante décadas. El desarrollo de soluciones técnicas a este problema se presenta como un reto debido a su complejidad. De hecho, muchos de los procesos que tienen lugar en el sistema auditivo humano son aún desconocidos. En términos generales, hay dos formas de tratar el problema de la atención selectiva, como un problema de reconocimiento del habla o como uno de separación. En esta tesis nos centraremos en la segunda clase problemas.

En lo concerniente a los problemas de separación, se puede encontrar en la bibliografía gran cantidad de metodologías. El tipo de solución de separación depende de las circunstancias en las que el problema tiene lugar, es decir, depende de factores, como por ejemplo, el número de hablantes y micrófonos, los niveles de ruido y reverberación, entre otros. Quizás, uno de los principales problemas para los algoritmos de separación del habla es la reverberación, provocando que muchos de ellos no funcionen correctamente. Y por tanto, parece bastante clara la necesidad de desarrollo de algoritmos de separación robustos frente a la reverberación. Para resolver el problema de separación en entornos reverberantes, se puede distinguir tres grupos importantes de técnicas. El primer grupo son las técnicas de análisis computacional de la escena auditiva (también conocidas como CASA, de sus siglas en inglés, *Computational Auditory Scene Analysis*), aunque no ha sido la línea de investigación más exitosa. Otro tipo de propuestas son las técnicas de conformado de haz que presentan algunos problemas importantes en presencia de reverberación. Las técnicas de conformado de haz se basan en el principio de filtrado espacial, es decir, están especialmente diseñadas para realzar y eliminar señales que provengan de unas determinadas direcciones. La reverberación se puede considerar como interferencias que pueden llegar de cualquier dirección, y por tanto, las técnicas de conformado de haz no son las más idóneas para eliminar los efectos de la reverberación. Además, en términos generales, estos métodos requieren conjuntos de sensores sofisticados y conllevan una carga computacional importante. Estas desventajas limitan el uso de las técnicas de conformado de haz en muchas aplicaciones. El tercer grupo de soluciones son las técnicas de separación ciega de fuentes (también conocidas como BSS, de sus siglas en inglés *Blind Source Separation*) que se basan en propiedades estadísticas y algunas otras propiedades de las señales. En esta tesis doctoral centraremos nuestra atención es este último grupo de técnicas.

Muchas de esas técnicas se utilizan conjuntamente con diversas herramientas para conseguir la separación de señales. De esas herramientas, las redes de sensores juegan un papel crucial. En la actualidad, el uso de redes inalámbricas de sensores es muy

popular debido a que conllevan numerosas ventajas. Sin embargo, estas redes poseen algunas peculiaridades que dan lugar a problemas para muchos algoritmos clásicos de separación diseñados para redes no inalámbricas.

Con esto en mente, esta tesis está dividida en dos partes principales. La primera está relacionada con el diseño de técnicas de separación del habla robustas frente a reverberación. Adicionalmente, será muy importante que estas técnicas de separación sean eficientes desde el punto de vista computacional y no requieran de redes de micrófonos complejas. La segunda parte trata, quizás, el problema más importante para los algoritmos clásicos de separación del habla cuando redes acústicas de sensores inalámbricas (también conocidas como WASNs, de sus siglas en inglés *Wireless Acoustic Sensor Networks Separation*) son utilizadas, el problema de la sincronización.

Con respecto a la primera parte, un nuevo método de separación se ha introducido. Los requisitos de este método son: debe funcionar mejor que ciertos algoritmos BSS clásicos en entornos reverberantes, debe utilizar la matriz de micrófonos más sencilla, es decir, solo dos micrófonos, y debe ser eficiente desde el punto de vista computacional. Para comparar nuestro método con algoritmos clásicos de separación, algún método comparable debe ser elegido. Uno de los tipos de métodos de separación más exitosos son aquellos basados en representaciones dispersas. Y dentro de estos métodos, el famoso algoritmo DUET ha sido seleccionado pues trabaja con solo dos micrófonos en entornos reverberantes, obteniendo resultados aceptables. Para hacer una comparación valida, se ha llevado a cabo un estudio para determinar las mejores configuraciones de matrices de micrófonos y tamaños de trama para DUET en nuestros problemas de separación. Usando esas configuraciones, el algoritmo DUET puede ser comparado con nuestra propuesta. Además, la etapa de separación del algoritmo DUET, que se basa en enmascaramiento binario en tiempo-frecuencia, ha sido comparada con otra técnica popular de enmascaramiento que se fundamenta en el uso de minimización con la norma L1. De este estudio, hemos observado que la máscara binaria supera a la otra.

Nuestro procedimiento de separación realiza la estimación de la matriz de mezclas basándose en una análisis geométrico del escenario de separación. Conociendo información disponible como la separación entre micrófonos, el ángulo mutuo o el tipo de micrófono, la matriz de mezcla es estimada. Los parámetros de mezcla tienen dos componentes, las diferencias de tiempo y de nivel. Usando nuestro método de estimación de la matriz de mezclas, solo se calculan las diferencias de tiempo, pues se ha establecido una relación entre ambas diferencias. Esto conlleva que nuestro método de separación sea menos costoso desde el punto de vista computacional. Además, el evitar el cálculo de las diferencias de nivel es importante pues es una tarea muy difícil cuando hay reverberación. Para estimar esas diferencias de tiempo, se utilizarán métodos de estimación de retardos o TDE (del inglés, *Time Delay Estimation*). En etse sentido, uno de los métodos TDE más robusto frente a reverberación se ha usado, el algoritmo GCC-PHAT. Un estudio ha demostrado su idoneidad en todos nuestros problemas de separación excepto cuando la separación de micrófonos es pequeña. Considerando esto, un algoritmo TDE nuevo ha sido desarrollado para matrices de micrófonos pequeñas, obteniendo muy buenos resultados. Finalmente, comparando nuestra solución de separación con DUET, se ha demostrado que supera a DUET en todos los escenarios de separación (niveles de reverberación, número de hablantes, etc.). También se debe mencionar que tanto DUET como nuestra propuesta poseen

una etapa de separación basada en enmascaramiento binario, el cual introduce un problema importante en aplicaciones acústicas, el ruido musical. Para minimizar este problema, un algoritmo de reducción del ruido musical se ha propuesto, demostrando buenos resultados.

En la segunda parte de la tesis, hemos abordado la introducción en WASNs de algoritmos BSS clásicos que utilizan herramientas de análisis de tiempo reducido. Quizás el principal problema de esos algoritmos BSS en WASNs es la desincronización de las señales capturadas en los distintos nodos. En este sentido, una nueva metodología de sincronización basada en procesado de señal ha sido propuesta. El primer aspecto a mencionar es la novedad de considerar las diferencias de los retardos de propagación, en contraste con las soluciones clásicas de sincronizado que solo abordan el problema de desincronización debido al reloj. Un análisis teórico se ha desarrollado para establecer el retardo teórico entre mezclas de voz. Además, dos nuevos métodos TDE acordes a nuestro retardo teórico han sido implementados. Estos métodos poseen la ventaja añadida de usar poca información para transmitir y no requieren muchos recursos computacionales. Nuestro estudio revela que con nuestra solución de sincronización, algoritmos clásicos BSS puedes ser usados en WASNs.

A mis padres, a mis hermanos.

"Empieza por hacer lo necesario, luego haz lo posible
y de pronto estarás logrando lo imposible."

San Francisco de Asís

Agradecimientos

Escribiendo las últimas líneas de esta memoria, me vienen a la cabeza todas las personas que de una u otra forma me han ayudado a acabar esta tesis doctoral. Por tanto, no quiero dejar pasar la oportunidad de agradecérselo.

En primer lugar quiero y debo agradecérselo a dos personas sin cuya ayuda esta tesis no hubiese sido posible, Manolo y Roberto, mis directores.

Roberto, ni que decir tiene que eres parte fundamental de esta tesis. Gracias por compartir conmigo el gran entusiasmo que tienes por tu trabajo, tu enorme capacidad, tu gran iniciativa, el torrente continuo de ideas que tienes. Agradecer también tu paciencia, tu disposición, la cantidad de tiempo y esfuerzo que me has dedicado, no todo el mundo puede decir que su director se sienta con él diariamente. Es un privilegio trabajar contigo.

Manolo, te estaré eternamente agradecido por haberme dado esta oportunidad. Gracias por permitirme entrar en tu grupo de investigación, por abrirme las puertas al procesado de la señal, por enseñarme lo que es la investigación, por transmitirme tu vasta experiencia, tus consejos e ideas. Gracias también por tu calidad humana, por la confianza que nos das a tus estudiantes y por estar siempre ahí para solucionarnos cualquier problema.

Agradecer a Manuel Utrilla su acogida en el grupo de investigación, sus consejos, el compartir su experiencia y sobre todo, esas clases de circuitos que tanto me sirvieron.

También mi gratitud al Dr. Martin Kleinsteuber de la Universidad Técnica de Munich, por permitirme trabajar en su grupo y por el trato recibido.

En general, a todos los docentes que han intervenido en mi formación.

A los numerosos compañeros que han pasado por el fondo S31, por tanto compartido, por vuestra ayuda, por todos los buenos y no tan buenos momentos que hemos pasado. De entre todos ellos, debo dar las gracias muy especialmente a Lorena, que has terminado siendo una amiga. Siempre que te lo he pedido me has ayudado, has sido una gran maestra. Te deseo lo mejor en tu carrera como docente.

Empezando con mis amigos y dado el contexto en el que estoy, debo empezar por los Telequines. Ha sido un verdadero lujo compartir tantos años de fatigas y alegrías con vosotros. En lo que se refiere a esta tesis me veo obligado a mencionar a dos de ellos. Por un lado, a Delia, gracias por hacer que mi estancia en Munich fuese fantástica, presentándome a gente estupenda. También, y como no podía ser de otra manera, a Raquel, mi eterna compañera de laboratorio y de penurias durante esta tesis. Ánimo y a por todas, que ya lo tienes.

Al Cofi parchí, que habéis hecho que las horas del café y la comida fuesen un verdadero espectáculo, se os sigue echando mucho de menos.

Y como no, al resto de mis amigos, al Tomillo 2, a Franelina, a las Churreras, a Power Herminia, a los 3 Fantásticos, a las muniquesas y a unos cuantos más a los que os tengo que poner mote ya mismo. Habéis hecho que la vida sea mucho mejor.

A Dios, por todo lo que me da sin pedir nada y por estar en todo momento conmigo.

El mayor de los agradecimientos a mi familia, por su apoyo incondicional. Especialmente a mis padres, Pío y Teresa, todo esto es por y para vosotros. Por la entrega total a vuestros hijos, por darme esta oportunidad y por ser modelo a seguir. A mis hermanos, Estrella, Francisco y Julio, por ser fundamentales en mi vida. Por vuestros apoyos y ánimos durante estos cuatro años. A Francisco, que también lo disfrutará. A mis abuelos, que tanto lucharon por darle estudios a mis padres, para que tuviesen una vida mejor. Esto lo podéis considerar una extensión de lo que vosotros empezastéis.

La presente Tesis Doctoral ha sido financiada por el Ministerio de Economía y Competitividad bajo el proyecto TEC2012-38142-C04-02, por el Ministerio de Educación, Cultura y Deporte bajo el proyecto TEC2015-67387-C4-4-R y por medio del programa de becas FPI de la Universidad de Alcalá.

En general, a todos aquellos que estuvistéis, que estáis y que estaréis.

Cosme

Contents

List of Figures

List of Tables

List of Symbols

$x[n]$	Discrete-time signal
f_s	Sampling rate
S	Number of sources
M	Number of mixtures
$x_s[n]$	Sound signal produced by the s-th source
$z_m[n]$	Signal received by the m-th microphone
$y_{m0}[n]$	Noise signal received by the m-th microphone
$a_{ms}[n]$	Acoustic impulse response between the s-th source and the m-th microphone
α_{ms}	Attenuation of the signal that travels from the s-th source to the m-th microphone
δ_{ms}	Time delay of the signal that travels from the s-th source to the m-th microphone
δ_K	Kronecker delta function
k	Frequency index
K	Number of frequency bands
l	Time frame index
L	Number of time frames
ω_k	Central frequency of the k-th frequency bin of the STFT
$X_s(k,l)$	STFT of the signal produced by the s-th source
$Z_m(k,l)$	STFT of the signal received by the m-th microphone
$Y_{m0}(k,l)$	STFT of the noise signal received by the m-th microphone
$\mathbf{A}[n]$	Mixing matrix in the time domain at the n-th time instant
$\mathbf{A}(k)$	Mixing matrix at the k-th frequency bin
$A_{ms}(k)$	Coefficient of the mixing matrix in the time-frequency domain
g_{ms}	Level difference between the m-microphone and the reference sensor for the s-th source
τ_{ms}	Time difference between the m-microphone and the reference sensor for the s-th source
ψ_s	Set of time-frequency points of the mixtures in which only the s-th source is working
$M_s(k,l)$	Time-frequency mask for the s-th source
θ	Azimuth angle
ϕ	Elevation angle
α	Mutual angle
m_m	Direction of the maximum sensitivity of the m-th microphone in Cartesian coordinates
EL_{ms}	Effective length of the acoustic impulse response between the s-th source and the m-th microphone
p_q	Direction of arrival of a incident sound wave due to the q-th source
α_m	Time delay due to clock differences in nodes
β_m	Time delay due to RF communication between the m-th node and the central one
Δ_m	Time delay between the m-th mixture and the reference one
$r_{pq}[\tau]$	Time correlation between the p-th and q-th speech mixtures
$(\cdot)^H$	Conjugate transpose operator

$(\cdot)^T$	Transpose operator
$(\cdot)^*$	Complex conjugation
$\|\|x\|\|_1$	l_1-norm

Glossary

A

ADC Analog-to-Digital Converter.

AED Adaptive Eigenvalue Decomposition.

AMNOR Adaptive Microphone arrays system for NOise Reduction.

ASA Auditory Scene Analysis.

ASDF Average Square Difference Function.

ATF Acoustic Transfer Function.

B

BASS Blind Audio Source Separation.

BSS Blind Source Separation.

C

CASA Computational Auditory Scene Analysis.

CDB Constant Directivity Beamformer.

D

DFT Discrete Fourier Transform.

DOA Direction Of Arrival.

DS Delay and Sum.

DUET Degenerate Unmixing Estimation Technique.

DWT Discrete Wavelet Transform.

E

ECG ElectroCardioGram.

EEG ElectroEncephaloGraphy.

EM Expectation-Maximization.

EMG ElectroMyoGraphy.

F

FBR Front-to-Back Ratio.

FFT Fast Fourier Transform.

FIR Finite-Impulse Response.

G

GCC Generalized Cross-Correlation.

GSC Generalized Sidelobe Canceller.

H

HOS Higher-Order Statistics.

I

ICA Independent Component Analysis.

IDFT Inverse Discrete Fourier Transform.

ILD Interaural Level Difference.

IPD Interaural Phase Difference.

ITD Interaural Time Difference.

L

LCMV Linearly Constrained Minimum Variance.

LMM Laplacian Mixture Model.

LMS Least Mean Squares.

LOST Line Orientation Separation Technique.

LS Least Squares.

LTI Linear Time-Invariant.

M

MAC Medium Access Control.

MAP Maximun A Posteriori.

MDCT Modified Discrete Cosine Transform.

MIPS Millions of Instruction Per Second.

ML Maximum Likelihood.

MSE Mean Square Error.

MVDR Minimum Variance Distorsionless Response.

N

NMF Non-negative Matrix Factorization.

P

PCA Principal Component Analysis.

PDF Probability Density Function.

R

RF Radio Frequency.

RIRG Room Impulse Response Generator.

RMSE Root Mean Squared Error.

S

SCOT Smoothed COherence Transform.

SDB SuperDirective Beamformer.

SIR Signal-to-Interference Ratio.

SNR Signal-to-Noise Ratio.

SRO Sampling Rate Offset.

SSS Sound Source Separation.

STFT Short-Time Fourier Transform.

STLE Short-Term Log-Energy.

STOI Short-Time Objective Intelligibility.

T

TDE Time Delay Estimation.

TDOA Time Difference Of Arrival.

TOA Time Of Arrival.

W

WASN Wireless Acoustic Sensor Network.

WDO W-Disjoint Orthogonality.

WSN Wireless Sensor Network.

Part I

Preliminary work

Chapter 1

Introduction and scope

1.1 Introduction

Humans have the ability to distinguish an individual sound source from a mixture of sounds reaching their ears. This ability is quite complicated and currently it is not fully understood. One particular case of the use of this ability is related to the perception of speech in complex acoustic environments. Considering a room where many people are speaking, a human listener is able to put the attention on a single speaker and understand what that speaker is saying. This problem is known as *Cocktail Party Problem*, which was first stated and popularized in [Cherry, 1953]. The author focused on the attentional component of the problem, observing if listeners can select one speech signal over another, how they can switch the attention between speech signals and whether listeners can retain anything about the non-selected signal. In order to develop technical solutions for this problem, Cherry suggested to consider the following factors:

- The direction whence the voices come from.
- Characteristics of speaking voices, such as speed, pitch, accents, among others.
- Transition probabilities.
- Visual information as gestures, lip-reading, etc.

Technically, to solve the cocktail party problem is not easy because there are many processes at play, such as the human speech production, the auditory systems and language processing. For instance, the ear obtains frequency decompositions on sound signals by means of sensory organs as the cochlea and after, the brain may use that information. Brain develops complex tasks as inferring the correct sounds or estimating the sound of primary interest. It must also be considered that the human auditory system carries out this task with only two sensors (ears) and without any prior knowledge of the sources or the acoustic environment. For these reasons, and many others, it is very difficult to find technical solutions for this problem. Cherry proposed that it is possible to solve the problem similarly from a technical point of view, but it must be considered:

> *"On what logical principles could one design a machine whose reaction, in response to speech stimuli, would be analogous to that of a human being? How could it separate one of two simultaneous spoken messages?"*

The cocktail party problem has been studied for more than six decades and many studies and experiments have been performed to understand how humans can attend to one speaker when

more people are speaking. Some questions like *how can we segregate speech sounds?* or *what cues in speech signal are used for separating one speech signal from others?* have been raised for many years. The objective of works in the literature has been to understand the human auditory system as well as possible aiming at imitating its behavior.

Many works have focused on characterizing different aspects of the human auditory system. For instance, in [Cherry, 1953] the time interval required to transfer attention between ears is determined, proving that a time over 170 ms is observed in many subjects. In [Spieth et al., 1954] the responses when simultaneous messages reach our ears are analyzed. In their experiments many configurations are tested, varying the number of loudspeakers, placing loudspeakers with different azimuths or separations between them, etc. Moreover, the authors conclude that some techniques as filtering the signals that arrive at ears, or spatial separation of sound sources help to better listen to either stream. The above-mentioned works study the response of a subject to one of two simultaneous messages. Other studies analyze the response of one subject to both of two simultaneous messages. This approach was presented in [Webster and Thompson, 1954] where the amount of information received when two overlapping messages reach the listener is formally studied.

Another relevant work is [Broadbent, 2013] which deals with the issue of selective listening. The author ensures that the role of the central nervous system is more important than the one played by sensory factors when information is discarded. Broadbent also determines some factors that help to eliminate irrelevant information. For instance, whether irrelevant information comes from different place or has not the same frequency components as the relevant one, the auditory system will discard this information more easily.

Nearly 20 years after proposing Cherry the concept of cocktail party problem, sound segregation that was termed auditory scene analysis (ASA) was introduced in [Bregman, 1990]. This line of research within the human cocktail party problem relates the behavior of the auditory system to vision. These techniques use properties of speech, but they do not employ language knowledge. ASA techniques must solve some problems, such as determining the number of speech sources, their characteristics or their locations. The success of Bregman is to discover many aspects of the human auditory system that are relevant to solve the cocktail party problem. For instance, Bregman observes two aspects: spatial cues play a key role and are stronger when they are combined with non-spatial cues, and humans tend to perceive sounds that come from locations of visual events. These ideas are for general sound sources, but Bregman also establishes some specific and important aspects of speech within auditory scene analysis. Pitch trajectory or spectral continuity are illustrative examples of useful properties of speech that are used to solve the cocktail party problem. Auditory scene analysis will be studied in greater detail in next sections of this chapter.

Despite the large number of studies, methods and algorithms aiming at imitating the human auditory system, it must be highlighted that there are not algorithms achieving similar results to the ones of our auditory system. The main reason is that its functioning is partially understood since many important aspects are unknown. For instance, it is not known how speech signals are grouped and segmented from mixtures, the influence of linguistics or the mechanisms of auditory attention. Nowadays, the development of machines that are able to listen in the same way as humans do is an important line of research.

Broadly speaking, the cocktail party problem can be understood as two groups of problems: 1) human speech recognition and 2) speech separation problems. In respect of the first group, it can be said that many speech recognition algorithms obtain excellent results when one speaker is alone in quiet rooms, but they do not perform correctly in noisy and reverberant rooms or when there are more talkers in rooms. Recognizing speech in the presence of competing speech-like

distortions is also considered as cocktail party problem [Huang et al., 2001]. In this thesis, we will focus on the second problem, that is, on separating speech sources from speech mixtures, what is not a trivial task. The engineering community has a strong interest in this issue since in real-world conditions speech separation is an unresolved problem.

This thesis deals with the cocktail party problem from different points of view. The development of speech separation algorithms in reverberant environments is challenging since the vast majority of them do not perform correctly in the presence of reverberation. In a first approach, we put our attention on designing new separation methods for solving the cocktail party problem under these circumstances. One important aspect to be considered is that if cheap and feasible solutions must be obtained, then the microphone array should not be vey complex. Some of the most important sound separation algorithms are based on the use of time-frequency binary masking which introduces musical noise artifacts in the separated sources. Solutions aiming at minimizing the presence of musical noise have also been studied in this thesis.

In general terms, many classical speech separation algorithms in the literature rely on the use of wired networks with central processing and microphone separations of the order of centimeters. Nowadays, there is a great deal of interest in using wireless sensor networks (WSN) and acoustic sensor networks are no exception. The use of wireless acoustic sensor networks (WASNs) is becoming very popular day by day since their characteristics involve a large number of advantages, while it is true that some aspects of WASNs may lead to some problems [Bertrand, 2011]. Factors such as the independency of the clocks of the different nodes, the acoustic effects of large distances between sensors, or the possibility of nodes changing their position, introduce important delay differences between the recordings at the different microphones, which causes that classical separation algorithms do not perform correctly. In order to minimize the effects of desynchronization between sensor nodes, a large number of synchronization algorithms can be found in the literature [Elson et al., 2002, Wu et al., 2011, Schmalenstroeer et al., 2015]. The vast majority of them are clock synchronization protocols and algorithms in the communication layer, but these alternatives do not tackle the lack of synchronization caused by the different distances between sources and nodes [Llerena-Aguilar et al., 2016]. In this thesis our objective has been to provide synchronization solutions from a new perspective based on signal processing techniques.

The remaining of this chapter contains: (I) a comprehensive review of the state of the art of sound source separation, which covers three important types of separation methodologies: computational auditory scene analysis (CASA), beamforming and blind source separation (BSS) techniques. (II) Another important review of the state of the art of synchronization solutions in wireless networks and finally, (III) the chapter ends with a description of the main goals and structure of this thesis.

1.2 Sound source separation: state of the art

The sound source separation (SSS) problem consists in decomposing a real auditory scene into individual sound sources. The human auditory system is able to easily carry out this task despite severe constraints, such as the use of only two sensors (ears) or its development in real time. Humans can put their attention on a specific sound source of a sound mixture, ignoring the rest of sound objects. Imagine a conversation in a crowded room where people can follow a speaker in spite of the presence of more people talking and background noise. From the point of view of technical solutions [Vincent et al., 2012], the problem has only been partially solved since it entails high complexity. Classical audio source separation algorithms do not perform correctly in the presence of certain factors as reverberation, noise or similar interfering sound sources to

the target one. Different solutions will be analyzed in the following paragraphs.

Audio source separation is an important line of research and is presented in a large number of applications:

- *Surveillance applications.* To extract a sound source from a mixture signal can be very useful in this type of applications. Representative examples of it are police applications where speech signals are extracted in order to get any kind of information as speaker identity or the contents of a conversation.

- *Medical applications.* Audio source separation is presented in different medical applications. For instance, sometimes breath sounds are eliminated when heartbeat is studied. Another separation problem occurs when foetus's heartbeat is analyzed by means of ultrasound based Doppler instruments and it is necessary to separate it from the mother's heartbeat that masks it.

- *Music transcription.* This task consists in separating the instruments that are playing in a recording to facilitate the work to music transcribers.

- *Remixing of studio recordings.* New tools appear aiming at extracting sounds from many type of recordings and then, they are available to be mixed in other recordings.

- *Noise suppression.* This process is presented in many devices such as mobile phones or hearing aids. Noise suppression can be understood as a part of the field of research known as speech enhancement. The objective is to eliminate undesirable noise components in order to achieve more intelligible sound signals.

- *Post-processing of recordings.* This type of post-processing stages are included in films, songs and other kind of recordings aiming at editing them, obtaining the expected results. For instance, it is very usual to extract actors' voice in films in order to dubbing in other languages.

- *Efficient coding of music.* Each sound signal due to a musical instrument has different characteristics from the ones of the rest of instruments. Examples of these characteristics are pitch or bandwidth. These particular characteristics demand different bandwidth for transmission. If sound sources are decomposed, each source signal can be independently encoded and compressed, what is clearly more efficient.

Sound source separation can be divided into different groups according to several criteria:

1. Depending on the type of mixture: one criteria classify mixtures as instantaneous, anechoic [Anemüller and Kollmeier, 2003] or echoic [Melia and Rickard, 2007] mixtures. In [O'grady, 2007] a BSS algorithm that assumes instantaneous mixtures is proposed. Other classification separates mixtures into time-invariant or time-variant mixtures.

2. Relative number of mixtures and sources: underdetermined, determined and overdetermined. For instance, different separation methods are applicable to underdetermined mixtures in [Abrard and Deville, 2005, Araki et al., 2007].

3. Available *a priori* information about the mixing process or the sources: computational auditory scene analysis techniques [Wang and Brown, 2006b], source-based models, semi-blind or blind source separation [Cao and Liu, 1996].

4. Domain in which sound source separation is developed: time domain [Murata et al., 2001], frequency domain, time-frequency domain [Yilmaz and Rickard, 2004, Araki et al., 2006], etc.

5. Number of channels: single-channel or multi-channel methods. An example of single-channel method is presented in [Hu and Wang, 2010] where segregation of voiced portions of target speech is performed. One of the most important separation methods with two channels can be found in [Yilmaz and Rickard, 2004].

In order to solve the sound source separation problem, many different algorithms have been implemented over the years. It is very difficult to divide them into classes but in general terms, they can be divided into three main classes: CASA methodology, beamforming techniques and blind source separation algorithms. Within CASA most of the algorithms are single-channel techniques. In contrast to it, in beamforming and BSS the vast majority of algorithms are multi-channel techniques, that is, they rely on the use of spatial information. These three classes will be explained in a more detailed way in Sections 1.2.1, 1.2.2 and 1.2.3.

1.2.1 Computational auditory scene analysis

CASA solutions have become very popular in the framework of sound source separation. Some of the most remarkable works dealing with this problem are presented in Section 1.2.1.1. The main goal of this thesis is to solve the cocktail party problem under different circumstances and so, the suitability of CASA methods for solving it is analyzed in Section 1.2.1.2.

1.2.1.1 Sound source separation within CASA

ASA defines the humans' ability to segregate an acoustic mixture and focus on a target sound, even with one ear. The way that humans separate signals is still not clear and so hard to model it. In [Bregman, 1990], Bregman assumes the idea that audition and vision have many similarities. In techniques based on ASA principles, the human auditory system is supposed to represent sound signals by means of a neural representation, which will be processed in a similar way to images.

Computational auditory scene analysis consists of a large amount of computational implementations of ASA principles that have been successfully applied in different frameworks. The main idea behind CASA solutions is the performance of sound source separation by means of using intrinsic acoustic features of sound signals, or in other words, they work guided by the mechanisms of the human auditory system instead of using statistical tools as BSS solutions. In [Wang and Brown, 2006b], CASA is defined as "the field of computational study that aims to achieve human performance in ASA by using one or two microphone recordings of the acoustic scene". To limit the number of microphones to two makes sense since the human auditory system should be imitated (at most two ears). According to the number of microphones, it can be distinguished two groups of CASA methods: monaural and binaural CASA solutions.

Concerning the first group, monaural CASA methods have also been used aiming at carrying out sound source separation despite of using only one microphone. Before CASA, many methods had the problem of only working with specific sound sources or interferences, losing the capacity to deal with general acoustic signals. For instance, many algorithms for speech enhancement based on single-channel systems seek to characterize target speech signals or interferences to perform speech enhancement [Virag, 1999, Ephraim et al., 2003] or noise reduction [Martin, 2001]. In contrast to them, ASA methods emerge with the objective of working in many situations,

that is, ASA solutions tend to segregate sound sources from interferences without making any assumptions about acoustic interferences, number or type of sound sources.

In the framework of CASA, the first monaural method for separating two speech signals of two simultaneous talkers is [Weintraub, 1985]. This work is based on a frequency analysis of the acoustic mixture by means of a bank of bandpass filters, of whose outputs, the interpeak interval is determined. From this information, pitch periods and the number of sources are estimated. Once these parameters are established, each speech signal is characterized by the state of a Markov model (onset, offset, silent, etc.). Then, a spectral estimation solution uses this information about the state of each sound source to establish how the energy should be distributed in each frequency band. Since this algorithm is founded upon segregating according to the fundamental frequencies, the main limitation of this proposal is the requirement of two speech signals with different average pitch. Another important contribution within CASA methods is [Cooke, 2005], where a time-frequency representation of the acoustic signal known as synchrony strand is used. This representation is obtained by applying continuity constraints and local similarity to the output of a cochlear model. Comparing with previous methods, it has the advantage of being able to separate speech sources with crossing fundamental frequencies. One of the main disadvantage of this method is that it is not able to group acoustic events from the same source when these events are separated by silent intervals.

A large number of CASA methods have been developed and many of them are based on transforming directly the output of a cochlear model into feature maps instead of discrete time-frequency representations. From these maps, features such as onsets, offsets, periodicity or frequency transition information are extracted. Different tools are needed in order to extract features from the maps and in this sense, the use of some tools has been proposed. Perhaps one of the most popular tools is the correlogram introduced in [Brown and Cooke, 1994] and from which, periodicity information is extracted, as was stated above.

In [Hu and Wang, 2006] a CASA approach to monaural speech segregation is introduced. This solution segregates both unvoiced and voiced speech. An algorithm that segregates voiced portions of target speech is presented in [Hu and Wang, 2010]. First, this proposal roughly estimates pitch and uses it for segregating speech sources. When they are segregated, it improves pitch estimation and speech segregation iteratively. A more recent work is [Yu et al., 2013] where it is introduced an improved algorithm for monaural speech segregation. It introduces improvements in the initial segmentation stage to extract energy feature more accurately. An interesting comparative study of CASA techniques for monaural speech segregation can be found in [Zeremdini et al., 2015]

The possibility of introducing two channels in CASA has given rise to a category of approaches called binaural CASA systems. The major advantage of binaural CASA over monaural CASA is that since more than one channel is used, non-spatial and spatial cues can be integrated to carry out sound separation. At a high level, these methods follow a common approach. When left and right mixtures signals are transformed to the time-frequency domain, interaural cues, such as interaural level differences (ILD), interaural time differences (ITD) or interaural phase differences (IPD), are calculated. Source locations are estimated using these interaural cues. This localization procedure may be very useful to group time-frequency samples according to target locations. For instance, a study of the role of ITD in perceptual grouping is presented in [Darwin and Hukin, 1999].

[Lyon, 1983] proposes the first binaural CASA system in which ITDs were used to calculate time-varying gains for recovering the desired sources. Specifically, this work was implemented aimed at separating and dereverberating two sounds that come from different directions. Subsequent works have followed the scheme of this approach but introducing different improvements.

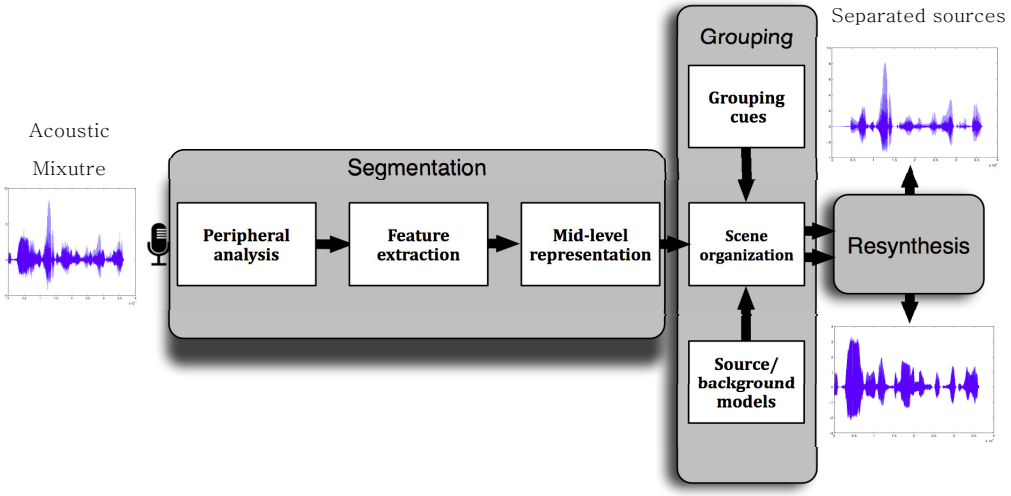

Figure 1.1: Scheme of a typical CASA system.

One work including new improvements is [Kollmeier and Koch, 1994], in which the first stage of the approach combines envelope modulation (non-spatial cue) with different spatial cues. Specifically, it calculates the low-frequency modulation that speech signals suffer and once it is determined, the amplitude modulation spectrum is weighted by a function calculated from the spatial cues. This process achieves than only energy coming from a specific direction passes, eliminating the rest of energy. [Roman and Wang, 2008] illustrates an important review of the state-of-the art in binaural CASA.

With respect to CASA-based speech segregation systems, more recent works may be mentioned. A speech segregation procedure which combines monaural grouping evidence and a binaural segregation system is presented in [Woodruff et al., 2010]. This combination is used to estimate the ideal binary mask. One of the most recent contributions [Jiang et al., 2014] considers the binaural segregation problem as binary classification. It also uses deep neural networks to develop the binaural segregation.

Besides speech segregation, CASA systems can be found in many other interesting applications, such as speech and speaker recognition [Li et al., 2010], speech enhancement in hearing aids [Gil-Pita et al., 2015], automatic music transcription [Wang and Brown, 2006a] or auditory scene reconstruction [Esquef et al., 2002].

1.2.1.2 CASA and the cocktail party problem

In this section the general structure of CASA solutions is studied aiming at determining their suitability to solve the cocktail party problem. Figure 1.1 represents a typical structure of a CASA system where different functional blocks, which will be explained in the next paragraphs, are depicted.

In [Bregman, 1990] the author establishes that ASA-based algorithms must contain at least two parts: the segmentation and grouping ones. Having a look at Figure 1.1 it can be observed that each part can also be divided into different stages:

1. Broadly speaking, the **segmentation procedure** consists in segmenting the acoustic mixtures into time-frequency regions (segments) which describe significant acoustic events. The stages included in this part are:

   - First, a mixture signal undergoes a *peripheral analysis* that consists in representing the acoustic mixture in a proper domain in which different features can be more easily extracted. Broadly speaking, speech mixtures are represented in the time-frequency domain by means of different available tools such as the short-time Fourier transform (STFT), the discrete wavelet transform (DWT) [Nakatani and Okuno, 1999, Liu et al., 2001] or the gammatone filter bank [Cooke, 2005]. This latter one has many advantages since it models the cochlear frequency selectivity, making the process more similar to the one performed by the human ear. In gammatone filter banks the center frequencies of the channels are not uniformly distributed, specifically having higher frequency resolution for low frequencies. Different systems modelling the different parts of the human auditory system can be found in the literature [Lyon, 1982, Deng and Geisler, 1987, Patterson et al., 1995].
   - In the second stage, *extraction of acoustic features*, such as onsets, offsets, harmonicity or amplitude and frequency modulations, is performed in the time-frequency domain.
   - Using those features, mid-level representations of the mixture signal can be obtained. In general, these mid-level representations are segments or other intermediate descriptions.

   There are different ways of developing the segmentation stage, and perhaps the most popular one uses cochlear-filter banks. But other interesting ways can be found in the literature, such as segregation by using the spectrogram as proposed in [Jordan and Bach, 2005], where physical and psychophysical properties of speech are combined with learning methods. Another remarkable segmentation stage that relies on the use of neural networks is implemented in [Hu and Wang, 2001]. Another interesting work [Hu and Wang, 2004] proposes to implement the segregation of low-frequency and high-frequency parts separately due to their different characteristics. [Hu and Wang, 2010] raises a tandem algorithm that uses pitch parameter to perform segregation of voiced portions of target speech.

2. In a second part, a **grouping stage** which combines the segments that are supposed to belong to a source is performed, forming a structure called stream. Scene organization determines the way the segments are combined. Grouping stages can be classified into two types:

   - On the one hand *schema-based ASA* uses trained models or schemas (learned patterns) for developing segregation. Auditory features that belong to the same learned pattern are grouped together, that is, it is a process based on prior knowledge.
   - On the other hand, *primitive ASA* segregates speech signals from acoustic interferences using intrinsic sound attributes, such as rhythm, periodicity [Hu and Wang, 2002a], frequency modulation [Brown and Cooke, 1994], spatial location, amplitude modulation [Hu and Wang, 2002b, Hu and Wang, 2004, Kim and Loizou, 2010], closeness in time and frequency, onsets and offsets [Hu and Wang, 2007] or continous or smooth

transitions. Primitive grouping is carried out following very similar mechanisms to the proposed ones by Gestal psychologist in the visual perception [Palmer, 1999].

3. Finally, the main task of the **resynthesis stage** is to perform the inversion of the time-frequency representation [Cooke, 1991] to obtain an audio waveform.

It must be highlighted that Figure 1.1 depicts a typical system architecture but over the years, researchers have implemented a large number of them. From classical architectures such as the one proposed in [Brown and Cooke, 1994], characterized as data-driven, to new architectures known as prediction driven approaches [Ellis, 1996]. The main aspect of this model is that the auditory scene is conceived as a hierarchy of sound elements.

Once the different parts of CASA methods are known, and considering the cocktail party problem, some drawbacks of these methods can be highlighted:

- To obtain robust CASA-based methods working under different conditions is difficult since they are heuristic and experimental.
- As mentioned in [van Der Kouwe et al., 2001], CASA techniques need in general that time-frequency representations of mixtures show well-defined regions corresponding to each source signal to perform segmentation. In real rooms, that is, where our cocktail party problem takes place, the presence of reverberation causes delayed versions of the sources and the presence of well-defined regions is less likely. In the feature extraction stage, depending on which features are used, reverberation affects CASA methods considerably. For instance, in the presence of reverberation, to extract pitch information is a difficult task.
- In general, CASA systems work poorly in the presence of wideband interference signals [van Der Kouwe et al., 2001] since it is more difficult to find well-defined time-frequency regions due to each source. Examples of wideband interferences are white noise and speech signals different from the target one, which are presented in the cocktail party problem.
- Many CASA algorithms tend to focus on one speech signal (the target one) and ignore the rest of them (interferences) as humans do. In the cocktail party problem, every source signals from the mixtures must be segregated.
- An important part that can be highlighted in segmentation and regrouping stages is the use of time-frequency masks as in other separation solutions like BSS algorithms [Wang, 2008]. Time-frequency masking entails that the perceptual quality of the separated sources decreases due to cross-talk and musical noise artifacts. Nowadays, musical noise artifacts is a problem presented many types of separation algorithms.

To sum up, different factors such as white noise, competing speech signals or reverberation make difficult the application of CASA in many cocktail party problems.

1.2.2 Beamforming

Beamforming is a challenging signal processing technique that has been developed for many years. It is presented in many fields, as for instance, radar, sonar or telecommunications. Beamforming techniques obtain separated source signals by using the principle of spatial filtering. This principle relies on the use of sensor arrays which allow enhancing the source signals that come from a determined direction and attenuating the interfering signals that come from other directions. To decide in which directions signals are attenuated or enhanced, the information

from all the sensors is combined. It is important to highlight that this technique assumes that target and interfering sources are not in the same place.

Beamforming techniques can be divided into two relevant categories:

1. *Conventional (data-independent) beamforming.* This type of beamformers use fixed weights and time delays aiming at combining the signals from the sensors in the array. Data-independent beamformers need to know the directions of interest in advance or to assume them.

2. *Adaptive (data-dependent) beamforming.* In contrast to conventional beamforming, this type of beamformers are able to adapt their beam pattern by means of analyzing the captured signals. This update of the beam pattern can be developed in different domains (*e.g.*, time or frequency domains) and it is made following different criteria. One criterion can be to consider the presence of noise, that is, the beam pattern is modified until the contribution of noise is minimum in the output signals. Adaptive beamforming is more effective to reject highly directive noise or noise sources enabled with mobility, but it entails higher computational cost than the conventional one [Krim and Viberg, 1996].

Sound source separation by means of beamforming techniques is presented in many applications in the literature [Mitianoudis and Davies, 2003, Benesty et al., 2008, Günel et al., 2008]. With respect to speech separation, many of these applications focus on solving the cocktail party problem. Due to the particularities of this problem (important reverberation effects in rooms, speech is a wideband signal, the assumption of far-field is not always valid, among others) many specific beamforming solutions have been proposed [Wang et al., 2010, Ayllón et al., 2013, Kidd et al., 2015]. The role of the microphone array will be of key importance in these solutions. In the following paragraphs a literature review of beamforming methods will be made since many of them are used to carry out SSS.

Firstly, beamforming techniques were designed to solve separation problems in narrowband applications, and later adapted to work with wideband signals, speech [Ayllón et al., 2013] being an example of wideband signal . Considering conventional beamforming, one of the most known and simplest beamformers is the delay and sum (DS) one. The operation of these beamformers is based on estimating the time difference of arrival (TDOA) between the signal arriving at a reference sensor and the rest of signals at the other sensors. When these time differences are calculated, the signals in these channels are delayed in such a way they are equally combined. Once the delays are established, the beam pattern of the array is fixed and will not be adapted to new changes. An important aspect to be considered is that a larger number of microphones and longer arrays involve more attenuation of noise components. DS beamformes are included in a more general class of beamformers known as filter and sum [Brandstein and Ward, 2001]. These beamformers apply a finite-impulse response (FIR) filter to each channel which are adjusted in order to obtain a specific beam pattern.

A specific beamforming technique for microphone arrays is the constant directivity beam former (CDB) as the one proposed by [Flanagan et al., 1985]. This beamforming solution takes into account that the frequency response of narrowband-designed arrays varies in a wide frequency range. It means that interfering signals are not equally attenuated in that range what disturbs the output speech. One solution is the use of harmonically-nested subarrays which are narrowband arrays that share some sensors among them.

Another important data-independent beamformer is the superdirective beamformer (SDB) which is defined for diffused noise. In this case, the filter coefficients are adjusted to maximize the

array gain, understood as the SNR improvement between the reference channel and the beamformer output. The array gain must be maximized without distorting the signal in the desired direction. To maximize the array gain implies that the noise variance is reduced. The narrowband adaptation of the minimum variance distortionless response (MVDR) beamformer or Capon filter [Capon, 1969] is a solution for this optimization problem. One requirement of MDVR is to know the noise covariance matrix. The covariance matrix of diffuse noise fields is constant and so, MVDR filter does not have to be updated. If noise field is non-stationary the noise covariance matrix must be updated and therefore, MVDR would be an adaptive beamformer.

SDB and MVDR have given rise to many useful beamformers but it is true that some problems appear when these techniques are used. Different factors such as channel mismatch, errors in microphone spacing or electrical sensor noise produce important gain of incoherent noise in initial implementations of SDB. Aiming at reducing uncorrelated noise due to self noise and phase error in microphones a gain constraint is imposed in [Gilbert and Morgan, 1955]. The work in [Cox et al., 1986] uses a sensitivity constrain to reduce uncorrelated white noise. Concerning MVDR, it is known that when the acoustic transfer functions (ATFs) between the target source and the microphones are known, MVDR performs dereverberation perfectly. The drawback of this solution is the difficulty of estimating the acoustic transfer function. Examples of this methodology can be found in [Benesty et al., 2007, Habets et al., 2010], in the latter one a study of reverberation cancellation and noise reduction for different noise fields is made. The authors conclude that the trade-off between noise reduction and dereverberation depends on the direct-to-reverberation ratio of the acoustic impulse response between the source and the desired microphone. A generalization of MVDR is linearly constrained minimum variance (LCMV) beamformers as for example the work in [Er and Cantoni, 1983]. In LCMV beamformers there are multiple linear constraints that are used to include relative ATFs rather than ATFs aiming at cancelling coherent interferences. A comparative study between LCMV and MVDR is done in [Habets et al., 2009].

With respect to adaptive beamforming, many successful methodologies have been used in a large number of applications. As previously mentioned, data-dependent beamformers are able to adapt to any change in the acoustic environment, as for instance, when noise or target sources move. This property involves important computational cost and sometimes, the occurrence of errors in the steering vector. Many adaptive beamformers minimize the mean square error (MSE) between the output signal and a reference signal highly correlated to the desired signal. To achieve this minimization the least mean squares (LMS) algorithm is used. The use of the LMS algorithm can introduce distortions in the target speech signal, what led the authors of [Frost III, 1972] to propose an algorithm to solve this problem. This proposal carries out a constrained LMS minimization, where the constraint is determined by a transfer function for the target signal. The generalized sidelobe canceller (GSC) introduced in [Griffiths and Jim, 1982] is an alternative to the Frost's algorithm. In GSC, two paths are implemented for the signal and by means of subtracting the signals at the end of both paths, the output signal is obtained. In the first path there is only a fixed beamforming. The other path consists of: a stage that removes the desired signal from this path and a set of adaptive filters to minimize the output noise power. Later, a generalization of GSC to include multiple constraints is implemented [Buckley, 1986], what is known as LCMV broadband beamformer.

An important problem for GSC is reverberation since in this technique it is assumed that signals at sensors are simple delayed versions of the source signals. This assumption causes that GSC would not be suitable for reverberant rooms. An adaptation of GSC for reverberant environments is presented in [Gannot et al., 2001], where GSC is adapted to work with any type of ATF by means of calculating ATF ratios exploiting the non-stationary of speech. LCMV

methods have the drawback of hard constraints that hinder a free election of the filters and so, the level of noise reduction is limited. One way of using a softer constraint is the adaptive microphone array system for noise reduction (AMNOR) of [Kaneda and Ohga, 1986] which lets to introduce some distortions in the target signal that are not perceived by the human ear. This technique requires the knowledge of the ATFs between the source and the microphones.

At this point it can be highlighted that some of the most relevant beamformers have been presented in this section, but it must be said that many others methods can be found in the literature. Concerning the cocktail party problem, many solutions can be mentioned. Relevant solutions oriented to hearing aids are [Ayllón et al., 2013, Kidd et al., 2015]. The work in [Wang et al., 2010] focuses on combining superdirective beamforming and BSS techniques to work with highly reverberant signals. Nowadays, one trend is to combine sound and image processing. A combination of image and beamforming techniques is shown in [Naqvi et al., 2012] aiming at separating moving sources. In this work the use of a circular microphone array is evaluated.

1.2.3 Blind source separation

Besides CASA and beamforming solutions, the third approach to solve the cocktail party problem is blind source separation techniques. A description of BSS methods and an extensive literature review will be presented in the following sections.

1.2.3.1 Introduction to BSS

Blind source separation has been a topic of interest in engineering for almost thirty years. BSS techniques [Cao and Liu, 1996] basically consist in estimating unobserved signals from a set of observed mixed signals. These mixtures of source signals are received at a set of sensors, in which each sensor receives a different combination of the sources. The term blind indicates that there is 1) neither information (or very little information) about the source signals, 2) nor information about the way the mixing process is carried out. This lack of knowledge about the mixtures is compensated by making some assumptions, such as the fact that the sources are assumed to be mutually statistically independent or decorrelated. This little prior information forces BSS techniques to exploit other signal properties. Just in this respect, BSS techniques take advantage of the spatial diversity provided by a set of sensors. The use of a set of sensor arrays allows BSS methods to have more available information that may be used to separate signals [Brandstein and Ward, 2001].

BSS was first suggested in [Jutten et al., 1985], and the first relevant work within BSS can be considered [Herault and Jutten, 1986]. The objective of this work is to separate an instantaneous linear even-determined mixture of non-Gaussian independent sources. It is based on an artificial neural network assuming that the source signals are independent. Basing on information theory, unsupervised learning rules that maximize the average mutual information between the inputs and outputs of an artificial neural network are introduced in another of the first works [Linsker, 1989].

BSS techniques can be classified according to different criteria, some of them are:

- One classification is made considering *the way the mixing process is carried out.* The mixing model can be classified into three types: instantaneous, anechoic or echoic. The instantaneous model is presented in many applications [O'grady, 2007] and it consists in considering that delays are not involved in the mixing process. The anechoic model [Omlor and Giese, 2011, Mirzaei et al., 2015] takes into account the delays suffered by the

source signals but not signals reflections. These two models are presented in many applications, but they are inadequate for solving the cocktail party problem. In contrast to them, the model used to deal with the cocktail party problem is the convolutive (echoic) one. This model considers both delays and reflections and it is presented in many applications [Parra and Spence, 2000, Melia and Rickard, 2007, Sawada et al., 2011].

The convolutive mixing problem in the frequency domain can be interpreted as an instantaneous mixing problem for each frequency bin [Smaragdis, 1998]. This interpretation reduces the computational burden but it entails some problems that degrade the signal reconstruction in the time domain. The permutation ambiguity problem in each frequency bins is an illustrative example of this degradation. The permutation ambiguity in each frequency bin should be aligned, achieving in time-domain signals the frequency components of the same source signal. Three main solutions to this problem can be mentioned: 1) the study of the interfrequency dependence of the amplitude of separated signals, 2) limitation of the filter length in the time domain, and 3) other solutions relies on the use of the position information about the sources [Saruwatari et al., 2003]. Works dealing with this problem are [Murata et al., 2001, Sawada et al., 2011, Wang et al., 2011].

- Depending on the *domain* in which the algorithm works. The first approach is time-domain BSS, being independent component analysis (ICA) methods [Comon, 1994] an important example of it. To avoid the problems of BSS algorithms in the time domain, numerous algorithms working in the frequency domain appear [Pearlmutter and Parra, 1996, Smaragdis, 1998]. Frequency representation of mixture signals can be obtained using tools as the Fast Fourier Transform (FFT) [Cooley and Tukey, 1965]. Later, different algorithms [Rickard and Yilmaz, 2002, Févotte and Doncarli, 2004, Févotte and Godsill, 2005] working in the time-frequency domain have emerged.

There are some ambiguities in BSS algorithms due to that they are only based on the assumption of mutual independence of source signals. Specifically, the ambiguities that can be found are: permutation and scaling ambiguities [Parra and Alvino, 2002]. Permutation ambiguity indicates that the order of the independent sources can not be determined, and scaling ambiguity entails that the estimated source signals can only be determined up to a scalar factor. With only the signals captured by sensors and without any additional a-prior information, it is not possible to correct these ambiguities since it can not be distinguish if the permutation and scaling ambiguities occur in the mixing system or in the source signals. Despite those and other problems, BSS techniques become very popular and applied to a wide variety of data types. Examples of applications of BSS are:

- *Biomedicine.* BSS has been applied to many medical applications because it is very common that signals exhibit crosstalk. For this reason, it is beneficial in many biomedical applications to separate the desired signal from the interference ones.

  Some of the earliest applications of BSS in medicine consist in separating the foetal electrocardiogram (ECG) from the maternal ECG. Using the foetal ECG the health of the baby can be determined, but it has been demonstrated that the maternal ECG is an important problem since it is a strong signal. A BSS procedure that is able to separate the two ECG signals is implemented in [Zarzoso et al., 1997]. BSS techniques are also used to analyze different ECG signals from different parts of the heart (atria, ventricles). For instance, it would be very helpful to separate the atrial components from the ventricular signal which

is stronger as it can be observed in the proposal [Raine et al., 2003]. Other interesting contributions are [Seoane et al., 2013, Seoane et al., 2014] where the ECG signal is separated from other interference signals aiming at evaluating the level of stress.

With respect to the study of the brain, different electroencephalography (EEG) signals can be analyzed. It is very usual to capture EEG signals that contain contaminating artifacts from other signals, artifacts from signals related to muscle activity being an example. Some BSS algorithms have focused on isolating EEG signals from these interferences. An illustrative work [Duda et al., 2000] which combines classical BSS techniques and support vector machines. Nowadays, BSS is being applied to functional magnetic resonance imaging which is a open line of research.

BSS is also applied to electromyography (EMG), or in other words, the study of electrical signals in muscles. EMG signals from entire muscles are separated to analyze the activity of individual muscles. In this sense, a study of possible BSS techniques to separate the EMG signal from interferences is made in [Chowdhury et al., 2013].

- *Econometrics.* The application of ICA to find underlying factors in perturbations of financial data is becoming very popular. ICA is used in conjunction with principal component analysis (PCA) as it can be observed in the work in [Duda et al., 2000].

- *Image processing.* There are two important procedures based on BSS methods in the framework of image processing: image denoising and compression. Example of a denoising procedure of natural images is the methodology followed in [Hyvärinen, 1999]. Currently, one representative technique for extracting elements from images is the non-negative matrix factorization (NMF).

- *Telecommunications.* In the proposal [Ristaniemi et al., 2002], BSS is used in mobile communications to separate the desired signal from the rest of them.

- *Blind audio source separation.* This is the framework of this thesis. To extract source signals from mixtures in audio applications is the so-called blind audio source separation (BASS)[Vincent et al., 2005, Vincent et al., 2006]. Nowadays, BASS is an important topic of research in a large number of fields, for instance, audio signal processing or cognitive psychology. BASS is presented in many applications, such as electronic music composition, voice cancellation, simultaneous translation for real-time speaker separation or speech enhancement.

  BASS can be addressed as a two-part problem. In the first part the number of sources must be estimated and then, a set of parameters is obtained from each source. In the second part, the separation is carried out by means of filtering the mixtures using the obtained parameters. Within the framework of BASS, sound sources are divided into two groups, music and speech sources. Both types of sources have many properties in common, for example, signals are nearly periodic, wideband noise and transient, but some differences are also found. Examples of these differences are the dependence between sources, the spectral envelope range or the fundamental frequency. For these reasons, depending on which type of signal is dealt, BASS methods have specific characteristics.

  Acoustic signals, such as music or speech, have some properties that can be exploited by BSS algorithms. For instance, in the particular case of speech it is known that this signal is feature-rich what has given rise to a wide variety of speech separation methods. Some properties of sound signals are:

- Signal distribution. The probability density function (PDF) describes the signal distribution, and speech or music signals have a PDF that is modeled by the Laplacian PDF. For instance, in [Lee et al., 1999] speech signals with silent time segments are approximated by a Laplacian model aiming at developing BSS.
- Stationarity. Two concepts of stationarity can be found in the literature, strict-sense and wide-sense stationarity. When higher-order moments only depend on the time-difference, the signal is considered as strict-sense stationary. Wide-sense stationary signals are those whose mean is constant and the second-order correlation only depends on the time-difference. In general, audio signals are non-stationary and they are considered as wide-sense or strict-sense stationary signals only for short periods [Rabiner and Schafer, 1978]. Pitch detection algorithms are an example of this consideration since they assume speech as quasi-stationary within periods from 10 ms to 30 ms. Speech is a highly non-sationary signal due to several factors such as amplitude modulations in the voiced portions of speech.
- Temporal dependencies. Broadly speaking, audio signals have temporal dependencies. For the case of speech these dependencies can be originated by the vocal tract. The second-order correlations have been subject of study for developing linear prediction as it can be studied in [Deller Jr et al., 1993].

The basic assumption in BSS algorithms is the statistical independence of the sources. This assumption can be combined with several BASS criteria that take advantage of the aforementioned acoustic signal properties. Following this idea, BASS approaches can be classified into two main classes depending if they exploit the nongaussianity or nonstationarity properties of sound source signals.

- Non-gaussianity. It was mentioned that acoustic source signals have a PDF that is not Gaussian. BSS methods using higher-order statistics (HOS) exploit nongaussianity to separate sound signals. Note that BSS algorithms using HOS are also termed ICA algorithms [Comon, 1994].
- Non-stationarity. As it was mentioned, the short-term dependencies/correlations of sound signals are time-variant. Then, many BSS applications exploit non-stationarity by simultaneous diagonalization of short-term output correlation matrices at different time instants. Example of this type methods are sparse-based methods which will be explained in Section 1.2.3.3.

1.2.3.2 Statistical-based BSS algorithms

During the 1990s, a large number of BSS methods were developed based on different tools, such as maximum likelihood information [Pahm et al., 1992], higher order statistics [Cardoso, 1998], mutual information [Torkkola, 1999], natural gradient [Amari, 1998], or information maximization [Torkkola, 1996, Lee et al., 2000]. In [Comon, 1994], Common presents a clear formulation of the BSS problem which is very popular since it establishes important concepts. For instance, it proposes that the most natural measure of independence is the mutual information. In this sense, it demonstrates that maximizing the non-Gaussianity of source signals is equivalent to minimizing the mutual information between them. Common is considered the person that introduced the concept of ICA, methodology in which source separation is formulated as a mixing matrix estimation problem, assuming that sources are statistically independent and non-Gaussian. The basic idea of ICA methods is to estimate the independent signals by using a procedure

that maximizes the signals' statistical independence either minimizing the mutual information or maximizing the non-Gaussianity [Cichocki and Amari, 2002]. Thus, independence measures are used to evaluate the level of independence. The main difference between ICA methods is in that metric. Before Common's work, separation techniques were based on the signal processing technique of PCA. This technique was used to extract source signals from speech mixtures using autocorrelation or decorrelation. PCA technique is still an important tool and it is used in many procedures, although it must be said that it was quickly replaced by ICA methods after the solution of Common.

ICA algorithms can separate sound sources in a room if there are multiple microphones positioned at different locations. If the number of microphones is equal to the number of sources, the assumption of non-Gaussian sources may be enough to extract sources from the mixtures. From the different ICA methods that have been published, the approach in [Bell and Sejnowski, 1995], known as BS-Infomax, is one the most popular because of its simplicity. In this work, until ten speakers are separated in order to solve the cocktail party problem. The FastICA algorithm [Hyvärinen and Oja, 1997] is another separation method that relies on mutual information. This algorithm is able to find all non-Gaussian independent components, regardless of their probability functions. FastICA is faster than gradient-based ICA algorithms and its convergence is guaranteed. Other work oriented to solve the cocktail party problem is [Lee et al., 1998], where a recorded voice is separated from a mixture signal with music in the background.

ICA has been an important approach in many applications but it has considerable limitations. ICA methods only perform properly when the mixing process is instantaneous, but their results separating convolutive mixtures are not as good as expected. Other problem in ICA is that the original design is not valid for underdetermined separation problems. More limitations of ICA are that the signal sources should come from different spatial directions, the number of sources must be known in advance or the mixing matrix must be stationary during a period of time. Aiming at solving these limitations, different ICA approaches have been implemented. An example of overcoming these limitations is the work in [Parra and Spence, 2000] which uses the non-stationarity of speech to estimate the multiple channels of echoic speech mixtures. Another proposal of ICA, efficient FastICA (EFICA), is presented in [Koldovsky et al., 2006] where it is assumed that the PDFs of the independent signals are generalized Gaussian distributions. Anyway these limitations have favored the emergence of new separation approaches. Numerous BSS methodologies are based on modeling the sources using Gaussian mixture models [Donoho and Elad, 2002] or Laplacian distributions [Lewicki and Sejnowski, 2000, Mitianoudis and Stathaki, 2005]. Other problem for many separation algorithms was the presence of moving speakers which was firstly addressed in [Rickard et al., 2001, Anemüller and Kollmeier, 2003].

1.2.3.3 Sparsity-based BSS methods

Besides ICA, there is a type of BSS methods based on sparse representations of source signals. Although there is no universally accepted definition of sparsity [Hurley and Rickard, 2009], however, most experts agree that a signal (or a distribution) with all its energy in one coefficient and all others zero (or almost zero) is the "most sparse" signal. Other equivalent view is that a signal is said to be sparse when it is zero (or nearly zero) more than might be expected from its variance. Thus, intuitively, a sparse representation is one in which a small number of coefficients contain a large proportion of the energy. From the point of view of source separation, we can notice that if sources are sparse, the probability of multiple sources being non-zero simultaneously is very low what can facilitate the task of separation.

In a large number of BSS algorithms one step consists in transforming the source signals into a

domain in which its sparsity degree is clear (e.g., wavelet or time-frequency domains) so that the separation can be carried out by a partition of the transformed space due to the sparsity level of this representation. Sparsity can be achieved using time-frequency representations but also multi-resolution signal processing [Daubechies et al., 1992, Proakis John and Manolakis Dimitris, 1996, Vaidyanathan and Hoang, 1988]. Important tools to achieve sparse representations are the discrete wavelet transform [Daubechies et al., 1992, Daudet and Torrésani, 2002] and the modified discrete cosine transform (MDCT) [Daudet and Sandler, 2004]. Optimization principles have been developed to represent source signals as sparse as possible, examples of these optimization principles can be found in [Chen et al., 1998, Elad and Bruckstein, 2002, Lee et al., 1999]. To evaluate the sparseness of a representation of a source signal, it is necessary to use any objective measure. A successful measure of sparsity has demonstrated to be the l_1-norm as it is presented in [Donoho and Elad, 2003, Saito, 2004]. Minimizing the l_1-norm, the number of non-zero entries in the sequences is minimized. This minimization can be easily carried out by a linear program. By 2000, the first separation method based on sparsity appeared, [Lewicki and Sejnowski, 2000]. The assumption of sparse sources is also used in the contribution of [Zibulevsky et al., 2001], where the mixing matrix is estimated by using a clustering method. Considering the previous method, the effects of sparsity on the calculation of a basis matrix when the clustering method is k-means is analyzed in [Li et al., 2004].

Sparsity-based BSS methods can be divided into two different types. In the first type of methods the estimation of the independent sources relies on maximum a posteriori (MAP) after estimating the mixing matrix either by using the maximum likelihood (ML) criterion or a clustering method. It is very usual to carry out the MAP estimation using l_1-minimization. The second class of algorithms are those based on time-frequency masking. Using these masks, that can be calculated in a variety of ways, time-frequency samples due to each source signal can be extracted. One of the main advantages of time-frequency masking is that it can solve the underdetermined separation problem. In [Bofill and Zibulevsky, 2001], authors introduce a notable example of sound source separation method based on time-frequency masking. This algorithm deals with speech and music separation problems with only two sensors and contains two different parts. In the first part, a clustering method is used in order to calculate the mixing matrix. Once the mixing matrix is calculated, the separation of the independent sources must be carried out. This second stage uses l_1-norm aiming at extracting as sparse sources as possible from the mixtures. This method has demonstrated good separation performance even with six source signals, despite only two microphones are used. The line orientation separation technique (LOST) proposed in [O'grady, 2007] is another interesting separation method for instantaneous mixtures. It provides a dual geometric interpretation in which the separation of sources in a speech mixture is equivalent to the separation of linear subspaces in a mixture of oriented lines. These lines are identified in a scatter plot by means of using an expectation-maximization (EM) procedure. After estimating the lines, the demixing stage is performed using l_1-norm minimization.

Numerous sound source separation methods work in the time-frequency domain because many acoustic sources can be considered sparse in this domain. This is the case, for instance, of speech, which exhibits sparsity in the T-F domain. Then, time-frequency representations have grown in importance in BSS as it can be observed in the works of [Févotte and Doncarli, 2004, Févotte and Godsill, 2005]. Perhaps the most important SSS algorithm based on sparsity working in the T-F domain is the degenerate unmixing estimation technique (DUET) proposed by [Rickard and Yilmaz, 2002], which was the first practical algorithm in anechoic environments. The performance of the DUET algorithm reduces when it is applied to echoic mixtures. In general, DUET is used with signals that are disjoint in the T-F domain, also known as W-disjoint orthogonal signals [Rickard and Dietrich, 2000]. W-disjoint orthogonality (WDO) is a

good indicator of the separation performance for SSS methods based on sparsity. Using two microphones, DUET uses spatial information (delay and level differences between the microphones) to separate speech sources. First, a weighted two-dimensional histogram from these differences is constructed. Peaks in the histogram correspond to each source and are identified by using unsupervised clustering. It is necessary to know how many peaks (and so sources) must be located, what is a limitation, since the number of sources must be known in advance. When these peaks are located, the mixing parameters of each source are estimated. In a second stage the unmixing procedure is performed by time-frequency binary masking that are generated based on a proximity criteria. Binary masks have the disadvantage of introducing residual musical noise in the separated speech sources. This algorithm will be studied in a more detailed way in Chapter 2.

Many separation methods based on DUET have been performed in the last years. The time-frequency ratio of mixtures (TIFROM) algorithm proposed in [Abrard and Deville, 2005] is an example of it. This method separates speech sources from instantaneous linear mixtures using two microphones. The advantage over DUET is that it can separate sources even whether the original source signals almost fully overlap in the T-F domain. To compensate the impact of this overlap on the separation performance, it requires the presence of slight differences in the time-frequency distributions of the original speech sources. Using these slight differences, the mixing matrix estimation is carried out. One important limitation of DUET is that only two channels are used, in this sense, some proposals can be found in the literature. A multi-channel DUET algorithm [Melia and Rickard, 2007] called DESPRIT works using two tools: the sparse assumption and the estimation of parameters via rotational invariance technique (ESPRIT). DESPRIT has the disadvantage of being limited to linear arrays. The work in [Araki et al., 2007] overcomes this limitation since it can be used with any geometry and number of sensors. It uses k-means to cluster normalized time and level differences between sensors.

1.2.4 Combined SSS solutions

In general, SSS algorithms have limitations that must be solved. In many contributions the solutions of those limitations are obtained by means of combining different types of SSS methods. Limitations can be related to many factors, such as the presence of noise and reverberation, or the fact that the techniques implemented in the algorithms introduce important distortions into the separated sources. In the following paragraphs some of the most remarkable combinations of SSS methods are presented.

As stated above, many sparsity-based BSS methods use time-frequency mask to separate sound sources. Analyzing the quality of the separated sources, it can be observed that one of the disadvantages of time-frequency masking is that it introduces important distortions into the separated sound sources. One representative example of those distortions are musical noise artifacts which are strong fluctuations in the time-frequency domain. This phenomena is very common when spectral subtraction type algorithms are applied [Goh et al., 1998]. Concerning source separation, different methods [Makino et al., 2005, Araki et al., 2006, Madhu et al., 2008] have been introduced to reduce musical noise.

Some successful solutions to minimize the presence of musical noise artifacts consist in combining sparsity-based BSS methods with other types of SSS techniques. For instance, in the work in [Araki et al., 2004], sparsity-based BSS methods are jointly used with ICA techniques. The procedure of this algorithm is as follows: first, the time-frequency points where only one source is active are identified and removed from the sound mixture using similar techniques to DUET. Then, frequency domain ICA is applied to the remaining mixture signals. Additionally, directiv-

ity pattern solutions are used to reduce the distortions of applying binary masking. Another way of combining ICA and time-frequency masking is to apply a time-frequency mask to the ICA outputs as it can be observed in [Kolossa and Orglmeister, 2004]. These masks are calculated from the ratio of the demixed signal energies and are applied to two frequency-domain ICA techniques. The quality of the signals recovered is higher from the point of view of the signal-to-noise ratio (SNR). There are more examples of combinations of ICA and sparsity-based BSS methods with different objectives. In [Pedersen et al., 2005] ICA and and binary time-frequency masking are iteratively used to separate underdetermined mixtures. Two ways of combining ICA and binary masks are studied in [Saruwatari et al., 2005].

It is very popular to use beamforming with other SSS techniques. Comparing BSS with beamforming, it can be said that BSS techniques focus on separating all the signals presented in the mixtures regardless of whether sources are desired signals or interferences, whereas beamforming obtains desired signals and eliminates interferences [Coviello and Sibul, 2004]. BSS has some disadvantages in comparison with beamforming, for instance, many BSS algorithms perform worse in high reverberant [Wang et al., 2010] and/or dynamic environments [Naqvi et al., 2010]. Another situation in which BSS is limited occurs when the number of microphones is lower than the number of sources (underdetermined problem) [Melia and Rickard, 2007]. However, BSS does not need to know the direction of arrival of any signal or the array geometry, what can be useful in many problems.

Many authors have analyzed the relationship between BSS and beamforming and in some works, these techniques have been combined aiming at improving source signal separation. Different ways of combining them can be found in the literature, as for instance, beamforming is used as preprocessor of BSS in order to reduce the effects of high reverberation as it is explained in [Wang et al., 2010]. In other works [Davies and Mitianoudis, 2004], beamforming solves the permutation problem of frequency-domain BSS.

Concerning CASA, this type of solutions is also jointly used with other techniques. For example, aiming at separating music signals, a combination of DUET and CASA techniques is presented in [Woodruff and Pardo, 2007]. The first part of this algorithm is very similar to the one of DUET where a cross-channel histogram is calculated. Concerning the part of CASA, the pitches of the original sources are extracted from the histogram and with these pitches harmonic masks are constructed. Finally, the harmonic amplitude envelopes are obtained from the pitch estimations.

To sum up, important contributions that combine different types of SSS techniques can be found in the literature, and many of them can be used to solve the cocktail party problem.

1.3 Synchronization in wireless acoustic sensor networks: state of the art

At present, there is a significant development of new types of sensor arrays. An important trend is to use wireless sensor networks since they introduce improvements over traditional sensor networks [Akyildiz et al., 2002], as for instance, they can cover much larger areas than conventional sensor networks, more sensors can be integrated or many power and size constraints are avoided. The use of WSNs also takes place in many sound applications where wireless acoustic sensor networks are used instead of classical networks. Despite the advantages of WSNs, some problems related to their use may also emerge. One of the main problems is the synchronization one, which can affect many signal processing algorithms. The emergence of WASNs aimed at solving the cocktail party problem is a challenging field of research due to the advantages and

disadvantages that they entail.

1.3.1 Microphone arrays in sound source separation problems

The use of microphone arrays in sound applications is presented in many problems, such as acoustic echo cancellation, sound source localization [Nishiura et al., 2000] and tracking, audio coding, speech enhancement [Ayllón et al., 2013] or sound source separation. Like in other fields of research, new acoustic sensor networks appear in SSS problems what will be discussed in the next paragraphs.

Broadly speaking, there are two traditional and significant trends of microphones arrays in SSS applications: microphone arrays used with beamforming techniques and the ones with BSS and CASA techniques. Microphone arrays in beamforming techniques are sophisticated since beam patterns must point towards the directions of the target sources. To achieve it, the constructive parameters (microphone separation, size, type of microphone, etc.) and the microphone array structure (linear, triangular, circular, spherical, etc.) must be considered. The objective in beamforming is to carried out a complex-weighted sum of the sensor data [Van Veen and Buckley, 1988]. Alternatively, many BSS and CASA algorithms utilize simple microphone arrays since these algorithms relies on different assumptions about the mixing process or the source signals, but they do not utilize any information concerning the geometry of the auditory scene [Mitianoudis and Davies, 2004].

With respect to speech separation, one of the first contributions about array processing and beamforming is [McCowan, 2001]. The main objective in this work is to enhance signals coming from a particular direction. In this work, the suitability of some array configurations are analyzed according to the characteristics of the speech signal. In [Kidd et al., 2015] some benefits of acoustic beamforming to solve the cocktail party problem are analyzed.

On the other hand, the use of very simple microphone arrays to perform speech separation can be found in some contributions. Speech extraction with small arrays has been extensively researched from different perspectives, such as focusing on biological models [Zeremdini et al., 2015] or on hearing aids [Gil-Pita et al., 2015]. Concerning the cocktail party problem, many works can be mentioned. A robust speech separation algorithm that recovers speech signals with a two-microphone array is explained in [Rennie et al., 2003]. Two, three and four speech sources corrupted by noise are separated using a model based on TDOA information. The work in [Pedersen et al., 2008] introduces a method for separating anechoic and instantaneous mixtures with an arbitrary number of speech signals with only two microphones. This method combines ICA and time-frequency binary masking. Another interesting proposal is [Cobos and Lopez, 2010] where a sparse method for speech separation using a two-microphone array is presented. The use of only two microphones entails many advantages since the number of microphones increases, cost and processing time both increase.

Nowadays the use of WASNs is becoming very popular in many applications, speech source separation being no exception. WANs entail many advantages but there is a very important problem for many signal processing algorithms, the synchronization problem. Some of the most important solutions in this respect will be presented in the following sections.

1.3.2 Wireless acoustic sensor networks

Wireless acoustic sensor networks are composed of spatially distributed autonomous nodes with acoustic sensors (microphones) to monitor sound sources and with wireless communication links to cooperatively pass data through the network. They represent the next-generation technology for audio acquisition and processing. In a WASN, each node wirelessly communicates with

other nodes transmitting information and provides in-network signal processing, for instance, to either separate or estimate a desired signal (*e.g.*, blind source separation of speech signals) or to extract certain parameters (*e.g.*, the location or identity of speakers [Bertrand, 2011, Griffin et al., 2015]).

Due to the own features of WASNs, they can be beneficial for many applications. Some examples of these applications are:

- *Acoustic monitoring.* Many works focus on acoustic monitoring of different environments by using WASNs. Two examples of this use of WASNs are acoustic surveillance applications and vehicle tracking.
- *Intelligent environments.* Currently, many intelligent environments rely on the use of WASNs due to their advantages. The main characteristic of these environments is that sensors communicate wirelessly with the user to satisfy its needs. Domotic systems are example of interaction between users and wireless sensor networks.
- *Speech enhancement*, presented in many applications, such as hearing aids, hands-free telephony or security applications. The use of microphone arrays may achieve important improvements in the quality of the captured speech signals by means of minimizing the presence of noise components. One of the advantages of WASNs is that their microphones can be placed at any noise environments, as for instance, airports, stations, markets or companies.

The use of WASNs in many applications is not trivial and in this sense, many challenges must be overcome. As in the case of WSNs, WASNs must comply with requirements such as:

- *Precision.* Depending on the application this requirement will be more or less relevant. For instance, some applications, such as signal processing on audio, need an accuracy on the order of a few microseconds what indicates that this requirement will be of key importance. In others, a coarse precision (over seconds) may be tolerable.
- *Energy efficiency.* The available energy in sensor nodes must be taken into account. Sensor nodes use batteries which have a limited service life.
- *Cost and size.* The cost of the application scheme must be reasonable. For instance, sensor nodes usually are very cheap devices and so, it does not make sense to use expensive hardware (e.g. a GPS receiver) in the nodes. Furthermore, sensor nodes are very small, what limits considerably the hardware components that can be inside them.
- *Robustness.* Due to different problems, some sensor nodes can fail. If the WASN is robust, it will be able to remain functional for the rest of the network.
- *Immediacy.* This requirement is associated with sensor networks that develop applications as emergency detection. In this case, the synchronization protocol can not spent a lot of time when a emergency situation happens. One solution consists in that nodes should be pre-syncrhonized at all times.
- *Lifetime.* It is the time that sensor nodes are synchronized and it depends on the carried out task. This time can be instantaneous or as long as the operation time of the network.
- *Scalability.* In some applications the required number of nodes is very large. In this sense, protocols must be able to scale well with increasing number of nodes.

- *Availability.* The coverage within a region must always be available.

Additionally there are specific requirements for WASNs. Two of these particular requirements are:

- *Minimizing input-output delay* [Bertrand, 2011]. This process is very typical in real-time audio streaming WASNs. For instance, in hearing aids a perceptible delay is not allowed since it will distort the perception of the listener. It seems intuitive that to comply with this requirement is difficult since many elements in the network can introduce important delays.
- *Microphone subset selection* [Bertrand, 2011]. It is a requirement presented in large-scale WASNs. Sometimes, good performance can be obtained by using only a subset of microphones of all of them and in this way, an important reduction of the computational cost can be achieved.

One of the main problems in WASNs is the desynchronization of the signals captured by different microphones, which worsens outcomes of numerous methods. Many signal processing algorithms suffer degradation performance due to the desynchronization of nodes since they require perfectly synchronized audio data. Very accurate synchronization protocols in order to mitigate the effects of desynchronization can be found in algorithms such as beamforming, localization and separation algorithms. In order to understand its importance, let us think about localization methods. For instance, in localization algorithms based on the estimation of the time difference of arrival, if the sampling frequencies of two nodes mismatch, the TDOA can not be properly estimated and so, the corresponding source is not located. Classical solutions to this problem are based on the use of clock synchronization protocols and algorithms in the communication layer [Elson et al., 2002, Wu et al., 2011, Schmalenstroeer et al., 2015]. In a WASN with dedicated and uniform hardware, synchronization of the sampling rates for analog-to-digital Converters (ADCs) is usually manageable [Wu et al., 2011] and is sometimes even unnecessary if the oscillators are of sufficient quality. Before studying the synchronization problem in WASNs, which are distributed systems and wireless sensor networks, a brief overview of synchronization solutions in distributed systems and WSNs is made in Sections 1.3.3 and 1.3.4, respectively.

1.3.3 Synchronization in distributed systems

Synchronization protocols in distributed systems have been developed for many years. Many networks are distributed systems what means that they do not have a global clock, that is, each node has its own clock. The absence of global clock results in synchronization problems between nodes that must be solved. Different methodologies are oriented to solve these problems and are known as time synchronization methods, clock synchronization protocols, etc. The objective of these techniques is to provide a common notion of time for local clocks of nodes. It is easy to think that delays between different clocks can appear since they are imperfects. In general, clocks are based on the use of quartz crystals which work at slightly different frequencies in each node. These differences in frequency cause some problems, such as accumulating errors over times of the clocks or clock drifts, what introduces divergences between the clock values. Depending on the application, these divergences can degrade its performance. For this reason, clock synchronization protocols are required in distributed networks in order to correct these deviations.

In general terms, any clock synchronization protocol must accomplish some requirements in distributed networks. Some of the most important requirements are: the computational

load associated with synchronization must not affect the system performance, time never runs backward, each node must be able to estimate the local time on the other node when two nodes are synchronized. The reader can sense that other requirements for synchronization protocols are directly related to the type of network.

Many clock synchronization protocols in distributed networks have been implemented for years. A large number of classifications can be found according to different criteria. Some important types of synchronization protocols are the Time Transmission protocol [Arvind, 1994], set-valued estimation methods [Lemmon et al., 2000] or the Network Time protocol [Mills, 1991].

The clock synchronization problem can be approached from a number of different perspectives. Many solutions propose a common time base in order to develop a task. In some networks this common time is provided by a master node. In these solutions, the master sends a timestamp to the rest of nodes to make clock corrections. It is evident that a solution that relies on the use of one master has many disadvantages. The computational load of the master increases when the number of nodes is high and if the master fails, the synchronization method too. Other solutions introduce distributed clock synchronization where one possibility is that only neighbor nodes exchange timestamps.

In general, time synchronization methods are based on exchanging messages between nodes. Some nondeterministic factors, such as physical channel access time or propagation time, make necessary synchronization in distributed systems. When a synchronization message is sent from one node to another one, the message faces different delays that prevent the receiver to know exactly the time of the sender. These sources of error can be classified into four types: send time, access time, propagation time and receive time. Send time is the time that the sender uses to construct a message. Access time refers to the delay at the Medium Access Control (MAC) layer before transmission. Propagation time is the time that the message expends between the network interfaces of the sender and the receiver. And finally, receive time is the required time for the the receiver aiming at transferring the message to the host. Factors like the above-mentioned times increase the complexity of the synchronization methods.

1.3.4 Synchronization in wireless sensor networks

Concerning sensor networks, many methodologies have been applied to wired sensor networks obtaining excellent results in terms of synchronization of the different devices in networks. At present, most of the principles of these synchronization methodologies are implemented in protocols for wireless sensor networks, even though there are many differences. WSNs present many peculiarities that have led to the emergence of specific protocols. A large variety of synchronization protocols aimed at working in WSNs can be found in the literature and some of the most important ones are studied in this section.

Sensor networks are a collection of wirelessly connected nodes that are capable of sensing, computation and communication. Each device in a sensor network, also known as node, is able to cooperate with other nodes in order to do a sensing task. This type of networks entails a large number of advantages with respect to wired networks, such as the fact that sensors are separated by greater distances and so larger areas are covered, higher number of sensors can be integrated, power and space constraints associated with wired networks are avoided, sensors are enabled with mobility, sensors can be placed at positions where it is difficult to place wired sensors, etc. These advantages lead to the use of WSNs in a wide variety of applications areas, as for instance, military applications, medical tools, industrial products or scientific applications. This new kind of wireless networks is directly related to the development of new type of sensors which have some advantages in respect of traditional sensors. Some examples of the advantages

of these new type of sensors are energy efficiency, small size, integrability, among others.

One of the main operations in WSNs is data fusion, or in other words, data captured by sensors are aggregated into a meaningful result. Due to desynchronization of sensors in WSNs, information from them can be wrongly combined and therefore, applications can fail. The problem of synchronization is caused by different factors, such as limitations of the communication protocols, deviation between the different clocks of the nodes, large distance between sensors, etc. Aiming at performing data fusion successfully many clock synchronization protocols for wirelessly networks can be found in the literature. The objective of these protocols is to provide a common time reference for all the nodes. These protocols entail high complexity since wireless networks have some specific characteristics. Factors such as high number of nodes, larger distances between nodes, limitation of energy and bandwidth cause that classical synchronization methods would not be appropriated for these networks. Then, aiming at implementing synchronization methods in WSNs, their most restrictive intrinsic properties for synchronization solutions must be identified. Some examples of it are:

- *Available bandwidth.* Bandwidth limitation causes that all the algorithms performed in nodes, and so synchronization protocols, must reduce the amount of information transmitted.

- The limitation of bandwidth is directly related to another one, the *limitation of energy.* The vast majority of energy consumed in this type of networks is due to information transmission. WSNs can use a large number of sensors, reaching thousands of them in certain applications. It makes very difficult to wired each sensor to a power source. Then sensors need batteries what implicates to have limit amount of energy. Considering it, nodes must work as efficiently as possible.

- Another and not less important limitation is related to *hardware.* It is known that sensors in this kind of networks are very small and then, hardware must comply with important space constraints. It gives rise to restrictions on computational power and storage space. The reader can think that to increase the size of sensors can be a solution, but bigger sensors involve more power consumption and raise their price.

- The use of *wireless mediums* to transmit information is also an important constraint. External interferences can lead to a high percentage of message loss what can limit the efficient of the synchronization protocol.

Considering the aforementioned properties of WSNs, the design of synchronization protocols can be carried out. The requirements on synchronization schemes for sensor networks depend on the application at hand, but it can be said that some principles are common for the vast majority of protocols. Likewise, not all the protocols must comply with all the requirements in the same way. Examples of these common requirements are explained in Section 1.3.2.

Since WSNs are presented in a large number of applications, many types of synchronization protocols can be found. Possible classifications of these protocols in WSNs are: 1) internal [Mock et al., 2000] versus external synchronization, 2) clock correction versus untethered clock [Römer, 2001], 3) sender-to-receiver versus receiver-to-receiver synchronization, 4) master-slave versus peer-to-peer synchronization, 5) probabilistic [Arvind, 1994] versus deterministic synchronization, 6) stationary versus mobile networks, 7) single-hop versus multi-hop networks, 8) mac-layer based versus standard approach, among others.

Synchronization protocols have been designed over many decades in different types of networks. Broadly speaking, the main objective of these algorithms is to maintain synchronization

of physical clocks [Elson and Römer, 2003]. Mainly, these protocols aim at compensating clock skews and offsets. Notwithstanding the variety of protocols, there are a lot of aspects in common. For example, the existence of messaging protocols, the change of clock information between the different elements of the network, the use of different tools in oder to mitigate the effects of non-derterminism in message delivery, etc. As previously mentioned, synchronization protocols in WSNs have their particular features since WSNs involves some limitations. Some of the most known and robust synchronization protocols are the reference broadcast synchronization protocol [Elson et al., 2002], continuous clock synchronization in wireless real-time applications [Mock et al., 2000], delay measurement time synchronization citeromer2001time, etc.

1.3.5 Synchronization in wireless acoustic sensor networks

The problem of synchronization in WASNs is mainly caused by two factors. First, since each node of a WASN has its own clock, there is an inevitable clock phase offset and clock frequency skew from a standard reference clock frequency, which raise a problem known as sampling frequency mismatch. Second, the different distances between nodes and sources give rise to different acoustic propagation delays, which in some configurations can affect the performance of the algorithms. Broadly speaking, synchronization protocols within the framework of distributed audio signal processing only deal with the first factor, that is, they aim at eliminating the delays that imperfect clocks cause.

Concerning the problem of sampling frequency mismatch in WASNs, two important type of synchronization methodologies are proposed in the literature. The first one relies on the exchange of timestamps between the different nodes of the sensor network. The second methodology focuses on analyzing the sampled signals to estimate the frequency mismatch. In respect of the first type of methodology, different works have been proposed and now, a brief overview of some of the most representative is presented. Sampling rate offsets (SROs) are estimated in works as [Wehr et al., 2004] in which a synchronization scheme is designed for distributed ad-hoc audio networks. In that work, a modulated radio frequency reference signal is broadcasted to all the nodes in the network. In [Bahari et al., 2015], SRO is estimated between nodes using the phase drift of the coherence function between two sound signals. Authors assume that SRO originates a linear increase of the time delay between these signals. The method works in the time-frequency domain in order to compensate this time delay. In [Schmalenstroeer and Haeb-Umbach, 2013, Schmalenstroeer et al., 2015] two solutions of sampling clock synchronization based on message exchange over a wireless communication link are presented. These proposals combine hardware and software approaches aiming at synchronizing nodes in a WASN. By means of a message exchange protocol between pairs of nodes, phase and frequency differences are estimated. This information between pairs of nodes is used to obtain network synchronization. Finally, providing hardware solutions, clock of slaves nodes are aligned to the clock of a master node with a considerable accuracy. To achieve global synchronization from only local messages is proposed in [Boyd et al., 2006]. In that paper, distributed algorithms, known as gossip algorithms, are used for exchanging information. The main characteristic of these algorithms is to be robust against changes in networks.

On the other hand, an alternative approach consists in only analyzing the sampled signal without using message exchange. Preventing message exchange makes unnecessary the use of communication channels, dedicated protocols and hardware to broadcast reference signals. This type of solution is useful, for instance, in non-uniform ad-hoc WASNs with devices from different manufacturers or without good management of the communication layer. In these WASNs synchronization of the ADCs may be hard (or impossible), and the resulting signal

drift must then be taken into account by the signal processing algorithms as it was done in the work [Bertrand, 2011]. An example of this way of synchronization of sound signals for beamforming techniques is the solution introduced in [Markovich-Golan et al., 2012] where sampling rate offsets between nodes are estimated and later compensated. This method is applied to speech-absent time segments with slow time-varying interference statistics. First, the sampling rate offsets are estimated by using the phase drift in the coherence between two sound signals captured by two microphones. Second, SRO is compensated by resampling with Lagrange polynomial interpolation methods. Obviously, it has the disadvantage of assuming the presence of speech-absent time segments with slow time-varying interferences. Another important approach to be mentioned is [Miyabe et al., 2013a] where the authors implement a novel method for compensation of sampling frequency mismatch for asynchronous microphone arrays. The estimation of the frequency mismatch involves the use of a maximum likelihood estimator based on TDOA observations. It compensates the sampling frequency mismatch in the time-frequency domain by means of implementing a linear phase shift. This method assumes that sound sources are stationary and motionless, what is a limitation. Other example of SRO estimation is explained in [Pawig et al., 2010] which uses a reference signal in an echo cancellation solution. The authors implement a least mean square adaptive algorithm in order to calculate the frequency offset. It is able to eliminate a frequency offset on the order of Hz.

In this thesis we put our attention on solving synchronization problems in WASNs when BSS algorithms are applied. It must be said that only a few synchronization solutions have been proposed focusing on BSS problems. An interesting study was done in [Wehr et al., 2004] which concludes that BSS performs well when the data of distributed acoustic sensors is synchronized. The first step of this proposal is the study of the sensitivity of one classical BSS algorithm to sampling rate deviations, demonstrating that only a few Hertz is enough to degrade its performance. In the solution one element of the network, that acts as master node, sends a digital synchronization signal by radio frequency to the rest of nodes to synchronize them. The WASN used is composed of personal computing devices instead of dedicated audio hardware since its use becomes very popular in many research fields. Several synchronization signals are analyzed to determine the most suitable one for its BSS application. When these signals are captured by the different nodes of the network, cross-correlations between the captured and their reference signals are calculated to determine and compensate the sampling rate deviations. This work demonstrates to achieve high precision but it does not address the synchronization problem due to differences in propagation delays between source signals and sensors.

1.4 Problem formulation and scope of the thesis

Audio source separation is an important field of research presented in a large number of applications such as real-time speaker separation for simultaneous translation, sampling of musical sounds for music composition, speech enhancement within hearing aids or voice cancellation. One important problem within audio source separation is the cocktail party problem in which competing speech sounds must be separated. The cocktail party problem is of undoubted interest for the scientific community since it is presented in many systems such as domotic systems, hands-free systems in cars, etc. In certain circumstances, as for instance, reverberant rooms or very noisy environments, it can still be considered an open and unsolved problem. Given these different aspects to consider, the main objective of this thesis is to provide new tools for developing speech separation in reverberant environments.

From the main objective of the thesis, the next particular problems are addressed:

- Speech separation in reverberant environments is not a easy task for many SSS solutions. Reverberation can be considered as interferences that arrive at the sensor array in many directions, what indicates that it is an important problem for beamforming techniques since they are designed to eliminate interferences that come from certain directions. Furthermore, beamforming has the disadvantage of using sophisticated microphone arrays what increases the complexity of the problem. In respect of CASA methods, reverberation hinders stages as segmentation since it causes overlapping between the regions of the speech sources in the time-frequency domain and so, acoustic features as pitch can not be correctly extracted. In the case of BSS methods, sparse-based methods perform worse in the presence of reverberation because there is more overlapping between sources. However, CASA and BSS techniques can use very simple microphone arrays what is an important advantage. In this respect, our objective is to establish a speech separation solution that works properly in the presence of reverberation effects but using a simple microphone array.

- Time-frequency masking is a very popular technique utilized both in CASA and BSS algorithms. In the case of BSS it entails some advantages, for instance, it is able to separate sound signals in underdetermined separation problems. However, since it is a subtraction type algorithm, it introduces important distortions in the separated sound sources. Considering it, in this thesis we aim at minimizing the presence of these distortions.

- WASNs are very popular aiming at developing acoustic monitoring of rooms for many applications. In the case of speech source separation they are further utilized, but with important drawbacks for classical separation methods. Since WASNs are wireless networks and cover large areas, important delays between the signals captured by the different microphones are introduced. With respect to wirelessly connections, not all the nodes have the same clock and then, the clock synchronization problem arises. In this respect, there are many solutions of the clock problem in the literature. Concerning covering larger areas than classical microphone arrays, important delays differences between each source and the microphones appear. It also causes desynchronization between the mixtures received at the microphones. In this regard, there are not works dealing with this problem and so, in this thesis we aim at solving this problem in order to use classical BSS with WASNs.

1.5 Structure of the thesis

This thesis has been divided into two main blocks, which are organized as follows:

- The first block contains two chapters in which the preliminary study of the problem has been carried out. The first chapter, which is the current one, contains a description of the cocktail party problem, a review of the state of the art of the problems addressed in this thesis and a description of the goals and structure of the thesis. The second chapter establishes the theoretical basis necessary to understand our separation problem as well as the material used in the experiments to evaluate our algorithms.

- The second block is the main block. It contains a description of the research conducted to fulfill the objective of this thesis. The experimental works and the outcomes obtained are also described in this block. The chapters of this second block correspond with each of the goals of this thesis and they are the next:

  - In Chapter 3 different proposals are introduced in order to improve speech separation when very simple microphone arrays are utilized. These proposals address both the

mixing matrix estimation and the separation stages of BSS algorithms in different ways.

- Chapter 4 deals with the synchronization problem that signals captured by the different nodes suffer in wireless acoustic sensor networks. First, this synchronization problem is analyzed and then, different solutions are proposed with the objective of using BSS methods with WASNs.
- Chapter 5 sums up the main contributions of the thesis by means of analyzing the most relevant results during this research. In a second part a description of future research lines has also been presented. Finally, a list of publications derived from this thesis are also included.
- The last section includes the bibliography used in this thesis.

Chapter 2

Materials and Methods

2.1 Introduction

This chapter intends to provide a formal description of the speech separation problems addressed in this thesis and to present the tools to evaluate the performance of the proposed algorithms. First, the mixing model is described mathematically to understand the separation problem to be solved. The circumstances under which speech sources are mixed must be modeled to find appropriate separation solutions. Different factors such as noise or reverberation should be considered to extract speech signals from the mixtures correctly. In a second part, the speech separation methods used as reference are explained. Third, the classical time delay estimation (TDE) methods that have been studied to synchronize speech mixtures captured by microphones in WASNs are discussed. Finally the methodologies used to evaluate the different solutions proposed in this thesis are presented. Examples of these methodologies include the mixing model to simulate the auralization of rooms, the sound database or the objective measurements to determine the quality of the separated speech sources, among others.

2.2 Mixing models

In oder to avoid any doubt, it must be clarified the notation used in this thesis. It is very usual to consider the variable (t) to represent continuous time and the variable (n) for discrete time. $[n]$ will be nT_s, where T_s is the sampling period, that is, the inverse of the sampling frequency $T_s = 1/f_s$.

The way that sound mixtures are generated depends on the type of mixing scenario. From a mathematical point of view, in a typical BSS problem, it is assumed that a set of S source signals $\{x_1[n], \ldots, x_S[n]\}$ reaches M different sensors/microphones, where $z_m[n]$ is the mixture signal at the output of the m-th microphone. The general expression for an additive mixing model is given by

$$z_m[n] = \sum_{s=1}^{S} a_{ms}[n] * x_s[n] = \sum_{s=1}^{S} y_{ms}[n], \ \forall m = 1, ..., M, \tag{2.1}$$

where $*$ is the linear convolution operator. The acoustic impulse response ($a_{ms}[n]$) represents the acoustic channel between the s-th source and the m-th microphone. $a_{ms}[n]$ is the impulse response of a filter, concretely, a linear time-invariant (LTI) filter. $a_{ms}[n]$ depends on the scenario where the mixtures are produced and is function, among others, of the position of the source, the position of the microphone and the physical properties of the paths between the source and the

sensor. These impulse responses from sources to sensors are used to model the different phenomena, like, for instance, delays of the signals which are considered in [Bofill and Zibulevsky, 2001], path loss or multipath effects [Fabrizio and Farina, 2011]. The transformation of the acoustic impulse response into the frequency domain is known as acoustic transfer function. The characteristics of the ATF will determine the mixing model to be used. Finally, $y_{ms}[n]$ is the signal captured by the m-th microphone due to the s-th source.

2.2.1 Instantaneous or linear mixing model

This model is the simplest one and assumes that only a scaled version of each source signal reaches each sensor. Then, a mixture signal is a linear combination of the original sources at every time instant. Putting in a mathematical way, each mixture is determined by the following expression

$$z_m[n] = \sum_{s=1}^{S} \alpha_{ms} \cdot x_s[n], \ \forall m = 1, ..., M. \tag{2.2}$$

Since source signals are only scaled, the acoustic impulse response is $a_{ms}[n] = \alpha_{ms} \cdot \delta_k[n]$, where $\delta_k[n]$ is the Kronecker delta function. The instantaneous model is applicable to many separation problems, such as the problem of identifying underlying components of brain from recordings of its activity [Duda et al., 2000]. It is also a very common model in image processing to extract independent features or to improve image quality [Hyvärinen, 1999]. Within audio applications, some examples of this class of problems aiming at separating sources from instantaneous mixtures are, for instance, compact disc recordings (stereo instantaneous mixtures) in which speakers are separated from instruments. An interesting BSS algorithm working with this model in BASS is [O'grady, 2007].

2.2.2 Anechoic mixing model

This second model also includes the delays that source signals suffer along the direct paths between sources and microphones. Each mixture can be expressed as

$$z_m[n] = \sum_{s=1}^{S} \alpha_{ms} \cdot x_s[n - \delta_{ms}], \ \forall m = 1, ..., M. \tag{2.3}$$

As in the instantaneous problem, α_{ms} represents the attenuation that a source signal suffers from the s-th source to the m-th microphone. δ_{ms} is the novelty of this model and is the delay introduced by the channel between the s-th source and the m-th microphone. Then, the acoustic impulse response is defined as $a_{ms}[n] = \alpha_{ms} \cdot \delta_k[n - \delta_{ms}]$.

2.2.3 Echoic or convolutive mixing model

This is the most complex model since it considers that each microphone captures the source signals following the direct path, but also the attenuated and delayed versions of these source signals due to reverberation. A mixture signal captured by one sensor can be expressed as

$$z_m[n] = \sum_{s=1}^{S} \sum_{p=1}^{N_p} \alpha_{msp} \cdot x_s[n - \delta_{msp}], \ \forall m = 1, ..., M, \tag{2.4}$$

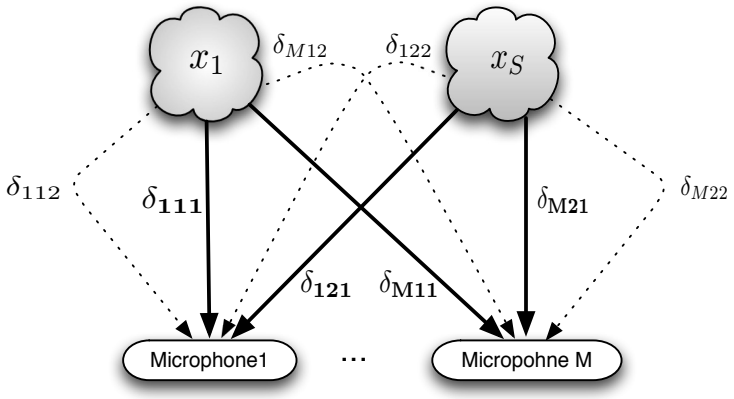

Figure 2.1: Illustrative example of the echoic model with S sources and M sensors (microphones). Solid lines indicate direct path and dashed lines non-direct paths.

where α_{msp} and δ_{msp} are the attenuations and delays introduced in the p-th path. N_p is the number of paths that signals take from the source to the microphones. Considering it, the acoustic impulse response is $a_{ms}[n] = \sum_{p=1}^{N_p} \alpha_{msp} \cdot \delta_k[n - \delta_{msp}]$. Figure 2.1 depicts a typical echoic model where there are S sources and M microphones. Each source signal follows N_p paths to arrive at each microphone. For simplicity, $N_p = 2$ has been chosen.

In this thesis separation problems of acoustic signals in rooms are dealt and so, the convolutive mixing model is the most suitable. The sound signals are reflected on the different surfaces and some of these reflections are captured by the microphones, what makes more difficult the separation task. There are many sound applications that consider the echoic model in rooms, as for instance, [Doclo and Moonen, 2003] where the echoic (reverberant) environment of a room is assumed to implement a time delay estimation method.

2.2.4 Noisy model

Bearing in mind our separation problems, the three previous models are incomplete since noise has not been considered. It is very usual to find additive noise in the signals captured by the microphones. This additive noise can be acoustic noise and/or noise caused by inaccuracies in the measurements performed by sensors. Then in each mixture, a noise term ($y_{m0}[n]$) is included, resulting

$$z_m[n] = \sum_{s=1}^{S} a_{ms}[n] * x_s[n] + y_{m0}[n] = \sum_{s=0}^{S} y_{ms}[n], \ \forall m = 1, ..., M, \tag{2.5}$$

where $y_{m0}[n]$ is usually considered as additive Gaussian noise with zero mean and σ^2 variance, which models the effects of a diffuse noise field.

For simplicity, it is very usual to use matrix notation to express the mixing models. Assuming M sensors, $\mathbf{z}$ is defined as a $M \times 1$ vector of mixtures, $\mathbf{z} = [z_1[n], \ldots, z_M[n]]^T$, where the operator $(\cdot)^T$ denotes matrix transposition. Let us define the $S \times 1$ vector of original sources as $\mathbf{x} = [x_1[n], \ldots, x_S[n]]^T$, the $M \times 1$ noise vector $\mathbf{y}_0 = [y_{10}[n], \ldots, y_{M0}[n]]^T$ and the mixing matrix as

$$\mathbf{A}[n] = \begin{bmatrix} a_{11}[n] & \dots & a_{1S}[n] \\ \vdots & \ddots & \vdots \\ a_{M1}[n] & \dots & a_{MS}[n] \end{bmatrix}. \tag{2.6}$$

Thus the BSS problem can be modeled by the expression

$$\mathbf{z}[n] = \mathbf{A}[n] \cdot \mathbf{x}[n] + \mathbf{y}_0[n], \tag{2.7}$$

where $\cdot$ denotes the element-wise convolution operation.

2.3 Classical separation algorithms

Different classical separation algorithms have been used as reference in this thesis. These methods work in the time-frequency domain since they rely on the assumption that speech sources are sparse in this transformed domain. Thus, the mixing model must also be formulated in the time-frequency domain. It leads to calculate time-frequency representations of the speech signal by means of different tools, where the STFT is one of the most popular.

With respect to the structure of these separation algorithms, as in the case of DUET, these algorithms contain two different parts: 1) the mixing matrix estimation and 2) the separation stages. Concerning the first stage, two solutions have been tested. One solution can be considered a generalization of the one proposed in DUET since it can work with more than two sensors. This proposal uses a multi-dimensional histogram. The other method is based on the use of expectation-maximization stages to estimate the mixing matrix. In the second stage, once the mixing matrix has been calculated, the original sources must be separated. In both solutions this separation process is developed by means of T-F masking, but the difference is that in one algorithm a binary mask is used (like in DUET [Rickard, 2007]), and in the other case, masks are generated using l_1-norm minimization, like in the LOST algorithm [O'grady, 2007].

2.3.1 The short-time Fourier transform (STFT)

Commonly, the discrete short-time Fourier transform (discrete STFT) is used for multi-channel source separation as it can be observed in different works such as [Araki et al., 2006, Rickard, 2007, De Fréin and Rickard, 2011]. From now on, we will use the acronym STFT instead of discrete STFT in the remainder of this thesis. This powerful tool will be explained in the following paragraphs.

Jean Baptiste Joseph Fourier (1768-1830) introduced the theory that all the signals are the sum of sinusoids. Applying a proper frequency analysis, a signal can be separated into its frequency components. The Fourier transform is a tool that allows to separate signals into their sinusoidal components in the frequency domain. Assuming a signal $x[n]$ in the discrete-time domain, its frequency representation $X(k)$ is expressed as

$$X(k) = \sum_{n=0}^{N-1} x[n] e^{-i\frac{2\pi}{N}kn}, \; k = 0, \dots, N-1, \tag{2.8}$$

where N is the total number of samples of $x[n]$ and $X(k)$ the discrete-frequency domain representation. In the discrete-time domain, this tool is known as discrete Fourier transform (DFT) and the inverse discrete Fourier transform (IDFT) is determined as follows

$$x[n] = \frac{1}{N}\sum_{k=0}^{N-1} X(k)e^{i\frac{2\pi}{N}kn},\ n = 0, \ldots, N-1. \tag{2.9}$$

The DFT is an orthogonal transformation which uses complex exponentials as basis functions. The normalized frequency of the basis functions is fixed and given by $\frac{kf_s}{N}$ with $k = 0, \ldots, N-1$ and the sampling frequency is f_s. Since the samples of speech signals are real numbers, the DFT is symmetric. It causes that only $\lfloor \frac{N}{2} + 1 \rfloor$ frequency bands are used. The DFT is a frequency localized transformation and has the disadvantage that changes in frequency over time can not be shown.

To observe changes in frequency over time the STFT is proposed, which can be considered a two-dimensional transformation. It consists in splitting the input signal into segments using a sliding time-limited window and then, the DFT of each segment is calculated. These DFTs of segments are introduced into the columns of the STFT. It is very usual to overlap the segments with each other aiming at creating smooth transitions between STFT frames. Furthermore, the discontinuities that may occur if signals are segmented into frames in which a sinusoid does not complete a full period. It causes frequency smearing (also known as spectral leakage) that appears as tails at the base of the peak that indicates the frequency. One solution in order to avoid spectral leakage is to multiply windows ($w(t)$) by the STFT frames. In this thesis, the Hanning window will be used since it takes into account frequency resolution and sidelobe behavior.

Now, the steps followed in the STFT are described mathematically. First, the discrete time input signal ($x[n]$) is segmented intro frames as Equation (2.10) shows

$$x_l[p] = w[p]x[p + lD],\ p = 0, \ldots, P-1. \tag{2.10}$$

In the above expression, $x_l[p]$ is the windowed l-th frame of the signal, P is the window length and p a local time index. The number of samples that the sliding windows move between two consecutive frames is D (hop size). Now, the DFT of each windowed frame must be calculated as Equation (2.11) presents

$$X(k, l) = \sum_{p=0}^{P-1} w[p]x[p + lD]e^{-i\frac{2\pi}{N}kp},\ k = 0, \ldots, N-1. \tag{2.11}$$

2.3.2 Time-frequency mixing model

Once the STFT has been explained mathematically, the time-frequency domain version of Equation (2.5) can be expressed as

$$Z_m(k, l) = \sum_{s=1}^{S} A_{ms}(k) \cdot X_s(k, l) + Y_{m0}(k, l) \tag{2.12}$$

where k is the frequency index, and l the time index. $Z_m(k, l)$ and $X_s(k, l)$ represent the STFTs of the m-th mixture $z_m[n]$ and of the s-th source signal $x_s[n]$, respectively. The mixing matrix coefficient in the T-F domain for the different frequency bins ω_k is $A_{ms}(k)$, where $\omega_k = \frac{2\pi}{K} \cdot k$ and K is the STFT frame size.

It is important to highlight that Equation (2.12) is only considered a good approximation of Equation (2.5) in the time-frequency domain under important restrictions, for instant, a

minimum STFT frame size is required [Wang et al., 2010]. These requirements will be studied in a more detailed way in Chapter 4.

For simplicity, Equation (2.12) will be rewritten as Equation (2.13),

$$Z_m(k,l) = \sum_{s=0}^{S} Y_{ms}(k,l), \tag{2.13}$$

where the component for $s = 0$ represents the contribution of noise and for $s \neq 0$, the contributions of the S speech sources.

The signals at the output of the M sensors can also be described using matrix notation, obtaining Equation (2.14),

$$\mathbf{Z}(k,l) = \mathbf{A}(k) \cdot \mathbf{X}(k,l) + \mathbf{Y}_0(k,l), \tag{2.14}$$

where $\mathbf{A}(k)$ is the $M \times S$ mixing matrix for the frequency bin ω_k. At each time-frequency point, the $M \times 1$ vector $\mathbf{Z}(k,l)$ and the $S \times 1$ vector $\mathbf{X}(k,l)$ are both complex. Similarly, the $M \times S$ matrix $\mathbf{A}(k)$ is also complex for each frequency point. $\mathbf{Y}_0(k,l)$ is a column vector of M i.i.d. white complex Gaussian noise signals with zero mean and variance σ^2. Since the problem is expressed as separate multiplications in each time-frequency bin, it is usual in some BSS algorithms to split the problem into smaller ones in each frequency bin of the STFT, that is, the frequency components of the original sources can be calculated separately.

Matrix forms are determined by the way the mixing process occurs, that is, it depends on several factors, like, the propagation of the source signals, the characteristics of the microphone array, etc. Therefore, mixing matrix coefficients $A_{ms}(k)$ respond to different expressions, depending on several situations:

1. *Instantaneous mixtures.* In this case, mixing matrix coefficients can be expressed as shown in Equation (2.15)

$$A_{ms}(k) = \alpha_{ms},\ \alpha_{ms} \in \mathbb{R}. \tag{2.15}$$

   It can be observed that it has no dependence on frequency and coefficients are real.

2. *Microphone array with flat response and non-reverberant room.* In this situation $A_{ms}(k)$ is determined by using expression

$$A_{ms}(k) = \alpha_{ms} e^{-i\omega_k \delta_{ms}},\ \alpha_{ms} \in \mathbb{R}^+. \tag{2.16}$$

   It is clear that the magnitude component α_{ms} has no dependence on frequency, but the phase of the mixing matrix coefficients varies linearly with frequency.

3. *Microphone array with non-flat response and non-reverberant room.* In this situation, the mixing matrix coefficients are expressed as Equation (2.17) indicates

$$A_{ms}(k) = \alpha_{ms}(k) e^{-i\omega_k \delta_{ms}},\ \alpha_{ms} \in \mathbb{R}^+. \tag{2.17}$$

In our case of study, its elements depend on the number and position of the scatters and therefore $A_{ms}(k)$ has a general expression.

2.3.3 Sparse assumption

The separation algorithms implemented in this thesis assume that speech sources are sparse in the time-frequency domain and in this sense, it must be analyzed whether speech signals are actually or not sparse in this domain. Theorically, thanks to the Laplacian PDF of speech signals, sparse representations of them can be achieved. Firstly, it makes sense to define what sparsity is. From a mathematical point of view, assuming two sparse signals in the time-frequency domain $(X_s(k,l), X_j(k,l))$, the following expression is satisfied

$$X_s(k,l)X_j(k,l) = 0, \forall k,l \text{ and } s \neq j, \tag{2.18}$$

but this assumption is not real for speech. Observing time-frequency representations of speech sources, we can notice that speech sources are active, that is, they have values significantly different from zero, only in some samples. It entails that for the vast majority of samples it will be accomplished

$$X_s(k,l)X_j(k,l) \simeq 0, s \neq j, \tag{2.19}$$

what can be considered an approximation of Equation (2.18), or in other words, it can be concluded that speech signals are quasi-sparse in this transformed domain. It means that the probability of two sources having large energy in the same time-frequency point is low.

2.3.4 The DUET algorithm

In this section DUET [Rickard, 2007] is described since, maybe, it is the most important SSS algorithm dealing with the underdetermined separation problem. Assuming an anechoic mixing model, it is able to recover any number of sources from only two mixtures. Its main characteristic is to exploit the sparseness of sources in the time-frequency domain. The mixing matrix is estimated by means of clustering the relative attenuation-delay pairs between the two microphones. Once the mixing parameters are obtained, time-frequency masks are generated to separate sound sources. These steps are described below.

Ignoring noise components, Equation (2.14) can be rewritten for only two mixtures and S sources as

$$\begin{bmatrix} Z_1(k,l) \\ Z_2(k,l) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 & \dots & 1 \\ g_{21}e^{-i\omega_k\tau_{21}} & \dots & g_{2s}e^{-i\omega_k\tau_{2s}} & \dots & g_{2S}e^{-i\omega_k\tau_{2S}} \end{bmatrix} \begin{bmatrix} X_1(k,l) \\ \vdots \\ X_s(k,l) \\ \vdots \\ X_S(k,l) \end{bmatrix}, \tag{2.20}$$

where g_{2s} and τ_{2s} are the level and phase differences between both microphones for the s-th source, respectively. In a mathematical form, $g_{2s} = \alpha_{2s}/\alpha_{1s}$ and $\tau_{2s} = \delta_{2s} - \delta_{1s}$. Considering valid the assumption that sources are sparse, that is, only the j-th source $(X_j(k,l))$ is active in a point (k,l), Equation (2.20) results

$$\begin{bmatrix} Z_1(k,l) \\ Z_2(k,l) \end{bmatrix} \simeq \begin{bmatrix} 1 \\ g_{2j}e^{-i\omega_k\tau_{2j}} \end{bmatrix} X_j(k,l). \tag{2.21}$$

Considering it, from the ratio of the STFT of the two sound mixtures at the point (k,l), $R(k,l)$, the mixing parameters can be obtained since this ratio does not depend on the active sources. In Equation (2.22) it can be observed the expression of this ratio

$$R(k,l) = \frac{Z_2(k,l)}{Z_1(k,l)} = g_{2j}e^{-i\omega_k\tau_{2j}}. \tag{2.22}$$

Then, the local mixing parameters for each time-frequency point are estimated from

$$\hat{g}(k,l) = |R(k,l)| \tag{2.23}$$

$$\hat{\tau}(k,l) = -\frac{1}{\omega_k}\angle(R(k,l)). \tag{2.24}$$

The local mixing parameters are the mixing parameters if the speech sources are strictly sparse, but really, this assumption is not true since speech sources are quasi-sparse. Thus, the local mixing parameters are an approximation of the mixing parameters and they will cluster around the mixing parameters.

But a problem arises when delays differences are obtained due to spatial aliasing. This problem occurs if the distance between microphones is large, specifically, the phase is not unique if $|\omega_k\tau_{2j}| < \pi$. From the point of view of the distance between microphones, to avoid spatial aliasing $d < \frac{c}{2f_{max}}$, where f_{max} is the highest frequency of interest and c is the speed of sound. One way to overcome this limitation can be to analyze the phase difference between adjacent time-frequency points as illustrated in the following expression

$$R'(k,l) = \frac{Z_2(k,l)}{Z_1(k,l)}\left(\frac{Z_2(k+\Delta k,l)}{Z_1(k+\Delta k,l)}\right)^* = g_{2j}^2 e^{i\Delta\omega_k\tau_{2j}}, \tag{2.25}$$

where the operator $(\cdot)^*$ represents complex conjugation. Thanks to it, the constrain ambiguity has been relaxed to $|\Delta\omega_k\tau_{2j}| < \pi$. The value $\Delta\omega_k$ can be made small by oversampling the frequency bands, since $\Delta\omega_k$ is controlled by Δk. Then, the delay estimator for large distances between microphones becomes

$$\hat{\tau}'(k,l) = -\frac{1}{\Delta\omega_k}\angle(R'(k,l)). \tag{2.26}$$

Other problem related to the attenuation estimator $\hat{g}(k,l)$ arises if the microphone signals are swapped. In order to solve it, a symmetric attenuation estimator $\hat{g}'(k,l)$ [Rickard, 2007] is used:

$$\hat{g}'(k,l) = \hat{g}(k,l) - \frac{1}{\hat{g}(k,l)}. \tag{2.27}$$

At this point a clustering process is carried out by means of a two-dimensional smoothed weighted histogram. $(\hat{g}'(k,l),\ \hat{\tau}(k,l))$ are weighted by a time-frequency dependent weight to construct the histogram. This weight is usually $|Z_1(k,l)Z_2(k,l)|$. A number of clusters equal to the number of sound sources in the mixtures appear, and they are centered on the actual mixing parameters. The centers of the clusters will be the mixing parameters associated to each source signal. If these cluster are reasonably separated, the histogram can be smoothed by a FIR filter. It must be considered that the DUET algorithm knows in advance the number of sound sources in the mixtures and so, the number of peaks that must be identified in the histogram.

Once the peaks have been located, time-frequency binary masks can be generated to separate the sources. Each time-frequency point of the mixture is assigned to the peak which is closest to the local mixing parameter of that point.In [Yilmaz and Rickard, 2004] it is proposed the use of a measure of closeness based on the likelihood function. Denoting the center of the s-th cluster

by $(\hat{g}'_s,\ \hat{\tau}_s)$, consequently $\hat{g}_s$ and $\hat{\tau}_s$ are the s-th source mixing parameter estimates. $\hat{g}_s$ will be obtained by the expression

$$\hat{g}_s = \frac{\hat{g}'_s + \sqrt{\hat{g}'^2_s + 4}}{2}. \tag{2.28}$$

Using Equation (2.29)

$$J(k,l) := \underset{s}{argmin} \frac{|\hat{g}_s e^{-i\omega_k \hat{\tau}_s} Z_1(k,l) - Z_2(k,l)|^2}{1+\hat{g}_s^2}, \tag{2.29}$$

each time-frequency point is assigned to a source.

The time-frequency mask for the s-th source is constructed as the following expression indicates

$$M_s(k,l) = \begin{cases} 1 & J(k,l) = s \\ 0 & \text{otherwise.} \end{cases} \tag{2.30}$$

Then, the original sources are extracted using the maximum likelihood (ML) estimator [Yilmaz and Rickard, 2004] and the binary masks in Equation (2.30), resulting

$$\hat{X}_s(k,l) = M_s(k,l) \left(\frac{Z_1(k,l) + \hat{g}_s e^{-i\omega_k \hat{\tau}_s} Z_2(k,l)}{1+\hat{g}_s^2} \right). \tag{2.31}$$

2.3.5 Classical multi-channel separation method: a generalization of the DUET algorithm

As mentioned in the previous section, the so-called DUET algorithm is one of the most representative methods for sound source separation and it represents a good solution when the mixtures are anechoic. This algorithm is able to recover any number of sources from only two mixtures, thus it deals with the undetermined problem. The mixing matrix estimation stage of DUET consists in determining the relative attenuation-delay pairs (spatial cues) between two sensors. Unlike DUET, our generalization can work with any number of microphones. Then, the mixing model expressed by Equation (2.20) can be rewritten in the time-frequecy domain for any number of sensors as Equation (2.32) indicates

$$\begin{bmatrix} Z_1(k,l) \\ \vdots \\ Z_m(k,l) \\ \vdots \\ Z_M(k,l) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 & \dots & 1 \\ g_{21}(k)e^{-i\omega_k\tau_{21}} & \dots & g_{2s}(k)e^{-i\omega_k\tau_{2s}} & \dots & g_{2S}(k)e^{-i\omega_k\tau_{2S}} \\ \vdots & & \vdots & & \vdots \\ g_{m1}(k)e^{-i\omega_k\tau_{m1}} & \dots & g_{ms}(k)e^{-i\omega_k\tau_{ms}} & \dots & g_{mS}(k)e^{-i\omega_k\tau_{mS}} \\ \vdots & & \vdots & & \vdots \\ g_{M1}(k)e^{-i\omega_k\tau_{M1}} & \dots & g_{Ms}(k)e^{-i\omega_k\tau_{Ms}} & \dots & g_{MS}(k)e^{-i\omega_k\tau_{MS}} \end{bmatrix} \begin{bmatrix} X_1(k,l) \\ \vdots \\ X_s(k,l) \\ \vdots \\ X_S(k,l) \end{bmatrix}, \tag{2.32}$$

where $Z_m(k,l),\ \forall\ m = 1,\dots,M$ are the STFTs of the mixture signals received by the M microphones. The reader can note that all the elements in the first row are set to one, the reason is that the first microphone is assumed to be the reference sensor, what does not entail loss of generality. Considering it, the mixture signals contain delayed and attenuated versions of the sources captured by the first microphone, $X_s(k,l),\ \forall\ s = 1,\dots,S$. Having a look at Equation (2.32), it can be observed that the mixing matrix coefficients $A_{ms}(k)$ have two components: g_{ms} and τ_{ms}, which are the level and time differences between the first microphone (reference sensor)

and the m-th microphone for the s-th source. Henceforth, in order to simplify the notation it is assumed $g_{ms}(k) = g_{ms}$ (flat response of the microphone asumption) and so, g_{ms} coefficients are considered independent of frequency, resulting Equation (2.32) as Equation (2.33),

$$\begin{bmatrix} Z_1(k,l) \\ \vdots \\ Z_m(k,l) \\ \vdots \\ Z_M(k,l) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 & \dots & 1 \\ g_{21}e^{-i\omega_k\tau_{21}} & \dots & g_{2s}e^{-i\omega_k\tau_{2s}} & \dots & g_{2S}e^{-i\omega_k\tau_{2S}} \\ \vdots & & \vdots & & \vdots \\ g_{m1}e^{-i\omega_k\tau_{m1}} & \dots & g_{ms}e^{-i\omega_k\tau_{ms}} & \dots & g_{mS}e^{-i\omega_k\tau_{mS}} \\ \vdots & & \vdots & & \vdots \\ g_{M1}e^{-i\omega_k\tau_{M1}} & \dots & g_{Ms}e^{-i\omega_k\tau_{Ms}} & \dots & g_{MS}e^{-i\omega_k\tau_{MS}} \end{bmatrix} \begin{bmatrix} X_1(k,l) \\ \vdots \\ X_s(k,l) \\ \vdots \\ X_S(k,l) \end{bmatrix}. \tag{2.33}$$

Now, how our speech separation algorithm estimates the mixing matrix must be explained. As in the case of DUET, this separation algorithm works under the assumption of sparse speech sources in the time-frequency domain, or in other words, it assumes that for a set of points Ψ_s of the transformed mixtures $Z_m(k,l),\ \forall\ m = 1,\dots,M$, only the $s-th$ source is working (sparse assumption). Mathematically it can be expressed by means of Equation (2.34)

$$\Psi_s = \left\{ (k,l) \;\middle|\; |X_p(k,l)| \simeq 0\ \forall\ p \neq s \right\}. \tag{2.34}$$

And then, the expression of the mixing model in points that belong to Ψ_s can be approximated to

$$\begin{bmatrix} Z_1(k,l) \\ \vdots \\ Z_m(k,l) \\ \vdots \\ Z_M(k,l) \end{bmatrix} \simeq \begin{bmatrix} 1 \\ g_{2s}e^{-i\omega_k\tau_{2s}} \\ \vdots \\ g_{ms}e^{-i\omega_k\tau_{ms}} \\ \vdots \\ g_{Ms}e^{-i\omega_k\tau_{Ms}} \end{bmatrix} X_s(k,l), \tag{2.35}$$

where $X_s(k,l)$ is the active source in Ψ_s. So in any point of Ψ_s, the values of the transformed mixtures only depend on the values of the s-th column of the mixing matrix and on the contribution of the s-th source, that is, only one source is contributing to the mixtures.

Bearing in mind that the objective of our algorithm is to estimate the mixing matrix, looking at Equation (2.35) it seems evident that if the contribution of the s-th source is eliminated from the mixtures in Ψ_s, the s-th column of the matrix will be obtained. And now the question is: *how can we estimate the columns of the mixing matrix?* Two alternatives have been considered in this thesis and they are explained in the two following sections.

2.3.6 Multi-dimensional histogram

The first solution relies on the use of a multi-dimensional histogram [Master, 2006] that can be considered a generalization of the two-dimensional one performed in DUET. The first step consists in calculating ratios between the first mixture (reference one) and the rest of them for all the time-frequency points. In order to simplify the mathematical formulation, these ratios are included in a vector $\mathbf{v}(k,l)$ that represents any time-frequency point

$$\mathbf{v}(k,l) = \left[\frac{Z_2(k,l)}{Z_1(k,l)} \dots \frac{Z_m(k,l)}{Z_1(k,l)} \dots \frac{Z_M(k,l)}{Z_1(k,l)} \right],\ \forall (k,l). \tag{2.36}$$

Assuming a point $(k,l) \in \Psi_s$, that is, considering Equation (2.35), then Equation (2.36) results

$$\mathbf{v}(k,l) = \left[\frac{g_{2s}e^{-i\omega_k\tau_{2s}}X_s(k,l)}{X_s(k,l)} \dots \frac{g_{ms}e^{-i\omega_k\tau_{ms}}X_s(k,l)}{X_s(k,l)} \dots \frac{g_{Ms}e^{-i\omega_k\tau_{Ms}}X_s(k,l)}{X_s(k,l)} \right]. \tag{2.37}$$

And simplifying,

$$\mathbf{v}(k,l) = \left[g_{2s}e^{-i\omega_k\tau_{2s}} \dots g_{ms}e^{-i\omega_k\tau_{ms}} \dots g_{Ms}e^{-i\omega_k\tau_{Ms}} \right]. \tag{2.38}$$

It can be observed that the elements of the vector $\mathbf{v}(k,l)$ contain the mixing parameters of the s-th column of $\mathbf{A}(k)$ if $(k,l) \in \Psi_s$. And so, identifying the points that belong to each set Ψ_s, $\forall s = 1, ..., S$, all the columns of the mixing matrix can be obtained using the $\mathbf{v}(k,l)$ representation of the time-frequency samples. If speech sources are strictly sparse, the mixing parameters will be in $\mathbf{v}(k,l)$, but speech sources are quasi-sparse and so, the elements of $\mathbf{v}(k,l)$ will contain local mixing parameters that are clustered around the mixing parameters.

To facilitate the mixing matrix estimation, the mixing matrix is calculated at each frequency bin independently. Then, the goal of our algorithm is to determine S sets of points $(\Psi_s, \forall s = 1, \dots, S)$ at each frequency bin. Each of these S concentrations of points corresponds to the points that belong to Ψ_s, or in other words, the points where one source is dominant. This procedure at each frequency bin can be carried out by means of a multi-dimensional histogram that groups the time-frequency points when they are represented by the vector $\mathbf{v}(k,l)$. It can be observed that the vector $\mathbf{v}(k,l)$ contains $M-1$ complex elements, where each of them has magnitude (attenuation) and phase (delay) components. Aiming at making easier this clustering process, it has been preferred to represent (k,l) points by a vector

$$\mathbf{w}(k,l) = \left[g_{2s} \quad \dots \quad g_{ms} \quad \dots \quad g_{Ms} \quad -\omega_k\tau_{2s} \quad \dots \quad -\omega_k\tau_{ms} \quad \dots - \omega_k\tau_{Ms} \right], \tag{2.39}$$

where $\mathbf{w}(k,l)$ is a vector that contains $2 \cdot (M-1)$ elements. And so, the multi-dimensional histogram technique has $2 \cdot (M-1)$ dimensions, one per element of $\mathbf{w}(k,l)$. Theoretically, representing all the time-frequency points of the mixtures by means of $\mathbf{w}(k,l)$ and calculating the multi-dimensional histogram, S concentrations of points may appear at each frequency bin. When these concentrations are determined, a vector $\hat{\mathbf{w}}_s(k)$ that represents the center of each set is used to determine the mixing parameters.

If speech sources are completely sparse, the estimation of the concentrations of points by using a multidimensional histogram can be easily performed. It must also be considered that sparsity is not always a valid assumption for speech signals, what implicates that the estimation of the mixing parameters will be approximated. On the other hand, important parameters such as the range of variation and the bin resolution of the histogram must be fixed, and the performance of the method highly depends on their values. For these reasons, different clustering solutions have been tested in this thesis [Llerena et al., 2012, Llerena et al., 2013a], and another one is presented in the next section.

2.3.7 Clustering technique based on expectation-maximization stages

This approach is based on the one performed in the LOST algorithm [O'grady, 2007], which uses expectation-maximization stages. The main difference is that LOST uses this solution for solving separation problems in which speech mixtures are instantaneous, while we have adapted it to work with convolutive speech mixtures. Some of the most important aspects will be described in the followings paragraphs.

Our procedure works directly with the time-frequency samples of the mixtures without applying any transformation to them, unlike in the case of the multi-dimensional histogram, where time-frequency points are represented by $\mathbf{w}(k,l)$. However, as in the previous case, the estimation of the mixing matrix is calculated at each frequency bin independently.

This method assumes that speech sources are sparse in the time-frequency domain and so, they can be characterized by a Laplacian PDF as shown in the expression below

$$\mathbf{P}(c) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|c|}. \tag{2.40}$$

These distributions are usually fat-tailed with sharp peaks at the origin. Since speech sources have this distribution, the mixture observations are generated by a linear combination of these Laplacian densities, commonly known as a Laplacian mixture model (LMM). The Laplacian density can be expressed by

$$\mathcal{L}(\nu|\beta, z) = \beta e^{-2\beta|z-\nu|} \propto e^{-\beta|z-\nu|} \tag{2.41}$$

where ν represents the centre of the Laplacian and β controls the boundary of the density. Note that the Laplacian density depends on the absolute difference from the centre.

The basic idea of this method is to determine the variance of a data set and its directions, as PCA methods do. S different directions ($\nu_s(k),\ \forall s = 1, \ldots, S$) will be extracted from the data set and they will correspond to the columns of $\mathbf{A}(k)$. The mixing matrix is generated as the following expression shows

$$\mathbf{A}(k) = [\nu_1(k)|...|\nu_S(k)] \tag{2.42}$$

We establish the data set as a vector of observations $\mathbf{Z}_k(l)$ for the k-th frequency bin as indicated below

$$\mathbf{Z}_k(l) = \begin{bmatrix} Z_1(k,l) \\ \vdots \\ Z_M(k,l) \end{bmatrix},\ \forall l = 1, \ldots, L. \tag{2.43}$$

Once the observations are ready, the directions $\nu_s(k)$ must be obtained. In this respect, the LOST algorithm randomly initializes these directions and then, after a number of signal processing EM steps, they are adjusted. It must be mentioned that we have modified the number of repetitions of these steps in order to improve the results of the separation algorithms.

In *E-step*, the method computes the probability that a sample of *any* mixture belongs to a direction. Thus, a metric that measures the distance between such directions is required. In this sense, an appropriate measure is achieved by calculating the difference between $\mathbf{Z}_k(l)$ and the projection of $\mathbf{Z}_k(l)$ onto ν, as illustrated below:

$$q_{sl}(k) = ||\mathbf{Z}_k(l) - (\nu_s(k) \cdot \mathbf{Z}_k(l)\nu_s(k))\,||, \tag{2.44}$$

where $\cdot$ denotes the dot product of the Euclidean space.

If observation and the Laplacian center are coincident, $q_{sl}(k)$ is at its minimum. By means of Expression (2.41), each linear subspace is characterized by the distribution

$$\mathcal{L}(q_{sl}(k), \beta) = e^{-\beta q_{sl}(k)}. \tag{2.45}$$

Then, it can be defined the mixture of multivariate Laplacian as

$$P(\mathbf{Z}_k(l)) = \sum_{s=1}^{S} \mathcal{L}\left(q_{sl}(k), \beta\right) = \sum_{s=1}^{S} e^{-\beta q_{sl}(k)}, \tag{2.46}$$

where speech is assumed to be identically and independently distributed, and β is assumed to be the same for each distribution.

Now, the mixture is defined. Since there must be S directions (one per speech source) in each frequency bin, the observations are segregated into sets associated with each direction. This segregation is achieved by estimating the probability of an observation belonging to a direction:

$$P(\mathbf{Z}_k(l)|\nu_s(k)) = \frac{e^{-\beta q_{sl}(k)}}{\sum_j e^{-\beta q_{jl}(k)}} \equiv \widetilde{q}_{sl}(k), \tag{2.47}$$

where $\widetilde{q}_{sl}(k)$ indicates the membership of the observation $\mathbf{Z}_k(l)$ to the direction $\nu_s(k)$.

At this point, it is known if a sample of a mixture belongs to one or another direction $\nu_s(k)$. In the next stage (**M-step**), the algorithm verifies if the estimations of the directions are correct or not.

In the effort of identifying the directions from a cloud of data that are already characterized, the principal eigenvector of its covariance matrix can be used. So, the covariance matrix of the weighted observations assigned to each direction is computed. The covariance matrix expression and assignment weightings are combined as follows:

$$\mathbf{C}_s(k) = \frac{\sum_l \widetilde{q}_{sl}(k) \left(\mathbf{Z}_k(l) - \mu(k)\right) \left(\mathbf{Z}_k(l) - \mu(k)\right)^T}{\sum_l \widetilde{q}_{sl}(k)}, \tag{2.48}$$

where $\mu(k)$ is a vector of the mean values of the rows of $\mathbf{Z}_k$, and $\mathbf{C}_s$ is the covariance of weighted observations associated with the s-th direction in the k-th frequency bin.

The principal eigenvector of the matrix is used as the new direction estimate. The expression of the covariance matrix can be expressed as

$$\mathbf{C}_s(k) = \mathbf{U}_s \mathbf{\Lambda}_s \mathbf{U}_s^{-1}, \tag{2.49}$$

where the columns of the matrix $\mathbf{U}_s$ contain the eigenvectors of $\mathbf{C}_s(k)$, and the diagonal matrix $\mathbf{\Lambda}_s$ contains their associated eigenvalues. The orientation of a linear subspace can be thought of as the direction of its greatest variance. The principal component with the largest variance λ_{max}, which corresponds to the principal eigenvector $\mathbf{u}_{max}$ of the variance matrix that is chosen as the new direction

$$\nu_s(k) = \mathbf{u}_{max} \tag{2.50}$$

These two steps (EM) are repeated until the values of the different directions $\nu_s(k)$ converge. It is worth mentioning that the new value of β in each step is chosen from the second greatest eigenvalue of the covariance matrix. Therefore, when the convergence of the vector of the lines is demonstrated, the mixing matrix can be estimated. This can be made with the structure expressed by Equation (2.42) where each column of $\mathbf{A}(k)$ is one of the estimated directions $(\nu_s(k))$.

In the LOST algorithm that has been implemented, the vectors are initialized by means of eigenvectors of the original data, which basically guarantees that there are not two initialized vectors in close proximity to each other and the best eigenvector will be that whose line direction is the one with higher variation of data. We have also repeated randomly the initialization of

vectors five times and selected the best case according to β values in the aim of reducing the effects of local minima in the process of optimization EM.

2.3.8 Separation stage

After performing mixing matrix estimation, the separation stage is carried out. In this stage speech sources are extracted from speech mixtures. Two of the most robust and known separation methods have been implemented and both of them use time-frequency masking techniques. The first one is based on time-frequency binary masking and the second one uses l_1-norm minimization to calculate smooth time-frequency masks. The reason why both classical proposals use time-frequency masking is that we assume that speech signals are sparse or quasi-sparse in the time-frequency domain and furthermore, time-frequency masking allows us to work with underdetermined separation problems, that is, the number of speech sources is greater than the number of microphones.

The dimensions of the estimated mixing matrix $\hat{\mathbf{A}}(k)$ are $M \times S$, where M is the number of mixtures and S is the number of speech sources. It must be taken into account that the dimensionality of $\mathbf{A}(k)$ determines the difficulty of the process of unmixing. With this idea in mind, let us suppose a set of sources which are presented in a time-frequency representation (k, l). Then, we model the mixtures as Equation (2.14) indicates. The term $\mathbf{Y}_0(k, l)$ models the error in the estimation of the mixing matrix at the time-frequency points (k, l). Within this model, we can employ a ML estimation to find the values of the sources $\mathbf{X}(k, l)$. So, we define the likelihood $L(k, l)$ of $\mathbf{X}(k, l)$, given the data $\mathbf{Z}(k, l)$, as the reader can see in Equation (2.51)

$$L(k,l) = \exp\left(\frac{-1}{2\sigma^2}\left(\mathbf{Z}(k,l) - \hat{\mathbf{A}}(k)\mathbf{X}(k,l)\right)^H\left(\mathbf{Z}(k,l) - \hat{\mathbf{A}}(k)\mathbf{X}(k,l)\right)\right), \tag{2.51}$$

where $(\cdot)^H$ denotes the conjugated transposed matrix. Thus, maximizing $L(k, l)$ is equivalent to minimize $J(k, l)$ as the expression below shows

$$J(k,l) = \left(\mathbf{Z}(k,l) - \hat{\mathbf{A}}(k)\mathbf{X}(k,l)\right)^H\left(\mathbf{Z}(k,l) - \hat{\mathbf{A}}(k)\mathbf{X}(k,l)\right). \tag{2.52}$$

Note that the minimization of this expression leads to the least squares solution. In order to minimize Equation (2.52), we can generate a system of equations $\partial J(k, l)/\partial X_s(k, l) = 0$, $s = 1, \ldots, S$. By solving this system of equations, we get the ML estimated $\hat{\mathbf{X}}(k, l)$ of the sources, for the l-th time frame and the k-th frequency bin, as illustrated in Equation (2.53),

$$\hat{\mathbf{X}}(k,l) = \left(\hat{\mathbf{A}}^H(k)\hat{\mathbf{A}}(k)\right)^{-1}\hat{\mathbf{A}}^H(k)\mathbf{Z}(k,l). \tag{2.53}$$

In those situations in which there is a determined case, $\hat{\mathbf{A}}(k)$ is square ($M = S$), the estimation of the mixtures can be calculated by using Equation (2.54),

$$\hat{\mathbf{X}}(k,l) = \hat{\mathbf{A}}^{-1}(k)\mathbf{Z}(k,l). \tag{2.54}$$

But underdetermined separation problems must also be solved since they are very common in cocktail party problems. In these situations where $M < S$, or in other words, $\hat{\mathbf{A}}^H(k)\hat{\mathbf{A}}(k)$ is not invertible, the solution of the unmixing problem must be estimated by other means, like for example using non-linear techniques. In this thesis we focus on the two methods based on time-frequency masking mentioned at the beginning of this section.

2.3.8.1 Separation stage using binary masking

Deepening a little more in the underdetermined case, roughly speaking, it can be said that different methods which make use of time-frequency domain work on the assumption that signals are sparse. Then, it is considered that the solution for a given time-frequency point (k, l) only depends on one source which is dominant. Probably, the most important algorithm based on this assumption is DUET [Rickard, 2007], which was already described in Section 2.3.4.

Considering that sparseness is a correct assumption, then, only a column of matrix $\hat{\mathbf{A}}(k)$ has non-zero values in a time-frequency point $(k, l) \in \Psi_s$ (see Equation (2.35)). Then, Equation (2.53) becomes Equation (2.55), where only the s-th source is emitting,

$$\hat{X}_s(k,l) = \frac{\sum_{m=1}^{M} Z_m(k,l)\hat{A}^*_{ms}(k)}{\sum_{m=1}^{M} \left|\hat{A}_{ms}(k)\right|^2}. \tag{2.55}$$

One important factor to take into account is that, observing Equation (2.55), we can see that the estimation for each source extracted from the mixtures comes from a linear combination, whose weights are proportional to the coefficients of the mixing matrix. This is similar to a weighted delay and sum beamformer, steered to each estimated source and, therefore, the application of the binary masking does not replaces the use of a beamformer for separating sources, but it compliments its use. Replacing $X_s(k, l)$ by its estimation $\hat{X}_s(k, l)$ in Equation (2.52) leads to Equation (2.56)

$$J(k,l)\Big|_{(k,l)\in\Psi_s} = \sum_{m=1}^{M} \left|Z_m(k,l) - \hat{A}_{ms}(k)\hat{X}_s(k,l)\right|^2. \tag{2.56}$$

So having a look at Equation (2.56), it is clear to note that the point (k, l) will be mainly influenced by the s-th source if $J(k, l)$ is minimized. It is important to notice that Equation (2.29) is a particularization of (2.56) in the case of M=2, as it was described in the DUET algorithm in Section 2.3.4. In the same way, Equation (2.55) generalizes Equation (2.31) for any number of microphones, as it was expected.

2.3.8.2 Separation stage using l_1-norm minimization

If signals are sparse, the probability of multiple signals being non-zero in an instant (k, l) is very low, but, in the case of speech signals and with a considerable number of speech sources (large value of S), this statement is not necessarily true. Since sparsity assumption may be questionable in the case at hand, a method aiming at overcoming this drawback is implemented.

Different separation methods to solve the underdetermined case can be found in the literature as the ones in [Rickard and Dietrich, 2000, Vielva et al., 2001]. The lack of information makes more complex techniques be necessary to separate the sources, but we have observed that the vast majority of them are focused on the instantaneous case. With binary masking, each sample of a mixture in an instant is assigned to only one source. Using l_1-norm minimization it is possible to assign each sample to more than one source, which is more realistic. This idea will be described in a more detailed way in the next paragraphs.

In order to extract the original signals from the mixtures, this method uses the l_1-norm minimization which is a linear operation that allows to assign the energy of $\mathbf{Z}(k, l)$ to M columns of $\hat{\mathbf{A}}(k)$, with the remaining $S - M$ columns assigned with zero coefficients. In other words, it supposes that at most M sources are non-zero valued in an instant. Therefore, the main

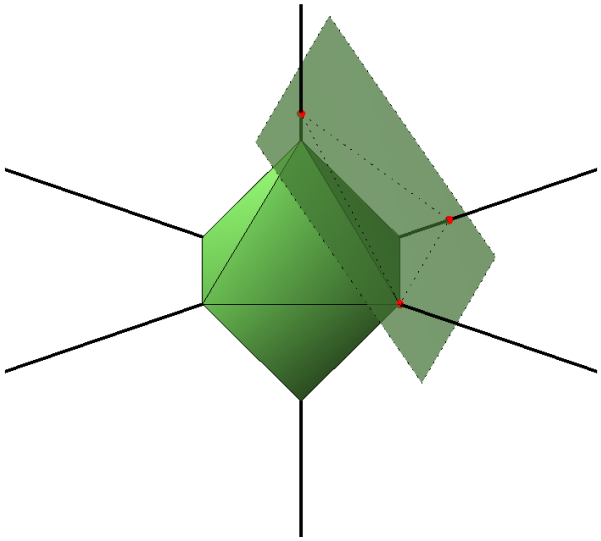

Figure 2.2: An illustrative example of the geometric of the sparse solution via l_1-norm minimization. Underdetermined case of one mixture ($S = 1$) and three sources ($N = 3$).

advantage of the proposed method in comparison with others is to consider that in an instant, more than one source could have non-zero value. l_1-norm is utilized in different works as for instance [Winter et al., 2005, Winter et al., 2007] aimed at obtaining the solutions $\hat{\mathbf{X}}(k,l)$ as Equation (2.57) indicates

$$min\|\hat{\mathbf{X}}(k,l)\|_1, \tag{2.57}$$

subject to $\hat{\mathbf{A}}(k)\hat{\mathbf{X}}(k,l) = \mathbf{Z}(k,l)$.

l_1-norm has already been successfully applied to instantaneous cases [Takigawa et al., 2004, O'grady, 2007] and in the problem at hand, it will be used to estimate $\hat{\mathbf{X}}(k,l)$ when mixtures are convolutive. It is very common to solve this problem with linear programming or quadratic programming methods [Chen et al., 1998].

This method is based on the idea that at most M sources are non-zero valued in an instant, and consequently, M columns can be selected from the matrix $\hat{\mathbf{A}}(k)$. Since M columns are selected, this new $\hat{\mathbf{A}}'(k)$ is square and its inverse can be calculated. One possibility is to calculate all the possible square matrices $\hat{\mathbf{A}}'(k)$ (brute force method), obtaining its respective results and choosing the best solution. Note that in our cases of study, it is feasible to prove all the $\hat{\mathbf{A}}'(k)$ if the number of possible $\hat{\mathbf{A}}'(k)$ is not too large. This number of square matrices is low whether the difference between S and M is not very large. Specifically, $\hat{\mathbf{A}}'(k)$ are submatrices with M columns of $\hat{\mathbf{A}}(k)$, which in turns contains S columns, then, $\frac{S!}{(M!(S-M)!)}$ combinations are calculated in the aim of assessing the possible solutions. Once, all the possible matrices are calculated, the problem $\hat{\mathbf{A}}(k)\hat{\mathbf{X}}(k,l) = \mathbf{Z}(k,l)$ is reduced to $\hat{\mathbf{A}}'(k)\hat{\mathbf{X}}(k,l) = \mathbf{Z}(k,l)$. The solutions which better accomplishes Equation (2.57) must be chosen.

In the aim of understanding the meaning of Equation (2.57), where the sources are estimated by using l_1-norm minimization, its geometric interpretation is explained with two illustrative examples. In the first example, we will evaluate a problem in which there are three sources ($S = 3$) and there is only a microphone ($M = 1$). Note that it is not a BSS problem strictly speaking since at least two sensors are needed but, it is very representative in order to demonstrate the sparse character of solutions. So in this case, the solution of the problem can be geometrically

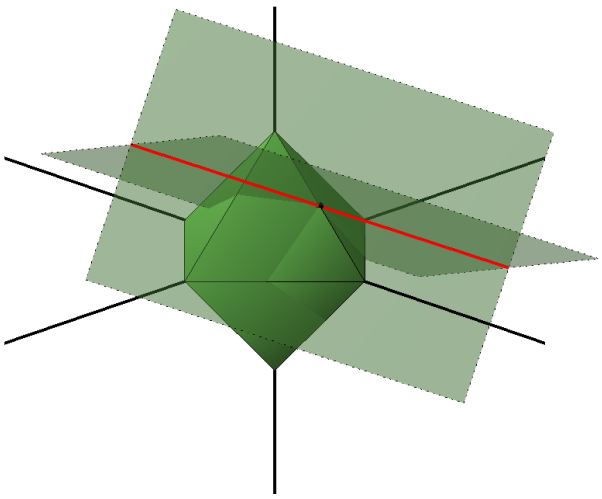

Figure 2.3: An illustrative example of the geometric of the sparse solution via l_1-norm minimization. Underdetermined case of two mixtures ($S = 2$) and three sources ($N = 3$).

interpreted as the point of the subspace defined by $\hat{\mathbf{A}}(k)\hat{\mathbf{X}}(k,l) = \mathbf{Z}(k,l)$ with least l_1-norm (short distance to the origin of coordinates). In this situation, that is, a three dimensional case, $S = 3$, the l_1-ball or cross-polytipe [Ziegler, 1995, Grünbaum, 2003] is the convex hull of an octahedron (see Figure 2.2 for further details). And with respect to the subspace defined by $\hat{\mathbf{A}}(k)\hat{\mathbf{X}}(k,l) = \mathbf{Z}(k,l)$, if only one mixture is available ($M = 1$), only one equation describes the subspace, what it defines a plane in a 3-dimensional space. Or in other words, since $\hat{\mathbf{A}}(k)$ is a 1×3 matrix, there are one equation and three variables. Summarizing, the solution will be the intersection between a l_1-ball of the least possible radius and the subspace defined by equations in $\hat{\mathbf{A}}(k)\hat{\mathbf{X}}(k,l) = \mathbf{Z}(k,l)$. It can be observed in Figure 2.2 that when the octahedron is expanded, the first and minimum solution is reached when the octahedron (minimum octahedron) first touches the plane, corresponding this solution to a vertex of the octahedron. Since the solution corresponds to a vertex, it will be over an axis, and thus, two components of the solution are zero (then, it is sparse). Note that it is coherent with our assumption, since in this case, one component ($M = 1$) of $\hat{\mathbf{X}}(k,l)$ must be non-zero valued and two ($S - M = 3 - 1 = 2$) must be zero valued.

For a more complete geometrical demonstration of how sparse solutions are obtained, another underdetermined case is studied. This case consists of three sources ($S = 3$), but now, the number of presented mixtures is two ($M = 2$). This case is more appropiated and useful for us, since it is a possible situation in the context of BSS. In this problem, $\hat{\mathbf{A}}(k)$ is a matrix whose dimensions are 2×3, and thus, $\hat{\mathbf{A}}(k)\hat{\mathbf{X}}(k,l) = \mathbf{Z}(k,l)$ is a system that has two equations and three variables. Geometrically, it means that the solution is produced when two planes intersect producing a line and therefore, there is an infinite number of possible solutions. Now, the solution of Equation (2.57) is determined by the point where octahedron first touches the line originated by the intersection of the planes.

Having a look at Figure 2.3 it is easy to see the solution of the problem at hand. The solution is located on an edge of the octahedron, for geometrical knowledges, it is known that the points on any edge of the octahedron have a zero valued component what indicates that the solution is sparse. Intuitively, we can predict that the number of zero valued components in the solution is $S - M = 3 - 2 = 1$.

It is important to highlight that Figures 2.2 and 2.3 represent situations in which the solutions are real but in our general case, they are complex, since convolutive mixtures are considered. For this reason, these figures are only considered for illustrative proposes in order to understand why solutions are sparse.

To sum up, it is clear to note that l_1-norm tends to produce a solution $\hat{\mathbf{X}}(k,l)$ with a large number of components equal to zero, what means l_1-norm helps to achieve *sparse solutions* of $\hat{\mathbf{A}}(k)\hat{\mathbf{X}}(k,l) = \mathbf{Z}(k,l)$, with no more than M non-zero valued components.

2.4 Time delay estimation in speech separation problems

One of the classical tool that is commonly used in speech separation related applications is source localization, which has been traditionally solved using TDE techniques. These techniques are included in several algorithms found in the literature with the aim of assisting to the main separation algorithm [Mirzaei et al., 2014], and they are also used in several proposals included in the thesis. For instance, in the next chapter we propose a novel methodology for sound source separation that makes use of these TDE techniques, and in Chapter 4 the use of TDE is studied in the synchronization of mixtures for WASN. This section includes an analysis of TDE, with the aim of introducing and revisiting the generalities of the classical approaches.

2.4.1 Time delay estimation

TDE is used in many applications, such as source identification, source detection and source location which are presented in different research fields as radar, communications, geophysics, ultrasonics and, of course, sound signal processing. According to the type of application, TDE is divided into two categories: time of arrival (TOA) and TDOA estimations. Relevant works in the first category are [Ehrenberg et al., 1978, Tremblay et al., 1987], and examples of the second category are the proposals [Knapp and Carter, 1976, Carter, 1987]. The first category refers to the calculation of the time delay of a signal between its transmission and the reception of its echo. In TDOA the travel time of a signal between separated sensors is obtained. In this second category it is necessary to have at least two sensors, that is, sensor arrays are required. The development of many TDE methods is closely related to the use of sensor arrays. The issue addressed in this chapter of the thesis is framed in the TDOA category, since speech mixtures at the different microphones (sensors) are compared in order to determine the delay differences between them.

Different phenomena can degrade the performance of TDE algorithms. Examples of these phenomena are noise and reverberation which are factors that make more difficult the task of TDE between signals captured at sensors. Both unavoidable phenomena in real scenarios distort the signal captured by the sensors in different ways. Noise introduces new components in the signal and reverberation produces multiple delayed and scaled versions of the source signals. Maybe, to minimize the effects of reverberation on TDE methods is one of the most difficult problem which still remains open. Another difficulty to be considered is that sources can move what varies the delays from time to time. Because delays change, TDE algorithms must be able to update these values. Another problem is directly related to the number of source signals that can be included in signals at sensors. In order to overcome these and other difficulties, there are numerous TDE algorithms that can be classified according to different criteria:

- Adaptive approaches as [Etter and Stearns, 1981, Ching and Chan, 1988] are able to update time delays when they change due to the movement of sources. If TDE algorithms can not update the delay they are known as non-adaptive solutions.

- Depending on the number of source signals that are presented in the signals captured by sensors, two types of algorithms can be distinguished. Single-source TDE algorithms, as the one proposed in [Knapp and Carter, 1976], which work with signals where there is only the contribution of one source or multiple-source TDE algorithms, when there are more than one source signal. A TDE method working with signal mixtures is introduced in [Feder and Weinstein, 1988]. Evidently, the complexity of the algorithms is higher when the number of source signals increases.

- A large number of TDE methods can be found and their characteristics depend on the environment where they estimate delays. To model the environment is of key importance since it helps us to understand a certain problem and to decide which TDE algorithm can be the most appropriated to solve it. These three models are: the single-path propagation model [Knapp and Carter, 1976], the multi-path propagation model [Fuchs, 1999] and the reverberation model [Doclo and Moonen, 2003]. These three models correspond with the three mixing models studied in Section 2.2. The reverberation (echoic) model is very popular in many audio applications since rooms are reverberant. Reflections due to many elements such as walls, floor, furniture or ceiling must be considered. In reverberant environments, source signals will be more degraded when they reach microphones than in the other cases.

There are also other classifications according to various criteria:

- One classification depends on the tools that are used in TDE algorithms. Some of the most representative types are those based on the generalized cross-correlation (GCC) or on the use of high-order-statistics [Wu, 2002]. It must be said that GCC methods are very popular when time delays between speech signals must be calculated.

- Signals can be represented in a certain basis and so, from the relation between signals in this basis, the delay is estimated. According to this basis, three representative classes of TDE methods can be found. First, time-domain approximation methods which work in the time domain using tools as the cross-correlation. Second, frequency-domain approximation methods where delays are estimated in the frequency domain. And finally, Laguerre domain approximation methods where Laguerre functions are used to determine delays between signals.

More classifications can be presented because the number of TDE methods is large. Many of these methods can be considered as classical TDE solutions since they have been widely used in many applications obtaining good outcomes. Moreover in the vast majority of these methods, different types of improvements have been introduced. One example of these improvements is to employ multiple sensors since it provides more information. Other improvements focus on modeling reverberation in order to minimize its effects. Some authors use *a priori* knowledge about the distortion sources to improve the performance of TDE methods. With this prior knowledge, distortion can be estimated and removed from the signals.

2.4.2 Types of TDE algorithms

There is a large variety of TDE algorithms which have been used in many sort of applications. In this section a brief overview of some of the most successful TDE techniques is presented.

1. **Cross-correlation methods**. The earliest TDE algorithms relies on the use of the cross-correlation and were mainly focused on the single-path propagation model. In oder to

obtain a delay between two signals, the cross-correlation of them is calculated and the position of its maximum corresponds with the delay between the signals.

2. **Generalized cross-correlation (GCC) methods**. This type of methods, which were first proposed in [Knapp and Carter, 1976], can be assumed as a follow-up of the cross-correlation method. They have the novelty of introducing weighting functions in the expression of the cross-correlation in the frequency domain. The role of weighting functions is to emphasize the maximum that corresponds with delays between signals. As in the previous case, the delay between two signals is extracted from the maximum of this modified cross-correlation. Some of these algorithms will be explained in a more detailed way in Section 2.4.3.

3. **Methods based on difference functions**. These methods employ difference functions between signals aiming at obtaining delays. Difference functions are calculated between delayed versions of the same signal. The location of the minimum of the difference function indicates the delay in which the similarity of the signals is maximum and therefore, this delay is the time delay between them. An important example of this algorithm is the AMDF method that can be found in the article [Jacovitti and Scarano, 1993].

4. **LMS-type adaptive TDE methods**. One of the most known of this type of methods was implemented in [Reed et al., 1981] and it also works under the assumption of ideal propagation. This proposed method obtains time delays by minimizing the mean square error between a signal and a filtered version of another signal. The used filter is FIR and the lag time associated with the largest component of this filter is the delay between the signals.

5. **Multi-channel cross-correlation method**. This methodology can be considered a generalization of the classical cross-correlation method. Unlike cross-correlation method, it calculates delays using a larger number of channels. Obviously, the complexity of the algorithm increases but it introduces some advantages due to the redundant information provided by the different channels. Some examples of these methods are the proposals [Chen et al., 2003, Benesty et al., 2004].

6. **Fusion algorithm based on multiple sensor pairs**. This methodology also takes advantage of the fact of having more than two sensors. The need to eliminate the effects of reverberation and noise gives rise to this type of algorithms. Examples of this proposal can be found in the works [Kirlin et al., 1981, Nishiura et al., 2000].

7. **Adaptive eigenvalue decomposition algorithm**. The main goal of the adaptive eigenvalue decomposition (AED) algorithm [Huang et al., 1999] is to eliminate the influence of reverberation. This method is not based on the cross-correlation function. The basic idea of this algorithm is to estimate the channel impulse response from sources to sensors and once they have been obtained, direct paths can be identified. The delays are calculated from the difference between direct paths. It has the disadvantage of calculating delays only between two channels.

8. **Adaptive multichannel time-delay estimation**. In the previous method two impulse responses are used to calculate the delay between signals captured by two sensors. But considering reverberant rooms which usually have long impulse responses, this methodology has some associated problems. If impulse responses are long, the probability of having simultaneous zeros or close to zero in both responses is high and so, the AED algorithm

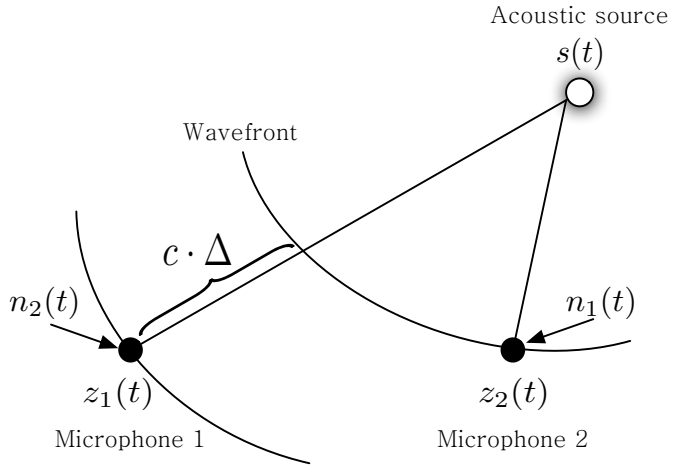

Figure 2.4: Illustrative example of the time delay (Δ) between two microphones (z_1 and z_2). c is the speed of sound in medium (m/s).

can fail. In adaptive multichannel time-delay estimation the idea is to use more channels since the probability of having common zero in all the channels is lower. An example of this algorithm is presented in [Huang and Benesty, 2003].

To summarize, classical algorithms are based on the use of the cross-correlation function. Later, improvements of these classical methods consist in introducing weighting functions in the cross-correlation expression, aiming at solving the problems associated with factors such as noise or reverberation. After these methods, novelties were related to the use of higher number of channels what is linked to the development of sensor arrays. To have more channels allows to use redundant information to increase the accuracy of TDE algorithms. Finally, new methodologies that are not based on the cross-correlation are implemented with the objective of overcoming them.

2.4.3 Study of classical TDE methods

The classical TDE algorithms studied and implemented in this thesis are explained in a detailed way in the paragraphs that follow. It must be said that they have been chosen because they are well-known and robust methods for estimating delays between different type of signals, speech signal being a possible example. We will explore how well these algorithms estimate delays between speech mixtures. For simplicity, in oder to explain these methods, a single speech source $s(t)$ and two signals, $z_1(t)$ and $z_2(t)$, captured by two microphones are supposed, assuming absence of reverberation. The delay between the signals at the microphones will be denoted by Δ. This scenario is depicted in Figure 2.4.

For a pair of separated sensors, the signals captured by the microphones are modeled as

$$\begin{aligned} z_1(t) &= s(t) + n_1(t) \\ z_2(t) &= s(t) + n_2(t), \end{aligned} \tag{2.58}$$

where $n_1(t)$ and $n_2(t)$ are noise terms. Two types of TDE algorithms have been implemented: six algorithms that are within GCC methods and one method which uses a difference function. The main difference between these six algorithms is the weighting function that is introduced in the expression of the cross-correlation to make easier the search of the maximum.

2.4.3.1 Cross-Correlation (CC) method

The cross-correlation between the two signals recorded by microphones is calculated in the first method. The expression of the cross-correlation is shown in Equation (2.59)

$$r_{12}(\tau) = E\left[z_1(t)z_2(t-\tau)\right] = \int_{-\infty}^{\infty} z_1(t)z_2(t+\tau)dt. \tag{2.59}$$

It is well-known that the delay between both signals can be obtained from the position of the maximum peak of the cross-correlation as it is explained in [Carter, 1987]. This maximum value can be obtained as follows

$$\Delta = \arg\underset{\tau}{max}\left[r_{12}(\tau)\right]. \tag{2.60}$$

2.4.3.2 Phase Transform (PHAT) method

This algorithm has been chosen since it has been widely used for estimating delays between acoustic signals arriving at spatially distributed microphones. PHAT method can be classified into the group of GCC methods, or in other words, a weighting function (ψ) is introduced in the expression of the cross-correlation. Equation (2.59) can be expressed in the frequency domain as

$$r_{12}(\tau) = \int_{-\infty}^{\infty} G_{z_1z_2}(f)e^{j2\pi f\tau}df, \tag{2.61}$$

where $G_{z_1z_2}(f)$ labels the cross-spectrum of the received signals. In GCC methods a weighting function is introduced in Equation (2.61), resulting

$$r_{12}(\tau) = \int_{-\infty}^{\infty} \psi(f)G_{z_1z_2}(f)e^{j2\pi f\tau}df. \tag{2.62}$$

In this algorithm, the weighting function is

$$\psi(f) = \frac{1}{|G_{z_1z_2}(f)|}. \tag{2.63}$$

This new weighting function can be very useful since it aims to sharpen the peaks of the cross-correlation by means of whitening the input mixtures, making easier to find the location of the maximum peak. Having a look at Equations (2.62) and (2.63), it seems clear to note that the information related to phase is preserved. The reader can note that in the case of the CC method, the weighting function is $\psi(f) = 1$.

2.4.3.3 Modified Phase Transform (PHAT-β) method

A modified version of the previous method is introduced in [Donohue et al., 2007]. It has been shown that this new proposal provides very good results when delays between signals corrupted by both independent noise and reverberation effects are calculated. The weighting function is very similar to the used one in the PHAT algorithm, but in this case a new parameter (γ) is introduced. The expression of this new weighting function can be observed in Equation (2.64)

$$\psi(f) = \frac{1}{|G_{z_1 z_2}(f)|^{\gamma}}. \tag{2.64}$$

This parameter allows us to control the degree of whitening and limit the amount of degradation from the independent noise. Please note that γ is a real number ranging from 0 to 1. If γ is equal to 0, the algorithm is equivalent to the CC method and if γ is set to be 1, the algorithm is equivalent to the PHAT method. In the case of intermediate values, a process of partial whitening occurs.

2.4.3.4 Maximum Likelihood (ML) method

The ML-based method [Saarnisaari, 1996] is also included within GCC methods and has been selected since it works in systems where multi-path effects should be considered. It tends to obtain maximum likelihood solutions for TDE problems. The weighting function of this method is shown in Equation (2.65)

$$\psi(f) = \frac{1}{|G_{z_1 z_2}(f)|} \frac{|\gamma_{z_1 z_2}(f)|^2}{1 - |\gamma_{z_1 z_2}(f)|^2}, \tag{2.65}$$

where $|\gamma_{z_1 z_2}(f)|^2$ is the magnitude squared coherency and it responds to Equation (2.66)

$$|\gamma_{z_1 z_2}(f)|^2 = \frac{|G_{z_1 z_2}(f)|^2}{G_{z_1 z_1}(f) \cdot G_{z_2 z_2}(f)}. \tag{2.66}$$

The ML function aims at increasing the accuracy of the calculation of the delay. It can be observed that the larger weight is assigned to frequency bands that give near-unity coherence. In the same line of reasoning as that in the previous methods, the maximum of the cross-correlation must be computed.

2.4.3.5 Roth Processor (ROTH)

This method [Roth, 1971] has been chosen since it has been proven to be very efficient in scenarios where additive noise is presented as it is discussed in [Knapp and Carter, 1976]. It suppresses frequency regions where noise is clearly presented. Within this algorithm, the weighting function has been found to be as follows:

$$\psi(f) = \frac{1}{G_{z_1 z_1}(f)}. \tag{2.67}$$

2.4.3.6 Smoothed Coherence Transform (SCOT)

The SCOT method [Carter et al., 1973] has been used in many TDE applications in which the presence of noise is important. In this case, the expression of the weighting function is as indicated in Equation (2.68)

$$\psi(f) = \frac{1}{\sqrt{G_{z_1 z_1}(f) \cdot G_{z_2 z_2}(f)}}. \tag{2.68}$$

It can be considered as a pre-whitening filter followed by a process of cross-correlation. Having a look at Equation (2.67), it seems clear to note that if $G_{z_1 z_1}(f) = G_{z_2 z_2}(f)$, the SCOT method is equivalent to the ROTH algorithm.

2.4.3.7 Average Square Difference Function (ASDF) method

The ASDF method [Jacovitti and Scarano, 1993] does not belong to GCC methods thus instead of using the cross-correlation function, it uses a difference function what involves lower usage of computational load. It can be observed that multiplications are not needed. This difference function is the square error between the signals as shown in Equation (2.69)

$$Di(\tau) = \frac{1}{T} \sum_{t=0}^{T-1} \left| z_1(t) - z_2(t-\tau) \right|^2 . \tag{2.69}$$

By searching the minimum of the previous function, the delay between the signals is determined from its corresponding τ. In a more mathematical way, it can be written as follows:

$$\Delta = \arg\min_{\tau} \left[Di(\tau) \right]. \tag{2.70}$$

2.4.4 Preliminary study of the implemented TDE algorithms

Our objective is to determine by means of comparative studies which TDE algorithm is the most suitable for estimating delays between speech mixtures in reverberant rooms. During these studies, it was taken into account that generally, these alignment methods are not designed to align mixture signals but single signals (only one source) with at most noise components. From different studies as it can be observed in our works [Llerena Aguilar et al., 2012, Llerena et al., 2013a, Llerena et al., 2013b, Llerena-Aguilar et al., 2015], it was concluded that the GCC-PHAT algorithm is the one that obtains the best results in reverberant environments. A good performance of the TDE algorithm in the presence of reverberation is of key importance since in real rooms this phenomenon is presented. For this reason the PHAT algorithm will be chosen as reference in Chapter 4 when new TDE methods would be implemented.

It must be mentioned that since our speech separation methods work in the time-frequency domain, a modified version of the PHAT method expressed by Equation (2.62) has been implemented, specifically, the weighted cross-correlation between two signals has been developed in the time-frequency domain. We denote this version by STFT-based GCC-PHAT in which the time correlation $r_{mp}[\tau]$ (in discrete time) is determined using Equation (2.71), which uses L frames of $Z_m(k,l)$ and $Z_p(k,l)$, which are the STFTs of the m-th and p-th mixture signals.

$$r_{mp}[\tau] \simeq \mathcal{F}^{-1} \left(\sum_{l=1}^{L} \frac{Z_m(k,l)}{|Z_m(k,l)|} \frac{Z_p^*(k,l)}{|Z_p(k,l)|} \right), \tag{2.71}$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform. This version of the PHAT method was first introduced in [Mandel, 2010].

2.5 Evaluation of speech separation methods

In this section, the material and methodologies used to evaluate the speech separation solutions implemented in this thesis are presented. First, in order to quantify the quality of the separated speech signals, different objective measurements are defined. Some of these objective measurements are widely used signal quality parameters, and others are more specific for determining speech intelligibility. Aiming at evaluating the proposals of this thesis, the use of a database of speech signals is required, and in this sense, the one used is described. This database is combined with a simulation program that allows us to generate speech mixtures in a controlled way. The main characteristics of this simulation program are also explained in this section. One limitation of the simulation program is to consider ideal microphone responses and aiming at overcoming it, real microphone responses have been introduced. The most important aspects of this type of real microphones and how their real responses have been characterized will also be presented.

2.5.1 Evaluation of the proposed separation solutions

In this thesis two main approaches within the framework of speech separation are proposed: 1) new speech separation algorithms and 2) alignment methods for speech mixtures when WASNs are used. Considering it, different objective measurements have been implemented. With respect to the evaluation of speech separation algorithms, measurements of speech quality and speech intelligibility have been calculated. Some of them are standard signal quality parameters and others focus on determining speech intelligibility. Concerning alignment methods, they are based on calculating delays between speech mixtures. The errors between theoretical and estimated delays that these alignment methods obtain have been analyzed.

2.5.1.1 Signal-to-interference ratio (SIR)

SIR is a measure very used in a large number of signal processing applications. This measure represents the ratio between the power of the desired signal and the power of the interference signals. In this work, this parameter evaluates the quality of our proposed speech separation methods by means of measuring the similarity between the original speech sources and the separated ones. SIR in dB is expressed as

$$\text{SIR (dB)} = 10\log_{10}\left(\frac{||s_{target}||^2}{||e_{interf}||^2}\right). \tag{2.72}$$

Considering the work in [Yilmaz and Rickard, 2004] and Equation (2.30), one estimated speech source can be extracted from the m-th mixture as Equation (2.73) indicates

$$\hat{X}_s(k,l) = M_s(k,l)Z_m(k,l). \tag{2.73}$$

And so, the SIR measure can indicate how the mask suppresses the interfering signals for the s-th speech source, what is given by

$$SIR_s = \frac{||M_s(k,l)X_s(k,l)||^2}{||M_s(k,l)I_s(k,l)||^2}, \tag{2.74}$$

where $I_s(k,l)$ is the STFT of the signal interfering with the s-th source.

2.5.1.2 Short-time objective intelligibility (STOI)

STOI [Taal et al., 2011] has been used in this thesis to measure the intelligibility of the separated speech sources. It is a very popular function to evaluate speech enhancement algorithms. It is based on a correlation coefficient between the temporal envelopes of the clean and degraded speech signals in short-time. STOI returns a scalar value is expected to have a monotonic relation with the average intelligibility.

STOI has several advantages with respect to other objective parameters for intelligibility prediction of TF-weighted noisy speech, such as the one in [Dau et al., 1996], the coherence speech-intelligibility index (CSII) [Kates and Arehart, 2005] or the frequency-weighted segmental signal-to-noise ratio (FWS) [Hu and Loizou, 2008]. Indeed, STOI shows better correlation with speech intelligibility than these parameters. One important difference between STOI and many of these parameters is that STOI works with short-time segments (386 ms), while the others are based on global statistics across entire sentences.

2.5.1.3 Correct words (%)

In order to measure the relation between intelligibility and the outcomes of STOI, a mapping procedure is used. From this procedure a logistic function is proposed in [Taal et al., 2011]. Considering this work, we propose the Correct words (%) function as

$$\text{Correct words (\%)} = \frac{100}{1 + e^{-17.4906 \cdot STOI + 9.6921}}. \tag{2.75}$$

This measurement scores (in %) the number of words correctly understood. It is important to highlight that the database used in our experiments is the same as the one used to estimate Equation (2.75).

2.5.1.4 Root mean square error (RMSE)

This parameter has been selected since it is a very good error metric for numerical predictions. The RMSE has been calculated to determine the difference between the delays estimated (Δ_m) by our delay estimation methods, and the theoretical delay $\overline{\Delta_m}$. Assuming N_t values of delays obtained, RMSE can be expressed mathematically as

$$RMSE = \sqrt{\frac{1}{N_t} \sum_{j=1}^{N_t} (\Delta_m(j) - \overline{\Delta_m}(j))^2} \tag{2.76}$$

2.5.2 Database

The NOIZEUS database [Hu and Loizou, 2007] is a noisy speech corpus employed to evaluate and compare speech enhancement algorithms. It contains thirty IEEE sentences presented in [Rothauser et al., 1969] and generated by six speakers (three males and three females), which were recorded in a sound-proof booth using Tucker Davis Technologies (TDT) recording equipment. The thirty sentences are phonetically-balanced with relatively low word-context predictability. They include all the phonemes in the American English language. The sentences were originally sampled at 25 kHz and after, downsampled to 8 kHz. They were stored in a wave format of 16 bits PCM.

2.5.3 Simulation of room acoustics

In this thesis different algorithms have been proposed to solve the cocktail party problem. The feasibility of our proposals has been studied using a room impulse generator (RIRG), which is based on the image method [Allen and Berkley, 1979]. The RIRG allows us to generate the acoustic impulse response between one sound source and one microphone inside a room. Once responses have been calculated, the software provides information about which sound components are incident at the position of the microphones, that is, combining the impulse responses generated by our RIRG and the speech signals in our database, echoic speech mixtures can be generated anywhere in the room. As mentioned in Section 2.2.3, echoic mixtures contain attenuated and delayed versions of the original sources due to reverberation, and this phenomenon is simulated by means of the image method.

The wave equation rules the propagation of waves through fluids (liquid or gas) in physics. Sound propagation can be described by the wave equation and so, to obtain the impulse response between two points entails to solve this wave equation. Due to the complexity of the wave equation of sound, its solution can be approximated by several ways. Different methods model this solution and they can be classified into three main categories:

1. *Wave-based models.* In general terms, these methods obtain approximated solutions of the wave equation due to its high complexity. Examples of these methods are the finite element method (FEM) proposed in [Kleiner et al., 1993] or the boundary element method (BEM) [Pietrzyk, 1998]. These solutions require important computational resources what limits their use in real-applications, unless some simplifications would be introduced. The main difficulty of these methods is to define the geometrical description of the objects and the boundary conditions.

2. *Statistical models.* This type of models is very common in aerospace or automotive industry. With respect to the simulation of auralization, they do not model the temporal behavior of a sound field and so, they are not appropriate for simulating it.

3. *Ray-based methods.* This methodology is very used in room acoustics, well-known examples are ray-tracing [Kulowski, 1985] or the above-mentioned image method. The way reflection paths are obtained is the main difference between these two methods. Ray-tracing methods model the sound power emitted by a sound source as a finite number of rays. Then, the impulse response is obtained when all the rays are processed. The image method is the one used by our RIRG for simulating reverberation in a room. The image method is widely used in the acoustic signal processing community. This method calculates a finite impulse response that models the acoustic channel between each pair of sound source and sensor in rectangular rooms. Its main characteristic is that for simulating reverberation it introduces virtual sources also known as images of the sources.

In order to test the proposals of this thesis, different rooms have been simulated with the RIRG. This RIRG allows us to vary several parameters of the simulation:

- The positions (coordinates) of microphones and talkers. For instance, in Chapter 4 synchronization problems associated with the use of WASNs are dealt. In that chapter, microphones have been placed in the superior corners of the room, that is, with important distances between them. On the other hand, classical microphone array configurations in which microphones are separated by only a few centimeters are used in Chapter 3. In both chapters, talkers have been placed randomly in any place of the room between 1.4 and 2.1 high.

- Aiming at evaluating the different proposals of the thesis, the room dimension has been randomly varied from 15 $m \times 15\ m \times 3\ m$ to 20 $m \times 20\ m \times 5\ m$.

- The reverberation conditions of the room, that is, the reverberation coefficient of the walls. Sometimes, instead of talking about the reverberation coefficient, the reverberation time RT_{60} (defined as the time required for reflections of a direct sound to decay 60 dB) is used. A formula, proposed in [Pierce et al., 1991], relates the reflection coefficient and RT_{60} as follows

$$RT_{60} = \frac{24ln(10)V}{c\sum_{i=1}^{6} Sup_i(1 - Cr_i^2)}, \tag{2.77}$$

  where V is the volume of the room, Cr_i is the reflection coefficient and Sup_i is the surface of the i-th wall.

  In this thesis RT60 has been estimated using the impulse responses between the sources and the microphones, and making a log-linear interpolation of the response from the point with an attenuation of 5 dB to the point with an attenuation of 35 dB.

  For each room, three reverberation conditions have been tested: a) absence of reverberation, that is, a reflection coefficient (Cr) equal to 0, b) low reberveration ($Cr = 0.3$) and c) high reverberation ($Cr = 0.6$). For these values of the reflection coefficient, RT_{60} has been calculated, obtaining average values of 0 ms, 334 ms and 498 ms with a standard deviation of 0 ms, 30 ms and 23 ms for each scenario, respectively. With respect to noise, a background noise of 40 dB has been introduced.

- Regarding microphones, their steering directions, the main lobe to back lobe ratio, and a parameter that configures the shape of the microphone response for different frequencies can be configured.

Different modifications have been introduced in the image method to introduce the directivity pattern of real microphones. The main characteristics of these microphones used are presented in the next sections.

2.5.4 Microphone directivity

The classical speech separation algorithms implemented in this thesis have been tested with different types of microphones. One classification of microphones depends on its directionality, also known as polar pattern. The signal attenuation $D(\theta, k)$ of one microphone can be expressed as

$$D(\theta, k) = R(k)|s(k)\cos(\theta) + 1 - s(k)| + G(k)\ \text{(dB)}, \tag{2.78}$$

where θ denotes the angle between the steering direction of the microphone and the direction of arrival of the incident wave, and k is the frequency index. The parameter $R(k)$ is the difference of gain (in dB) between the direction of maximum sensitivity and the direction of minimum sensitivity. Finally, parameter $G(k)$ indicates the intensity registered in the steering direction (in dB).

Then, modifying the parameter $s(k)$, the polar pattern is controlled. In Table 2.1 some of the most common polar patterns and their corresponding values of $s(k)$ are mentioned.

Directivity pattern	$s(k)$
Bidirectional (Dipole)	1
Hypercardioid	0.75
Cardioid	0.5
Subcardioid	0.25
Omnidirectional	0

Table 2.1: Values of s for the different polar patterns.

In oder to perform our simulations, four types of microphones have been considered: omnidirectional, bidirectional, cardioid and hypercardioid. For illustrative purposes, their polar patterns are depicted in Figure 2.5.

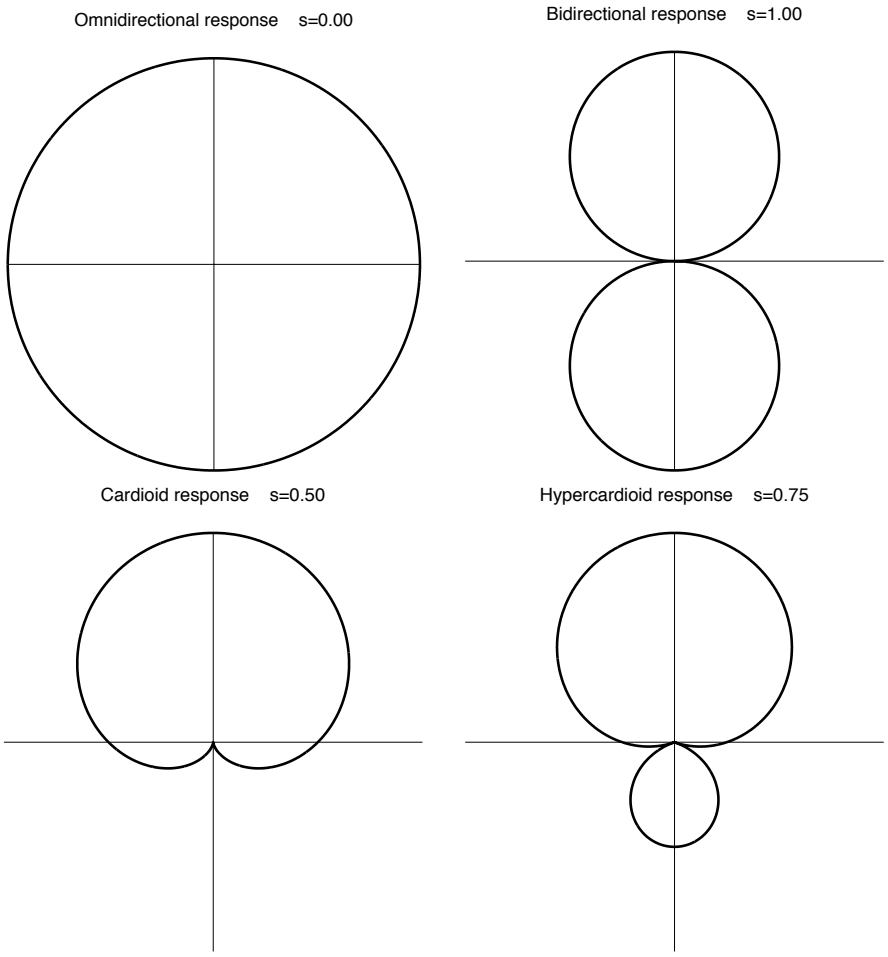

Figure 2.5: Prototype microphone responses.

The front-to-back ratio (FBR) is an interesting parameter in the aim of describing the directivity of some types of microphones, for instance, hypercadioid microphones. It indicates the gain of microphone for signal propagation to the front of the microphone relative to the rear portion.

It seems evident that these microphone responses are ideal and significantly different from the ones that can be found in real experiments. Bearing it in mind, real microphone responses have also been included in our work, with the objective of implementing suitable separation methods

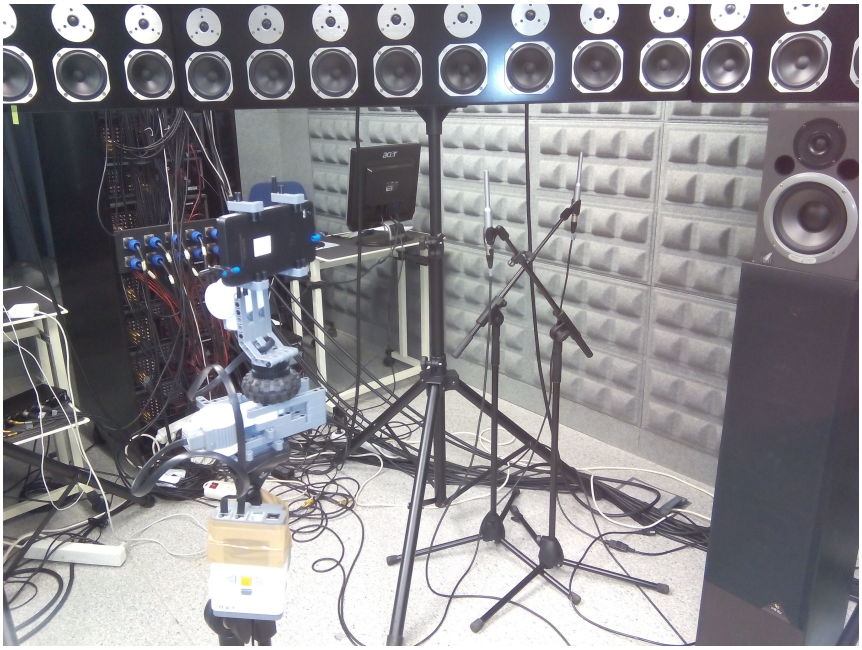

Figure 2.6: Anechoic chamber of the Audio and Communications Signal Processing Group (GTAC), Institute of Telecommunications and Multimedia Applications (iTEAM), Universitat Politecnica de Valencia (UPV), where the mems frequency response estimation has been carried out.

for real separation problems. In the next section the real microphones used in this thesis are described.

2.5.5 Microelectromechanical (MEMS) microphones

MEMS technology is a miniaturization technology where structures below one micron are designed. There are many MEMS systems that can be divided into two groups, actuators and sensors. Within sensor systems, microphones are an important example. A large number of MEMS microphones, with very different implementations, can be found on the market. Nowadays, more than four billions of MEMS microphones are shipped since they have demonstrated to be very useful for sound signal acquisition. The basic operation of many of these microphones is to use the microphone membrane as a capacitor at constant charge. Air pressure changes modify the capacity, which results in voltage changes. Due to their wide use in many devices, such as smartphones, tables, ultrabooks, laptops or headsets, MEMS microphones have been chosen for our speech separation applications. The use of real microphone responses will be vey useful in oder to implement solutions for real cocktail party problems.

The two microphones of the *BQ Aquaris E4.5* mobile phone have been studied. This mobile phone has two MEMS microphone: one is used for audio acquisition and the second one is used to develop noise cancellation. The study of their responses has been made in an anechoic chamber that is depicted in Figure 2.6. This room has been generously provided by the Audio and Communications Signal Processing Group (GTAC), which belongs to the Institute of Telecommunications and Multimedia Applications (iTEAM) of the Universitat Politecnica de Valencia (UPV). To identify the directional properties of these microphones, a fixed sound source emits different specific frequencies and the microphones rotates in front of that sound source.

For illustrative purposes, Figure (2.7) represents an example of the microphone response and its estimation using the model described by Equation (2.78), measured in the anechoic chamber

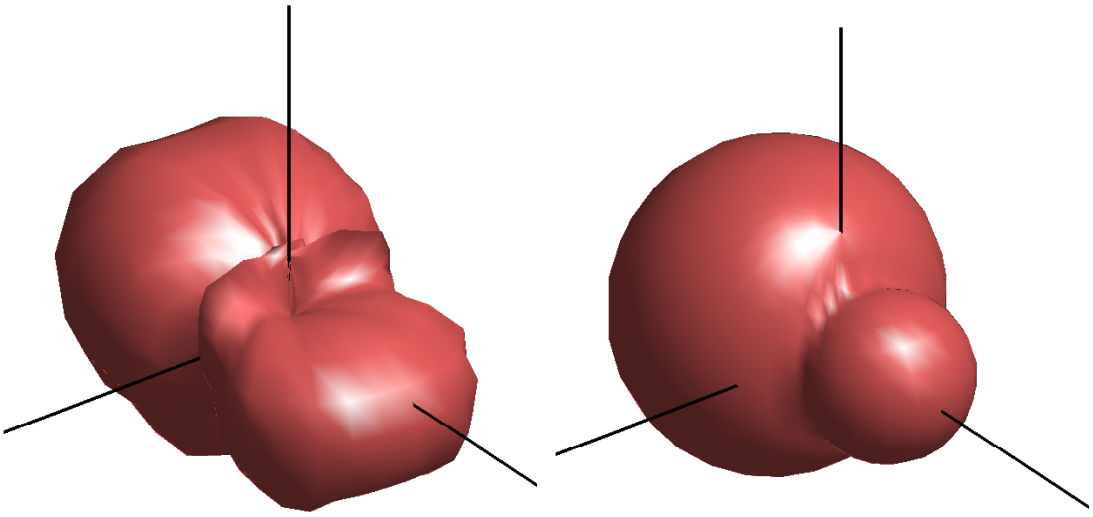

Figure 2.7: Examples of the MEMS microphone response (left) and its estimation (right) for a frequency of 2 kHz.

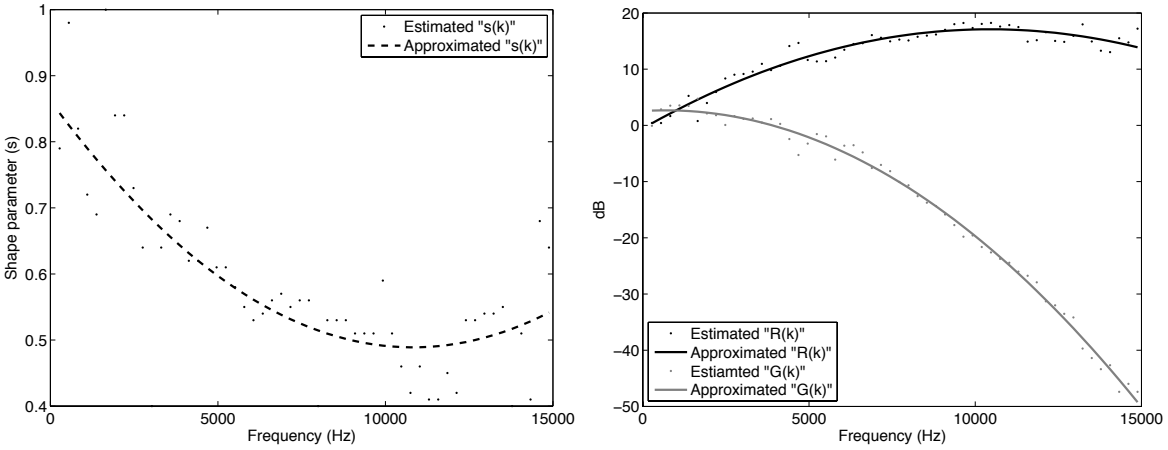

Figure 2.8: Estimation of the parameters s, r and G of the real MEMS microphones along the frequency (dots) and their proposed approximations (lines).

for a frequency of 2 kHz. Looking carefully, it can be said that the microphone response for that frequency can be classified as hypercardioid since they pick up a small amount of sound from directly behind.

It is worth mentioning that in general, microphones have a wide frequency range (from 20 Hz to 20 KHz). It means that the polar pattern will vary (more or less) for different frequencies. Looking at Equation (2.78), it can be observed that parameters $s(k)$, $r(k)$ and $G(k)$ depend on frequency and so, if we want to introduce these real microphone responses in our simulations, their values for the different frequencies must be known. Considering it, Figure 2.8 shows the values of these parameters obtained experimentally in the anechoic chamber (dots in the figure).

In order to determine the values of the parameters $s(k)$, $G(k)$ and $R(k)$ that best match the performance of the real MEMS, we first have look for the values for each frequency that minimize the difference between the measured response and the response generated by Equation (2.78). These values are depicted as dots in figure 2.8. After that, polynomial approximation of

these values have been carried out, for making smoother the estimations. These approximations are also depicted in Figure 2.8, and they havee been used to model the MEMS response in the simulations carried out in the thesis.

Part II

Proposed methods and results

Chapter 3

Design of speech separation algorithms for classical microphone arrays

3.1 Introduction

After doing an extensive literature review of BSS methods, separation algorithms based on sparsity have been used as reference. This type of methods is chosen since they have received a great deal of attention in recent years and have demonstrated their usefulness in a large number of applications, being speech source separation a representative example of it. The main advantage of these methods is to tackle underdetermined source separation issues, this aspect being really important since the underdetermined case still remains an open problem. Although it is true that some limitations appear within speech separation tasks. As mentioned in Section 2.3.3, speech signals in the time-frequency domain can only be considered as quasi-sparse due to their characteristics, making more difficult their separation. Moreover in rooms, reverberation gives rise to attenuated and delayed versions of the original speech sources at the different microphones, what produces many time-frequency samples with contributions of more than one speech source. It makes us wonder if the assumption of sparsity is still valid or not in the presence of reverberation. Indeed, the vast majority of these separation algorithms do not perform correctly in medium- and high-reverberant environments. Another important issue in most of these algorithms is musical noise artifacts, which are associated with time-frequency masking techniques. These aspects, among others, led us to develop new speech separation algorithms.

Broadly speaking, sound source separation is dealt with three types of methodologies: beamforming, CASA and BSS techniques. Perhaps aiming at solving the cocktail party problem, the most popular ones are beamforming and BSS techniques. Beamforming techniques approach the problem from a spatial point of view, that is, spatial filtering based on microphone arrays is used to extract signals from some specific directions and to reduce the contamination of signals from the rest of them. These techniques require to know the characteristics of the sensor array and the positions of the sound sources. On the other hand BSS algorithms work making some assumptions about the nature of the sources without any *a priori* knowledge on these mixtures. It means that there is not any information about the sources, the environment in which mixtures are generated or the sensor array. Focusing on the cocktail party problem, it seems reasonable to think that some information about the microphone array can be available. If a microphone array is used in a room for solving the cocktail party problem, the number and type of microphones, their locations and their orientations may be known in advance. Taking advantage of this information that can be easily available, a new semi-blind separation technique is proposed in this

chapter, concretely, a novel estimation procedure for estimating the mixing matrix is introduced. As it was mentioned in Chapter 2, the elements of the mixing matrix consist of magnitude and phase components which correspond to level and time differences. Bearing it in mind, a novel way of estimating the mixing matrix based on the development of a geometric model is presented. The main aspects of this proposal will be explained in a more detailed way in Section 3.3. It must be mentioned that we aim to develop a separation technique able to work with very simple sensor arrays, in fact, results will be presented for the simplest one, the two-microphone array.

In principle our proposal focuses on the mixing matrix estimation stage but in respect of the separation stage, a classical procedure that relies on time-frequency masking will be used. Time-frequency masking techniques show good results when speech separation is carried out, but they have some associated problems. One of the most important is the presence of musical noise artifacts. Aiming at minimizing the presence of those artifacts, a simple but effective solution will be proposed.

3.2 Experimental setup

In order to demonstrate the suitability of our proposals, we must compare with classical mixing matrix estimation methods. In this sense, an extensive literature review has been done aiming at choosing appropriate classical speech separation algorithms that will be use as reference. The implemented classical methods, which are based on sparsity, are explained in Section 2.3. They have been selected because they are able to solve the underdetermined source separation problem and have demonstrated to achieve acceptable outcomes in the presence of noise and reverberation effects.

With the goal of carrying out a valid comparison between the chosen classical methods and our proposal, a study of the behavior of the classical algorithms has been previously conducted. This study has two parts:

1. In the first part the objective is to find for each classical separation algorithm the two-microphone array (type of microphones, microphone separation and mutual angle) and the frame length with which the best speech separation is performed.

2. Once the best configuration has been selected for each separation algorithm, both classical algorithms are compared and the one that achieves better estimation of the separated sources is selected to be compared with our proposal.

The experiments with which the classical algorithms have been tested are explained in Sections 3.2.1 and 3.2.2. Additionally, in the aim of quantifying how the classical algorithms work, the quality of the speech mixtures before applying the separation algorithms is presented in Section 3.2.3. These quality values will be compared with the ones obtained by the classical methods. The study of both traditional methods is carried out in Sections 3.2.4 and 3.2.5.

3.2.1 Reverberant rooms

In order to analyze the classical separation methods, different experiments have been carried out. The speech signals used in the experiments have been extracted from the IEEE database [Hu and Loizou, 2007] at a sampling frequency $f_s = 8000$ Hz. At most, four seconds of the speech signal have been taken into account in the separation process. Both the synchronization and the separation have been implemented using information from four-second time recordings. These signals have been simulated at 60 dB in 45 different experiments in which two or three

speech sources ($S = 2, 3$) have been located (in different locations) in rooms of different size and reverberation conditions with a background noise of 40 dB. In this sense, we consider that all of the speakers are placed randomly in any place of the room between 1.4 and 2.1 high. The room dimension has been randomly varied from 15 $m \times 15$ $m \times 3$ m to 20 $m \times 20$ $m \times 5$ m, with different reflection coefficients ($Cr = 0$, $Cr = 0.3$ and $Cr = 0.6$). These three values have been selected in order to consider three different groups of scenarios: one without reverberation ($Cr = 0$), one with low reverberation ($Cr = 0.3$) and one with high reverberation ($Cr = 0.6$). With these values, the reverberation time RT_{60} has been calculated, obtaining average values of 0 ms, 334 ms and 498 ms with a standard deviation of 0 ms, 30 ms and 23ms for each scenario, respectively.

The frame length of the separation algorithm has been varied from 32 to 512 ms in the experiments, in order to determine the most suitable in each case.

3.2.2 Microphone array

The objective of this chapter is to provide a feasible solution for the cocktail party problem, that is, we focus on separating speech source signals when different people are talking simultaneously in a room. The vast majority of speech separation algorithms rely on the use of microphone arrays because to have more than one input channel entails many advantages. It is known that nowadays very sophisticated microphone arrays can be found, many of them to perform beamforming techniques. In contrast to it, we propose the use of a simple (two-microphone) array to perform semi-blind speech separation. The only requirement of our proposal is to know very elemental information about the mentioned array, that is, the number and type (polar pattern) of microphones, their relative orientation and the distance between them. This requirement is realistic and not strong since if we install a microphone array, those characteristics are known. For illustrative purposes in Figure 3.1 is depicted a scheme of the used two-microphone array.

Having a look at Figure 3.1, it can be observed that the array has a par of microphones mounted at a mutual angle of α degrees. The capsules of the microphones are in the same plane (X-Y) spaced by a distance equal to D. The values of α and D are determined according to the study developed in the following sections, where the configuration that achieves the highest quality of the separated sources has been chosen. The sensor array is placed on a wall of the room around 1.5 meters above the floor. We assume the wall (y-axis) through the centers of the polar patterns of the microphones. In this way the speech sources placed around the microphone array are captured with both level and phase differences between the two channels. This array is very simple since there are only two microphones what involves some advantages, such as the amount of information captured to perform speech separation is not very large, the same processor can work with the information captured by the two microphones what avoids synchronization problems, among others.

Two-microphone arrays have been used, and the microphone separation has been varied from 0.05 to 1 m. With respect to the microphones, four different microphones have been tested: omnidirectional (labeled "Omn"), cardioid with 10 dB of FBR (labeled "C10"), cardioid with 20 dB of FBR (labeled "C20") and MEMS microphones (labeled "Mms"). The description of the responses of these microphones is included in Sections 2.5.4 and 2.5.5. The relative orientation between them has also been varied, valuing 0°, 30°, 60°, 90°, 120° and 180°. Note in this point that an angle of 0° means that the microphones are oriented in the same direction perpendicularly to the wall in which the array is placed, and an angle of 180° implies that the microphones are oriented in opposite directions parallel to the wall.

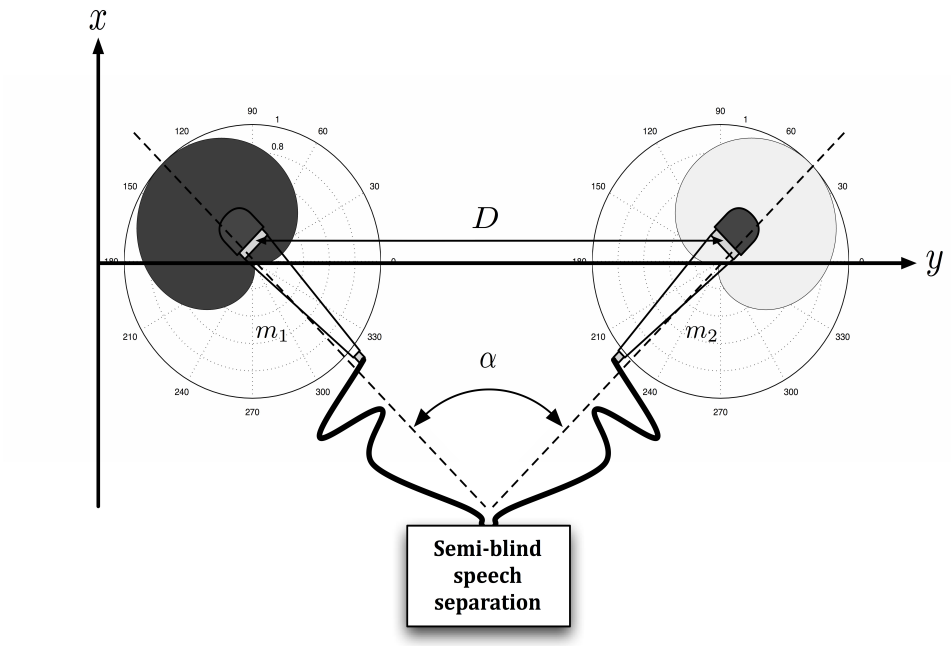

Figure 3.1: Scheme of the clssical two-microphone array used for solving the cocktail party problem dealt in this thesis.

3.2.3 Reference values without separation stage

The evaluation of speech separation algorithms involves to determine the quality of the separated speech sources in terms of different objective measurements. Aiming at quantifying the improvements introduced by the separation algorithms, some reference values must be considered. We have established that those reference values may be the quality of the speech mixtures before applying the separation algorithms. When the quality of the speech mixtures is obtained, it can be compared with the quality of the separated sources and so, it will be clear the differences between the use or non-use of the separation algorithms. Following with this idea, in Figure 3.2 and Tables 3.1 and 3.2 the quality values of the speech mixtures before applying the separation algorithms are shown.

Figure 3.2 presents a study of the influence of distance between microphones on the quality of the speech mixtures. This study has been performed for different values of reverberation (lack of it, medium reverberation and high reverberation) and for two or three speech sources. The evaluated values of distance between microphones are 0.05, 0.1, 0.15, 0.20, 0.25, 0.30, 0.40, 0.50, 0.75 and 1 m. The quality is shown in terms of three objective measurements: SIR, STOI and Correct words (%).

Firstly, it is observed that the quality of the speech mixtures is low in terms of SIR regardless of the number of sources or the distance between microphones, specifically, values lower than 4 dB are obtained in all the cases. These low values are expected since the speech sources have not yet been separated. In respect of the influence of distance, it is clear that SIR decreases when the distance increases until 0.2 m and from it, the quality can be considered more or less constant. Concerning STOI, the influence of the distance between microphones is minimum. Unlike the two mentioned quality parameters, the Correct words (%) measurement increases slightly when the distance between microphones also does, but it must be said that these improvements are lower than 1%.

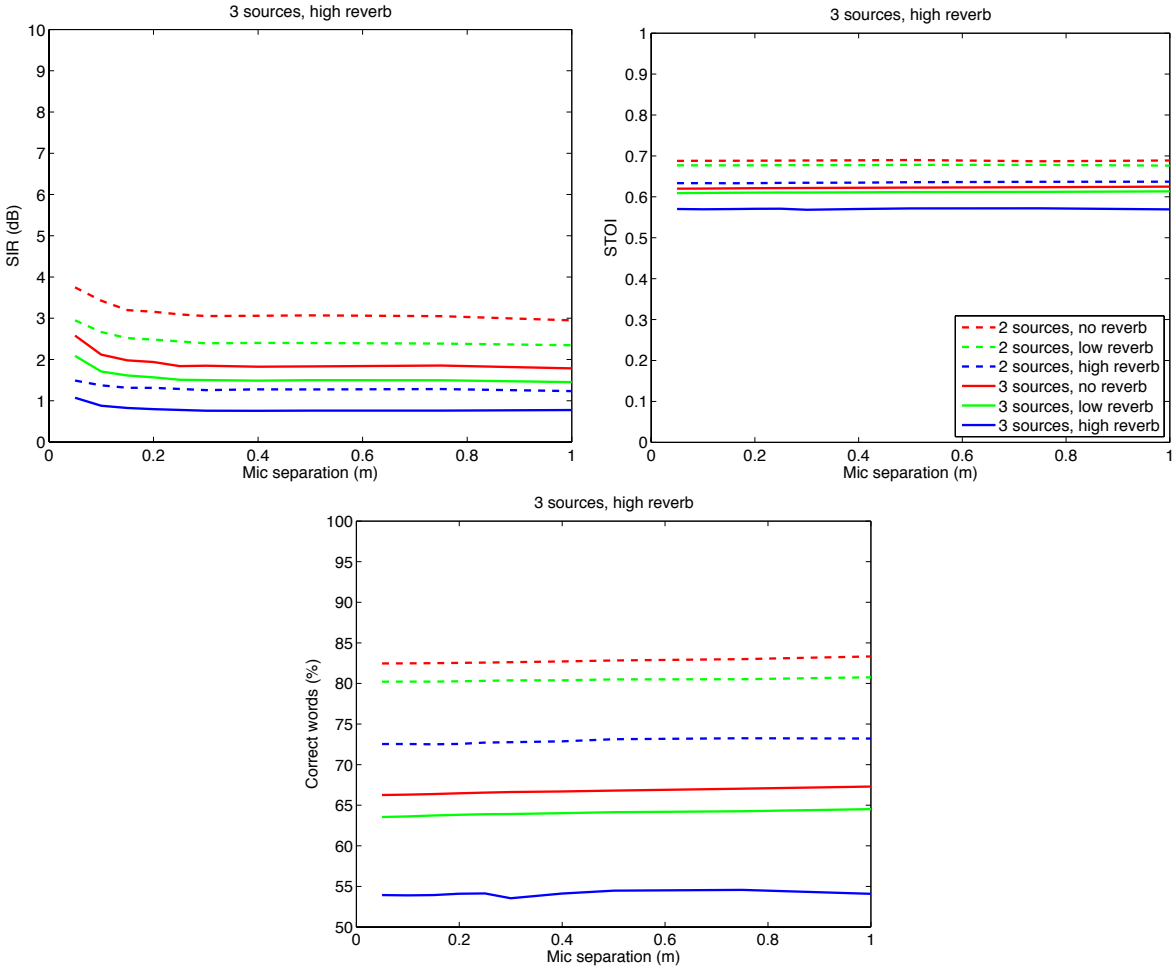

Figure 3.2: SIR (dB), STOI and Correct words (%) before applying the speech separation algorithms.

In Table 3.1 the configurations that achieve the highest quality values are shown. In these configurations it is considered the type of microphones, the angle between their directions of maximum sensitivity (mutual angle) and the microphone separation. Three types of microphones were tested: one omnidirectional, one MEMS microphone and several versions of a cardioid microphone, these versions depend on the FBR used. The results are presented for two or three speech sources in the mixtures and different values of reverberation.

Looking at Table 3.1, different aspects can be highlighted. First, it is clear that the cardioid microphone with a FBR equal to 20 dB (labeled as C20) has demonstrated to be the most suitable one in all the cases. Concerning the mutual angle, an angle $\alpha = 90°$ is the most appropriate in practically all the configurations, but in the case of two speech sources and high reverberation, an angle of 60° provides the best outcomes.

From the point of view of the SIR, the quality is extremely low for all the configurations, what indicates important differences between the original speech sources and the mixtures. This important distortion is expected since speech sources are still mixed and reverberation and noise components degrade the signal quality significantly. Considering speech intelligibility, STOI and Correct words (%) indicate medium intelligibility. From all these results it seems evident that it is necessary to apply speech separation algorithms.

Best Case	2 sources											
	No reverberation				Low reverberation				High reverberation			
D (m)	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75
SIR (dB)	3.7	3.2	3.1	3.1	2.9	2.5	2.4	2.4	1.5	1.3	1.3	1.3
STOI	0.69	0.69	0.69	0.69	0.68	0.68	0.68	0.68	0.63	0.63	0.63	0.64
Words (%)	82.5	82.5	82.7	83.0	80.2	80.3	80.4	80.5	72.5	72.6	72.9	73.2
Angle (α)	90	90	90	60	90	90	90	90	60	60	60	60
Microphones	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20
Best Case	3 sources											
	No reverberation				Low reverberation				High reverberation			
D (m)	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75
SIR (dB)	2.6	1.9	1.8	1.9	2.1	1.6	1.5	1.5	1.1	0.8	0.8	0.8
STOI	0.62	0.62	0.62	0.62	0.61	0.61	0.61	0.61	0.57	0.57	0.57	0.57
Words (%)	66.3	66.5	66.7	67.0	63.5	63.8	64.0	64.3	53.9	54.1	54.1	54.6
Angle (α)	90	90	90	90	90	90	90	90	90	90	90	90
Microphone	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20

Table 3.1: Results obtained for the original experiments

	Angle (α)					
Microphones	180°	120°	90°	60°	30°	0°
Mems (Mms)	11.7 (2.0)	10.0 (1.5)	9.9 (1.7)	10.2 (1.9)	10.8 (2.2)	11.5 (2.4)
Omnidirectional (Omn)	20.1 (3.4)	20.1 (3.4)	20.1 (3.4)	20.1 (3.4)	20.1 (3.4)	20.1 (3.4)
Cardioid 10dB (C10)	10.0 (3.0)	7.7 (1.4)	7.9 (1.2)	9.1 (1.2)	11.5 (1.5)	14.6 (1.7)
Cardioid 20dB (C20)	11.8 (4.6)	3.3 (3.1)	0.2 (0.5)	0.8 (0.7)	4.5 (1.8)	10.9 (2.6)

Table 3.2: Average loss (and Standard Deviation) in the rate of Correct words (%) obtained for the original experiments, in function of the selected microphone and the angle between both microphones.

A comparative study of the different microphones can be found in Table 3.2 . This table shows the average loss (and the standard deviation) in the rate of correct words obtained for different microphones and mutual angles. The first aspect to be considered is that the omnidirectional microphone obtains the worst results (average loss of 20.1 and a standard deviation of 3.4) in all the cases. The MEMS microphone is worse than the cardioid one but it improves the results in comparison with the omnidirectional one. The best outcomes are obtained using the C20 microphone except when the mutual angle is 180°, where the C10 microphone obtains the lowest losses. With an angle of 90° the $C20$ microphone achieves the best result, obtaining an average loss equal to 0.2.

3.2.4 DUET with l_1-norm minimization

In this section the outcomes obtained when the mixing matrix estimation stage of DUET is combined with a separation stage based on l_1-norm minimization are presented. Under the same conditions in Section 3.2.3, it can be observed in Figure 3.3 a study of the influence of the distance between microphones on the quality of the separated speech sources.

A priori, the outcomes for the three quality parameters show that the larger the distance is, the higher the quality is, but it will be discussed in more detail. Considering reverberation, different conclusions can be extracted. Without reverberation, the separation method obtains higher quality (in terms of the three parameters) when the distance between microphones increases. With low-reverberation conditions, the quality increases up to 0.40 m approximately. For larger distances, there is not improvements or even, the quality decreases. Finally, for high-

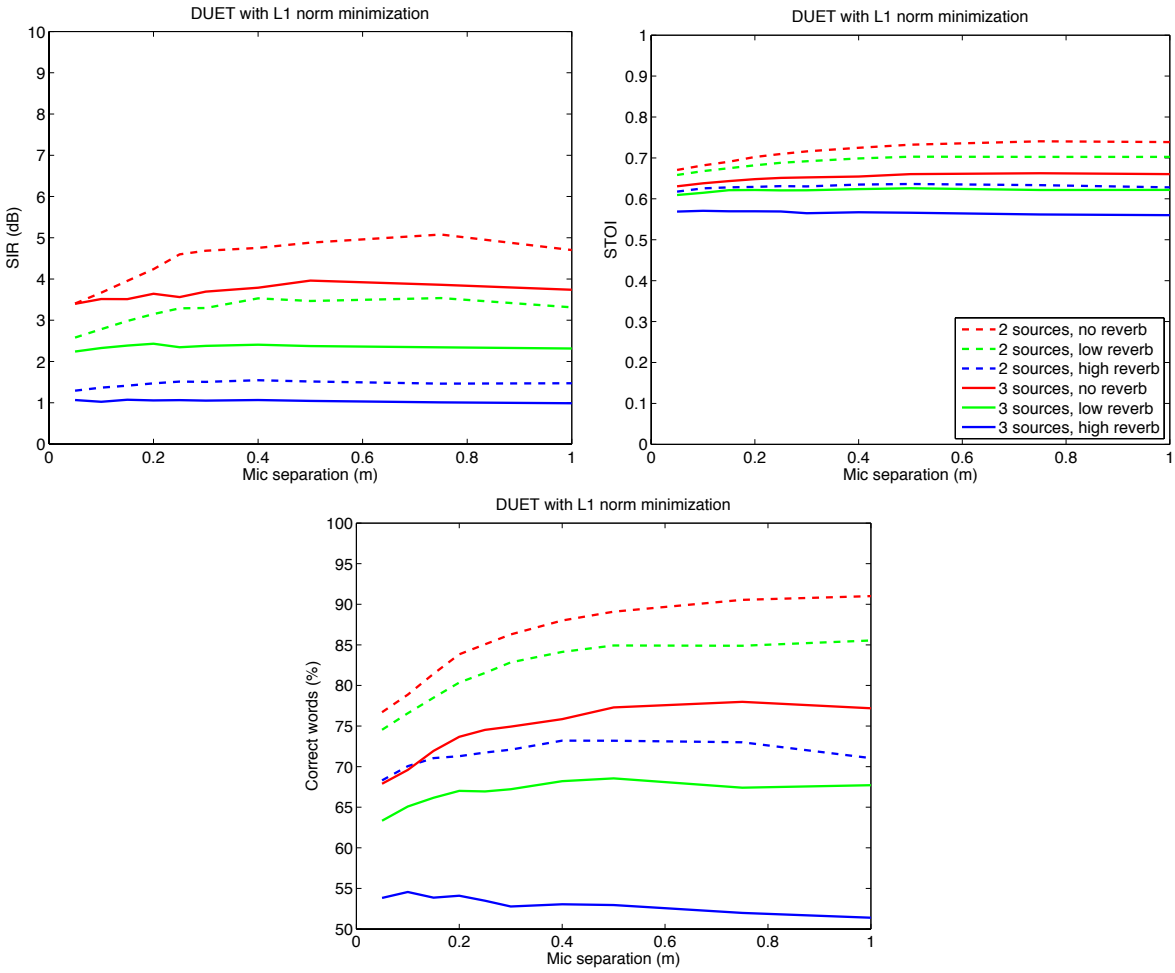

Figure 3.3: SIR (dB), STOI and Correct words (%) for the DUET method with l_1-norm minimization is used.

reverberation it can be observed that the outcomes of the separation algorithm are really poor and so, the effects of distance on quality are irrelevant.

The configurations with which the best speech separation is performed are presented in Table 3.3. This table is similar to Table 3.1 but it includes a new element, the frame length that the separation algorithm employs (labeled "window" in the table). Putting our attention on the table it can be highlighted that the cardioid microphone (C20) has demonstrated to be the most suitable. With mutual angles of 60° or 90° the best results are obtained. Broadly speaking, 90° is the best for the shortest distance (0.05 m) and 60° for the rest them. The values of window obtained (32, 64 and 128 ms) are typical values when there are not large distances between microphones.

Comparing Tables 3.1 and 3.3, we note that DUET with a separation stage based on l_1-norm minimization does not achieve important improvements of the quality of the separated sources. For instance, for the non-reverberant case with two sources, the best SIR is 3.7 (for $D = 0.05$) when there is not separation and 5.1 (for $D = 0.75$) with this separation solution. In terms of STOI the improvement is little since it goes from 0.69 to 0.74. For the rest of cases very similar situations can be found. For this reason, DUET with BM will be studied in the next section.

Other factor to be considered in a separation case is the type of microphone employed. With

Best Case	2 sources											
	No reverberation				Low reverberation				High reverberation			
D (m)	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75
SIR (dB)	3.4	4.2	4.8	5.1	2.6	3.1	3.5	3.5	1.3	1.5	1.5	1.5
STOI	0.67	0.70	0.72	0.74	0.66	0.68	0.70	0.70	0.62	0.63	0.63	0.63
Words (%)	76.7	83.8	88.0	90.5	74.5	80.4	84.1	84.9	68.3	71.3	73.2	73.0
Window (ms)	64	64	64	64	64	64	32	32	128	64	32	32
Angle (α)	90	60	60	60	90	60	60	60	90	60	60	60
Microphones	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20
Best Case	3 sources											
	No reverberation				Low reverberation				High reverberation			
D (m)	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75
SIR (dB)	3.4	3.6	3.8	3.9	2.2	2.4	2.4	2.3	1.1	1.1	1.1	1.0
STOI	0.63	0.65	0.65	0.66	0.61	0.62	0.62	0.62	0.57	0.57	0.57	0.56
Words (%)	67.9	73.7	75.9	78.0	63.3	67.0	68.2	67.4	53.8	54.1	53.0	52.0
Window (ms)	64	64	32	64	128	64	64	64	128	64	64	64
Angle (α)	90	60	60	60	90	60	60	60	60	60	90	60
Microphone	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20

Table 3.3: Best results obtained for the DUET with l_1-norm minimization method.

Microphones	Angle (α)					
	180°	120°	90°	60°	30°	0°
Mems (Mms)	20.8 (3.6)	17.3 (3.5)	17.8 (3.6)	19.4 (3.9)	22.6 (4.2)	27.8 (5.4)
Omnidirectional (Omn)	36.3 (6.9)	36.4 (7.0)	36.4 (6.9)	36.3 (6.9)	36.4 (6.9)	36.4 (6.9)
Cardioid 10dB (C10)	11.8 (3.6)	8.7 (1.8)	8.8 (1.5)	10.8 (2.0)	16.2 (3.2)	30.8 (5.4)
Cardioid 20dB (C20)	14.4 (2.9)	4.9 (2.3)	0.7 (0.6)	0.1 (0.4)	4.3 (2.0)	26.1 (6.1)

Table 3.4: Average loss (and Standard Deviation) in the rate of Correct words (%) obtained for the DUET with l_1-norm minimization method, in function of the selected microphone and the angle between both microphones.

regard to the microphone response, the three types mentioned in the previous section have been tested. In Table 3.4 it is shown the average loss (in terms of Correct words (%)) in function of the selected microphone and the mutual angle. From the table different conclusions can be extracted. First, the omnidirectional microphone obtains the worst results in all the cases (over 36% of loss), what is expected since due to its polar pattern, many reverberant components are picked up. The MEMS microphone achieves poor results (in the best case 17.3% of loss) in comparison with the cardioid one, what is logical since the microphone response is not ideal. Finally, the most suitable microphone is the cardioid when the FBR is 20 dBs. It indicates that it is better a directional microphone. Concerning the mutual angle, 60° is the best value for the C20 microphone.

3.2.5 DUET with Binary Mask

This section includes the results obtained with the DUET algorithm using binary mask in the separation stage. We must observe that the use of binary masking is more beneficial than the use of l_1-norm minimization. As in the case of a separation base on l_1-norm minimization, the influence of the distance between microphones on the quality of the separated sources is depicted in Figure 3.4.

At first glance it is observed that the quality of the estimations is higher than in the case of l_1-norm minimization, what suggests us to use a separation stage based on BM instead of

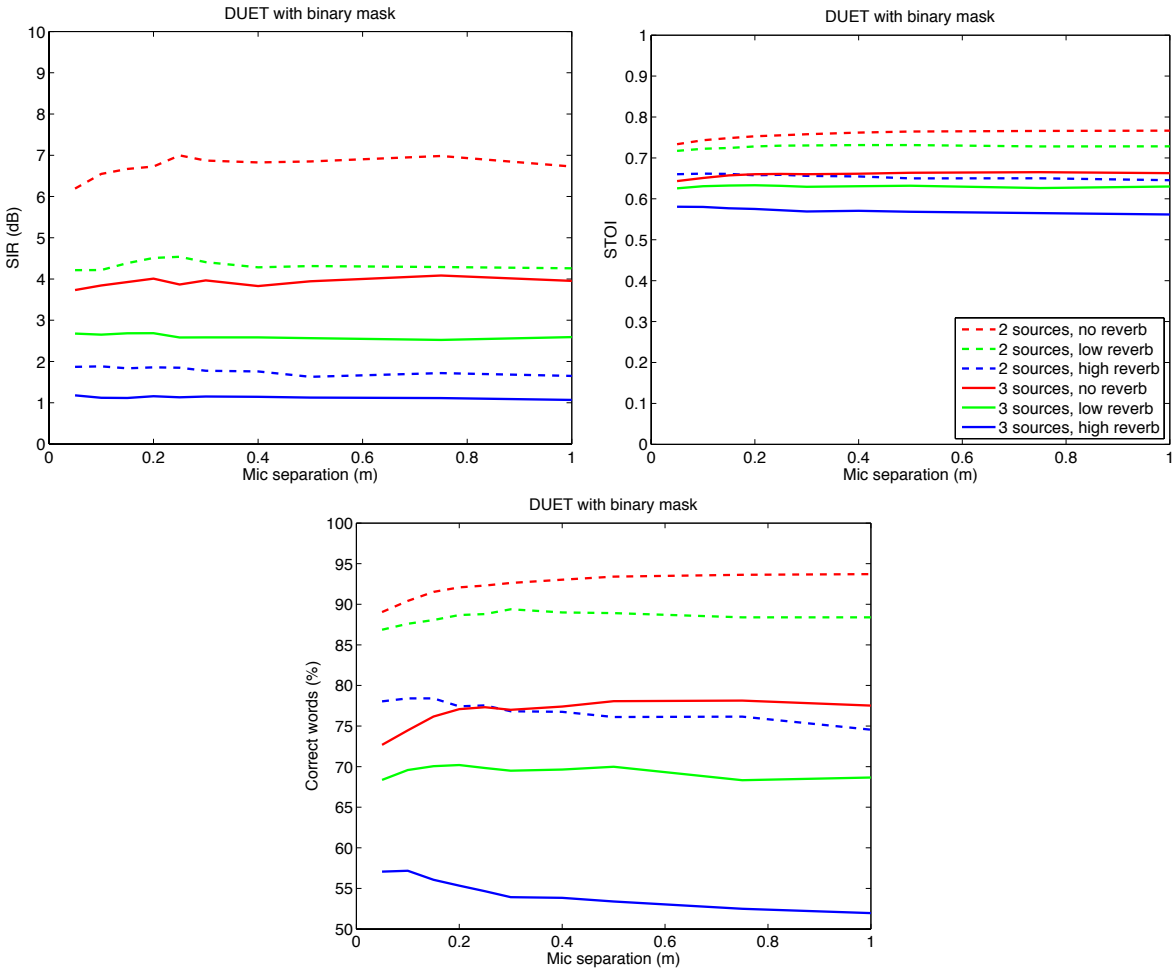

Figure 3.4: SIR (dB), STOI and Correct words (%) for the DUET method with binary mask.

l_1-norm minimization. As in the previous section, the influence of the distance on the quality depends on the reverberation. In the absence of reverberation, according the distance between microphone increases, the quality also does for both number of speech sources. In case of low reverberation, the highest quality is obtained for distances that range between 0.2 m and 0.3 m. For high reverberation conditions, the best separation is performed for a microphone separation over 0.05 m.

The best configurations are presented in Table 3.5. Very similar conclusions to the ones in the case of l_1-norm minimization can the extracted. The most suitable microphone proves to be C20 when the angle between microphones is 60° in almost all cases.

The average losses associated with the different microphones are shown in Table 3.6. As in the previous cases, the microphone that presents worst behavior is the omnidirectional one. The MEMS microphone also obtains poorer quality than the cardioid one in all the cases, achieving the best result (16.1) for $\alpha = 120°$. With respect to the cardioid microphone, the best results occur with $C20$ in all the cases, except for the one in which the microphones are oriented towards opposite sides, that is, $\alpha = 180°$. The best configuration is to use the C20 microphone and $\alpha = 60°$.

Finally, the separation stages based on l_1-norm minimization and BM are compared. In this sense, Table 3.7 contains the highest values obtained by both methods, regardless the configura-

Best Case	2 sources											
	No reverberation				Low reverberation				High reverberation			
D (m)	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75
SIR (dB)	6.2	6.7	6.8	7.0	4.2	4.5	4.3	4.3	1.9	1.9	1.8	1.7
STOI	0.73	0.75	0.76	0.77	0.72	0.73	0.73	0.73	0.66	0.66	0.65	0.65
Words (%)	89.0	92.1	93.0	93.6	86.9	88.7	89.0	88.4	78.0	77.4	76.7	76.2
Window (ms)	64	64	64	64	64	64	64	64	64	64	64	64
Angle (α)	60	60	60	60	60	60	60	60	60	60	60	60
Microphones	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20
Best Case	3 sources											
	No reverberation				Low reverberation				High reverberation			
D (m)	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75
SIR (dB)	3.7	4.0	3.8	4.1	2.7	2.7	2.6	2.5	1.2	1.2	1.1	1.1
STOI	0.64	0.66	0.66	0.67	0.63	0.63	0.63	0.63	0.58	0.58	0.57	0.57
Words (%)	72.7	77.1	77.4	78.1	68.4	70.2	69.6	68.3	57.1	55.3	53.8	52.5
Window (ms)	64	64	64	64	64	64	64	64	128	64	64	64
Angle (α)	60	60	60	60	60	60	60	60	60	60	90	90
Microphone	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20

Table 3.5: Best results obtained for the DUET with binary mask method

	Angle (α)					
Microphones	180°	120°	90°	60°	30°	0°
Mems (Mms)	19.6 (3.8)	16.1 (3.8)	16.6 (4.0)	18.0 (4.1)	20.7 (4.2)	25.3 (4.8)
Omnidirectional (Omn)	33.9 (6.6)	33.9 (6.6)	33.9 (6.7)	33.9 (6.7)	33.9 (6.6)	33.9 (6.6)
Cardioid 10dB (C10)	11.3 (3.2)	7.9 (1.5)	7.9 (1.4)	9.5 (2.0)	14.5 (3.3)	28.1 (4.7)
Cardioid 20dB (C20)	15.3 (2.4)	5.4 (2.5)	1.0 (0.7)	0.0 (0.1)	3.4 (1.8)	23.4 (5.2)

Table 3.6: Average loss (and Standard Deviation) in the rate of Correct words (%) obtained for the DUET with binary mask method, in function of the selected microphone and the angle between both microphones.

Case	No reverberation				Low reverberation				High reverberation			
Best	2 sources		3 sources		2 sources		3 sources		2 sources		3 sources	
D (m)	l_1	BM	l_1	BM	l_1	BM	l_1	BM	l_1	BM	l_1	BM
SIR (dB)	5.1	7.0	3.9	4.1	3.5	4.5	2.4	2.7	1.5	1.9	1.0	1.1
STOI	0.74	0.77	0.66	0.67	0.7	0.73	0.62	0.63	0.63	0.66	0.56	0.58
Words (%)	90.5	93.6	78.0	78.1	84.9	89.0	68.2	70.2	73.2	78.0	52.0	57.1

Table 3.7: Best results obtained for the DUET with l_1-norm and binary mask methods.

tions. It is easy to see that BM outperforms l_1-norm minimization in all the cases. For instance, in the absence of reverberation, DUET with BM obtains values of SIR equal to 7 dB and 4.1 dB for two and three speech sources, respectively. With a separation stage that uses l_1-norm minimization, values of SIR equal to 5.1 dB and 3.9 dB are obtained, which are lower than in the case of using BM.

3.3 A new mixing matrix estimation procedure

In this section the geometric analysis on which our mixing matrix estimation is based will be presented. This geometric model is directly related to the type of microphone array that is used and for this reason, a simple model with two microphones placed in a wall will be used. This

model will allow us to make estimations of the mixing matrix from time delay estimations, that will be used to propose a novel separation method. The quality separation obtained with this method and different TDE algorithms will also be presented to determine the feasibility of our proposal.

3.3.1 Geometric model

In this section the theory underlying our method is explained. The main idea is to obtain the components of the mixing matrix elements (level and time differences) in a new way. We provide a mixing matrix estimation approach based on the geometrical analysis of the problem, considering different information about the microphones, such as the mutual angle between them, the microphone separation and how their polar patterns are. Using this information our proposal consists in determining a geometric relationship between level and time differences that depends on the direction of arrival (DOA) of the speech source. Thus, being able to determine the time difference of each source between the two microphones, their corresponding level differences are obtained automatically by means of using this theoretical relationship. It means that our method only needs to obtain time differences for developing the estimation of the mixing matrix. This aspect is very important, since in general terms to estimate the values of the level differences correctly is very difficult. Factors, such as noise, reverberation effects or the fact that the assumption of sparsity is not always valid, distort their estimation.

From now on, our objective is to find the geometric expressions of both level and time differences and to establish the relationship between them. First the geometric expression of the level differences will be obtained. To determine the expression of the level difference between the two microphones we must consider that it depends on the DOA of the incident wave. Let us assume that both microphones have identical polar pattern which can be expressed as Equation (2.78) indicates. The polar patterns of the microphones vary with frequency, but in order to simplify the notation this dependence will not be represented in the following mathematical development without loss of generality. Then, Equation (2.78) results

$$D(\beta_{mq}) = R\,|s\cos(\beta_{mq}) + 1 - s| + G, \tag{3.1}$$

where β_{mq} is the angle between the direction of maximum sensitivity of the m-th microphone and the DOA of the q-th source. And so, the level difference (in dB) for the q-th source can be calculated as

$$L_q = D(\beta_{2q}) - D(\beta_{1q}) = R\,|s\cos(\beta_{2q}) + 1 - s| + G - R\,|s\cos(\beta_{1q}) + 1 - s| - G, \tag{3.2}$$

and simplifying,

$$L_q = R\left(|s\cos(\beta_{2q}) + 1 - s| - |s\cos(\beta_{1q}) + 1 - s|\right). \tag{3.3}$$

Then, $\beta_{mq},\ m = 1, 2$ must be calculated. Assuming the microphone array placed on a wall (coordinate $X = 0$) and having a look at Figure 3.1, the direction of maximum sensitivity can be expressed in Cartesian coordinates as

$$\mathbf{m}_1 = \left[\cos\left(\frac{\alpha}{2}\right), -\sin\left(\frac{\alpha}{2}\right), 0\right]^T, \tag{3.4}$$

$$\mathbf{m}_2 = \left[\cos\left(\frac{\alpha}{2}\right), \sin\left(\frac{\alpha}{2}\right), 0\right]^T. \tag{3.5}$$

The DOA of a incident sound wave due to the q-th speech source is denoted by $\mathbf{p}_q$ and can be expressed in Cartesian coordinates as

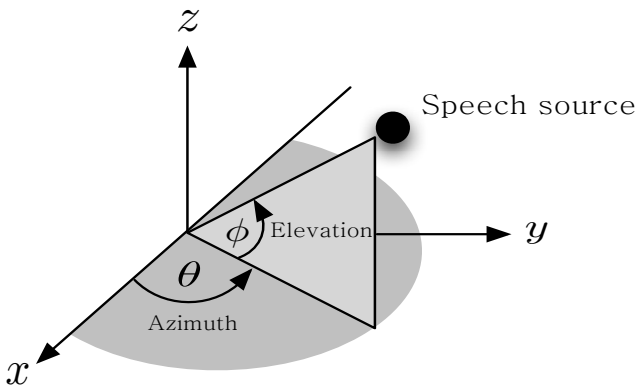

Figure 3.5: This diagram illustrates the coordinate systems.

$$\mathbf{p}_q = [\cos\phi\cos\theta, \cos\phi\sin\theta, \sin\phi], \tag{3.6}$$

where θ and ϕ represent azimuth and elevation (measurements in a spherical coordinate system), respectively. By way of illustration, it can be observed in Figure 3.5 the relation between Cartesian coordinates, θ and ϕ.

The sound propagation of a point source in the far field can be described by a plane wave, and then, the angles between the direction of the incident wave and the directions of maximum sensitivity of the microphones can be calculated as the following expressions indicate,

$$\beta_{1q} = \arccos\left(\mathbf{m}_1 \cdot \mathbf{p}_q\right) \tag{3.7}$$

and

$$\beta_{2q} = \arccos\left(\mathbf{m}_2 \cdot \mathbf{p}_q\right), \tag{3.8}$$

where $\cdot$ denotes the scalar product. Introducing these equations in Equation (3.3), we obtain

$$L_q = R\left(|s\cos(\arccos\left(\mathbf{m}_2 \cdot \mathbf{p}_q\right)) + 1 - s| - |s\cos(\arccos\left(\mathbf{m}_1 \cdot \mathbf{p}_q\right)) + 1 - s|\right), \tag{3.9}$$

and simplifying,

$$L_q = R\left(|s(\mathbf{m}_2 \cdot \mathbf{p}_q) + 1 - s| - |s(\mathbf{m}_1 \cdot \mathbf{p}_q) + 1 - s|\right). \tag{3.10}$$

Now the dot products are calculated:

$$\mathbf{m}_1 \cdot \mathbf{p}_q = \left[\cos\left(\frac{\alpha}{2}\right), -\sin\left(\frac{\alpha}{2}\right), 0\right] \cdot [\cos\phi\cos\theta, \cos\phi\sin\theta, \sin\phi]. \tag{3.11}$$

Operating, it results

$$\mathbf{m}_1 \cdot \mathbf{p}_q = \cos\left(\frac{\alpha}{2}\right)\cos\phi\cos\theta - \sin\left(\frac{\alpha}{2}\right)\cos\phi\sin\theta. \tag{3.12}$$

Removing common factor

$$\mathbf{m}_1 \cdot \mathbf{p}_q = \cos\phi\left[\cos\left(\frac{\alpha}{2}\right)\cos\theta - \sin\left(\frac{\alpha}{2}\right)\sin\theta\right]. \tag{3.13}$$

Taking into account the sum difference formula $\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)$, the equation above can be rewritten as

$$\mathbf{m}_1 \cdot \mathbf{p}_q = \cos\phi \cos\left(\theta + \frac{\alpha}{2}\right). \tag{3.14}$$

Repeating the same process for $\mathbf{m}_2 \cdot \mathbf{p}_q$, it is obtained:

$$\mathbf{m}_2 \cdot \mathbf{p}_q = \cos\phi \left[\cos\left(\frac{\alpha}{2}\right)\cos\theta + \sin\left(\frac{\alpha}{2}\right)\sin\theta\right]. \tag{3.15}$$

At this point a expression like $\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)$ must be achieved, but in Equation (3.15) + instead of − can be found. Knowing that $\cos(-a) = \cos a$ and $\sin(-a) = -\sin a$, we modify Equation (3.15), obtaining

$$\mathbf{m}_2 \cdot \mathbf{p}_q = \cos\phi \left[\cos\left(-\frac{\alpha}{2}\right)\cos\theta - \sin\left(-\frac{\alpha}{2}\right)\sin\theta\right]. \tag{3.16}$$

Finally the dot product results

$$\mathbf{m}_2 \cdot \mathbf{p}_q = \cos\phi \cos\left(\theta - \frac{\alpha}{2}\right). \tag{3.17}$$

Introducing Equations (3.14) and (3.17) in Equation (3.10), the level difference is

$$L_q = R\left(\left|s\cos\phi\cos\left(\theta - \frac{\alpha}{2}\right) + 1 - s\right| - \left|s\cos\phi\cos\left(\theta + \frac{\alpha}{2}\right) + 1 - s\right|\right). \tag{3.18}$$

Once the geometrical expression of the level difference has been calculated, another one for the time difference between the two microphones must be obtained. If the sound source is sufficiently far away from the microphone array, the problem is labeled as far-field problem. In this situation, it is assumed that the wave received at the array is a planar wave that propagates through the non-dispersive medium-air. An example of it is depicted in Figure 3.6 where it can be observed that the signal received at the microphone 2 ($z_2[n]$) is a delayed version of the signal at the reference sensor ($z_1[n]$). θ is the angle between the normal to the line joining the microphones and the direction of the incident wave.

As we mentioned, D is spacing between the two sensors and so, the time required for the plane wave to propagate through $D\cos(\theta)$ (from microphone 1 to microphone 2) is the time difference. This time difference for the q-th source will be denoted by t_q and can be calculated as the following expression indicates

$$t_q = \frac{D}{c}\cos(\phi)\sin(\theta), \tag{3.19}$$

where c is the sound velocity in air. It can be observed that estimating the time difference is essentially identical to the estimation of the DOA. This basic triangulation problem is the base of most of the source-localization techniques. Following this reasoning, let us define the relative time difference as

$$T_q = \frac{t_q c}{D} = \cos\phi\sin\theta. \tag{3.20}$$

At this point the geometric expressions of both level and relative time differences have been proposed (Equations (3.18) and (3.20)). Now our goal is to find a expression that relates them and in this sense, we aim at introducing Equation (3.20) in Equation (3.18). We consider that θ (azimuth) ranges between $-\pi/2$ and $\pi/2$ radians, since the microphone array is placed on a wall and the speech sources are always located in the frontal plane. By way of illustration, Figure

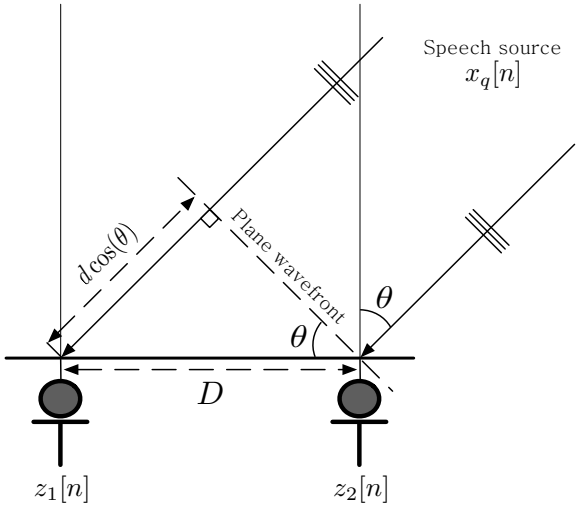

Figure 3.6: Illustration of the time-difference estimation using θ with the two-microphone array. Source $x_q[n]$ is located in the far-field, D is the microphone separation and θ is the angle between the normal to the line joining the microphones and the direction of the incident wave.

3.7 shows possible examples of talkers at different θ in the addressed problem. Note that in this figure the elevation (ϕ) has been omitted. Considering these values of θ, $\cos\theta$ ranges between 0 and 1 and it can be said that $\cos\theta = \sqrt{1-\sin^2\theta}$ without any uncertainty. Thus, knowing that $\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)$, the product $\cos\phi\cos\left(\theta \pm \frac{\alpha}{2}\right)$ can be expressed as

$$\cos\phi\cos\left(\theta \pm \frac{\alpha}{2}\right) = \cos\phi\left[\cos\theta\cos\left(\frac{\alpha}{2}\right) \mp \sin\theta\sin\left(\frac{\alpha}{2}\right)\right], \tag{3.21}$$

and also as

$$\cos\phi\cos\left(\theta \pm \frac{\alpha}{2}\right) = \cos\phi\cos\theta\cos\left(\frac{\alpha}{2}\right) \mp \cos\phi\sin\theta\sin\left(\frac{\alpha}{2}\right). \tag{3.22}$$

Taking into account Equation (3.20) and $\cos\theta = \sqrt{1-\sin^2\theta}$, it results

$$\cos\phi\cos\left(\theta \pm \frac{\alpha}{2}\right) = \mp T_q\sin\left(\frac{\alpha}{2}\right) + \cos\phi\sqrt{1-\sin^2\theta}\cos\left(\frac{\alpha}{2}\right). \tag{3.23}$$

Inserting $\cos\phi$ into the radical,

$$\cos\phi\cos\left(\theta \pm \frac{\alpha}{2}\right) = \mp T_q\sin\left(\frac{\alpha}{2}\right) + \cos\left(\frac{\alpha}{2}\right)\sqrt{\cos^2\phi - (\sin\theta\cos\phi)^2}, \tag{3.24}$$

and considering again Equation (3.20), the product of cosines is expressed as

$$\cos\phi\cos\left(\theta \pm \frac{\alpha}{2}\right) = \mp T_q\sin\left(\frac{\alpha}{2}\right) + \cos\left(\frac{\alpha}{2}\right)\sqrt{\cos^2\phi - T_q^2}. \tag{3.25}$$

Substituting and considering that the parameter $s \geq 0$, the following relation for the q-th source is obtained between L_q and T_q:

$$L_q = sR\left(\left|T_q\sin\left(\frac{\alpha}{2}\right) + \cos\left(\frac{\alpha}{2}\right)\sqrt{\cos^2\phi - T_q^2} - 1 + \frac{1}{s}\right| - \right. \tag{3.26}$$

$$\left. \left|-T_q\sin\left(\frac{\alpha}{2}\right) + \cos\left(\frac{\alpha}{2}\right)\sqrt{\cos^2\phi - T_q^2} - 1 + \frac{1}{s}\right|\right). \tag{3.27}$$

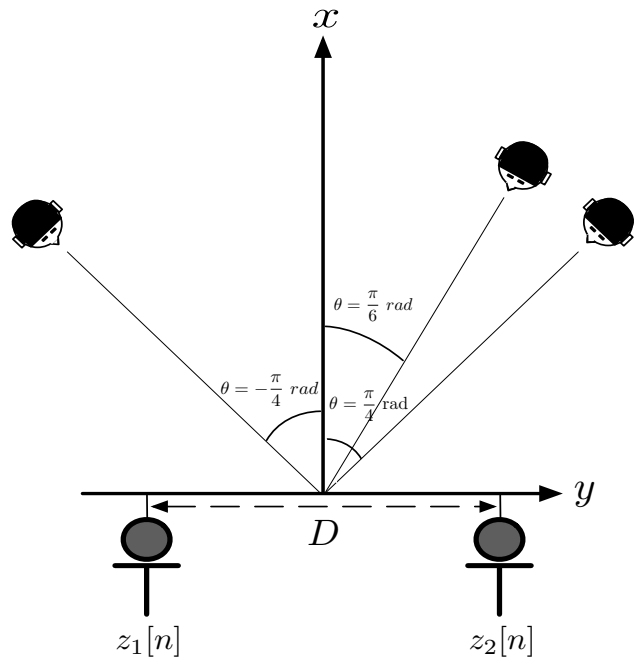

Figure 3.7: Example of three different talkers in the frontal plane of the microphone array for $\theta = -\pi/4$ rad, $\pi/6$ rad and $\pi/4$ rad.

In order to study this equation, the sign of the absolute values must be considered and so, three possible cases can be found:

$$L_q = \begin{cases} 2sR\left(1 - \frac{1}{s} - \cos\left(\frac{\alpha}{2}\right)\sqrt{\cos^2\phi - T_q^2}\right) & \text{if } T_q < -\frac{\cos\left(\frac{\alpha}{2}\right)\sqrt{\cos^2\phi - T_q^2} - 1 + \frac{1}{s}}{\sin\left(\frac{\alpha}{2}\right)} \\ -2sR\left(1 - \frac{1}{s} - \cos\left(\frac{\alpha}{2}\right)\sqrt{\cos^2\phi - T_q^2}\right) & \text{if } T_q > \frac{\cos\left(\frac{\alpha}{2}\right)\sqrt{\cos^2\phi - T_q^2} - 1 + \frac{1}{s}}{\sin\left(\frac{\alpha}{2}\right)} \\ 2sRT_q \sin\left(\frac{\alpha}{2}\right) & \text{other case.} \end{cases} \tag{3.28}$$

Denoting $s' = 1 - \frac{1}{s}$ and simplifying the above equation, we obtain the following equation,

$$L_q = \begin{cases} 2sR\left(s' - \cos\left(\frac{\alpha}{2}\right)\sqrt{\cos^2\phi - T_q^2}\right) & \text{if } T_q < s'\sin\left(\frac{\alpha}{2}\right) - \cos\left(\frac{\alpha}{2}\right)\sqrt{\cos^2\phi - s'^2} \\ -2sR\left(s' - \cos\left(\frac{\alpha}{2}\right)\sqrt{\cos^2\phi - T_q^2}\right) & \text{if } T_q > -s'\sin\left(\frac{\alpha}{2}\right) + \cos\left(\frac{\alpha}{2}\right)\sqrt{\cos^2\phi - s'^2} \\ 2sRT_q \sin\left(\frac{\alpha}{2}\right) & \text{other case.} \end{cases} \tag{3.29}$$

In the specific case of $s \leq 0.5$, then $s' \leq -1$, and the threshold is never reached, resulting a linear relation between L_q and T_q, that is, $L_q = 2sRT_q \sin\left(\frac{\alpha}{2}\right)$.

In Figure 3.8 the relationship L_q vs T_q is depicted for $\alpha = 60°$, $R = 10$ and two values of s, specifically, $s = 0.5$ (red curve) and $s = 0.75$ (black curve). Addittionaly, the relationship is represented for different values of elevation (ϕ). In the particular case of $\phi = 0$, $\cos^2\phi = 1$ and the Equation (3.29) is simplified.

$$L_q = \begin{cases} 2sR\left(s' - \cos\left(\frac{\alpha}{2}\right)\sqrt{1 - T_q^2}\right) & \text{if } T_q < s'\sin\left(\frac{\alpha}{2}\right) - \cos\left(\frac{\alpha}{2}\right)\sqrt{1 - s'^2} \\ -2sR\left(s' - \cos\left(\frac{\alpha}{2}\right)\sqrt{1 - T_q^2}\right) & \text{if } T_q > -s'\sin\left(\frac{\alpha}{2}\right) + \cos\left(\frac{\alpha}{2}\right)\sqrt{1 - s'^2} \\ 2sRT_q \sin\left(\frac{\alpha}{2}\right) & \text{other case.} \end{cases} \tag{3.30}$$

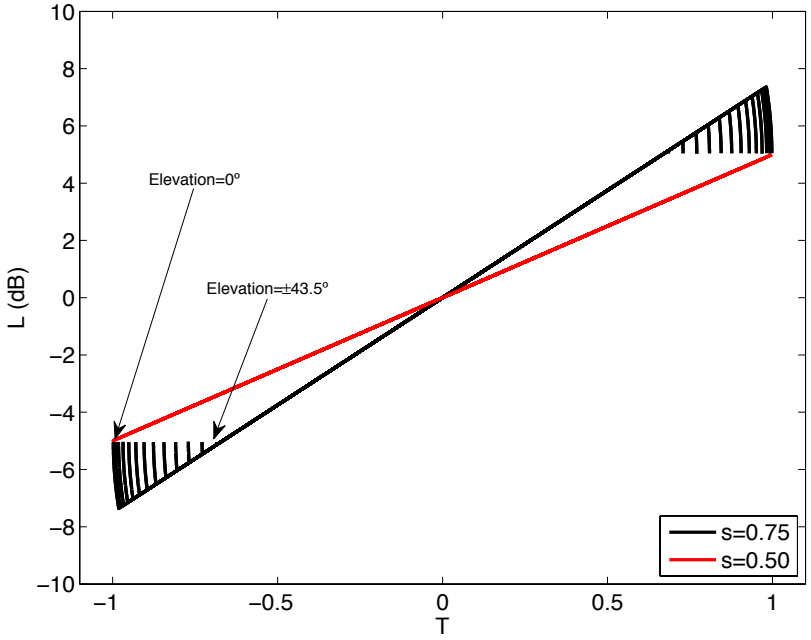

Figure 3.8: Relationship L_q vs T_q for $\alpha = 60°$, $R = 10$ and two values of s.

In the figure, the case $\phi = 0$ corresponds to the most extreme cases in the curves.

In order to solve the uncertainty produced for some elevations when $s > 0.5$, in this thesis we propose to use the most extreme case, that is, to suppose that $\phi = 0$. Given the distribution of the sources in the room, this supposition will lead to an error, but, as will be studied in the next sections, we will expect this error to have a low influence in the final separation system.

To sum up, in this section a new tool (a geometric model) for developing mixing matrix estimation is proposed. This geometric model provides a relationship between level and time differences. Considering the objective of estimating the mixing matrix, the main advantage of this model consists in that it only requires to calculate the time differences because using our L_q vs T_q relationship, the level differences can be automatically obtained. This is an important advantage since many speech separation methods do not estimate level differences as well as it can be expected.

3.3.2 True time differences with Binary Mask

In this section the suitability of our proposal to perform speech separation will be studied. Aiming at conducting this study, we compare its performance to one of the classical separation algorithms. Considering the above sections, it can be said that the application of l_1-norm minimization is not recommendable since BM obtains the highest quality in all the cases. For this reason, DUET with binary masking will be our reference separation method.

The main idea of this section is to test the validity of our approach. As mentioned earlier, the advantage of our method is that only time differences must be calculated to estimate the mixing matrix since level differences are obtained using our L_q vs T_q relationship. In this sense, the true delays of the simulations are used and their respective values of L are calculated by means of our relationship (supposing an elevation of $\phi = 0$, which is the most extreme case of Equation (3.29)). Once the time differences of the simulations and the level differences are available, the mixing matrix is estimated. When the mixing matrix is obtained, it will be combined with a

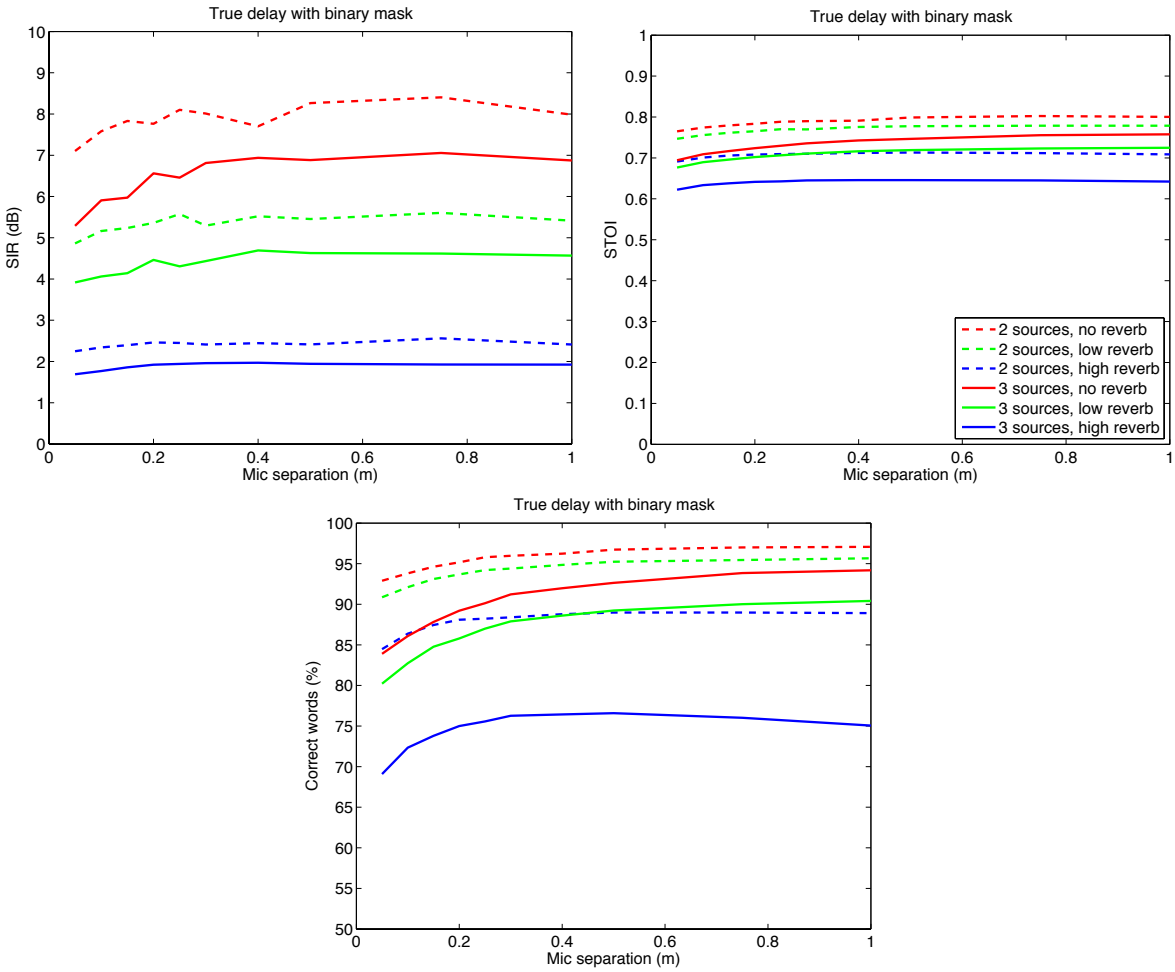

Figure 3.9: SIR (dB), STOI and Correct words (%) for the true time differences with binary mask.

classical separation stage based on BM and the separation will be carried out. Despite not being realistic since true delays are used, with these experiments we can assess whether estimating time differences with a considerable accuracy our mixing matrix estimation is suitable for performing speech separation. The reader can note that in Sections 3.3.3 and 3.3.4 different time delay estimation algorithms will be used to estimate the time differences from the simulated speech mixtures.

At this point, the performance of our mixing matrix estimation method based on a geometric analysis will be presented. First, the detailed results from running our proposal with different distances between microphones are illustrated in Figure 3.9.

A wealth of information is obtained from many measures that are included in the previous figure, and in following figures and tables. We use the same scenarios in the previous sections to evaluate our proposal, that is, how the frame length or the different type of arrays (different polar patterns of the microphones, mutual angle or spacing) affect the separation is studied.

It can be observed in Figure 3.9 that the vast majority of quality values are higher than when the classical speech separation method is used. This should come as no great surprise since our proposal has the advantage of not requiring the estimation of the level differences, what is important since in general, this estimation is poorly performed by speech separation algorithms. Since the values of time differences are the values used in the simulations, naturally this improves

Best Case	2 sources											
	No reverberation				Low reverberation				High reverberation			
D (m)	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75
SIR (dB)	7.1	7.8	7.7	8.4	4.9	5.4	5.5	5.6	2.2	2.5	2.4	2.6
STOI	0.76	0.78	0.79	0.80	0.75	0.77	0.78	0.78	0.69	0.71	0.71	0.71
Words (%)	92.9	95.2	96.2	97.0	90.9	93.7	94.8	95.4	84.5	88.1	88.8	89.0
Window (ms)	128	128	128	64	128	128	64	128	128	128	128	128
Angle (α)	60	30	0	30	60	30	30	30	30	30	30	30
Microphones	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20
Best Case	3 sources											
	No reverberation				Low reverberation				High reverberation			
D (m)	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75
SIR (dB)	5.3	6.6	6.9	7.1	3.9	4.5	4.7	4.6	1.7	1.9	2.0	1.9
STOI	0.69	0.72	0.74	0.76	0.68	0.70	0.72	0.72	0.62	0.64	0.65	0.65
Words (%)	83.9	89.2	92.0	93.8	80.2	85.8	88.6	90.0	69.1	75.0	76.4	76.0
Window (ms)	128	64	64	64	64	64	64	64	128	128	128	128
Angle (α)	60	30	30	30	60	30	30	30	30	30	30	60
Microphone	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20

Table 3.8: Best results obtained for the True delay with binary mask method.

the separation but also indicates that our proposal is completely valid. From this figure, many noticeable facts may be underlined. We can see by inspection that in almost any situations the quality of the separation, in terms of all the objective measurements (SIR, STOI and Correct words (%)), increases according spacing increases regardless the level of reverberation, the number of speech sources or the quality parameter.

In Table 3.8 the scores obtained with the configurations that achieve the highest quality are shown. First it must be highlighted that high quality values are obtained in terms of the three parameters. Examples of it are high SIR values for lack of reverberation (8.4 dB or 7.1 dB for two or three speech sources, respectively) or for low reverberation (5.6 dB or 4.7 dB for two or three speech sources, respectively). In the case of high-reverberant rooms, not very good results are obtained but it must be said that they are better than the ones of the classical algorithm and these reverberant conditions are not very common in the vast majority of rooms. Similar conclusions can be extracted from the point of view of speech intelligibility. For instance, considering Correct words (%), over 90% of words can be understood for lack of reverberation and low reverberation. In the case of high-reverberant environments, Correct words (%) equal to 89% and $76,4\%$ is obtained for two and three speech source, what can be considered as quite good quality.

In respect of the frame length, 64 ms and 128 ms are the most common values. For two sources 128 ms is the best frame length in virtually all cases, while for three sources, 128 ms is the best frame length for high-reverberant rooms, being 64 ms for the rest of cases. On the other hand the type of microphone. It is evident that the most suitable one is the cardioid microphone with a FBR equal to 20 dB. Studying the angle between the directions of maximum sensitivity, $\alpha = 30^\circ$ has demonstrated to achieve the highest separation quality.

Other aspect to be taken into account is the type of microphone array. As in previous sections, the three polar patterns and the mutual angle have been studied as it is illustrated in Table 3.9. At first glance, it is clear that the omnidirectional microphone obtains the worst results in all the cases. In contrast to previous studies, the MEMS microphone achieves higher separation quality than the C10 microphone for all the mutual angles. The best result for the MEMS microphone

Microphones	Angle (α) 180°	120°	90°	60°	30°	0°
Mems (Mms)	4.6 (2.2)	3.2 (1.9)	2.6 (1.8)	2.1 (1.7)	1.9 (1.7)	2.0 (1.8)
Omnidir. (Omn)	10.5 (3.3)	10.5 (3.3)	10.5 (3.3)	10.5 (3.3)	10.5 (3.3)	10.5 (3.3)
Cardioid 10dB(C10)	9.3 (1.8)	5.9 (1.0)	4.7 (1.0)	4.1 (1.2)	4.2 (1.6)	4.6 (2.1)
Cardioid 20dB(C20)	19.1 (7.2)	7.7 (4.8)	2.5 (1.2)	0.4 (0.3)	0.1 (0.3)	1.1 (1.3)

Table 3.9: Average loss (and Standard Deviation) in the rate of Correct words (%) obtained for the True delay with binary mask method, in function of the selected microphone and the angle between both microphones.

is obtained for $\alpha = 30°$, where an average loss equal to 1.9 and a standard deviation equal to 1.7 are achieved. Furthermore, the MEMS microphone obtains lower losses for $\alpha = 180°$ and $90°$ and similar losses for $\alpha = 90°$ than the C20 microphone. However, the C20 microphone is the one that gets the best result, an average loss of 0.1 and a standard deviation of 0.3 when the mutual angle is 30°, what can be considered very good results.

To conclude, the results of these experiments show that the use of our proposal is highly recommended because it obtains better separation results than the classical separation method. Although it is true that true time differences have been used. Considering it, in the following sections we must propose the use of time delay estimation algorithms to calculate these time differences. The effects of using TDE algorithms on the separation quality will also be studied.

3.3.3 GCC-PHAT based mixing matrix estimation method with Binary Mask

In Section 3.3.2 the suitability of our proposal for estimating the mixing matrix was studied. In order to evaluate our method, it was combined with a separation stage based on BM and the quality of separation was analyzed when different array configurations and frame lengths were employed. In the experiments, the true time differences of the simulations were introduced in the L_q vs T_q relationship aiming at obtaining the level differences and so, the estimation of the mixing matrix. In order to be realistic, a TDE algorithm must be used to determine the time differences from the speech mixtures. Following this idea, a literature review has been made in the aim of selecting the most appropriate TDE algorithm for our separation problem, that is, when time differences are extracted from speech mixtures recorded in reverberant rooms. From the different TDE methods in the literature, GCC-PHAT has been chosen since its performance is slightly degraded in the presence of reverberation.

Overall, the GCC-PHAT function is used to detect delays between the same single source signal at different sensors. In this sense, calculating the GCC-PHAT function of two signals received at different sensors, a peak can be found in a position corresponding to the delay (time difference). In our case of study, delays due to two or three speech sources ($S = 2, 3$) must be estimated and therefore, more than one peak must be located. It has been observed that GCC-PHAT function shows peaks in positions corresponding to the differences in the delays of each source to the considered pair of sensors. Bearing it in mind, a method that locates these peaks has been implemented. This method works as follows to detect S peaks:

- The first step consists in obtaining the maximum value of the GCC-PHAT function. Probably, the first time that this step is carried out, this peak corresponds to the dominant source in the mixture. The position of this peak is related to the time difference due to the source.

- Aiming at obtaining a better estimation of the time difference, an interpolation of the

maximum value of the GCC-PHAT function with its two neighbors is also considered. Or in other words, the maximum of the function is located and using the previous and next point, the function is approximated to a parabolic function in that interval. Finally, from this approximation, the position of the new maximum is located, being this new position the final value of the time difference.

- Once this peak has been estimated, other peaks must be calculated. Considering it, the values of the GCC-PHAT function in the location corresponding to the maximum and in the adjacent locations are discarded, that is, they are set to 0. This is justified by the fact that these adjacent values also correspond to the dominant source.

This process is repeated until the S peaks are extracted. It has been observed that the GCC-PHAT function does not present many peaks due to reverberation, in contrast to other TDE methods based on cross-correlation which are seriously affected by the reverberation.

It is evident that our objective is to have an accurate method for calculating time differences. In order to evaluate the accuracy of the GCC-PHAT method a study has been made. This study consists in calculating the RMSE between the true time differences of the simulations and the estimated ones. In order to evaluate its performance in several scenarios, RMSE has been calculated for the rooms and reverberation values describer in the experimental setup in Section 3.2.

Figure 3.10 displays the RMSE obtained by GCC-PHAT for the aforementioned scenarios. It is easy to see that the higher the reverberation is, the higher the error is. The number of speech sources in mixtures also affect the accuracy of GCC-PHAT. In all the situations, the RMSE when three sources are considered ($S = 3$) is higher than when there are two speakers ($S = 2$). Finally, the influence of the microphone separation is also analyzed. When the number of sources is two, the lowest error is obtained for $D = 0.05$ m, while the lowest error for three speech sources occurs when D is over 0.5 m.

Another aspect to be mentioned for the case of three speech sources is that the method obtains lower RMSE for small microphone separations because the possible time differences obtained are bounded by the microphone separation and so, the RMSE will be lower than for the largest microphone separations. Then, we note that the method does not work properly when three speech sources are presented in the mixtures.

In the next experiments a separation solution that contains three parts is evaluated. In the first part, GCC-PHAT method estimates the time differences, subsequently these time differences are introduced in our L_q vs T_q relationship to estimate the level differences and the mixing matrix. Finally, a classical separation stage that relies on binary masking is used. We denote this separation solution by Proposal 1.

At this point we can conclude that microphone separation has influence on GCC-PHAT and our proposed mixing matrix estimation method (see Section 3.3.2 for further information). Considering it, Figure 3.11 depicts how the distance between microphones modifies the quality separation of our solution. At first glance it should be pointed out that since the microphone separation increases, the quality also does, regardless the reverberation effects, the number of speech sources in the mixtures or the quality parameter. For instance in non-reverberant environments, improvements in terms of SIR of nearly 4 dB and 2 dB for two and three speech sources are achieved, respectively. Another example can be the increase in Correct words (%) between $D = 0.05$ m and $D = 0.75$ m for high-reverberant environments, it varies from 73 % to 88.88 % or from 58.6 % to 72.6 % when there are two or three speech sources in the mixtures, respectively.

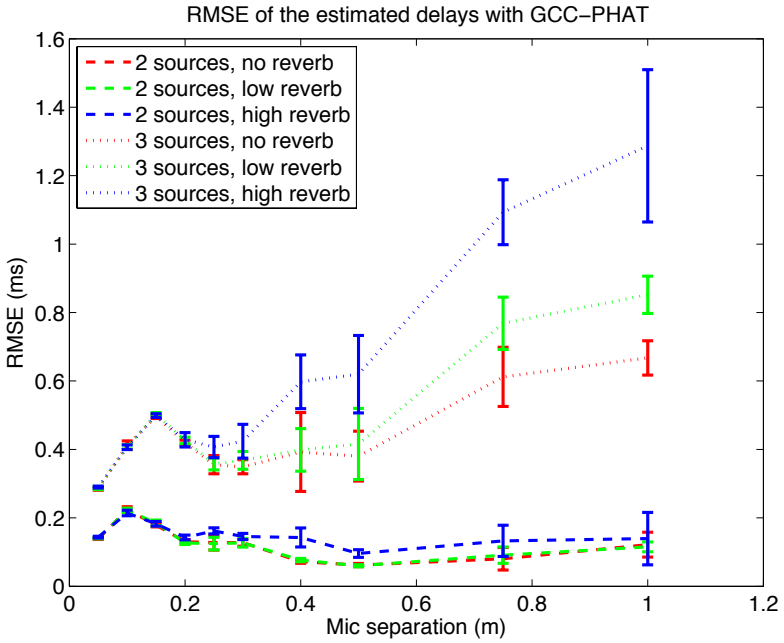

Figure 3.10: RMSE of the estimation of the delays of the sources with the standard GCC-PHAT, including the mean value and the range of variation at one standard deviation from the mean..

In respect of GCC-PHAT, from Figure 3.10 we conclude that the RMSE increases considerably along with the distance when the number of speech sources is three. Despite it, the quality increases when the microphone separation increases, what lead us to think that the accuracy of GCC-PHAT is enough to perform a good time difference estimation. When comparing the results obtained by DUET with binary masking and the ones achieved by Proposal 1 (GCC-PHAT+Geometric model + BM), Figures 3.4 and 3.10, we can notice that higher quality values are obtained with our solution.

To explore this issue further, Table 3.10 presents the best configurations from the point of view of the separation quality. It can be said that very good separated sources are estimated since quality parameters present high values for non- and low-reverberant environments. For example, for non-reverberant environments, values of SIR = 8.4 dB, STOI = 0.8 and Correct words (%) = 97 % are reached for two speech sources. Or the values of SIR = 4.5 dB, STOI = 0.72 and Correct words (%) = 87.7 % in low-reverberant environments when there are three speech sources in the mixtures. In respect of the frame length, 64 ms and 128 ms are the values more repeated, what was expected when microphone arrays with these microphone separations are used.

Considering the microphone array, the C20 microphone has demonstrated to achieve the best results with a mutual angles of 30° and 60° in almost all the cases. Deepening a little more in the type of microphone array, Table 3.11 shows the average loss and standard deviation obtained with different types of microphone and microphone separations. As in previous cases, the omnidirectional microphone obtains the worst results, surely, factors such as reverberation cause that omnidirectional microphones would not be suitable for speech separation tasks. The MEMS microphone, that is, the one whose response has been measured in an anechoic room, outperforms the C10 microphone for all the values of α. When comparing the MEMS microphone and the cardioid one with a FBR equal to 20 dB, for $\alpha = 180°$ and 120° the MEMS one obtains the lowest scores of average losses (4.3 and 3, respectively). For the rest of values of the mutual

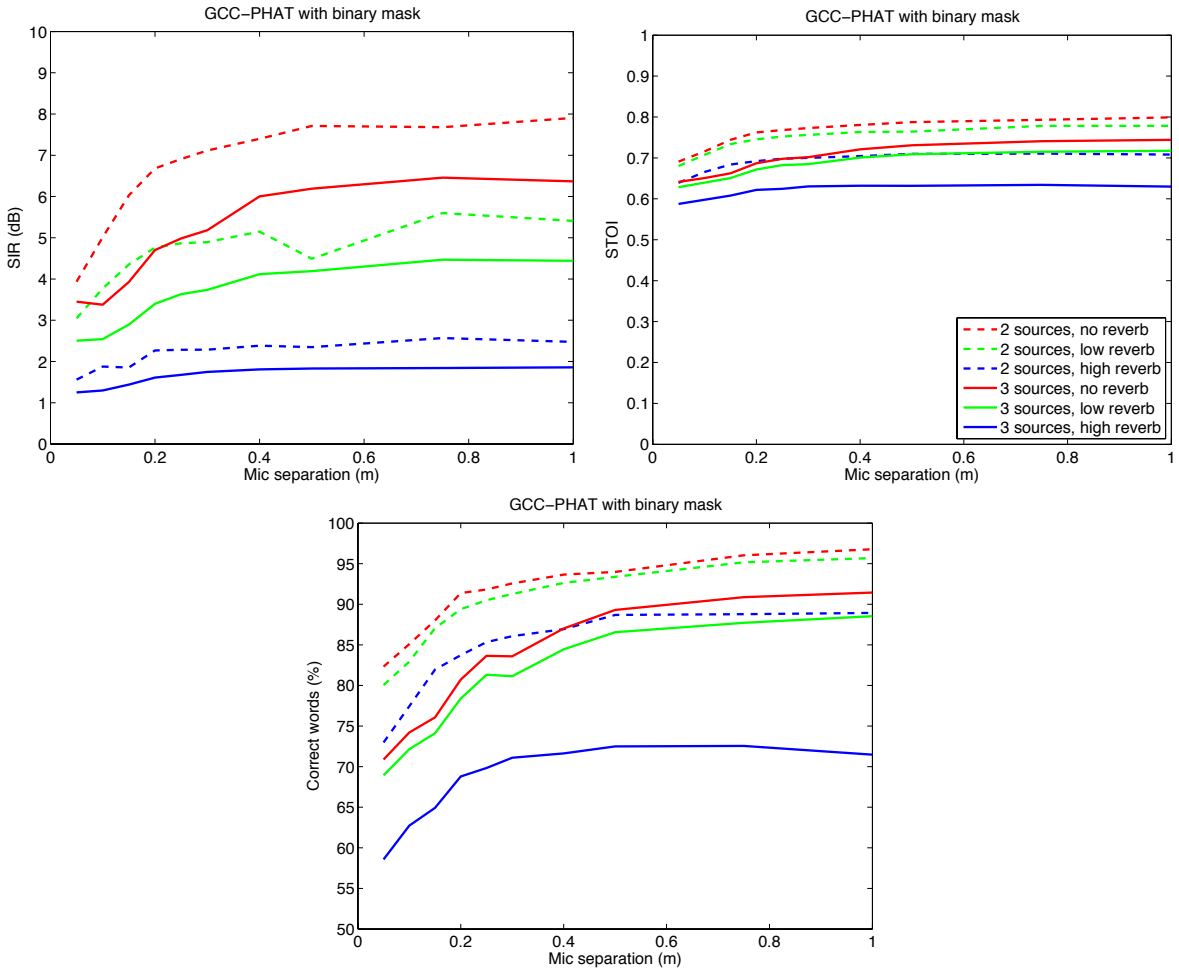

Figure 3.11: SIR (dB), STOI and Correct words (%) for Proposal 1 (GCC-PHAT + Geometric model + BM)

angle, the C20 microphone reaches the lowest losses, where the best result is an average loss equal to 0.2 and standard deviation (0.3) for $\alpha = 60°$.

To demonstrate that our solution works properly, in Table 3.12 the highest values obtained for DUET and Proposal 1 (GCC-PHAT+ Geometric model + BM, labeled Pro1 in table) are depicted. Both alternatives use the same separation stage based on BM. Looking at Table 3.12 it can be easily seen that our proposal outperforms the classical method in all the cases, from both the point of view of the quality separation (SIR) and the point of view of the speech intelligibility (STOI and Correct words (%)). Perhaps it must be mentioned the case of three sources in non- and low-reverberant environments. Improvements in terms of SIR of 2.4 dB and 1.8 dB has been achieved. The values of STOI has also increased, reaching 0.74, 0.72 and 0.63 when three sources are presented in the mixtures. In respect of Correct words (%), high values of % of words understood are achieved. For instance, it can be mentioned the increase of 25.5 % for three sources and high reverberation.

Finally, comparing the results in Tables 3.8 and 3.10, it is clear that for $D = 0.05$ m, our geometric model using GCC-PHAT performs much worse than when the true time differences are included. From these results we confirm that GCC-PHAT does not perform correctly for short distances and so, it is necessary to use another TDE method for small microphone separations.

Best Case	2 sources											
	No reverberation				Low reverberation				High reverberation			
D (m)	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75
SIR (dB)	3.9	6.7	7.4	7.7	3.0	4.8	5.1	5.6	1.6	2.3	2.4	2.6
STOI	0.69	0.76	0.78	0.79	0.68	0.75	0.76	0.78	0.64	0.69	0.70	0.71
Words (%)	82.3	91.4	93.7	96.0	80.1	89.4	92.6	95.2	73.0	83.7	86.9	88.8
Window (ms)	128	128	128	64	128	128	128	64	128	128	128	64
Angle (α)	60	30	30	0	60	30	30	30	60	30	30	30
Microphones	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20
Best Case	3 sources											
	No reverberation				Low reverberation				High reverberation			
D (m)	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75
SIR (dB)	3.5	4.7	6.0	6.5	2.5	3.4	4.1	4.5	1.2	1.6	1.8	1.8
STOI	0.64	0.69	0.72	0.74	0.63	0.67	0.70	0.72	0.59	0.62	0.63	0.63
Words (%)	70.9	80.7	87.0	90.9	68.9	78.4	84.5	87.7	58.6	68.8	71.6	72.6
Window (ms)	64	128	64	64	128	128	128	64	128	128	128	128
Angle (α)	90	60	60	30	60	60	60	60	90	60	60	60
Microphone	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20

Table 3.10: Best results obtained for Proposal 1 (GCC-PHAT + Geometric model + BM).

	Angle (α)					
Microphones	180°	120°	90°	60°	30°	0°
Mems (Mms)	4.3 (2.1)	3.0 (1.7)	2.5 (1.6)	2.3 (1.6)	2.3 (1.7)	2.6 (1.9)
Omnidirectional (Omn)	10.3 (3.1)	10.3 (3.1)	10.3 (3.1)	10.3 (3.1)	10.3 (3.1)	10.3 (3.1)
Cardioid 10dB (C10)	8.8 (2.0)	5.6 (1.1)	4.6 (1.1)	4.2 (1.3)	4.4 (1.7)	4.9 (2.1)
Cardioid 20dB (C20)	17.0 (7.0)	6.5 (4.4)	1.8 (1.2)	0.2 (0.3)	0.4 (0.6)	1.7 (1.5)

Table 3.11: Average loss (and Standard Deviation) in the rate of Correct words (%) obtained for Proposal 1 (GCC-PHAT + Geometric model + BM), in function of the selected microphone and the angle between both microphones.

Case	No reverberation				Low reverberation				High reverberation			
Best	2 sources		3 sources		2 sources		3 sources		2 sources		3 sources	
D (m)	BM	Pro1	BM	Pro1	BM	Pro1	BM	Pro1	BM	Pro1	BM	Pro1
SIR (dB)	7.0	7.7	4.1	6.5	4.5	5.6	2.7	4.5	1.9	2.6	1.1	1.8
STOI	0.77	0.79	0.67	0.74	0.73	0.78	0.63	0.72	0.66	0.71	0.58	0.63
Words (%)	93.6	96	78.1	90.9	89.0	95.2	70.2	87.7	78.0	88.8	57.1	72.6

Table 3.12: Best results obtained for our solution Proposal 1 (GCC-PHAT + Geometric model + BM) and DUET, both alternatives using a separation stage based on BM.

3.3.4 Proposed solution for small arrays: frequency domain optimization

Proposal 1 outperforms DUET with binary masking in terms of both speech quality and speech intelligibility. At this point it can be said that a novel and good mixing matrix estimation procedure (TDE method + geometric model) has been proposed in this thesis. The introduced L_q vs T_q relationship has demonstrated to be valid for all the addressed separation problems, that is, for different levels of reverberation, number of speech sources, distance between sensors or polar patterns. Nevertheless, it is clear that whether time difference estimation is not correctly performed, then the mixing matrix will not be properly calculated. In this respect, a reinterpretation of GCC-PHAT has been used in Section 3.3.3 since this TDE algorithm demonstrates good behavior in the presence of reverberation. Analyzing this TDE method, we have realized that there is a problem due to its accuracy when the distance between microphones is short. For those short distances, the time difference due to each speech source is also short. For instance, assuming a microphone separation $D = 0.05$ m, then, the maximum difference in time delay of one speech signal between the two microphones is $\Delta_{max} = D/c$, where c is sound velocity in air. If $c = 340$ m/s, the ratio $D/c = 0.05/340 \simeq 0.147$ ms. In other words, the maximum time difference with a microphone distance equal to 0.05 m is 0.147 ms. On the other hand the accuracy of the GCC-PHAT function must be considered. Assuming that speech mixtures are sampled at a frequency f_s equal to 8 Khz, the accuracy of the GCC-PHAT function is $1/f_s = 0.125$ ms, that is, 0.125 ms is the shortest delay that can be calculated without applying any approximation between the peak of the GCC-PHAT function corresponding to a time difference and the neighbor samples. Comparing the maximum delay for the microphone array (0.147 ms) and the accuracy of the GCC-PHAT function (0.125 ms) it is clear that a problem occurs. In order to overcome this limitation with short distances between microphones, a new TDE method with higher accuracy should be implemented.

For this purpose, the design of a new TDE method (labeled "optimization proposal") to avoid the problems of GCC-PHAT associated with the use of small microphone arrays is presented. Like GCC-PHAT, this proposal is also based on the idea of whitening the input signals to better calculate delays. First, our method calculates the sum of the normalized crosspower-spectrums of two mixtures ($Z_m(k,l)$ and $Z_p(k,l)$) of L STFT frames as Equation (3.31) indicates

$$R_{mp}(k) = \sum_{l=1}^{L} \frac{Z_m(k,l)}{|Z_m(k,l)|} \frac{Z_p^*(k,l)}{|Z_p(k,l)|}. \tag{3.31}$$

Comparing the above equation with Equation (2.71), the reader can note that the only difference is that in Equation (3.31) the inverse Fourier transform is not computed. Our objective is to avoid working in the time domain due to the problem of accuracy of the GCC-PHAT function. The idea underlying our method is: assuming S sources in the mixtures, if $r_{mp}[\tau]$ is formed by a set of deltas concentrated at the positions corresponding to the time differences due to each source between the m-th and p-th microphones (these time differences denoted by τ_{mps}), then, $R_{mp}(k)$ is formed by a set of exponentials, $e^{-i\omega_k \tau_{mps}}, \forall s = 1, \ldots, S$. Therefore, our method will attempt to determine these exponentials instead of localizing the set of deltas in the time domain. And now the question is: *how?*

$R_{mp}(k)$ will be modelled by a function as Equation (3.32) shows, where a set of delays and magnitudes, $\hat{\tau}_{mps}$ and a_{mps}, $\forall s = 1, \ldots, S$ that minimize the RMSE between $R_{mp}(k)$ and $\hat{R}_{mp}(k)$ will be estimated.

$$\hat{R}_{mp}(k) = \sum_{s=1}^{S} a_{mps} e^{-i\omega_k \hat{\tau}_{mps}} \tag{3.32}$$

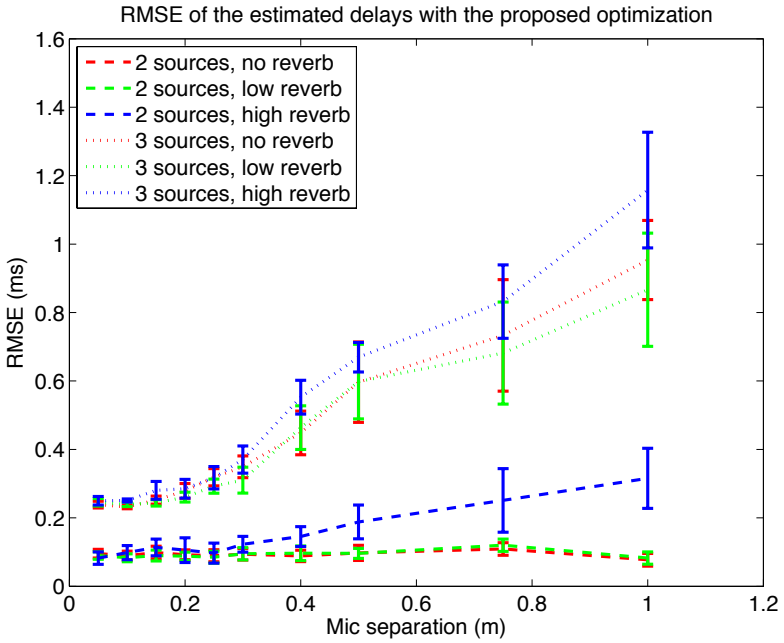

Figure 3.12: RMSE of the estimation of the delays of the sources with the proposed optimization, including the mean value and the range of variation at one standard deviation from the mean.

Equation (3.33) presents the function to be optimized in which it has been included a term (second addend in equation) that helps to discard one estimated source if it is very close to another one. The addition of this second term in the optimization serves a similar purpose than the removal of the neighbors, described in the last section. Thus, the optimization function results

$$RMSE_{mp} = E_k\{|\hat{R}_{mp}(k) - R_{mp}(k)|^2\} + ae^{-b\sum_{s=2}^{S}\sum_{t=s+1}^{S}(\hat{\tau}_{mps}-\hat{\tau}_{mpt})^2}. \tag{3.33}$$

It has been experimentally proven that $a = 0.02$ and $b = 0.3$ obtain the best results for the available speech database. The steps followed in the algorithm are:

- For each pair p and m, a Δ_τ-sweep from $-\Delta_{max}$ to Δ_{max} is made for all the possible values of τ_{mps}, $\forall s = 1, \ldots, S$. Two aspects must be mentioned: (a) our method will have more accuracy than GCC-PHAT if $\Delta_\tau < 1/f_s$, and (b) not all the combinations of delay must be tested due to symmetric properties. The initial value of Δ_τ will be $1/(2f_s)$.
- From each $\hat{\tau}_{mps}$, its corresponding a_{mps} is obtained and the RMSE is calculated.
- The candidates ($\hat{\tau}_{mps}$, a_{mps}) that obtain the lowest RMSE are selected. Then, Δ_τ is divided by 2, and the search range is consequently reduced. These three steps are repeated for values close to the selected candidates. This makes it possible to increase the accuracy of the estimations gradually.

Figure 3.12 shows the RMSE of the estimation of the delays of the sources with our method.

Having a look at this figure, the RMSE obtained by our proposal must be analyzed. Comparing these values of RMSE with the ones achieved by GCC-PHAT (Figure 3.10), it can be said that for distances between microphones shorter than 0.3 m our TDE method achieves lower RMSE in all the cases, as it was expected. Therefore, our objective of developing a new TDE

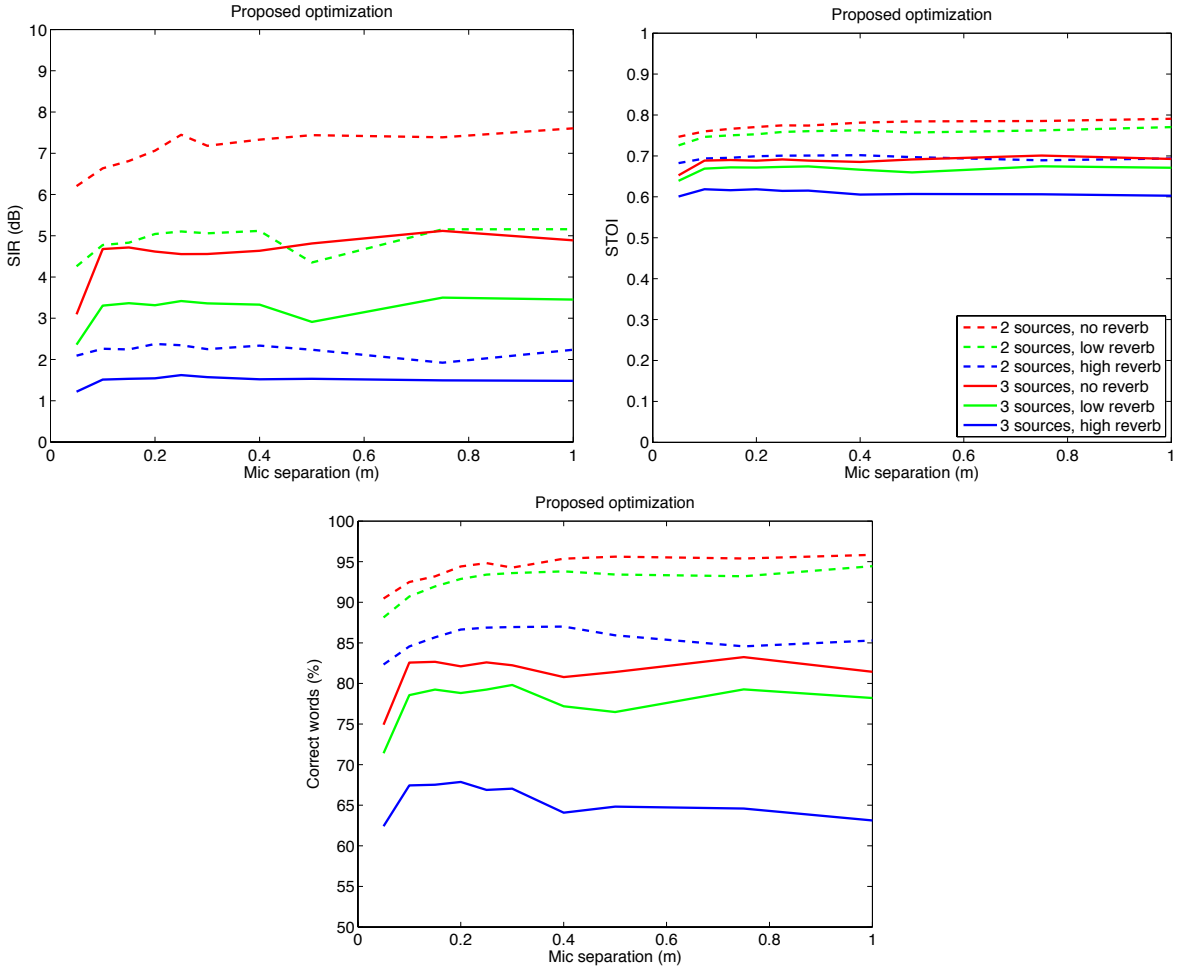

Figure 3.13: SIR (dB), STOI and Correct words (%) for Proposal 2 (Optimization proposal + Geometric model + BM)

algorithm with higher accuracy for short microphone distances has been met. With respect to distances larger than 0.3 m, our proposal obtains slightly worse results, this aspect being more evident in high-reverberant rooms.

A study of the use of our TDE method for separating speech sources will be carried out. The quality results achieved by the combination of our TDE method (optimization proposal), the geometric model and BM are presented. We will denote this separation solution as Proposal 2 in the following figures and tables. The influence of the microphone separation on speech quality and speech intelligibility in terms of SIR, and STOI and Correct words (%) is depicted in Figure 3.13. The first impression is that our TDE method helps to achieve higher scores than GCC-PHAT for the shortest microphone distances (0.05 m and 0.2 m). It is very significant that there are less differences between the values obtained for short and large distances when our optimization proposal is used.

In Table 3.13 the most suitable frame lengths and array configurations are presented. Considering the type of microphone, the C20 microphone has achieved the highest values in all the cases except for $D = 0.75$ m, two sources and high reverberation, where the MEMS microphone has been the selected one. The best frame length is 128 ms without exception, and mutual angles of 30° and 60° provide the best outcomes for two and three speech sources, respectively.

Best Case	2 sources											
	No reverberation				Low reverberation				High reverberation			
D (m)	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75
SIR (dB)	6.2	7.1	7.3	7.4	4.3	5.0	5.1	5.2	2.1	2.4	2.3	1.9
STOI	0.75	0.77	0.78	0.79	0.73	0.75	0.76	0.76	0.68	0.70	0.70	0.69
Words (%)	90.5	94.4	95.4	95.4	88.1	92.9	93.8	93.2	82.3	86.6	87.0	84.6
Window (ms)	128	128	128	128	128	128	128	128	128	128	128	128
Angle (α)	60	30	30	30	30	30	30	60	30	30	30	30
Microphones	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	MEM
Best Case	3 sources											
	No reverberation				Low reverberation				High reverberation			
D (m)	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75
SIR (dB)	3.1	4.6	4.6	5.1	2.4	3.3	3.3	3.5	1.2	1.5	1.5	1.5
STOI	0.65	0.69	0.69	0.70	0.64	0.67	0.67	0.67	0.60	0.62	0.61	0.61
Words (%)	74.9	82.1	80.8	83.2	71.4	78.8	77.2	79.3	62.4	67.9	64.1	64.6
Window (ms)	128	128	128	128	128	128	128	128	128	128	128	128
Angle (α)	60	60	60	30	60	60	60	30	60	60	60	60
Microphone	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20

Table 3.13: Best results obtained for Proposal 2 (Optimization proposal + Geometric model + BM).

Microphones	Angle (α)					
	180°	120°	90°	60°	30°	0°
Mems (Mms)	5.7 (2.2)	3.8 (1.9)	3.0 (1.8)	2.7 (1.8)	2.7 (1.8)	2.9 (2.0)
Omnidirectional (Omn)	12.2 (3.5)	12.2 (3.5)	12.2 (3.5)	12.2 (3.5)	12.2 (3.5)	12.2 (3.5)
Cardioid 10dB (C10)	10.2 (2.2)	6.7 (1.5)	5.6 (1.4)	5.1 (1.5)	5.3 (1.8)	5.9 (2.2)
Cardioid 20dB (C20)	18.2 (6.0)	7.3 (4.0)	2.1 (1.0)	0.3 (0.4)	0.4 (0.5)	1.9 (1.5)

Table 3.14: Average loss (and Standard Deviation) in the rate of Correct words (%) obtained for Proposal 2 (Optimization proposal + Geometric model + BM), in function of the selected microphone and the angle between both microphones.

Table 3.14 represents the average losses and standard deviations obtained by different types of microphones and mutual angles. As usual, the omnidirectional microphone obtains the worst results. The best microphone are the MEMS one for $\alpha = 180°$ and 120°, getting average losses equal to 5.7 and 3.8, and $C20$ for the rest of cases. The mininum losses are obtained by C20 where the average loss is equal to 0.3 for $\alpha = 60°$.

In order to show that Proposal 2, that is, the inclusion of our TDE method, obtains the best results for the shortest distances in the vast majority of cases, Table 3.15 shows the results of the separation for $D = 0.05$ m. Looking carefully it is evident that from the point of view of speech quality it obtains higher or similar results and from the point of view of speech intelligibility, best outcomes are achieved in all the cases, what leads us to recommend the use of our TDE method in small devices.

3.4 Reducing musical noise

In this chapter a new mixing matrix estimation stage based on a geometric interpretation of the separation problem has been introduced. However, in referring to the separation stage, a classical solution that relies on the time-frequency binary masking technique has been used. This technique, which has been explored for decades, is presented in many research fields, such as in

Case	No reverberation				Low reverberation				High reverberation			
Best	2 sources		3 sources		2 sources		3 sources		2 sources		3 sources	
D (m)	Pro1	Pro2	Pro1	Pro2	Pro1	Pro2	Pro1	Pro2	Pro1	Pro2	Pro1	Pro2
SIR (dB)	3.9	6.2	3.5	3.1	3.0	4.3	2.5	2.4	1.6	2.1	1.2	1.2
STOI	0.69	0.75	0.64	0.65	0.68	0.73	0.63	0.64	0.64	0.68	0.59	0.60
Words (%)	82.3	90.5	70.9	74.9	80.1	88.1	68.9	71.4	73.0	82.3	58.6	62.4

Table 3.15: Results obtained for $d = 0.05$ m by the GCC-PHAT based (Pro1) and the proposed optimization based (Pro2) mixing matrix estimation methods.

the field of speech enhancement [Li and Loizou, 2008], in hearing aid processing [Wang, 2008], in binaural systems to perform multi-band spatial separation [Kollmeier et al., 1993] or in CASA techniques [Wang et al., 2009], among others. In the framework of sound source separation, time-frequency binary masking is very popular since it has some advantages, for example, it deals with the underdetermined problem or it is computationally fast-paced. Nowadays, many studies about T-F binary masking address two important concerns: 1) the effects of room reverberation and 2) the musical noise problem. With respect to the latter one, different solutions have been proposed in the literature since this problem may be important in some applications. In this section we will focus on developing a simple but effective solution to minimize the distortion due to musical noise when T-F binary masking is used. First, the musical noise phenomenon and the most relevant solutions that have been proposed in this respect will be presented. In a second part, our solution will be explained and the most important results in terms of speech quality and speech intelligibility will be shown.

3.4.1 Musical noise

Musical noise is a perceptual phenomena that takes place when strong fluctuations in the time-frequency domain appear after processing with a spectral subtraction type algorithm. An example of work dealing with this problem is [Cappé, 1994]. Musical noise exhibits random and isolated peaks that are much larger than the surrounding energy and/or short ridges in the spectrum. The name of musical noise comes from the fact that isolated noise energies sound like narrowband tones to our ears.

Musical noise is a problem to be considered in many algorithms that work with sound signals. Illustrative examples of it are some single-channel speech enhancement algorithms with spectral subtraction. In our case of study, we focus on speech separation algorithms that use time-frequency binary masks to perform the separation, what can be considered as a sort of speech enhancement algorithms. Time-frequency binary masking can be considered as a subtraction type algorithms since it uses binary gains, which give rise to the musical noise artifacts.

There are different solutions for removing musical noise and some of the most known are: a) methods that use non-binary masks (e.g., masks estimated using l_1-norm minimization), b) solutions based on a finer shift Hanning window for the STFT and, c) techniques that use overlap-add methods for reconstructing the separated signals. Examples of works using non-binary masks are [Li et al., 2001, Araki et al., 2006], specifically, the objective of these two proposals is to soften the output sound signal by means of constant cutoffs at both low and high ends. On the other hand an example of temporal smoothing using smaller frame shifts is the proposal in [Makino et al., 2005]. The work in [Madhu et al., 2008] uses temporal smoothing in which the smoothing procedure is developed in certain frequency regions in the cepstral domain. Nowadays, the proposals tend to combine smoothing and sigmoidal masking.

Spectral smoothing techniques have demonstrated to reduce the amount of musical noise

successfully. Nevertheless, these techniques do not eliminate it and they cause other distortions. A possible solution is to combine smoothing techniques with other approaches. One of these approaches is parametric spectral subtraction which allows us to extract the noisy signal while the musical noise is masked. Another relevant approach is musical noise filtering. In this approach, it is necessary to estimate the speech presence or absence at each fame, in this way, a window is derived to filter the gain function to reduce the musical noise. The method proposed in this chapter of the thesis can be classified as spectral smoothing procedure technique since it works with the time-frequency representation of the separated speech sources to reduce the presence of musical noise artifacts.

It must be said that the vast majority of solutions within speech separation focus on suppressing the musical noise considering the quality of the separated sources, but not speech intelligibility. For instance, in [Makino et al., 2005] subjects are conducted to evaluate the amount of perceived musical noise. The evaluation of methods for attenuating the musical noise using speech quality subjective tests is very difficult since many factors related to the listener must be considered. The contribution in [Anzalone et al., 2006] considers different aspects of the listeners. Moreover, sometimes, solutions that improve speech quality degrade speech intelligibility. Hence, the solution to reduce the musical noise proposed in this chapter has been evaluated considering their impact both on speech quality and speech intelligibility. For this reason, its performance is tested using SIR (quality parameter) and Correct words (%) (speech intelligibility measurement).

3.4.2 Mixing different windows

The basic idea of our proposal is to carry out speech separation using different window sizes in the separation stage of our separation algorithms, that is, T-F binary masking with different window sizes and averaging the outputs of the separated sources obtained for each window size. Since the shape of the musical noise depends on the window size, each separation system will obtain a slightly different added noise signal. Thus, mixing different musical noises makes the final perceived noise more gaussian-like (due to the Central Limit Theorem [Rice, 2006]). It means that the separation algorithms will be executed several times (one per window size). This proposal is of particular interest with regard to mixing matrix estimation methods that use time-differences since they do not have source permutation problems. Carrying out a large number of experiments, two versions of our method have been chosen, depending on which windows are used:

- **MIX 3L**. In this version window sizes equal to 512, 1024 and 2048 samples have been used, corresponding to 64, 128 and 256 ms ($f_s = 8$ Khz). These widows have been chosen since they obtain the best results in the cases studied in this chapter.
- **MIX 7L**. Window sizes L equal to 512 (64 ms), 640 (80 ms), 704 (88 ms), 1024 (128 ms), 1280 (160 ms), 1408 (176 ms) and 2048 (256 ms) samples have been used. As mentioned in MIX 3L, these window sizes get the highest outcomes in terms of speech quality.

By way of illustration, different results when our proposal for reducing musical noise is used are presented in the following figures and table. Aiming at considering speech quality, Figure 3.14 displays values in terms of SIR, but in oder to do a complete evaluation of our musical noise reduction method, results in terms of speech intelligibility (Correct words (%)) are also shown in Figure 3.15 and Table 3.16.

In Figure 3.14 results in terms of SIR for DUET with binary mask, Proposal 1 and Proposal 2 are shown.

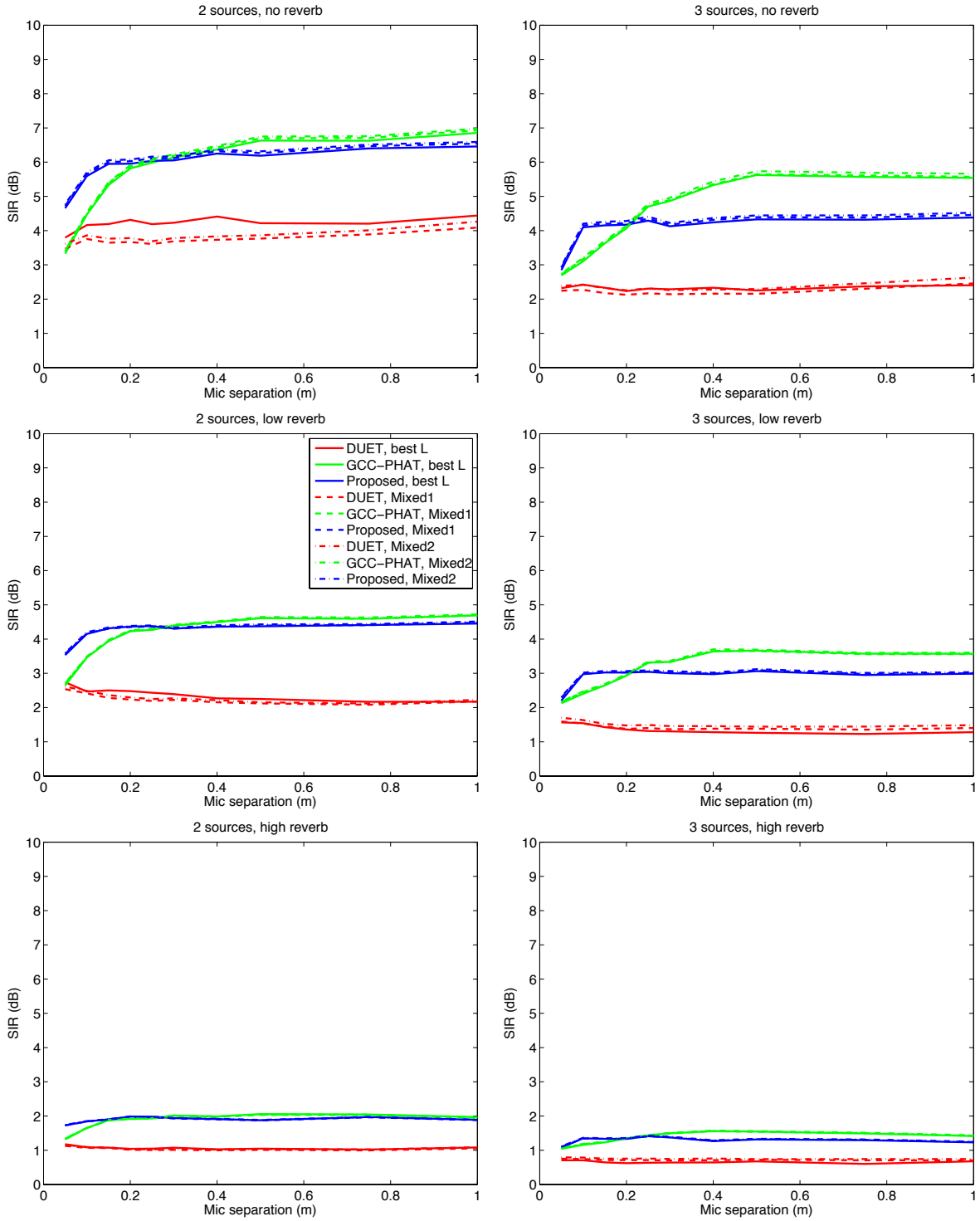

Figure 3.14: SIR (dB) for the DUET with binary mask, the GCC-PHAT based and the proposed optimization based methods with MEMS microphones at 60° in function of the window strategy.

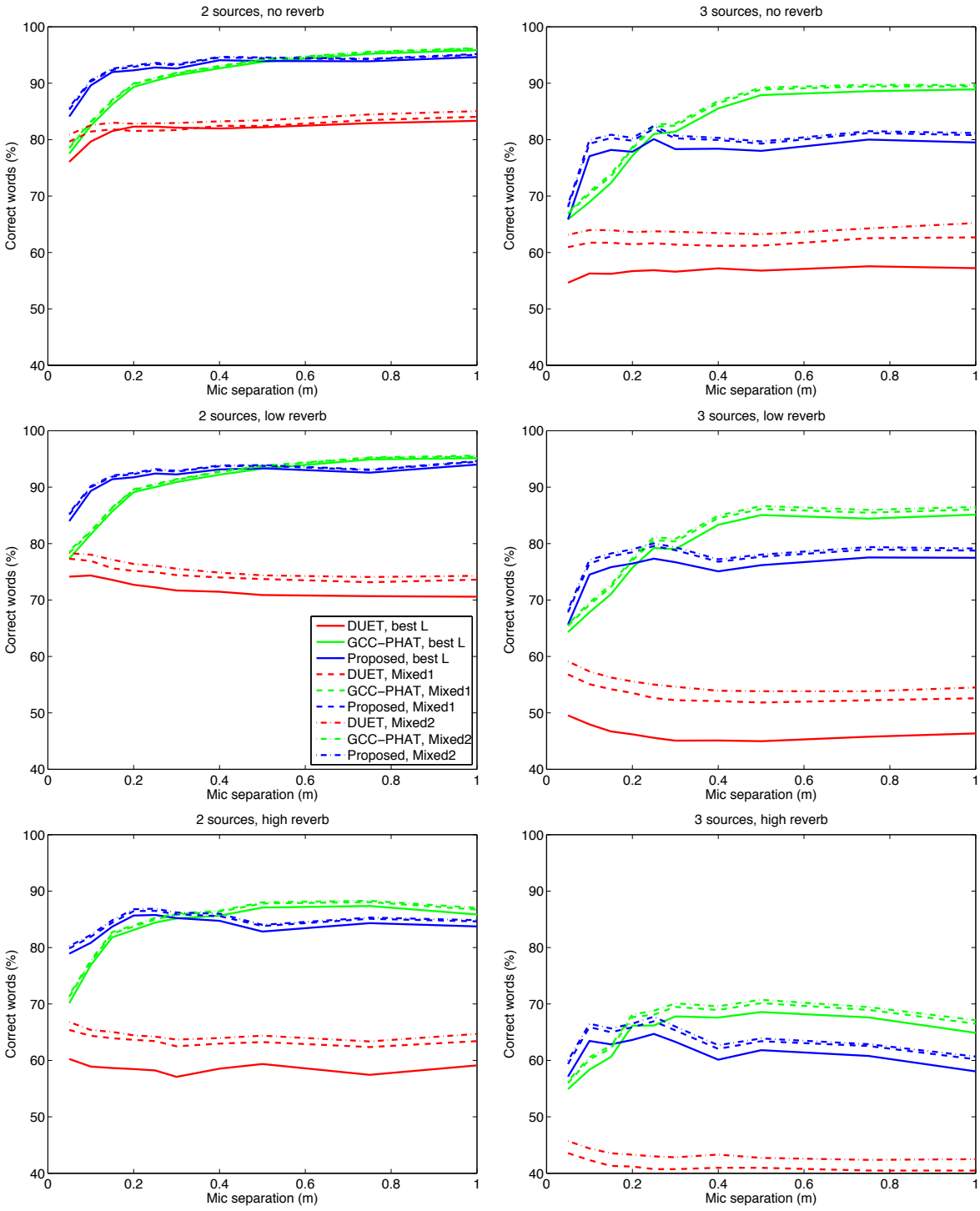

Figure 3.15: Percentage of correct words(%) for the DUET with binary mask, the GCC-PHAT based and the proposed optimization based methods with MEMS microphones at 60° in function of the window strategy.

Method	2 sources											
	No reverberation				Low reverberation				High reverberation			
D (m)	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75
DUET best L	76.1	82.3	82.0	82.9	74.1	72.7	71.5	70.7	60.3	58.5	58.5	57.4
DUET Mix 3L	79.7	81.5	82.5	83.5	77.3	75.2	74.0	73.1	65.4	63.6	63.0	62.4
DUET Mix 7L	80.9	82.8	83.2	84.5	78.3	76.4	74.9	74.1	66.8	64.4	64.0	63.4
GCC best L	77.5	89.4	92.6	95.2	77.4	89.1	92.2	94.9	70.1	83.1	85.6	87.4
GCC Mix 3L	78.5	89.8	93.0	95.5	78.4	89.5	92.7	95.2	71.3	83.7	86.4	88.0
GCC Mix 7L	78.8	90.0	93.1	95.6	78.7	89.6	92.8	95.3	71.6	83.9	86.5	88.3
Prop best L	84.1	92.3	94.1	93.9	84.0	91.7	93.1	92.6	78.9	85.7	84.7	84.3
Prop Mix 3L	85.3	92.9	94.6	94.2	85.1	92.3	93.7	93.0	79.8	86.5	85.6	85.1
Prop Mix 7L	85.7	93.2	94.7	94.3	85.4	92.5	93.9	93.1	80.2	86.8	86.0	85.3
Best Case	3 sources											
	No reverberation				Low reverberation				High reverberation			
D (m)	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75	0.05	0.20	0.40	0.75
DUET best L	54.6	56.7	57.2	57.6	49.5	46.2	45.1	45.8	36.5	35.0	35.2	34.2
DUET Mix 3L	60.9	61.4	61.2	62.5	56.8	53.5	52.1	52.2	43.6	41.2	41.0	40.5
DUET Mix 7L	63.1	63.6	63.4	64.3	59.1	55.6	53.9	53.8	45.7	43.3	43.3	42.4
GCC best L	65.9	77.1	85.5	88.6	64.3	75.7	83.3	84.4	54.9	66.1	67.6	67.6
GCC Mix 3L	66.8	78.4	86.5	89.4	65.4	77.0	84.5	85.5	56.0	67.7	68.9	68.9
GCC Mix 7L	67.1	78.7	86.9	89.7	65.7	77.4	84.9	86.0	56.3	68.1	69.6	69.4
Prop best L	65.8	77.8	78.4	80.0	65.6	76.5	75.1	77.5	57.1	63.6	60.1	60.8
Prop Mix 3L	68.1	79.8	79.9	81.2	67.8	78.5	76.8	78.9	59.4	65.9	62.0	62.5
Prop Mix 7L	68.5	80.3	80.3	81.5	68.3	79.0	77.2	79.4	60.0	66.5	62.7	62.9

Table 3.16: Percentage of Correct words (%) estimated for different methods in function of the window strategy, using Mems microphones and an angle of 60°.

The reader can note that these separation algorithms are applied with the window size that provides best results (labeled "best L" in figures) or with the two versions of our musical noise reduction method (labeled as Mixed 1 and Mixed 2 in figures). Beginning with the analysis of speech quality, it can be observed in Figure 3.14 that our method (dashed lines) increases the quality of the separated speech sources slightly regardless the level of reverberation, the number of sources or the distance between microphones. Indeed, sometimes DUET with binary mask achieve worst results when this technique is used, but it is not a problem since the use of our separation methods is preferable because they obtain better results than DUET with binary mask in all the situations.

Considering Correct words (%), that is, speech intelligibility, it can be observed in Figure 3.15 that our musical noise reduction method introduces improvements for all the reverberant environments, separation methods and distances between microphones. The most meaningful improvements occur when the number of sources is three, what it is important since it is a underdetermined problem and so, the separation task is more difficult. These improvements are more evident for the case of DUET with binary mask. In oder to see it in a more detailed way, these results are presented in numeric form in Table 3.16. Having a look at this table, it seems clear that the best results are reached for the version Mix 7L of our musical noise reduction method, that is, the larger the number windows used is, the higher the speech intelligibility is.

3.5 Discussion

This chapter deals with the cocktail party problem when classical microphone arrays are used, that is, with microphone distances in the order of centimeters. In this sense, a new mixing matrix estimation method has been proposed. This method has been designed to perform the separation issue with only two microphones, the simplest microphone array. Not very sophisticated microphone arrays are needed, in contrast to many BSS and beamforming techniques. By means of establishing a theoretical relationship between time and level differences, it is only necessary to obtain time differences for estimating the mixing matrix. This aspect is very important because in many speech separation algorithms to determine the level differences is not a trivial task. Several factors, as for instance reverberation or the fact that speech sources can only be considered as quasi-sparse, make more difficult to estimate level differences correctly. Results have demonstrated that our proposal outperforms the DUET algorithm in many situations in terms of SNR, STOI and Correct words (%). Examples of the suitability of our proposal are the results when the true delays are used, where SIR reaches values higher than 7 dB in absence of reverberation or STOI scores are higher than 0.7 in high-reverberant environments.

The correct estimation of time differences plays a crucial role in our mixing matrix estimation method. In this sense, GCC-PHAT has demonstrated its suitability in reverberant environments but our proposed TDE method outperforms it with short distances between microphones. For instance, for $D = 0.05$ m, the value of SIR increases from 3.9 dB to 6.2 dB for absence of reverberation, from 3.0 to 4.3 dB for low reverberation and from 1.6 to 2.1 dB for high reverberant environments. For example, improvements in Correct words (%) higher than 9 % have been obtained in high-reverberant environments.

Finally, in order to reduce the presence of musical noise artifacts a smoothing method is proposed. The influence of its use on speech intelligibility has been studied. Using the classical separation method, DUET, improvements in terms of Correct words (%) higher than 8 %, 9 % and 9 % for absence of reverberation, low and high reverberation have been achieved, respectively. In the case of Proposal 2, Correct words (%) has increased by more than 8 % in all the considered reverberant environments.

3.6 Summary of contributions

The main contributions described in this chapter of the thesis are the following:

- The performance of the so-called DUET algorithm has been evaluated in a variety scenarios including anechoic and echoic rooms, determined and underdetermined separation problems with different array configurations. With respect to the two-microphone array, elements such as the distance between microphones, the mutual angle or the polar pattern of the microphones (omnidirectional, cardioid with different values of FBRs, one MEMS microphone) have been varied.

- The polar pattern of real microphones (MEMS) has been analyzed in an anechoic chamber and introduced in our separation problems. The fact of introducing real microphone responses leads us to think about the application of our proposal in real scenarios.

- Considering the aforementioned scenarios, a comparative study of two of the most popular separation stages based on time-frequency masking has been carried out. One of these stages is based on time-frequency binary masking and the other option calculates time-frequency masks using l_1-norm minimization. The choice of separation stages based on

time-frequency masking is due to they are able to separate sources when the number of sensors is lower than the number of sources. It has been demonstrated that the separation stage based on binary masking outperforms the other methodology in all the studied scenarios. It must be mentioned that this comparative study has been made considering both speech quality and speech intelligibility. Furthermore, the type of two-microphone array with better separation performance has also been studied.

- The main aspect of this chapter is the proposal of a novel mixing matrix estimation procedure based on a geometric model. This new methodology requires to know some information about the microphone array such as the polar pattern of the microphones, their mutual angle or the microphone separation. This requirement is very realistic and it can be easily fulfilled since whether a microphone array is used in a real room, their characteristics will be known in advance. Several experiments have shown that our mixing matrix estimation method outperforms the classical one (DUET) in almost all the scenarios. Additionally, it only requires to estimate time differences and not level differences, what involves lower computational cost.

- Our mixing matrix estimation stage needs a good time difference estimation. In this sense, different experiments have been developed comparing classical TDE methods. Finally, we have concluded that GCC-PHAT is the most suitable one within cocktail party problems since it shows a special robustness in the case of echoic mixtures.

- Another important contribution in this chapter is the implementation of a new TDE algorithm aiming at overcoming the limitations of GCC-PHAT when distances between microphones are short. It estimates the delays in the frequency-domain instead of in the time domain. Different experiments have been carried out and it has been proven that our TDE method obtains slightly better outcomes in terms of speech quality and noticeable improvements with respect to speech intelligibility for small arrays (less than 0.2 m between microphones).

- Perhaps, the third most important contribution of this chapter is related to the separation stage in sound source separation algorithms. A novel musical noise reduction method that has the advantage of considering both speech quality and speech intelligibility has been introduced. The outcomes obtained confirm that our proposal helps to minimize the presence of musical noise.

In [Llerena et al., 2013a, Llerena et al., 2013b] some of these contributions can be found. Furthermore, two articles are currently being prepared as explained in Section 5.3.

Chapter 4

Synchronization based on mixture alignment for sound source separation in wireless acoustic sensor networks

4.1 Introduction

Desynchronization degrades the performance of many signal processing algorithms in wireless acoustic sensor networks. This lack of synchronization in WASNs is mainly caused by 1) the different distances between the source and each node, and 2) by the clock phase offset and frequency skew. As mentioned in Chapter 1, classical solutions use clock synchronization protocols and algorithms in the communication layer, but these classical alternatives do not tackle the lack of synchronization caused by the distances between sources and nodes.

In this chapter, we present a novel study of the synchronization problem in acoustic sensor networks from a signal processing point of view. First, we propose a theoretical framework that allows us to study the effects of misalignment over any short-time based algorithm, focusing on the requirements of the effective length of the analysis time frame. From this framework, a theoretical synchronization delay is established aimed at reducing the required length of the time frame. Second, different classical time delay estimation methods are studied in order to analyze its suitability for aligning speech mixtures. Finally, two novel alignment methods are developed and are tuned up to reduce the amount of synchronization information required for transmission. The results obtained demonstrate that the latter two proposed methods represent a good solution in terms of performance over the quality of a standard BSS algorithm, allowing us to reduce the transmission bandwidth required for synchronization data.

4.2 Proposal of synchronization in BSS problems

As previously mentioned, wireless communication-based solutions only consider the clock phase offset and clock frequency skew (the clock problem), but they omit the effects of unknown propagation delays on the performance of separation and localization algorithms. In contrast, our solution also deals with this problem in BSS from a different perspective.

In the case of non-stationary sources such as speech, signal processing algorithms are usually implemented using short-time analysis tools. These tools are based on segmenting the analyzed signals into time frames because it is assumed that the signal is considered nearly stationary in short-time frames. Desynchronization of mixtures entails that there would not be the same source contributions in simultaneous time frames at the different microphones, which can decrease the

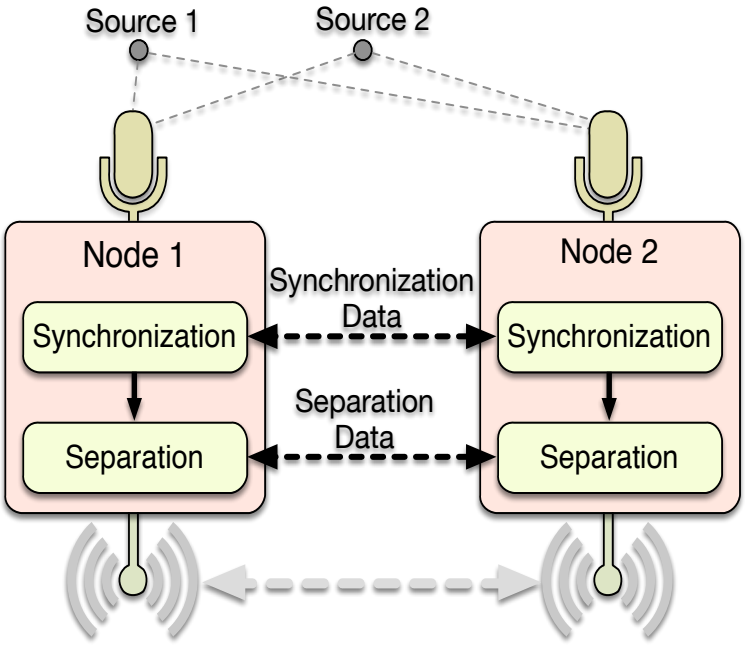

Figure 4.1: Scheme of a WASN for sound separation, including the mixture alignment stage proposed in this chapter.

performance of the algorithms. One possible solution consists of increasing the length of the time frames as it was proposed in [Sur et al., 2014], but in many cases, the required length is not suitable for analyzing non-stationary sources.

Some approaches suggest that it is possible to synchronize the mixtures using signal processing techniques by analyzing and comparing the signals received at the microphones. This idea has been explored for long recordings to correct the clock skew and offset in the time-frequency domain as it was done in [Miyabe et al., 2013b]. However, the design of systems working in wide areas with moving sources and the analysis of WASN particularities for these algorithms are still open issues. Discussions about these problems within signal separation can be found in the contribution [Ono et al., 2013].

BSS algorithms usually assume that the mixing process is modeled as a linear and time-invariant system. The impulse responses from the sources to the sensors are used to model the different phenomena, for instance, delays of the signals [Bofill and Zibulevsky, 2001], path loss or multipath effects [Fabrizio and Farina, 2011]. We are interested in BSS in WASNs. Figure 4.1 shows a block diagram of a WASN with two nodes, detailing the synchronization block, source separation block, and communication links.

As we can see in Figure 4.1, the first block is the synchronization block, which is the block this chapter of the thesis addresses. In a traditional approach, this block is mainly focused on clock and sampling synchronization, which can be solved using wireless network timing. A representative example of this kind of traditional solutions is [Wu et al., 2011]. In those applications in which source localization is required, clock timing synchronization is important because it controls the precision of the results. In the case of BSS algorithms, precise positioning of sources and sensors is not needed, but a good characterization of the mixing process is required. Thus, in the BSS case, clock synchronization can be replaced by any other type of synchronization that guarantees a good separation performance.

To obtain an alternative approach, in this chapter, we present a novel theoretical study of this block under a novel perspective based on a signal processing framework. Our solution addresses the synchronization problem of WASN nodes from a signal processing point of view, which can be divided into different important parts:

1. First, the characteristics of the different scenarios used to compare the different strategies are described, with the aim of studying the influence from the propagation delays, independently of the use of clock synchronization.

2. Second, a theoretical analysis is developed and presented, which allows a study of the requirements of the analysis frame length in a BSS system as a function of the delays caused by clock desynchronization and acoustic propagation. Specifically, a lower bound of the time frame length is obtained, which helps us to determine how alignment of the mixtures must be performed, minimizing the length constraint.

3. In a third part and according to this theoretical analysis, classical synchronization algorithms based on the use of time-delay estimation methods are studied. The most suitable of them will be chosen as reference synchronization method.

4. In the fourth part, once a lower bound of the frame length is established, two novel synchronization methods are proposed based on this theoretical analysis, which allows us the use of shorter analysis time frames for BSS without decreasing the performance. One method is based on the cross correlation of the mixtures, and the second method uses the short-term log-energy (STLE). These solutions are specifically designed for WASNs, analyzing the distributed and collaborative processing in terms of both computational cost and transmission bandwidth. In some proposals such as [Tatlas et al., 2015], compressing algorithms are proposed to reduce bandwidth usage in WASNs. The proposed solutions can avoid the problem of clock desynchronization, and they also improve the performance of separation algorithms over clock-synchronized systems for WASNs with larger distances between nodes.

4.3 Experimental setup for WASN scenarios

This section presents the description of the set of experiments performed to validate and compare the performance of the proposed mixture alignment methods. These experiments have been specifically designed to illustrate scenarios in which there will be influence from the propagation delays, independently of the use of clock synchronization. For this purpose, wide area WASNs are simulated where sources are placed close to one of the nodes.

The speech signals used in the experiments have been extracted from the IEEE database constructed in [Hu and Loizou, 2007] at a sampling frequency $f_s = 8000$ Hz. These signals have been simulated at 60 dB in 45 different experiments in which two or three speech sources ($S = 2$ or $S = 3$) have been located (in different locations) in rooms of different size and reverberation conditions with a background noise of 40 dB. The room dimension has been randomly varied from 15 $m \times 15$ $m \times 3$ m to 20 $m \times 20$ $m \times 5$ m, with three different reflection coefficients ($Cr = 0$, $Cr = 0.3$ and $Cr = 0.6$). These three values have been selected in order to consider three different groups of scenarios: one without reverberation ($Cr = 0$), one with low reverberation ($Cr = 0.3$) and one with high reverberation ($Cr = 0.6$). With these values, the reverberation time RT_{60} has been calculated, obtaining average values of 0 ms, 334 ms and 498 ms with a standard deviation of 0 ms, 30 ms and 23 ms for each scenario, respectively.

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
RMSE (ms)	41.9	34.7	41.9	34.7	42.0	34.7	41.6	34.9	41.6	34.9	41.7	34.9
SIR (dB)	2.60	2.60	2.64	2.64	2.18	2.18	1.61	1.61	1.61	1.61	1.36	1.36
STOI	0.67	0.67	0.69	0.69	0.68	0.68	0.57	0.58	0.59	0.59	0.58	0.58
Words (%)	82.2	82.2	86.5	86.6	84.6	84.6	55.8	56.2	59.7	60.1	56.6	56.7

Table 4.1: Parameters of the different scenarios considered in the thesis for WASN.

In every room, two ($M = 2$) or three ($M = 3$) microphones have been placed in the superior corners of the room, and each microphone is pointing toward the center of the room. Aiming at simulating scenarios as real as possible the microphones used in the experiments have a MEMS response, as described in Section 2.5.5. These MEMs responses were measured and characterized in an anechoic chamber.

As was stated above, we focus on the cases where the differences in the acoustic delays between speech mixtures influence the performance of the BSS. Specifically, these areas correspond to situations where speakers are relatively close to one microphone and far away from the others. If speakers were close to the middle of the room, the difference in delays between speech mixtures would not be very relevant. In this sense, we consider that all of the speakers are placed randomly in an area close to one of the microphones. The distance between this microphone and sources ranges from 2.70 to 8 m, and the distance between both speech sources varies from 1.7 to 4.8 m.

At most, four seconds of the speech signal have been taken into account in the separation process. Both the synchronization and the separation have been implemented using information from four-second time recordings.

Once the signals are synchronized, they are separated using standard BSS methods that segment speech signals into time frames. The frame length $f_s \cdot K$ of the separation algorithm has been varied from 32 to 512 ms in the experiments. To measure the quality of the separated signals, we have considered SIR, STOI and the estimated percentage of correct words, Correct words (%). The RMSE of the delays caused by the variations in the position of the sources with respect to the microphones has also been evaluated.

Table 4.1 shows the values of the different scenarios, and the quality values obtained over the mixtures without applying any separation algorithm. This table may be very useful since comparing their results with the ones obtained by the different separation solutions, the improvements introduced by these solutions can be quantified. Considering both the speech quality parameter (SIR) and the speech intelligibility parameters (STOI and Correct words(%)), it is easy to see that they reach low values regardless the number of sources or the level of reverberation.

Concerning the algorithms used for separating the audio sources, in this chapter we focus on two different algorithms:

- DUET with binary mask which was proposed in [Rickard, 2007]. This will be the chosen classical separation algorithm since from the study carried out in Chapter 3, it was demonstrated that it obtains quite good separation quality in some scenarios and higher quality than, for instance, separation stages based on the use of l_1-norm minimization to calculate smooth time-frequency masks.

- Proposed GCC-PHAT based mixing matrix estimation method with BM. To evaluate this proposal of the thesis, described in Section 3.3.3, makes sense since its suitability has been studied when microphone separations of a few centimeters are used but not for larger

distances between microphones. It is important to highlight that since the microphones are not close to each other, time delay estimation is carried out in time domain, using the methodology described in Section 3.3.3 (in this case, the proposed frequency domain methodology is not suitable). Furthermore, the estimation of the level differences is not carried out using the approximation proposed in Chapter 3 (geometric model), since again its validity requires small arrays. In this chapter, we have not considered level differences in the matrix estimation for the proposed CGG-PHAT and therefore all the estimated matrix coefficients have modulus equal to one. On the other hand, the proposed method was described to work with only two mixtures. The generalization of the method for three or more mixtures is not trivial, and the next procedure have been carried out for that purpose:

1. First, sources are estimated for any given pair of microphones using the described GCC-PHAT with a window of 256 ms. As it was described, level differences are not considered.
2. The obtained sources are then correlated one each other, and the combination that gets a higher average correlation is selected.
3. Source delays from each pair of sources is combined, taking into account the combination selected in the last step. Signals are then separated using the estimated mixing matrix, again without considering level differences.

Additionally, the strategy for window selection will follow a similar schema to the one described in Chapter 3. Thus, the use of a unique window length will be compared to the combination of three different windows (labeled as MIX 3L) and to the combination of seven different windows (labeled as MIX 7L). For further information, the reader can found a description of these combinations of windows in Section 3.4.2.

In the following four tables the behavior of the two separation algorithms is analyzed when microphones are placed in the superior corners of the rooms, or in other words, when there are microphone separations larger than a few centimeters. Since the speakers are close to one microphone and far from the rest of them, the information extracted from the tables will be very useful because they show how the desynchronization due to different propagation delays affect the algorithms. In the experiments it is assumed that the clock synchronization problem has been resolved. Tables 4.2 and 4.3 show the separation quality achieved by DUET with binary mask in terms of SIR and Correct words (%), respectively. The values obtained correspond to experiments with different number of sources ($S = 2$ or 3), different number of nodes/microphones ($M = 2$ or 3) and different frame lengths. Having a look at Table 4.2, at first glance, it can be mentioned that DUET with binary masking obtains lower values of SIR than in the cases depicted in Table 3.5, what is expected due to the synchronization problem. Using only one frame length, the best results are reached for $Kf_s = 128$ ms and 256 ms, in contrast to the results in Table 3.5, where 64 ms is the most suitable frame length. Larger frame lengths are required since desynchronization prevents the same source contributions in simultaneous time frames at the different microphones. The only way to have the same source contributions is to use larger frame lengths, but there is a limit related to the non-stationarity of speech sources. This situation lead us to develop a new synchronization methodology. As in Chapter 3, our proposal of combining different window sizes outperforms the rest of cases. Mainly, the best results are obtained for the case of mixing seven window configurations, case MIX 7L in table. These results confirm that our proposal minimizes the presence of musical noise artifacts in the separated speech sources.

With respect to the results in Table 4.3, very similar conclusions can be extracted from them. The most suitable frame length is 256 ms for all the cases except for the case of three sources and

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	2.38	2.37	2.44	2.43	2.07	2.02	1.42	1.44	1.39	1.42	1.19	1.22
$Kf_s = 64$ms	2.50	2.74	2.51	2.65	2.11	2.14	1.52	1.67	1.49	1.62	1.24	1.29
$Kf_s = 128$ms	2.74	3.24	2.67	3.05	2.08	2.12	1.61	1.83	1.61	1.68	1.24	1.27
$Kf_s = 256$ms	2.82	3.42	2.69	3.25	2.00	2.15	1.70	1.93	1.58	1.74	1.16	1.22
$Kf_s = 512$ms	2.72	3.17	2.62	3.07	1.93	2.09	1.57	1.72	1.46	1.56	1.10	1.14
MIX 3L	3.11	3.76	3.00	3.53	2.25	2.40	1.94	2.26	1.85	2.04	1.40	1.47
MIX 7L	3.20	3.94	3.08	3.69	2.29	2.44	2.02	2.39	1.93	2.15	1.47	1.54

Table 4.2: SIR (dB) for the DUET algorithm with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of using clock synchronization.

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	75.0	73.3	80.8	78.5	79.6	77.6	40.7	40.7	45.0	45.5	42.9	43.8
$Kf_s = 64$ms	78.0	79.3	82.4	82.7	81.4	79.4	47.5	51.0	51.6	55.0	49.0	50.4
$Kf_s = 128$ms	80.0	85.9	84.4	88.3	80.9	81.8	51.3	57.0	55.9	59.3	51.9	52.7
$Kf_s = 256$ms	81.5	87.1	84.9	89.5	80.9	84.0	53.8	58.9	56.2	60.5	50.6	51.6
$Kf_s = 512$ms	78.6	83.7	83.4	87.3	79.6	82.1	49.9	52.5	52.8	55.6	47.5	49.0
MIX 3L	84.7	89.9	88.0	91.9	84.2	86.1	59.3	65.4	62.9	67.4	58.0	59.6
MIX 7L	86.1	91.2	89.3	92.7	85.2	87.0	62.2	69.0	65.9	70.6	60.7	62.8

Table 4.3: Percentage of correct words (%) for the DUET algorithm with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of using clock synchronization.

high reverberation, where using a frame length equal to 128 ms the best outcomes are achieved. As for the case of SIR, the highest scores of Correct words (%) are reached combining seven windows. Values higher than 90 % and 70 % are obtained for two and three talkers in the room, respectively.

Tables 4.4 and 4.5 present the outcomes obtained by our GCC-PHAT based mixing matrix estimation method with Binary Mask. Comparing with the two previous tables, very similar conclusions can be extracted. A frame length equal to 256 ms is the most appropriate in the vast majority of cases in terms of both SIR and Correct words (%). Unlike DUET with binary mask, our MIX 7L achieves the highest scores in many situations but not in all. Observing SIR and Correct words (%) values, our separation method obtains important improvements with respect to DUET for $Kf_s > 128$ ms when $S = 2$, and for $Kf_s > 128$ when $S = 3$.

4.4 Theoretical analysis of the frame length

In this section, limits for the frame length are established by considering our novel theoretical analysis of the presented problem. BSS algorithms based on the STFT ideally work with STFT samples of the available mixtures that contain information on the same parts of the sources. Because the sources are affected by different delays in their paths to the microphones, they should be aligned to maximize the amount of common information for the sources at the frames used as inputs of the BSS algorithm.

With this objective in mind, let us now focus on the mixing process in the time domain expressed with Equation (2.5), and its approximation in the T-F domain given by Equation

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	1.68	1.64	1.76	1.71	1.42	1.32	0.98	0.95	0.99	0.97	0.83	0.77
$Kf_s = 64$ms	2.14	3.73	2.24	3.95	1.73	2.80	1.25	2.65	1.28	2.67	0.99	1.78
$Kf_s = 128$ms	4.29	6.12	4.35	6.54	2.81	4.14	2.97	5.29	2.84	5.25	1.73	2.97
$Kf_s = 256$ms	5.19	6.53	4.96	6.89	2.79	4.02	3.86	5.75	3.33	5.64	1.75	2.89
$Kf_s = 512$ms	5.02	6.18	4.66	6.50	2.58	3.71	3.61	5.20	3.04	5.05	1.59	2.58
MIX 3L	4.57	6.23	4.56	6.58	2.91	4.20	3.33	5.45	3.10	5.38	1.88	3.10
MIX 7L	4.75	6.47	4.75	6.82	3.06	4.37	3.50	5.72	3.29	5.66	2.00	3.26

Table 4.4: SIR (dB) for the GCC-PHAT based mixing matrix estimation method with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of using clock synchronization.

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	60.4	59.3	67.6	67.6	62.9	62.7	28.3	28.0	30.9	31.2	26.7	25.3
$Kf_s = 64$ms	75.4	91.0	82.0	94.5	77.5	91.5	44.5	74.9	49.2	79.1	42.1	67.5
$Kf_s = 128$ms	93.5	97.7	96.0	98.8	92.2	97.1	79.9	94.4	83.5	96.2	69.6	87.8
$Kf_s = 256$ms	95.6	98.0	97.2	99.0	92.3	97.0	85.4	94.9	87.3	96.6	70.7	87.4
$Kf_s = 512$ms	94.3	97.2	96.2	98.5	90.6	96.0	81.3	92.4	83.4	94.4	66.4	83.4
MIX 3L	94.1	97.8	96.3	98.9	92.8	97.3	82.3	94.6	85.1	96.2	72.7	88.7
MIX 7L	94.8	98.1	96.8	99.0	93.8	97.6	84.4	95.4	87.2	96.9	75.8	90.2

Table 4.5: Percentage of correct words (%) for the GCC-PHAT based mixing matrix estimation method with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of using clock synchronization.

(2.12). As we can see, the amount of common information in the T-F domain increases with the length of the frames (window length), but at the same time, the performance of BSS algorithms based on the STFT decreases if the frames are too long.

In our case of study, the length of the frames is conditioned by the STFT, which is computed using the DFT of the windowed signal. The length of the DFT should be greater than or equal to the length of the signal in the time domain to avoid time-domain aliasing when computing the inverse DFT. If the analyzed signal is the result of filtering some source ($a_{ms}[n]$ represents the impulse response of this filter), the length of the DFT should be greater than the filter length. Assuming that the length of $a_{ms}[n]$ is equal to P_{ms}, the STFT frame size (K) must be greater than the maximum size of the resulting convolutions, that is, $K >> \max_{ms}\{P_{ms}\}$. This problem is known as the Circular Convolution Approximation Problem and was addressed in different works as for instance [Parra and Spence, 2000].

If we analyze in depth the meaning of $a_{ms}[n]$, we can see that it will contain a large number of coefficients close to zero. The position of the first term of $a_{ms}[n]$ that is clearly distinct from zero (denoted $d_{ms} \cdot f_s$) will be related to the propagation time δ_{ms} of the acoustic signal from the source to the microphone (direct path). At this point, it is important to highlight that for the separation algorithm, the important issue is not the global delay a given mixture suffers from but the difference between the delays of the different sources. Placing this condition into a mathematical form, the minimum values of δ_{ms} can be compensated by time shifts γ_s in the source signals, and therefore, they will not have an effect on the final separation system. Hence, the constraint required for the frame length is rewritten as:

$$K >> K_{min} = \max_{ms}\{P_{ms}\} - \min_{ms}\{\delta_{ms}\} f_s, \tag{4.1}$$

where f_s is the sampling frequency and δ_{ms} is related to the propagation delay of the direct path in the samples. Furthermore, we can define EL_{ms} as the effective length of $a_{ms}[n]$, so that $P_{ms} = EL_{ms} + \delta_{ms} f_s$. Thus, Equation (4.1) can be rewritten as Equation (4.2),

$$K >> K_{min} = \max_{ms}\{EL_{ms} + \delta_{ms} f_s\} - \min_{ms}\{\delta_{ms}\} f_s. \tag{4.2}$$

Assuming that the reverberation time is approximately the same in all positions of the room, the effective length is approximately constant ($EL_{ms} \simeq EL$). Then, we can replace EL_{ms} by EL in Equation (4.2), simplifying the constraint and obtaining:

$$K >> K_{min} \simeq EL + \left(\max_{ms}\{\delta_{ms}\} - \min_{ms}\{\delta_{ms}\}\right) f_s. \tag{4.3}$$

As we can see, the constraint is composed of two terms: one directly related to the reverberation time and another related to the range of variation of the values of δ_{ms}. Note that in this equation, δ_{ms} includes the effects of all of the time delays that can appear in our separation problem before applying the BSS method. It is easy to identify the most important delays that can be found in our case of study, so that $\delta_{ms} = \tau p_{ms} + \alpha_m + \beta_m$. The first one is the propagation delay (τp_{ms}) or, in other words, the required time for the signal to propagate from the s-th source to the m-th sensor. The second delay term (α_m) is caused by clock differences in the nodes of the WASN. Finally, the third one (β_m) is the delay due to the radio frequency (RF) communication between the microphones and the fusion center, which sends the reference signal to the other nodes for synchronization. In this paper, RF channels are considered because they do not interfere with the audio scene. Because this communication is delivered by electromagnetic waves, the propagation time is small and nearly constant. So, in our separation problem, β_m is negligible.

The constraint over the analysis window might lead us to think that the larger the frame length is, the better the performance of the separation systems is. Unfortunately, due to the non-stationarity of most of the sound sources, for instance, speech signals, the use of large windows cause a reduction in performance because speech can only be considered stationary in relatively short frames.

On the other hand, the constraint can be relaxed with a suitable alignment of the mixtures, which might reduce in some cases the value of K_{min}, allowing the use of a smaller window length. The delays between speech mixtures are defined as $\Delta_m, \forall m = 2, \ldots, M$ and are obtained with respect to a reference mixture (typically the first one), so that all of the mixtures are synchronized with respect to this mixture. Figure 4.2 shows an example of the comparison between the required K_{min} when two speech mixtures (of two sources arriving at two sensors) are synchronized. Once the mixtures are properly time shifted by Δ_m, the constraint of Equation (4.3) can be relaxed, as shown in Equation (4.4):

$$K'_{min} = EL + \left(\max_{m}\{\max_{s}\{\delta_{ms}\} - \min_{s}\{\delta_{ms}\}\}\right) f_s. \tag{4.4}$$

Therefore, the second term of the constraint is relaxed to the maximum range of variation of the delays in a mixture, instead of the maximum range of variations of the delays considering all of the sources and all of the mixtures (see Equation (4.3)). This new constraint that is obtained once the mixtures are properly aligned is less restrictive than the original one, which allows the use of a shorter window and could lead to an improvement in the separation performance. This fact makes synchronization useful for distributed WASN-based sound separation problems.

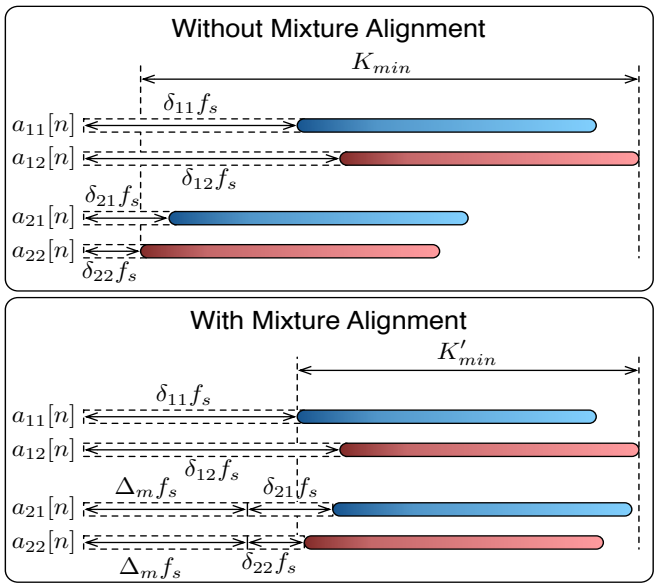

Figure 4.2: Examples of the effects of the alignment of the impulse responses over the minimum value of the frame length K_{min} in a case with two mixtures and two sources.

In order to propose a reliable estimator of the delays of the mixtures Δ_m, in this thesis we propose a Least Squared (LS) estimator. Let us define a squared error term that considers the degree of alignment of the information in the mixed sources once they have been shifted by 1) a source dependent alignment term γ_s and a mixture dependent alignment term Δ_m, so that $d_{ms} = \delta_{ms} + \Delta_m + \gamma_s$. So, the mean squared differences of the starting position of the sources can be expressed using (4.5).

$$E = \frac{1}{M^2 S^2} \sum_{s=1}^{S} \sum_{m=1}^{M} \sum_{r=1}^{S} \sum_{q=1}^{M} (d_{ms} - d_{rq})^2 \tag{4.5}$$

The term E measures the dispersion of the sources in the aligned mixtures, so that a low value of E implies a high level of synchronization of the information and, therefore, a low value of $K'min$. So, replacing d_{ms} in Equation (4.5), we obtain Equation (4.6).

$$E = \frac{1}{M^2 S^2} \sum_{s=1}^{S} \sum_{m=1}^{M} \sum_{r=1}^{S} \sum_{q=1}^{M} (\delta_{mr} - \delta_{qs} + \Delta_m - \Delta_q + \gamma_r - \gamma_s)^2 \tag{4.6}$$

So, our objective is to determine the values of Δ_m that minimize E. With this aim, we can now differentiate Equation (4.6) with respect to each mixture shifting value Δ_m and equaling the resulting to zero, obtaining a set of linear equations that can be resolved. After some simplifications, we can then obtain Equation (4.7).

$$\frac{\partial E}{\partial \Delta_m} = \frac{4}{M^2 S^2} \sum_{r=1}^{S} \sum_{s=1}^{S} \sum_{q=1}^{M} \delta_{mr} - \delta_{qs} + \Delta_m - \Delta_q + \gamma_r - \gamma_s = 0 \tag{4.7}$$

Thus, again simplifying and regrouping terms, we obtain a system of M equations, which m-th equation can be expressed using (4.8).

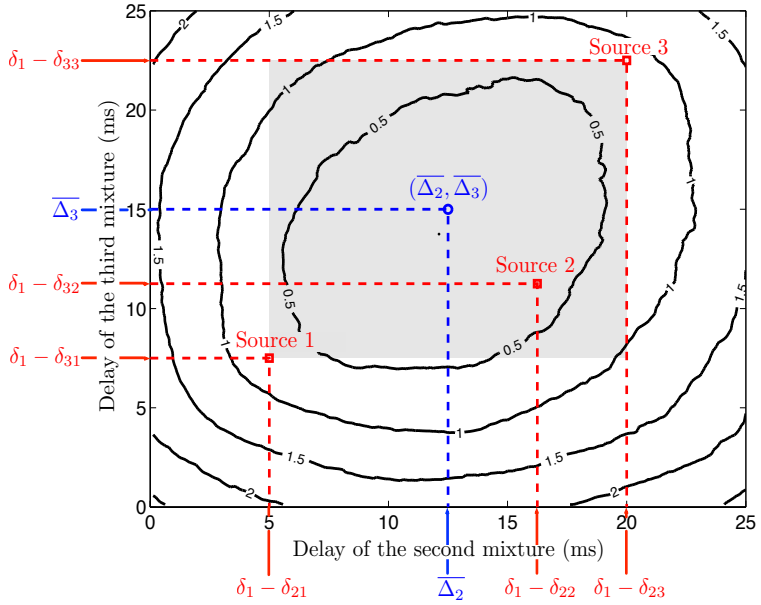

Figure 4.3: Contour plot of the loss in Signal-to-Interference Ratio (SIR) in dB obtained by the theoretical separation using binary masking with the true mixing matrixes in function with the delays introduced in the second and third mixture, Δ_2 and Δ_3. A window size of 64 ms has been used.

$$\Delta_m - \left(\frac{1}{M}\sum_{q=1}^{M}\Delta_q\right) + \left(\frac{1}{S}\sum_{r=1}^{S}\delta_{mr}\right) - \left(\frac{1}{MS}\sum_{q=1}^{M}\sum_{s=1}^{S}\delta_{qs}\right) = 0 \tag{4.8}$$

If we force now the delay of the first mixture to zero $\Delta_1 = 0$ and we focus on the first equation of the system of equations, we obtain Equation (4.9).

$$\frac{1}{M}\sum_{q=1}^{M}\Delta_q = \left(\frac{1}{S}\sum_{s=1}^{S}\delta_{1s}\right) - \left(\frac{1}{MS}\sum_{q=1}^{M}\sum_{s=1}^{S}\delta_{qs}\right) = 0 \tag{4.9}$$

And, finally, replacing Equation (4.9) in Equation (4.8) and rearranging the terms, we obtain the LS estimator of the mixture delays $\overline{\Delta_m}$, given by Equation (4.10).

$$\overline{\Delta_m} = \left(\frac{1}{S}\sum_{s=1}^{S}\delta_{1s}\right) - \left(\frac{1}{S}\sum_{s=1}^{S}\delta_{ms}\right) \tag{4.10}$$

As it was expected, the estimation of the LS delays depends exclusively on the average values of the terms δ_{ms}.

To perform an analysis of the consequences of selecting the difference of the average values, a set of experiments with three anechoic mixtures of three different sources has been performed. Figure 4.3 shows a contour plot of the loss in terms of SIR (dB) obtained by the theoretical separation using binary masking, as in the work [Rickard, 2007], with the true mixing matrixes in function with the delays introduced in the second and third mixtures, Δ_2 and Δ_3. A window size of 64 ms has been used, and the remaining features of the sources and signals are those described in the previous section. For the sake of simplicity, in this experiment, we used constant delays in the reference mixture, that is, $\delta_{1s} = \delta_1$. So, $\min_s\{\delta_{1s}\} = \max_s\{\delta_{1s}\} = \delta_1$. As we can see, the range of delays that achieves the minimum constraint (depicted as the dark area) matches the

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	4.00	5.02	3.30	4.25	2.26	2.58	2.08	2.68	1.84	2.22	1.35	1.51
$Kf_s = 64$ms	4.47	5.44	3.40	4.20	2.31	2.58	2.46	3.24	1.99	2.42	1.40	1.53
$Kf_s = 128$ms	4.08	4.96	3.27	4.07	2.12	2.41	2.35	3.03	1.86	2.27	1.27	1.36
$Kf_s = 256$ms	3.54	4.21	2.99	3.63	2.04	2.30	2.00	2.33	1.66	1.96	1.13	1.24
$Kf_s = 512$ms	2.87	3.39	2.74	3.18	1.94	2.12	1.58	1.77	1.44	1.59	1.04	1.12
MIX 3L	4.62	5.45	3.58	4.38	2.42	2.71	2.82	3.49	2.22	2.67	1.52	1.66
MIX 7L	4.75	5.60	3.64	4.47	2.48	2.75	3.00	3.71	2.31	2.79	1.58	1.74

Table 4.6: SIR (dB) for the DUET algorithm with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of aligning with the theoretical delay.

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	87.9	93.7	89.4	93.3	83.8	87.4	54.1	64.5	54.8	63.2	45.9	52.7
$Kf_s = 64$ms	91.4	95.0	92.4	93.8	86.8	88.7	63.8	76.6	65.2	71.2	55.2	58.4
$Kf_s = 128$ms	89.6	93.4	91.2	93.2	85.6	87.9	63.3	74.5	64.5	71.2	55.4	59.5
$Kf_s = 256$ms	85.7	90.2	88.6	91.4	83.7	86.5	57.3	64.4	59.1	64.8	52.3	54.8
$Kf_s = 512$ms	79.3	84.4	84.4	88.3	80.3	82.8	50.6	54.0	52.9	56.4	47.5	49.9
MIX 3L	92.4	95.1	93.3	94.6	89.2	90.8	70.6	80.0	71.2	76.3	63.4	66.5
MIX 7L	93.0	95.6	93.7	94.9	89.7	91.2	73.7	82.0	73.3	78.3	65.7	68.6

Table 4.7: Percentage of correct words (%) for the DUET algorithm with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of aligning with the theoretical delay.

region of the lowest SIR values theoretically obtained in the experiments. Furthermore, we can intuitively observe that the minimum SIR value is placed quite close to the point defined by the average delays. Repeating this experiment 45 times for all of the signals in the database with random delays of the sources, the root mean squared value of the difference between the delays corresponding to the minimum SIR and the proposed delays $\overline{\Delta_m}$ is 1.17 ms.

In order to further determine the effect of the proposed mixture alignment over the sound source separation experiments carried out in this chapter, Tables 4.6, 4.7, 4.8 and 4.9, and Figures 4.4 and 4.5 show the outcomes of DUET and of our GCC-PHAT based mixing matrix estimation method (both of them combined with a separation stage based on binary masking) when speech mixtures are aligned using our proposed theoretical delay. Note that the separation experiments are the same as the ones performed in Section 4.3, with the only difference of a mixture synchronization stage before applying the separation algorithms. In these experiments the theoretical delay will be used.

Tables 4.6 and 4.7 illustrate the separation quality obtained by DUET with binary mask. Comparing Tables 4.2 and 4.6 it is easy to see that using our mixture alignment (synchronization) stage, DUET with binary mask achieves highest quality results in almost all the situations. Considering the frame length, $Kf_s = 64$ ms is the most suitable one when our alignment stage is used, what leads us to think that our mixture synchronization is correct since shorter frame lengths are required. An example of these improvements with two sources and three microphones in low-reverberant conditions is the 4.20 dB ($Kf_s = 64$ ms) reached using our synchronization stage in contrast to the 3.25 dB ($Kf_s = 256$ ms) without applying it. As in the previous experiments, our MIX 7L proposal obtains the best outcomes. From the point of view of speech

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	5.25	6.76	5.15	7.08	3.04	4.30	3.50	5.70	3.21	5.53	1.87	3.09
$Kf_s = 64$ms	5.88	7.10	5.55	7.55	3.07	4.42	4.54	6.55	3.88	6.36	2.03	3.36
$Kf_s = 128$ms	5.91	7.01	5.40	7.45	2.88	4.23	4.77	6.47	3.82	6.25	1.90	3.17
$Kf_s = 256$ms	5.61	6.71	5.07	7.07	2.71	3.97	4.33	5.98	3.45	5.75	1.72	2.88
$Kf_s = 512$ms	5.11	6.23	4.68	6.55	2.54	3.67	3.70	5.24	3.05	5.02	1.58	2.59
3 windows	6.17	7.20	5.64	7.63	3.04	4.38	5.07	6.76	4.09	6.52	2.06	3.36
7 windows	6.29	7.26	5.72	7.69	3.06	4.42	5.23	6.88	4.17	6.63	2.09	3.41

Table 4.8: SIR (dB) for the GCC-PHAT based mixing matrix estimation method with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of aligning with the theoretical delay.

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	95.0	98.0	96.9	99.1	91.8	96.5	81.6	94.5	84.3	95.4	67.5	86.2
$Kf_s = 64$ms	96.9	98.6	98.1	99.4	93.1	97.2	89.6	96.7	90.7	97.3	73.7	90.0
$Kf_s = 128$ms	97.1	98.5	98.1	99.4	92.8	97.2	90.9	96.9	91.3	97.4	74.3	90.0
$Kf_s = 256$ms	96.3	98.2	97.5	99.1	91.8	96.7	88.2	95.7	88.5	96.4	71.1	87.8
$Kf_s = 512$ms	94.5	97.4	96.2	98.7	90.3	95.8	82.0	92.7	83.4	94.1	66.6	83.9
MIX 3L	97.4	98.7	98.3	99.4	93.5	97.5	92.0	97.3	92.3	97.7	76.7	91.1
MIX 7L	97.6	98.7	98.4	99.4	93.8	97.5	92.7	97.5	92.9	97.8	77.6	91.4

Table 4.9: Percentage of correct words (%) for the GCC-PHAT based method with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of aligning with the theoretical delay.

intelligibility (Correct words (%)) similar conclusions can be extracted from Table 4.7. Using a frame length equal to 64 ms the highest speech intelligibility is obtained, and our MIX 7L solution reaches the highest results. From the results in Tables 4.6 and 4.7 the reader can note that the classical separation algorithm, designed to work with microphone separations of the order of centimeters, can work with WASNs whether our synchronization stage is used. In this way, the synchronization problems due to different propagation delays are minimized.

Putting our attention on our GCC-PHAT based mixing matrix estimation method, the separation results are presented in Tables 4.8 and 4.9. Bearing in mind the results without our alignment stage (Tables 4.4 and 4.5), it can be mentioned that important improvements are achieved using our synchronization solution. In the case of SIR, higher values are obtained in practically any situations. For instance, these improvements are very evident in the case of $Kf_s = 32$ ms, where SIR values are lower than 2 dB and 1 dB for two and three speech sources, respectively, when our synchronization stage is not used (Table 4.4), and reaching values higher than 7 dB (two sources) and 5 dB (three sources) when our alignment stage is introduced. The separation algorithm performs best when $Kf_s = 64$ ms in all the cases except when $S = 2$ and $M = 2$, where $Kf_s = 128$ ms is the most suitable frame length. As in the previous cases, the combination of seven frame lengths (MIX 7L) gives rise to the best results.

In respect of speech intelligibility, the combination of the mixture alignment stage with the GCC-PHAT based mixing matrix estimation method with binary mask gets very good results for the different combinations of number of sources and microphones. As in the case of the results in terms of SIR, the highest values are obtained for $Kf_s = 64$ ms in the vast majority of cases.

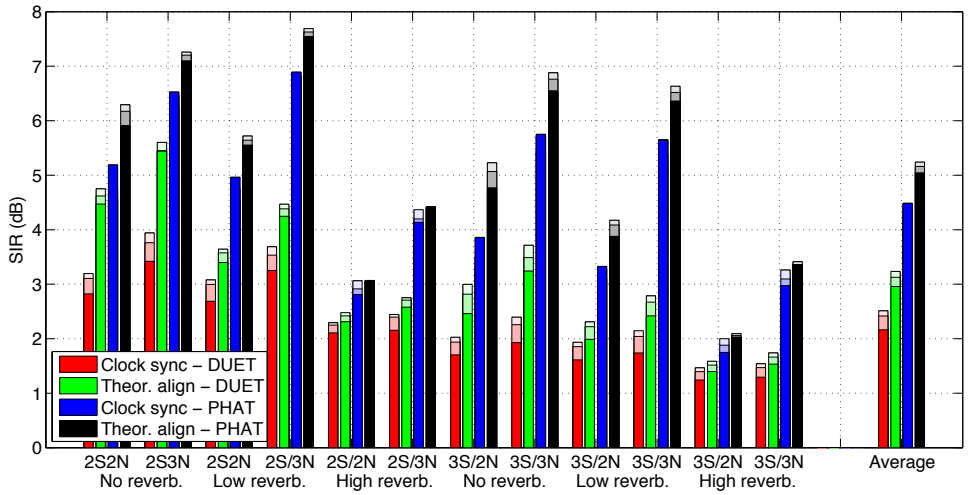

Figure 4.4: SIR (dB) of the DUET with binary mask and the proposed GCC-PHAT based mixing matrix estimation method with binary mask for the different scenarios included in the thesis, comparing the clock synchronization with the theoretical synchronization. The darker colors represent results with the best window length (best L), the mid colors represent results mixing three window lengths (Mix 3L) and the lighter colors represent results mixing seven window lengths (MIX 7L).

The solution MIX 7L minimizes the presence of musical noise artifacts successfully as it can be observed in Table 4.9. This proposal obtains Correct words (%) scores higher than 90 % for all the separation problems except in the case of three sources, two mixtures and high-reverberant conditions, where 77.6 % of words are understandable.

In order to summarize the information of the previous tables, Figures 4.4 and 4.5 depict their most relevant results for the different separation scenarios (number of sources and microphones, level of reverberation). Furthermore, these figures represent the results corresponding to the best frame length (darker colors), to MIX 3L (mid colors) and to MIX 7L (lighter colors). As previously mentioned, before applying our mixture alignment stage, it is considered that the clock problem has been resolved using one of the many clock synchronization protocols that can be found in the literature and for this reason, some values are denoted by Clock sync.

Looking at both figures different aspects can be highlighted. The first aspect to be mentioned is that the use of our mixture alignment stage increases the values of SIR and Correct words (%) independently of the separation algorithm. For the case of DUET, better results are obtained with the synchronization stage (green bars) than whithout it (red bars). In the case of the GCC-PHAT based method with binary mask the same conclusions can be extracted. Comparing DUET with our separation method it is clear that our method achieves better results than DUET in all the separation scenarios. Taking into account the frame length, it can be said that in general terms the best solution is obtained when the MIX 7L proposal is used (lighter colors).

To sum up, from the experiments carried out in this section it can be said that our alignment stage based on the theoretical analysis of the frame length helps separation algorithms to perform better speech separation. It must be considered that in the experiments theoretical delays have been introduced, what is not real since in real situations these delays will be unknown. Bearing in mind it, time delay estimation methods will be used.

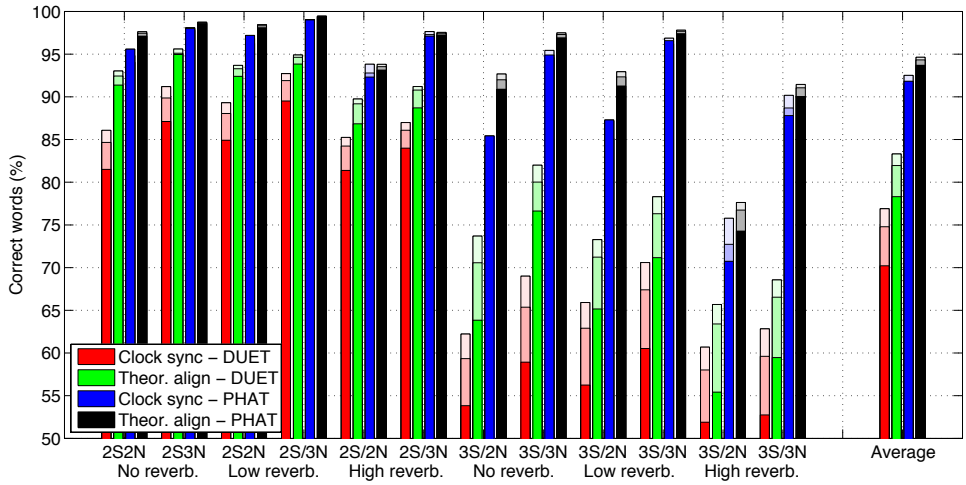

Figure 4.5: Estimated percentage of correct words(%) of the DUET with binary mask and the proposed GCC-PHAT based mixing matrix estimation method with binary mask for the different scenarios included in the thesis, comparing the clock synchronization with the theoretical synchronization. The darker colors represent results with the best window length (best L), the mid colors represent results mixing three window lengths (MIX 3L) and the lighter colors represent results mixing seven window lengths (MIX 7L).

4.5 Mixture alignment using the GCC-PHAT

Delving into the synchronization of speech mixtures from a signal processing point of view, different alignment solutions will be introduced. The alignment process that we propose consists of two parts: first, time delay estimation algorithms calculate delays between one speech mixture used as reference and the rest of them. Second, once these delays are calculated, all the speech mixtures are aligned with the reference one. In general terms, it must be considered that classical TDE methods are not specially designed to calculate delays between speech mixtures but rather delays between single speech signals, with at most background noise. Bearing it in mind, it may be analyzed if classical methods working with speech mixtures obtain delays according our theoretical analysis aiming at being suitable for our speech separation problem.

Sections 4.5 and 4.6 together with our theoretical analysis of the frame length are important contributions in this chapter of the thesis. Inasmuch as there is not specific delay estimation methods for speech mixtures, two methods have been designed and proposed. The first one is described in this section and it is a GCC-PHAT based estimation method. The second one, described in the next section, implies an important novelty since it uses a classical parameter which has never been used to calculate delays.

4.5.1 Adapting the GCC-PHAT to mixture alingment

Generally, alignment methods are not designed to align mixture signals, but they focus on synchronizing single signals (only one source) in the presence of noise. In these circumstances, cross-correlation based methods allow the alignment of signals by determining the position of the maximum of the CC function. As mentioned in Section 2.4.4, one of the most highlighted TDE methods is the generalized cross correlation with phase transform (GCC-PHAT) algorithm proposed in [Knapp and Carter, 1976] that renders good results in reverberant environments as it was demonstrated in [Zhang et al., 2008].

In the STFT-based GCC-PHAT, the time correlation $r_{mp}[\tau]$ is given by Equation (2.71),

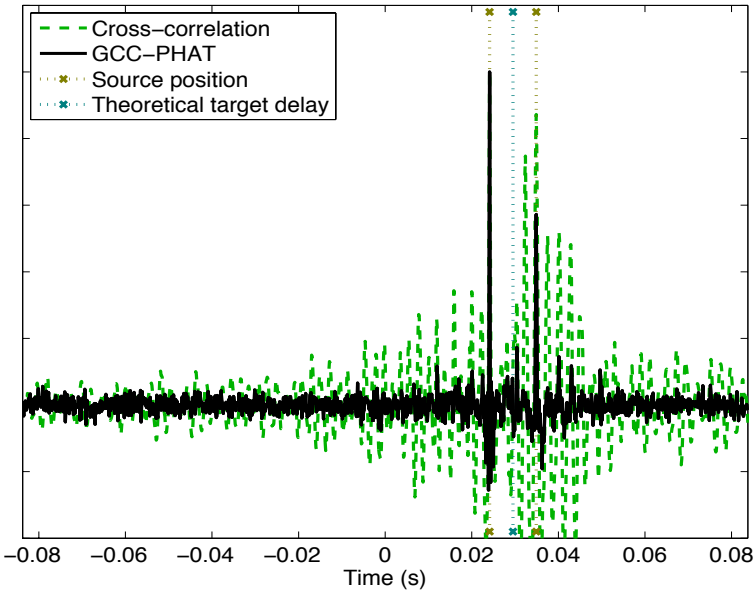

Figure 4.6: Comparative use of the time correlation between mixtures using the different correlation techniques described in this paper. Room size 18.3 x 18.1 x 4.9 m, reverberation time $RT_{60} = 470$ ms. For comparison purposes, the delay of the two sources (24 ms and 35 ms) and the theoretical target delay $\overline{\Delta}_m$ are also included.

which uses the representation of L frames for two mixtures in the STFT domain, $Z_m(k,l)$ and $Z_p(k,l)$. It can be seen in Equation (2.71) that the cross spectrum of each frame is whitened by using a normalization factor. The GCC-PHAT function shows peaks in positions corresponding to the differences in the delays of each source to the considered pair of sensors [Carter, 1993]. In Figure 4.6, we can see GCC-PHAT compared to the standard CC for two mixtures and two speech sources (4 seconds of speech signal in 50% overlapped frames for 4096 samples) recorded in a reverberant room. As we can appreciate, standard CC produces a high number of "false" peaks, and in the case of the GCC-PHAT, this number of significant peaks is reduced, so that the main two peaks correspond to the relative delays of the two sources.

As it was stated above, classical signal alignment methods propose the use of the position of the maximum of $r_{mp}[\tau]$ to determine the shift. Because our signals are mixtures of several sources with different delays, this method does not guarantee that the obtained solution follows the minimum constraint described in Equation (4.4). Therefore, to align mixtures, the position of the main peak should be considered and also a set of the most important peaks in $r_{mp}[\tau]$. The idea of using several peaks of the GCC-PHAT was first introduced in [Bechler et al., 2003] to measure the reliability of the information in speaker localization applications. In this chapter, we determine the P main peaks and we evaluate the average position of these peaks, which will lead to a better solution for the case of mixture alignment. Thus, by using the first mixture as a reference, the delay needed to align the mixtures is given by:

$$\Delta'_m = \frac{1}{S}\sum_{s=1}^{S} k_m[s], \tag{4.11}$$

where $k_m[s]$ is the time index of the s-th peak of $r_{m1}[\tau]$.

One way of demonstrating the suitability of this proposal consists of studying whether the value of Δ'_m is close to $\overline{\Delta_m}$ (Equation (4.10)). In this sense, if we suppose that $k_m[s]$ is placed

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	5.26	6.76	5.14	7.09	2.97	4.26	3.46	5.61	3.20	5.43	1.85	3.00
$Kf_s = 64$ms	5.88	7.10	5.56	7.55	3.05	4.40	4.50	6.52	3.84	6.32	2.00	3.30
$Kf_s = 128$ms	5.91	7.01	5.40	7.45	2.88	4.21	4.76	6.47	3.80	6.24	1.89	3.16
$Kf_s = 256$ms	5.62	6.71	5.07	7.07	2.71	3.96	4.33	5.99	3.44	5.75	1.72	2.87
$Kf_s = 512$ms	5.11	6.23	4.68	6.55	2.54	3.67	3.69	5.24	3.05	5.02	1.58	2.59
MIX 3L	6.18	7.20	5.64	7.63	3.03	4.37	5.05	6.76	4.07	6.50	2.04	3.34
MIX 7L	6.29	7.26	5.72	7.69	3.06	4.42	5.22	6.87	4.16	6.62	2.08	3.40

Table 4.10: SIR (dB) for the GCC-PHAT based mixing matrix estimation method with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of aligning with the two peaks ($P = 2$) of the GCC-PHAT.

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	95.0	98.0	96.9	99.1	90.7	96.1	80.0	93.8	83.5	95.0	65.7	84.9
$Kf_s = 64$ms	96.9	98.6	98.1	99.4	92.9	97.1	89.1	96.7	90.5	97.2	73.2	89.6
$Kf_s = 128$ms	97.1	98.5	98.1	99.4	92.7	97.1	90.7	96.9	91.3	97.4	74.1	89.8
$Kf_s = 256$ms	96.3	98.2	97.5	99.2	91.8	96.8	88.2	95.7	88.6	96.5	71.0	87.8
$Kf_s = 512$ms	94.5	97.4	96.2	98.7	90.3	95.8	82.0	92.6	83.4	94.1	66.5	83.8
3 windows	97.4	98.7	98.3	99.4	93.4	97.4	91.9	97.3	92.3	97.7	76.5	90.9
7 windows	97.6	98.7	98.4	99.4	93.7	97.5	92.3	97.4	92.6	97.7	76.4	91.0

Table 4.11: Percentage of correct words (%) for the GCC-PHAT based mixing matrix estimation method with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of aligning with the two peaks ($P = 2$) of the GCC-PHAT.

approximately in the value corresponding to the difference in the delays of the s-th source from the reference mixture and the m-th mixture so that $k_m[s] \simeq \delta_{1s} - \delta_{ms}$, then Equation (4.11) can be approximated by Equation (4.12),

$$\Delta'_m \simeq \frac{1}{S}\sum_{s=1}^{S}\delta_{1s} - \frac{1}{S}\sum_{s=1}^{S}\delta_{ms}. \tag{4.12}$$

Equation (4.12) shows that Δ'_m approximates the difference between the mean values of the delays of the sources at the reference and m-th microphones, which matches the result of the theoretical LS estimator $\overline{\Delta_m}$ (see Equation (4.10)).

At this point the suitability of using the STFT-based GCC-PHAT function to synchronize speech mixtures before applying speech separation will be studied. Since it is a huge amount of information, only the results of the best separation algorithm, that is, our proposed GCC-PHAT based mixing matrix estimation method with binary mask, are shown in the following tables and figures.

As in previous tables, Tables 4.10 and 4.11 show the results in terms of both SIR and Correct words (%) of our proposed separation method when there is a synchronization stage based on the STFT-based GCC-PHAT function. Note that STOI scores are not included because it is a huge amount of information and STOI is indirectly included in the expression of Correct words (%) (see Equation (2.75)).

Table 4.10 presents the values of SIR obtained in different separation scenarios using different

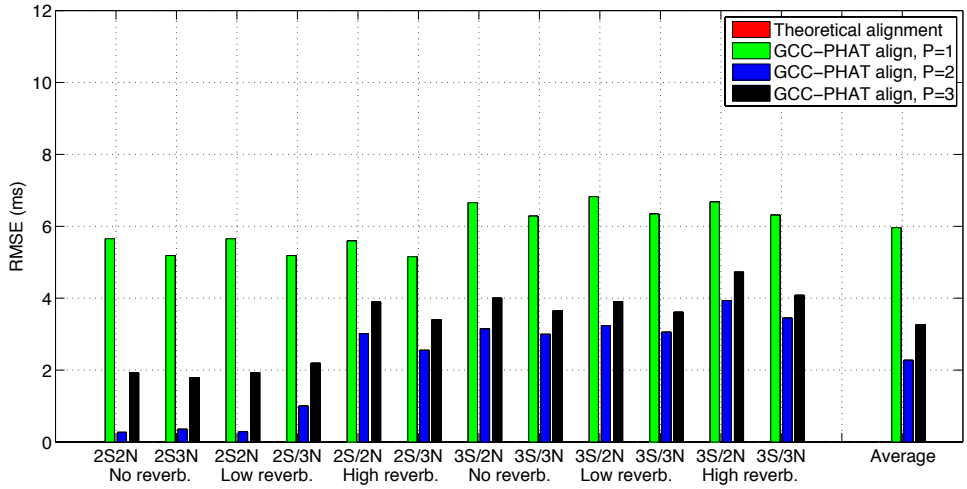

Figure 4.7: RMSE (ms) of the estimated alignment compared to the theoretical alignment for the different scenarios included in the thesis, comparing the performance of the GCC-PHAT based alignment method with different values of number of selected peaks P.

frame lengths. Comparing to Table 4.8 where the theoretical delay is used, very similar results (slightly worse) are obtained with this alignment method based on the GCC-PHAT function. From this comparison, we can conclude that this method estimates the delay needed to align the mixtures correctly. With respect to the values without mixture alignment stage (Table 4.4) different aspects must be mentioned. First, higher separation quality is obtained using this alignment method in the vast majority of scenarios, being this improvement very clear for short frame lengths ($Kf_s = 32$ ms and 64 ms). Second, using only one window, $Kf_s = 64$ ms is the most suitable one for all the separation cases except for the case of two mixtures (2M) and two sources (2S), where $Kf_s = 128$ ms helps to achieve the highest scores. Without synchronization stage, the best separation results are obtained for $Kf_s = 128$ ms. Finally, our proposal of using seven windows (MIX 7L) outperforms the rest of solutions regardless the separation scenario.

From the point of view of speech intelligibility, very similar conclusions to the case of SIR can be extracted. It must be mentioned the high Correct words (%) scores (more than 90 %) obtained by MIX 7L in all the separation scenarios except in the case of three sources and two microphones, where a Correct words (%) equal to 76.4 is reached.

Additionally RMSE between the theoretical delays and the estimated ones using our STFT-based GCC-PHAT function is depicted in Figure 4.7. Aiming at making a complete study, the RMSE obtained is shown when different number of peaks (P) of the STFT-based GCC-PHAT function are used. As it was expected, if there are two speech sources (2S), the RMSE reaches their lowest values when the number of selected peaks is two ($P = 2$). In contrast to it, when the number of talkers is three (3S), the lowest error is also obtained when $P = 2$ (blue bars in figure), what leads us to think that to find a number of peaks greater than two is not easy. The reader can note that the classical use of the GCC-PHAT function, or in other words, to select only the predominant peak ($P = 1$) gives rise to the highest values of RMSE in all the situations (green bars in the figure below).

In the aim of evaluating the effects of using different number of peaks of the GCC-PHAT function over the separation quality, Figures 4.8 and 4.9 depict the values of SIR and Correct words (%) obtained by our separation method when different number of peaks are selected. Furthermore, the values obtained when the theoretical delay is introduced are also represented. The first impression is that there is little difference between using the theoretical delay (red

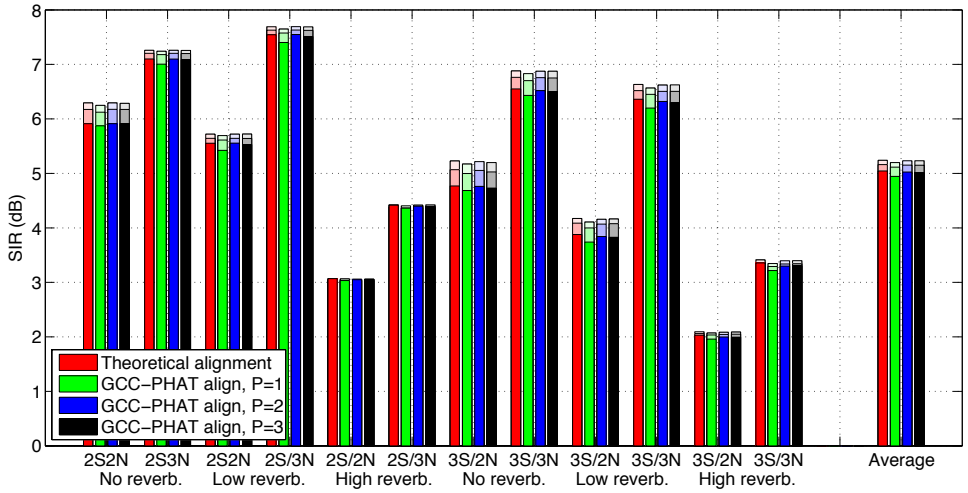

Figure 4.8: SIR (dB) of the estimated alignment compared to the theoretical alignment for the different scenarios included in the thesis, comparing the performance of the GCC-PHAT based alignment method with different values of number of selected peaks P. The darker colors represent results with the best window length (best L), the mid colors represent results mixing three window lengths (MIX 3L) and the lighter colors represent results mixing seven window lengths (MIX 7L).

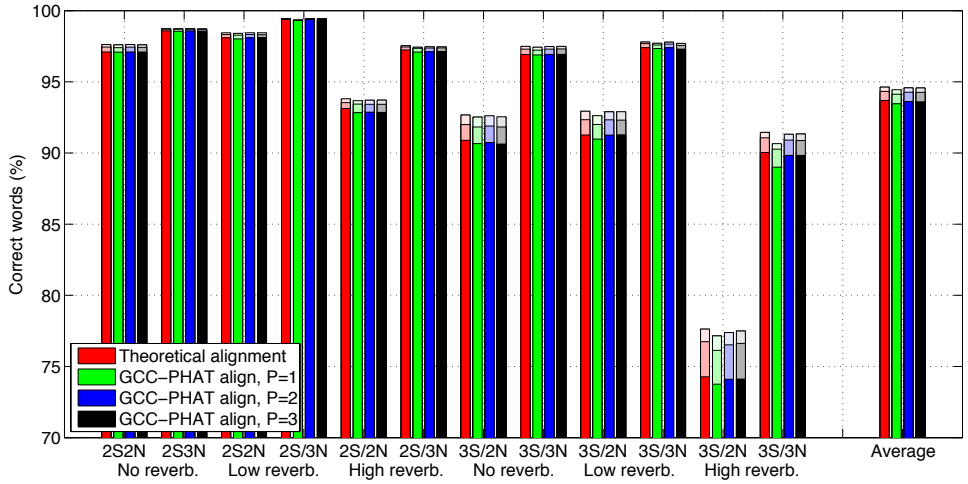

Figure 4.9: Estimated percentage of correct words(%) of the estimated alignment compared to the theoretical alignment for the different scenarios included in the thesis, comparing the performance of the GCC-PHAT based alignment method with different values of number of selected peaks P. The darker colors represent results with the best window length (best L), the mid colors represent results mixing three window lengths (MIX 3L) and the lighter colors represent results mixing seven window lengths (MIX 7L).

bars) and the alignment method based on the the GCC-PHAT function. These little differences suggest us that our alignment method works properly aiming at synchronizing speech mixtures. In respect of the alignment method, slightly worse outcomes are obtained when only one peak is considered (green bars). In the case of two speech sources, the best results are obtained for $P = 2$ and for three sources, comparable scores are obtained for two and three peaks ($P = 2, 3$). Moreover, it can be observed that our separation solutions that rely on combining more than one window (mid and light colors) outperform the use of only one window (dark colors).

4.5.2 Analysis of the computational cost and required transmission data

Another aspect that must be taken into account is the computational cost (number of necessary operations) and amount of synchronization information to be transmitted, associated with the implementation of the synchronization method in a distributed WASN following the scheme depicted in Figure 4.1.

As we can see in Equation (2.71), the GCC-PHAT is performed in the STFT domain. For this purpose, the signal is divided into frames of K_{max} samples, and the value of K_{max} is larger than two times the maximum difference of delays between mixtures to be considered (we must take into account that the generalized CC approximates the linear convolution by the circular convolution, and we must consider both positive and negative shifts). Then, the value of K_{max} will be conditioned by the maximum distance between microphones and the maximum differences in time synchronization.

For instance, if we suppose perfect clock timing of the nodes in a room of $20\ m \times 20\ m \times 5\ m$ and if $f_s = 8000$ Hz, the value of K_{max} is calculated as follows. The worst case (extreme case) in our room is assumed, that is, the maximum distance between microphones (longer diagonal of the room) and sources do not suffer any propagation delay in one mixture (all the sources in the same place as one of the microphones). Then, $K_{max} = \lceil 2 \cdot \frac{\sqrt{a^2+b^2+c^2}}{v} \cdot f_s \rceil$ samples, $v = 340$ m/s is the speed of the sound in the air and $a \times b \times c$ are the dimensions of the room (in meters). Substituting values, $K_{max} = 1340$ samples.

The main term of computational complexity of the GCC-PHAT-based alignment method is the evaluation of the DFT for each frame to obtain the STFT. This evaluation process requires $2K_{max}\lceil \log_2(K_{max}) \rceil$ complex products and $K_{max}\lceil \log_2(K_{max}) \rceil$ complex additions, where the operator $\lceil\ \rceil$ indicates that the first higher integer. Considering a typical 50% overlapped windowing of the STFT and taking into account that each complex product implies four products and two additions, the number of operations (products and sums) per second is approximately $16 f_s \lceil \log_2(K_{max}) \rceil$. For instance, for the same room and f_s described above, the number of operations per second and per node is approximately 1.4 million of instructions per second (MIPS).

Concerning the amount of information to be transmitted from the reference node (it transmits the synchronization signal to the rest of nodes), it is important to highlight that only the phase of the STFT is required to implement the GCC-PHAT because Equation (2.71) is equivalent to $e^{j\omega_k(\phi_m - \phi_p)}$, where ϕ_m and ϕ_p are the phases of $Z_m(k,l)$ and $Z_p(k,l)$, respectively. Using N_b bits to code the phase, the amount of data required for transmission per second is approximately $N_b f_s$ bits/s. Furthermore, this amount of information needed to obtain the synchronization can be easily reduced by neglecting the high frequency bands in the GCC evaluation (speech signal has its greatest components at low frequencies). Thus, if we only consider the $K_{max}/(2D)$ lower frequency bins of $Z_m(k,l)$, we reduce the amount of information needed for transmission. The effect is similar to downsampling the mixtures by a factor of D to determine the synchronization, and the amount of information transmitted will be therefore reduced by the same factor, totaling $N_b f_s / D$ bits/s (a typical value is $D = 8$). The use of a D factor greater than one will cause

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	5.20	6.67	5.05	7.00	2.90	4.10	3.37	5.48	3.11	5.31	1.71	2.77
$Kf_s = 64$ms	5.88	7.08	5.54	7.54	3.01	4.33	4.50	6.49	3.83	6.32	1.97	3.28
$Kf_s = 128$ms	5.91	7.01	5.38	7.44	2.86	4.21	4.75	6.46	3.81	6.24	1.88	3.15
$Kf_s = 256$ms	5.61	6.71	5.07	7.08	2.71	3.96	4.34	5.98	3.44	5.75	1.72	2.87
$Kf_s = 512$ms	5.11	6.23	4.67	6.55	2.54	3.67	3.70	5.24	3.05	5.03	1.58	2.60
3 windows	6.18	7.20	5.64	7.63	3.01	4.35	5.05	6.75	4.07	6.51	2.04	3.34
7 windows	6.29	7.25	5.71	7.69	3.04	4.40	5.21	6.87	4.16	6.62	2.07	3.40

Table 4.12: SIR (dB) for the GCC-PHAT based mixing matrix estimation method with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of aligning with the GCC-PHAT based method with two peaks, $N_b = 2$ and $D = 8$.

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	94.8	97.9	96.7	99.1	89.9	95.1	80.6	93.6	82.8	94.7	61.5	82.3
$Kf_s = 64$ms	96.9	98.6	98.1	99.4	92.8	96.9	89.1	96.6	90.4	97.2	71.7	89.4
$Kf_s = 128$ms	97.1	98.5	98.1	99.4	92.6	97.1	90.7	96.9	91.2	97.4	73.6	90.0
$Kf_s = 256$ms	96.3	98.2	97.5	99.2	91.8	96.7	88.2	95.7	88.5	96.5	70.9	87.8
$Kf_s = 512$ms	94.5	97.4	96.2	98.7	90.2	95.7	82.0	92.6	83.3	94.1	66.4	84.1
3 windows	97.4	98.7	98.3	99.4	93.4	97.3	91.9	97.3	92.3	97.6	76.1	90.9
7 windows	97.6	98.7	98.4	99.4	93.7	97.4	92.6	97.5	92.9	97.8	77.0	91.3

Table 4.13: Percentage of correct words (%) for the GCC-PHAT based mixing matrix estimation method with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of aligning with the GCC-PHAT based method with two peaks, $N_b = 2$ and $D = 8$.

lower resolution in the correlation and the obtained peaks in $r_{mp}[\tau]$ will be smoothed, but their maximum position will remain unaltered.

Considering a version of our GCC-PHAT function that helps to minimize the computational cost and the amount of information needed for transmission, its most relevant separation quality results are depicted in the following tables and figures. Tables 4.12 and 4.13 show the outcomes of separation quality in terms of SIR and Correct words (%), respectively. Comparing these tables with Tables 4.10 and 4.11 it can be observed that the outcomes of this customized version of the STFT-based GCC-PHAT function are slightly worse but it allows us to reduce the computational cost and the information for transmission.

Different values of the number of bits to code the phase (N_b) and of the downsampling factor (D) have been analyzed to achieve as high separation quality as possible. In this sense, Figure 4.10 presents the RMSE obtained with several combinations of N_b and D in the different separation scenarios studied in this thesis. It seems evident that the combination of $N_b = 8$ and $D = 1$ (green bars in figure) achieves the lowest RMSE in all the separation cases. This difference is clear when the number of talkers is two (2S) and for three speech sources when the level of reverberation is high.

Observing Figure 4.10 it is clear that $D = 1$ (as it was expected) obtains the highest accuracy when delays are estimated, but this value does not entail a reduction in the data required for transmission in the WASN. For this reason, the influence of these combinations (N_b, D) over the separation quality is shown in the Figures 4.11 and 4.12. Taking into account SIR, it is evident

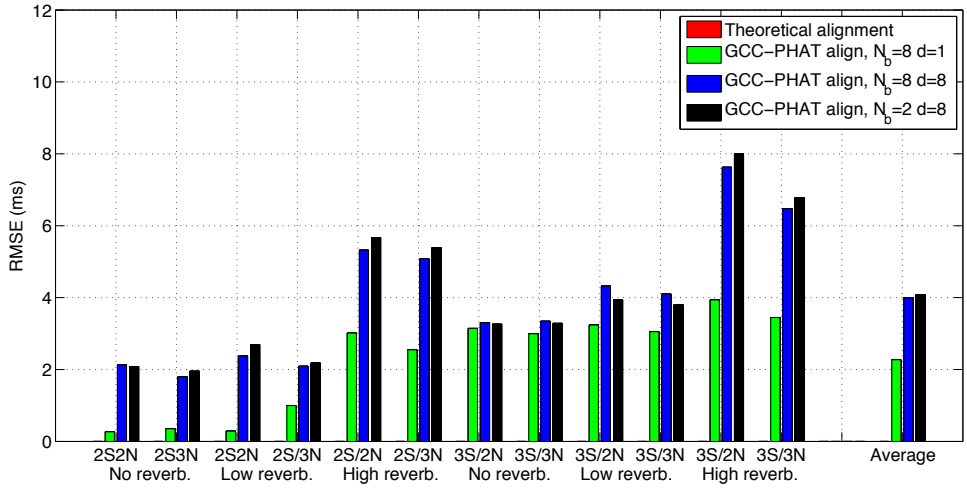

Figure 4.10: RMSE (ms) of the estimated alignment compared to the theoretical alignment for the different scenarios included in the thesis, comparing the performance of the GCC-PHAT based alignment method with $P = 2$, with different values of N_b and D.

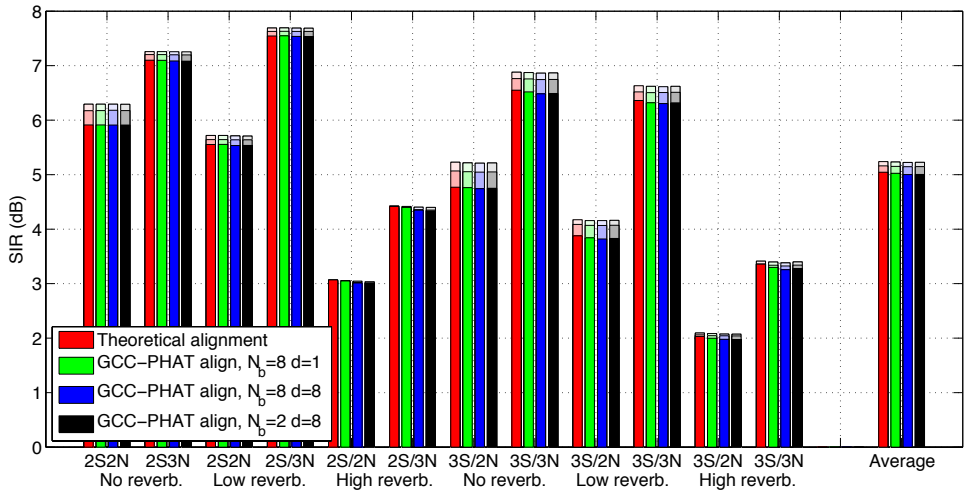

Figure 4.11: SIR (dB) of the estimated alignment compared to the theoretical alignment for the different scenarios included in the thesis, comparing the performance of the GCC-PHAT based alignment method with $P = 2$, with different values of N_b and D. The darker colors represent results with the best window length (best L), the mid colors represent results mixing three window lengths (Mix 3L) and the lighter colors represent results mixing seven window lengths (Mix 7L).

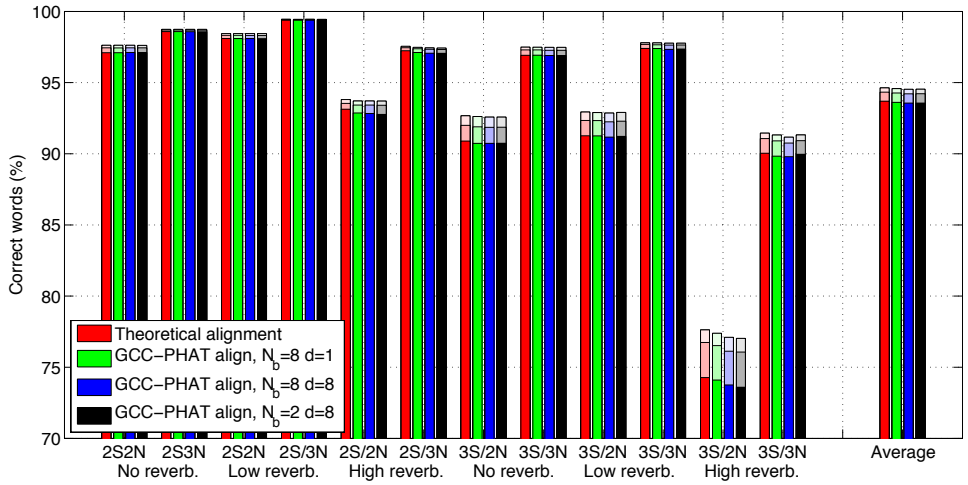

Figure 4.12: Estimated percentage of correct words(%) of the estimated alignment compared to the theoretical alignment for the different scenarios included in the thesis, comparing the performance of the GCC-PHAT based alignment method with $P = 2$, with different values of N_b and D. The darker colors represent results with the best window length (best L), the mid colors represent results mixing three window lengths (Mix 3L) and the lighter colors represent results mixing seven window lengths (Mix 7L).

that the difference of using these combinations of N_b and D is minimum, what leads us to use a downsampling factor equal to 8 because it entails a considerable reduction of the amount of data for transmission. From the point of view of speech intelligibility, the same idea is foregrounded, that is, the outcomes change little except in the case of three sources and two mixtures as it can be observed in Figure 4.12.

4.6 Mixture alignment using the Short-Term Log-Energy Cross-Correlation for non-stationary sources

In this section a new alignment method for speech mixtures is presented. First, the underlying theoretical basis is explained in a detailed way aiming at demonstrating that it calculates delays between speech mixtures in line with our theoretical analysis of the frame length. In a second part, it will be shown that this method is efficient in respect of the amount of data for transmission in our WASN.

4.6.1 Short-term Log-Energy Cross-Correlation-based method for delay estimation

In this section, we evaluate the use of the STLE sequence, which represents an alternative to standard CC-based solutions in the particular case of non-stationary sources such as speech. In the speech signal processing framework, STLE is used in different applications, such as voice activity detection [de Veth et al., 2001], speech recognition [Couvreur and Couvreur, 2004] or the estimation of reverberation time in rooms [Couvreur et al., 2001]. In our work, the full STLE of the m-th mixture $z_m[n]$ is denominated by $S\{z_m[n]\}$ and can be determined using the following equation:

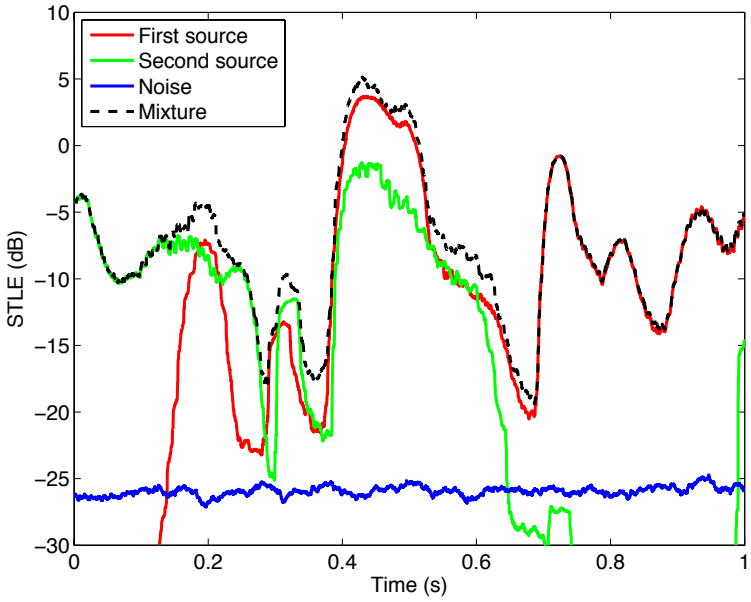

Figure 4.13: Example of the full STLE of a mixture of two sources using $N/f_s = 32$ ms and the full STLE of the components of the mixture. Room size 17.6 x 15.2 x 3.4 m, reverberation time $RT_{60} = 470$ ms. N is the window size to determine the short-term energy.

$$S\{z_m[n]\} = 10\log_{10}\left(\sum_{k=0}^{N-1} z_m^2[n+k]\right), \tag{4.13}$$

where N defines the window size to determine the short-term energy. Note that the term "full" is used to distinguish from decimated and interpolated versions of the STLE that are used and explained in the following paragraphs.

Substituting Equation (2.5) into Equation (4.13) and supposing that the different sources are uncorrelated with zero mean, which makes the cross terms negligible ($\sum_{k=0}^{N-1} y_{mr}[n+k]y_{ms}[n+k] \simeq 0$ for all $r \neq s$ and for the noise components, that is, $r = 0$ and $s = 0$), the full STLE can be approximated using Equation (4.14),

$$S\{z_m[n]\} \simeq 10\log_{10}\left(\sum_{s=0}^{S}\sum_{k=0}^{N-1} y_{ms}^2[n+k]\right). \tag{4.14}$$

Figure 4.13 shows an example of the full STLE of a mixture of two sources using $N/f_s = 32$ ms and the full STLE of the components of the mixture. The non-stationarity of speech signals and the fact that there is usually a dominant source in the mixture makes the approximation of the STLE by the following equation possible:

$$S\{z_m[n]\} \simeq \sum_{s=0}^{S} K_{ms}[n]S\{y_{ms}[n]\}, \tag{4.15}$$

where $K_{ms}[n] = 1$ if the s-th source is the dominant one and 0 in the other case. In other words, $S\{z_m[n]\} \simeq S\{y_{ms}[n]\}$ (dBs) if the s-th source is the dominant one for a particular n.

To determine the alignment of the mixtures, we propose to evaluate the CC of the full STLEs of the m-th and the reference mixture, which is given by

$$s_{m1}[\tau] = E\{S\{z_m[n]\}S\{z_1[n+\tau]\}\}, \tag{4.16}$$

where E is the expectation operator. Introducing Equation (4.15) into Equation (4.16) yields

$$s_{m1}[\tau] \simeq \sum_{r=0}^{S}\sum_{s=0}^{S} E\{K_{mr}[n]K_{1s}[n+\tau]S\{y_{mr}[n]\}S\{y_{1s}[n+\tau]\}\}. \tag{4.17}$$

Considering again the uncorrelation of the sources, we can suppose that the contribution terms from different sources ($E\{K_{mr}[n]K_{1s}[n+\tau]S\{y_{mr}[n]\}S\{y_{1s}[n+\tau]\}\}$ with $r \neq s$) and from the noise sources ($r = 0$ or $s = 0$) will not highly depend on the value of τ, and, therefore, can be assumed as approximately constant. Furthermore, the terms $E\{K_{ms}[n]K_{1s}[n+\tau]S\{y_{ms}[n]\}S\{y_{1s}[n+\tau]\}\}$ can be expected to be proportional to the expression $E\{S\{y_{ms}[n]\}S\{y_{1s}[n+\tau]\}\}$. Thus, we propose to approximate Equation (4.17) by a linear combination of the terms $E\{S\{y_{ms}[n]\}S\{y_{1s}[n+\tau]\}\}$, according to

$$s_{m1}[\tau] \simeq b_m + \sum_{s=1}^{S} w_{ms} E\{S\{y_{ms}[n]\}S\{y_{1s}[n+\tau]\}\}, \tag{4.18}$$

where b_m and w_{ms} are constants. In other words, the CC of the full STLE of the mixtures is a linear combination of the full STLEs for the contributions of each source at the different microphones.

To verify the suitability of these approximations, we have measured the difference between the values obtained by the CC of the full STLEs (Equations (4.13) and (4.16)) and the values obtained by the linear combination proposed in Equation (4.18), obtaining a mean absolute relative error of 0.95% for the database described in Section 4.3.

$E\{S\{y_{ms}[n]\}S\{y_{1s}[n+\tau]\}\}$ will be centered on the corresponding difference of the delays of the source ($\delta_{1s} - \delta_{ms}$). On the other hand, we can see that due to the use of the integration of N terms, the full STLE is a low pass signal. Thus, $E\{S\{y_{ms}[n]\}S\{y_{1s}[n+\tau]\}\}$ and, therefore, $s_{m1}[\tau]$ will vary smoothly and the smoothness will be controlled by the value of N. The maximum of $s_{m1}[\tau]$ (blue line in Figure 4.14) will be located on a point close to the terms $\delta_{1s} - \delta_{ms}$ (source positions in Figure 4.14), so that it can be used to estimate the values of Δ_m. Therefore, taking the first mixture as a reference, $\Delta''_m = (\arg\max_\tau\{s_{m1}[\tau]\})/f_s$.

To demonstrate the suitability of the proposal (for estimating delays close to the value $\overline{\Delta_m}$), let us now approximate the values of $E\{S\{y_{ms}[n]\}S\{y_{1s}[n+\tau]\}$ in the range of values close to Δ''_m by a quadratic function, so that $E\{S\{y_{ms}[n]\}S\{y_{1s}[n+\tau]\} \simeq b'_{ms} + w'_{ms}(\tau - f_s k_m[s])^2$ and $k_m[s] = \delta_{1s} - \delta_{ms}$. So, Equation (4.18) particularized for the evaluation of $s_{m1}[\tau]$ can be approximated by the following equation:

$$s_{m1}[\tau] \simeq b_m + \sum_{s=1}^{S} w_{ms}(b'_{ms} + w'_{ms}(\tau - f_s k_m[s])^2) = b''_m + \sum_{s=1}^{S} w''_{ms}(\tau - f_s k_m[s])^2. \tag{4.19}$$

Again, b''_m and w''_{ms} are constants. The maximum value of $s_{m1}[\tau]$, which is used to estimate the alignment, can be obtained by forcing the derivative of Equation (4.19) to zero. This value is expressed using Equation (4.20)

$$\Delta''_m = \frac{1}{f_s}\arg\max_\tau\{s_{m1}[\tau]\} \simeq \frac{\sum_{s=1}^{S} w''_{ms} k_m[s]}{\sum_{s=1}^{S} w''_{ms}}. \tag{4.20}$$

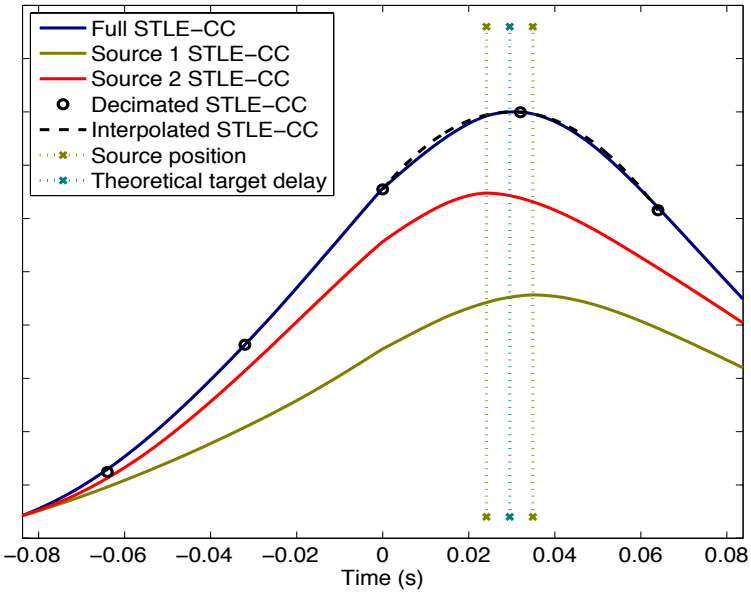

Figure 4.14: Comparison among the CCs of the STLEs: full CC-STLE ($s_{m1}[\tau]$), partial terms of the CC-STLEs for each source ($E\{S\{y_{ms}[n]\}S\{y_{1s}[n+\tau]\}\}$), decimated CC-STLE ($\hat{s}_{m1}[\tau]$ in Equation (4.21)) and interpolated CC-STLE (equation (4.22)). Room size 18.3 x 18.1 x 4.9 m, reverberation time $RT_{60} = 470$ ms. For comparison purposes, the delay of the two sources (24 ms and 35 ms) and the theoretical target delay $\overline{\Delta}_m$ are also included.

.

As we can see, the proposed alignment value approximates a weighted version of Equation (4.11), and therefore its value will also be close to the theoretical value $\overline{\Delta_m}$. Note that the weighting values w''_{ms} are a combination of w'_{ms}, b'_m, w_{ms} and b_m. The w'_{ms} and b'_m result approximates the values of $E\{S\{y_{ms}[n]\}S\{y_{1s}[n+\tau]\}$ by a quadratic function. w_{ms} and b_m come from the approximation expressed in Equation (4.18) and are related (and therefore w''_{ms}) to the contribution of each source in the different mixtures. In other words, w''_{ms} terms depend on the relative energy of the source in the mixture and on the total time in which the s-th source is the predominant one.

4.6.2 Proposed method for efficient data transmission

Another detail derived from the slow variation of the full STLE is that it can be decimated by a factor of N without loosing information. This fact is important because it can highly reduce both the amount of information (STLE of the reference mixture) needed for transmission from the reference node to the rest of them and the computational cost of the alignment method. Specifically in this work, to use this decimation of the full STLE in distributed processing, we define the decimated CC of the decimated STLE (developed in the different nodes except in the reference one) using Equation (4.21).

$$\hat{s}_{m1}[\tau] = E\{S\{z_m[nN]\}S\{z_1[(n+\tau)N]\}\} \tag{4.21}$$

Decimation of CC does not reduce the alignment information and allows a computational complexity reduction and a reduction of the transmission bandwidth by a factor of N from the central processor to the rest of the nodes. Figure 4.14 also represents the decimated CC of the STLE ($N = 256$), labeled as Decimated STLE-CC) for comparison with the CC of the full STLE (la-

beled as Full STLE-CC). It can be observed that Decimated STLE-CC reasonably approximates the Full STLE-CC.

Once the decimated STLE of the reference mixture (computed at the central node) arrives at each node, the CC of them can be performed, resulting in $\hat{s}_{m1}[\tau],\ \forall m = 2, \dots M$. Then, at each node, interpolation can be developed to determine the position of the maximum value of the CC from the decimated version with higher accuracy. This process applies a quadratic interpolation to fit a generic parabola using the largest sample of $\hat{s}_{m1}[\tau]$ and its two neighbors (dashed black line in Figure 4.14). This interpolation is described mathematically in the following paragraphs.

Denoting the position of the maximum sample of $\hat{s}_{m1}[\tau]$ as $\hat{k}_m$, that is, $\hat{k}_m = \arg\max_\tau \hat{s}_{m1}[\tau]$, and the maximum value of the interpolated CC as Δ'''_m (value to be determined), $\hat{s}_{m1}[\tau]$ can be approximated in the range $[\hat{k}_m - 1, \hat{k}_m + 1]$ by a parabola:

$$\hat{s}_{m1}[\tau] \simeq b'''_m + w'''_m \left(\tau - \frac{\Delta'''_m f_s}{N}\right)^2, \tag{4.22}$$

where there are three unknowns (b'''_m, w'''_m and Δ'''_m). Evaluating Equation (4.22) for $(\hat{k}_m, \hat{s}_{m1}[\hat{k}_m])$ and its neighbors in $\hat{k}_m - 1$ and in $\hat{k}_m + 1$, a system of three equations is obtained:

$$\begin{cases} \hat{s}_{m1}[\hat{k}_m - 1] = b'''_m + w'''_m \left(\hat{k}_m - 1 - \frac{\Delta'''_m f_s}{N}\right)^2 \\ \hat{s}_{m1}[\hat{k}_m] = b'''_m + w'''_m \left(\hat{k}_m - \frac{\Delta'''_m f_s}{N}\right)^2 \\ \hat{s}_{m1}[\hat{k}_m + 1] = b'''_m + w'''_m \left(\hat{k}_m + 1 - \frac{\Delta'''_m f_s}{N}\right)^2. \end{cases} \tag{4.23}$$

The value of Δ'''_m is obtained by solving this system of equations:

$$\Delta'''_m = \frac{N}{f_s}\left(\hat{k}_m + \frac{\hat{s}_{m1}[\hat{k}_m - 1] - \hat{s}_{m1}[\hat{k}_m + 1]}{2(\hat{s}_{m1}[\hat{k}_m - 1] + \hat{s}_{m1}[\hat{k}_m + 1] - 2\hat{s}_{m1}[\hat{k}_m])}\right). \tag{4.24}$$

Taking into account Equation (4.24), the proposed values of Δ'''_m will ideally be placed close to the values defined by Equation (4.10). For illustrative purposes, Figure 4.14 shows that the maximum value of the interpolated STLE (interpolated STLE-CC) is very close to the theoretical target delay, as we expected.

The computational cost required when the proposed interpolated STLE-CC in a distributed system is very low. In the case of $N > 64$, the number of instructions at each node is mainly caused by the sum of the squared values in Equation (4.13), which requires f_s multiplications and accumulations per second. This value is drastically lower than the one obtained by the GCC-PHAT method described in the previous section, which is approximately $8\lceil \log_2(K_{max}) \rceil$ times lower. For instance, for the same room and f_s mentioned in Section 4.5.2, our proposed method based on CC implies 1.4 MIPS and the proposed interpolated STLE-CC is 0.016 MIPS, which represents an important reduction.

Concerning data transmission, we only need to transmit (by broadcast) the values of the decimated STLE of the reference mixture from the reference node to each node, to calculate the delays. So, using N_b bits to codify the STLE value, we only require $N_b f_s/N$ bits/s, which is N times lower. Finally, delays between each mixture and the reference one are calculated at each node for applying the synchronization.

At this point the suitability of this proposal of alignment method that relies on the use of STLE must be analyzed. First of all, we will study how it affects the separation quality obtained by our GCC-PHAT based mixing matrix estimation method with binary mask. The outcomes achieved when this synchronization stage is included before applying the separation algorithm

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	4.20	5.90	5.05	6.92	2.90	4.07	2.89	5.03	3.13	5.38	1.75	2.76
$Kf_s = 64$ms	5.38	6.91	5.52	7.50	3.00	4.33	4.14	6.34	3.86	6.31	1.96	3.18
$Kf_s = 128$ms	5.76	6.98	5.44	7.44	2.86	4.19	4.61	6.44	3.82	6.23	1.87	3.09
$Kf_s = 256$ms	5.56	6.70	5.06	7.07	2.70	3.95	4.31	5.98	3.44	5.75	1.72	2.84
$Kf_s = 512$ms	5.10	6.23	4.68	6.55	2.53	3.67	3.69	5.24	3.05	5.02	1.58	2.56
MIX 3L	6.00	7.16	5.66	7.62	3.01	4.34	4.90	6.71	4.09	6.50	2.02	3.26
MIX 7L	6.14	7.23	5.73	7.69	3.03	4.38	5.08	6.84	4.18	6.62	2.05	3.32

Table 4.14: SIR (dB) for the GCC-PHAT based mixing matrix estimation method with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of aligning with the STLE based method with decimation.

are presented in Tables 4.14 (SIR) and 4.15 (Correct words (%)). Concerning SIR, comparing these results to the ones obtained without synchronization stage (Table 4.4), several aspects must be mentioned. As in earlier instances, shorter frame lengths are required to reach the highest scores of SIR. Depending on the reverberation, two frame lengths are the most suitable ones. For absence of reverberation it can be observed that $Kf_s = 128$ ms helps to obtain the highest quality values, e.g., 5.38 dB and 6.91 dB for a number of speakers equal to two, and for two or three microphones, respectively. On the other hand, $Kf_s = 64$ ms is the most suitable window size when reverberation effects are presented. The improvements introduced by the use of this synchronization stage are very clear for the shortest windows sizes when comparing to Table 4.4. Illustrative examples are the improvements of 5.21 dB and 4.41 dB in low-reverberant rooms when there are three microphones and two or three speech sources, respectively. It must also be mentioned that our proposal of combining different window sizes gets the highest quality values in all the situations. Some values to be highlighted are the 7.69 dB and 6.62 dB reached for low reverberation (very common reverberation level in real rooms). It can also be mentioned the value of SIR equal to 4.38 dB when there are two speakers in the presence of high reverberation since in these circumstances, the vast majority of separation algorithms do not perform correctly.

The main advantage of using aligning with the STLE based method with decimation is that it reduces the amount of information for transmission drastically, what is of key importance in WASNs. Establishing a comparison between the quality values of Table 4.14 and the ones obtained by the aligning with the GCC-PHAT based method that minimizes the computational cost and the amount of information (Table 4.12), it is easy to see that the results are slightly worse but comparable. In some applications, it may be very useful to reduce the amount of information for transmission at the expenses of reducing the separation quality slightly.

The information related to speech intelligibility is in Table 4.15, from which very similar conclusions to the case of SIR can be extracted. Comparing to the separation case without mixture alignment stage (Table 4.5), it is clear that the alignment process helps to obtain better results in virtually all the situations. These improvements are evident for the shortest frame lengths, as for instance the separation case with three sources, three microphones, high reverberant conditions and $Kf_s = 32$ ms, where Correct words (%) ranges from 25.3 % without synchronization stage, to 81.7 % using our STLE based method. As in the case of SIR, the scores of Correct words (%) are slightly worse than the ones obtained by the aligning with the GCC-PHAT based method (see Table 4.13).

One important aspect to be considered is the accuracy of our STLE based method in order to estimate differences of delay between speech mixtures. In Figure 4.15 a comparative study of

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	91.7	97.1	96.8	99.1	90.4	95.6	75.7	92.2	83.4	95.1	62.4	81.7
$Kf_s = 64$ms	96.2	98.5	98.1	99.4	92.6	97.0	87.9	96.4	90.5	97.3	72.2	88.3
$Kf_s = 128$ms	97.0	98.5	98.1	99.4	92.7	97.1	90.5	96.9	91.3	97.4	73.6	89.0
$Kf_s = 256$ms	96.3	98.2	97.5	99.1	91.8	96.7	88.0	95.6	88.5	96.5	70.9	87.1
$Kf_s = 512$ms	94.5	97.4	96.2	98.7	90.3	95.8	81.9	92.6	83.4	94.1	66.6	83.1
MIX 3L	97.3	98.7	98.3	99.4	93.4	97.3	91.5	97.2	92.3	97.7	76.1	90.1
MIX 7L	97.5	98.7	98.5	99.4	93.6	97.4	92.3	97.4	92.9	97.8	76.9	90.4

Table 4.15: Percentage of correct words (%) for the GCC-PHAT based mixing matrix estimation method with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of aligning with the STLE based method with decimation.

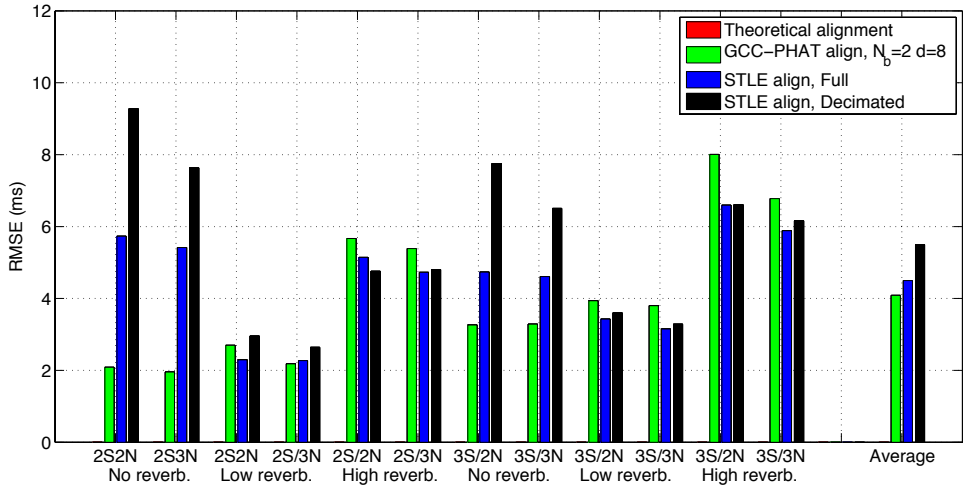

Figure 4.15: RMSE (ms) of the estimated alignment compared to the theoretical alignment for the different scenarios included in the thesis, comparing the performance of the STLE based alignment method with and without decimation.

the RMSE obtained by the different alignment methods is depicted. As it can be observed, the best configuration of the GCC-PHAT based method (green bars in figure) outperforms our STLE based method (blue and black bars) in absence of reverberation, this difference being important when the number of speech sources is two. In contrast to it, in the presence of reverberation, our solution that relies on Full STLE obtains lower RMSE than the GCC-PHAT based method.

Comparing both versions of the STLE aligning, it is easy to see that the decimated version obtains much worse outcomes in absence of reverberation, while that for reverberant environments, these differences are smaller. It must be highlighted the fact that our STLE based method outperforms the GCC-PHAT based one in presence of reverberation, what is very important since in general, the GCC-PHAT function is one of the most popular and robust tools in the literature.

But since this thesis deals with the cocktail party problem, we must put special attention on the quality of the separated sources. In this sense, Figures 4.16 and 4.17 show the separation quality obtained when the theoretical delay and our alignment solutions are used.

Having a look at Figure 4.16, many aspects can be mentioned. In absence of reverberation, the GCC-PHAT based alignment method obtains slightly better outcomes if the number of microphones is two (2N). For the rest of separation cases these differences are negligible. These similar results lead us to apply the STLE based alignment method because it has the advantage

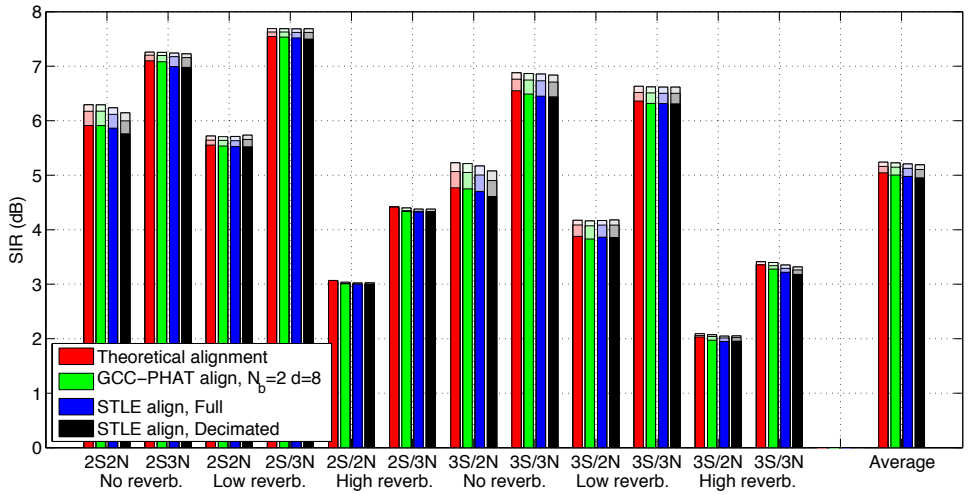

Figure 4.16: SIR (dB) of the estimated alignment compared to the theoretical alignment for the different scenarios included in the thesis, comparing the performance of the STLE based alignment method with and without decimation. The darker colors represent results with the best window length (best L), the mid colors represent results mixing three window lengths (Mix 3L) and the lighter colors represent results mixing seven window lengths (Mix 7L).

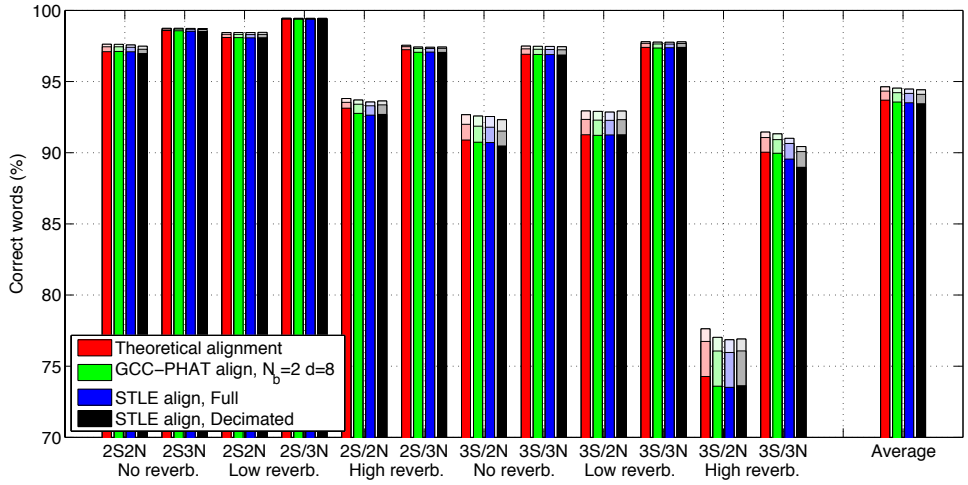

Figure 4.17: Estimated percentage of correct words(%) of the estimated alignment compared to the theoretical alignment for the different scenarios included in the thesis, comparing the performance of the STLE based alignment method with and without decimation. The darker colors represent results with the best window length (best L), the mid colors represent results mixing three window lengths (Mix 3L) and the lighter colors represent results mixing seven window lengths (Mix 7L).

of lower computational cost and it requires less information for transmission. With respect to the different versions of the STLE based alignment method, that is, the full version and the decimated one, it can be said that the differences between them are minimum. Additionally, these alignment methods have been evaluated using different frame lengths, specifically, the results for the frame lengths that achieve better separation (dark colors) are depicted. As in prior studies, our proposals of using several frame lengths (MIX 3L and MIX 7L) are also depicted in the figure. From it, we can see that all the alignment methods achieve the highest quality scores when they are combined with the MIX 7L proposal (light colors). On the other hand, Figure 4.17 shows the percentage correct words(%) that are obtained in the same experiments. From

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	4.09	5.70	5.05	6.87	2.89	4.05	2.80	4.90	3.10	5.31	1.75	2.74
$Kf_s = 64$ms	5.31	6.86	5.53	7.48	3.00	4.34	4.09	6.29	3.83	6.30	1.95	3.18
$Kf_s = 128$ms	5.72	6.97	5.43	7.45	2.85	4.20	4.59	6.42	3.82	6.23	1.86	3.09
$Kf_s = 256$ms	5.55	6.70	5.07	7.08	2.70	3.95	4.31	5.98	3.45	5.74	1.72	2.83
$Kf_s = 512$ms	5.10	6.23	4.68	6.55	2.54	3.66	3.69	5.24	3.05	5.02	1.58	2.56
MIX 3L	5.96	7.15	5.66	7.62	3.00	4.34	4.89	6.70	4.08	6.50	2.02	3.26
MIX 7L	6.11	7.22	5.74	7.69	3.03	4.38	5.06	6.83	4.17	6.62	2.05	3.32

Table 4.16: SIR (dB) for the GCC-PHAT based mixing matrix estimation method with binary mask for the different scenarios considered in the thesis in function of the window configuration, in the case of aligning with the STLE based method with decimation and $N_b = 2$.

No. sources	2S						3S					
Reverberation	No rev.		Low rev.		High rev.		No rev.		Low rev.		High rev.	
No. nodes	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M	2M	3M
$Kf_s = 32$ms	90.4	96.1	96.7	99.0	90.3	95.5	74.6	91.2	82.6	94.9	61.4	81.1
$Kf_s = 64$ms	96.0	98.4	98.0	99.4	92.7	97.0	87.7	96.4	90.4	97.2	71.8	88.3
$Kf_s = 128$ms	96.9	98.5	98.1	99.4	92.7	97.1	90.4	96.8	91.2	97.4	73.4	89.0
$Kf_s = 256$ms	96.3	98.2	97.5	99.1	91.8	96.7	87.9	95.6	88.5	96.4	70.9	87.0
$Kf_s = 512$ms	94.5	97.4	96.2	98.7	90.3	95.8	81.9	92.6	83.4	94.1	66.4	83.2
MIX 3L	97.2	98.7	98.3	99.4	93.4	97.3	91.4	97.2	92.3	97.7	75.9	90.0
MIX 7L	97.4	98.7	98.4	99.4	93.6	97.4	92.3	97.4	92.9	97.8	76.8	90.4

Table 4.17: Percentage of correct words (%) for the GCC-PHAT based mixing matrix estimation method with binary mask mask for the different scenarios considered in the thesis in function of the window configuration, in the case of aligning with the STLE based method with decimation and $N_b = 2$.

this figure the same conclusions can be extracted because the outcomes are consistent with the ones in Figure 4.16.

Since our synchronization solutions have been designed to be used in WASNs, it is very important to use the smallest amount of information for transmission as possible. Considering it, it has been studied the use of different number of bits (N_b) to code the information for transmission. The effects of varying N_b on the quality of the separated sources have been analyzed as it is shown in the following tables and figures. Tables 4.16 and 4.17 show the separation quality when the decimated CC-STLE alignment method is used and $N_b = 2$ bits, or in other words, the minimum number of bits is used. Comparing the results in Table 4.16 with the ones in Table 4.14, in which there is no restriction on the number of bits, it is clear that the scores of SIR have decreased in all the situations, but these decreases are in the order of hundredths of dB for the most suitable frame lengths. Despite of using a minimum number of bits, the synchronization solution is able to align the mixtures without affecting the separation algorithm considerably. For instance, in the presence of three speech sources, two microphones and low reverberant rooms, our MIX 7L separation solution obtains 4.18 dB (without restriction on N_b) and 4.17 dB using $N_b = 2$ bits.

With respect to the percentage of understood words, it can be said that comparing Tables 4.15 and 4.17 the influence of using only two bits on this speech intelligibility measurement is minimum. For example, performing the separation with our MIX 7L proposal, in the vast majority of separation scenarios the same scores have been reached. In the separation problems

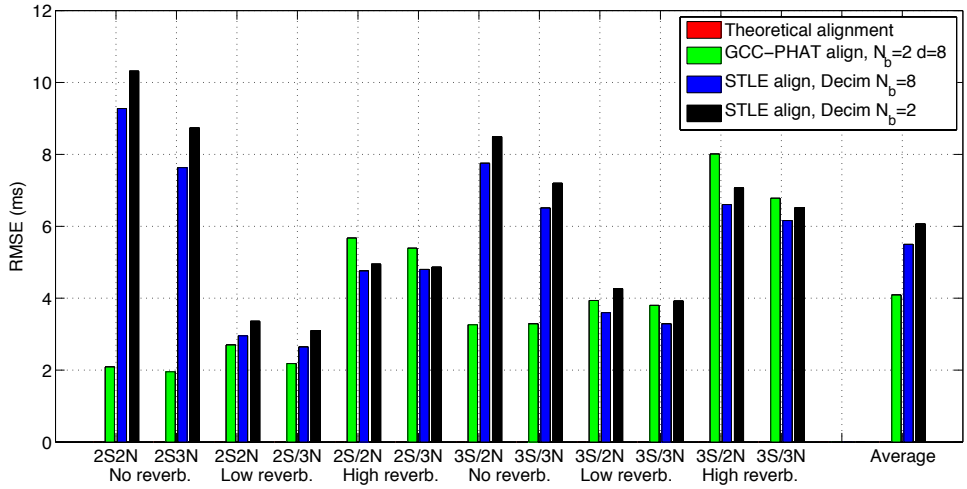

Figure 4.18: RMSE (ms) of the estimated alignment compared to the theoretical alignment for the different scenarios included in the thesis, comparing the performance of the STLE based decimated alignment method with different values of N_b.

where the obtained values of Correct words (%) are smaller, a maximum decrease of 0.1 % has occurred, what is negligible.

In order to make a comprehensive study of the influence of the number of bits, Figures 4.18, 4.19 and 4.20 present different quality measures for two possible values of this number of bits, $N_b = 2$ and 8. Moreover, the RMSE and the separation quality obtained by the GCC-PHAT based alignment method are also included. In Figure 4.18 the values of RMSE for the different alignment methods are depicted. Considering the GCC-PHAT based alignment method, three possibles situations directly related to the level of reverberation can be found. In absence of reverberation, the GCC-PHAT method outperforms the rest of synchronization proposals clearly. For example, for the two mixtures and two sources separation problem, the GCC-PHAT method achieves a RMSE close to 2 ms, while the STLE based alignment method obtains a RMSE equal to 8.3 ms and 10.2 ms for $N_b = 8$ and 2, respectively. In the case of low reverberant environments, the results of the three alignment methods can be considered comparable. If the number of sources is two, the GCC-PHAT based method achieves slightly better results, and for the case of three talkers in the room, the STLE based method for $N_b = 8$ gets the lowest RMSE. Finally, in the presence of high reverberation, the highest RMSE values are reached by the GCC-PHAT based method regardless the number of microphones and speech sources.

Figures 4.19 and 4.20 show the separation quality in terms of SIR and Correct words(%). Having a look at both figures, it is easy to see that when the GCC-PHAT based alignment method in the synchronization stage better results are obtained in all the separation problems, independently of the number of sources, number of microphones and level of reverberation. Using N_b equal to 2, the worse results are obtained, but it must be said that the differences are of the order of tenths of dB in terms of SIR and of the order of 1 % considering the percentage of correct words. Therefore, in some applications the STLE based method can be very useful, in spite of getting worse quality results, due to require lower amount of information for transmission.

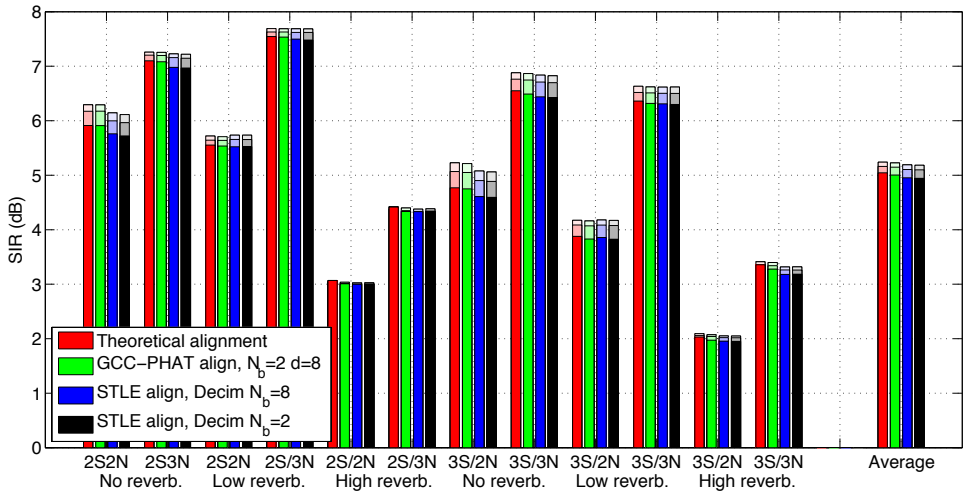

Figure 4.19: SIR (dB) of the estimated alignment compared to the theoretical alignment for the different scenarios included in the thesis, comparing the performance of the STLE based decimated alignment method with different values of N_b. The darker colors represent results with the best window length (best L), the mid colors represent results mixing three window lengths (Mix 3L) and the lighter colors represent results mixing seven window lengths (Mix 7L).

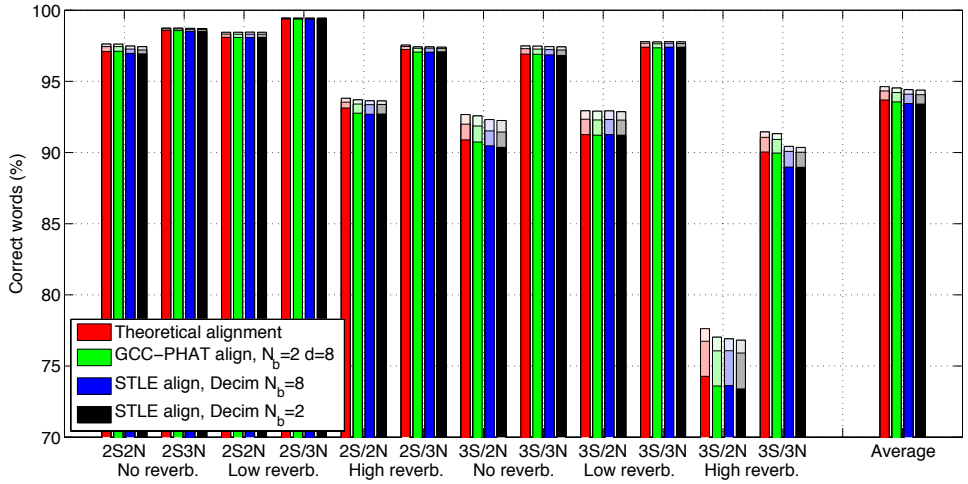

Figure 4.20: Estimated percentage of correct words(%) of the estimated alignment compared to the theoretical alignment for the different scenarios included in the thesis, comparing the performance of the STLE based decimated alignment method with different values of N_b. The darker colors represent results with the best window length (best L), the mid colors represent results mixing three window lengths (Mix 3L) and the lighter colors represent results mixing seven window lengths (Mix 7L).

4.7 Comparison of the results in terms of quality vs. bandwidth vs. computational complexity

In the previous sections two novel alignment methods have been proposed to synchronize the speech mixtures captured by the different microphones of a WASN. These two alignment methods have been designed to be part of a synchronization stage before applying speech source separation. Since the objective of these alignment methods is to facilitate the performance of different speech separation algorithms, the quality of the separated speech sources has been analyzed when these

Stage	Parameter	2-peak GCC-PHAT $N_b = 8$, $D = 1$	2-peak GCC-PHAT $N_b = 2$, $D = 8$	Interpolated STLE-CC $N_b = 2$
Alignment	RMSE (ms)	3.44 (1.88)	6.78 (4.37)	6.51 (3.22)
	Bandwith (bps)	128000	4000	125
	CPU load (%)	0.62 (0.12)	0.29 (0.09)	0.17 (0.04)
Separation GCC-PHAT based $Kf_s = 128ms$	SIR (dB)	3.16 (1.02)	3.15 (1.01)	3.09 (1.04)
	Words (%)	89.83 (14.01)	89.96 (14.68)	88.96 (15.89)
	CPU load (%)	27.27 (1.89)	27.27 (1.89)	27.27 (1.89)
Separation GCC-PHAT based Mix 3L	SIR (dB)	3.34 (1.06)	3.34 (1.06)	3.26 (1.09)
	Words (%)	90.91 (13.36)	90.92 (14.20)	90.02 (15.18)
	CPU load (%)	45.24 (4.39)	45.24 (4.39)	45.24 (4.39)
Separation GCC-PHAT based Mix 7L	SIR (dB)	3.40 (1.08)	3.40 (1.07)	3.32 (1.10)
	Words (%)	91.32 (13.02)	91.32 (13.95)	90.36 (15.08)
	CPU load (%)	73.31 (5.85)	73.31 (5.85)	73.31 (5.85)

Table 4.18: Summary of the most significative results (mean (standard deviation)) for $S = 3$ sources and $N = 3$ nodes, in a high reverberating scenario, including synchronization data required for transmission in the WASN to align the speech mixtures for the different alignment methods, and computational complexity measured in time in a distributed set of M nodes, each of them with a single core 1GHz CPU.

synchronization stages are used. However, it must be said that to evaluate the separation quality is not enough aiming at determining the suitability of these alignment algorithms. Since these methods will be used in WASNs, which are sensor networks with important resource constraints, other requirements must be accomplished. Hence, the computational cost associated with these methods and the amount of information required for transmission must be taken into account. Following this idea, Table 4.18 presents a comparative study of the most suitable separation solutions (alignment method + separation method) that considers the separation quality but also the needed bandwidth and the CPU load.

Different versions of the GCC-PHAT and STLE based alignment methods have been studied during this chapter. Two of these versions have been selected to be compared in our study. The chosen versions are briefly described below:

- **2-peak GCC-PHAT**. As proposed in this chapter of the thesis, this method uses the main $S = 2$ peaks of the GCC-PHAT and averages their position to obtain the proposed alignment.

- **Interpolated STLE-CC**. In this method, the decimated STLE with a decimation factor $N = 256$ is evaluated, and the CC is determined using Equation (4.21). Interpolation is then performed using the two neighbors of the most significant point of the CC and a quadratic curve as Equation (4.24) illustrates. Thus, the alignment can be determined with a reduced number of operations.

Looking at Table 4.18 the reader can see that the results of the 2-peak GCC-PHAT based alignment method are presented for different number of bits per sample (N_b) and downsampling factor (D). In the case of the interpolated STLE-CC alignment algorithm, the chosen value of N_b is two, since in the prior sections it was demonstrated that the use of $N_b = 2$ obtains comparable results (slightly worse) to the ones reached with other values of N_b, as for instance $N_b = 8$. Although it is true the amount of information for transmission is lower using $N_b = 2$, what is very important in wireless networks.

As previously mentioned, the chosen separation algorithm is based on two parts: our GCC-PHAT based mixing matrix estimation method, described in Chapter 3 (Proposal 1), and a

separation stage that relies on time-frequency binary masking. In this chapter we have observed that in general terms, $Kf_s = 128$ ms is the most suitable frame length for the separation stage. Aiming at minimizing the presence of musical noise artifacts we introduced in Section 3.4 a method that combines several window sizes, considering it, two of these solutions, MIX 3L and MIX 7L, are shown in Table 4.18.

Table 4.18 presents mean and standard deviation of different results for the three sources ($S = 3$) and three mixtures ($M = 3$) separation problem. For the rest of separation cases studied in this thesis, similar conclusions can be extracted but they are not shown since it is a huge amount of information. Considering the RMSE between the theoretical delays and the estimated ones by the different alignment methods, it is clear that the 2-peak GCC-PHAT based alignment one gets the lowest RMSE (3.44 ms) when $N_b = 8$ and $D = 1$. Bearing in mind these RMSE values, it is clear that avoiding the downsampling ($D = 1$), the accuracy of the method is higher with the disadvantage of requiring much higher bandwidth (128000 bps) and computational load (0.62 %) than the other options. The required bandwidth and CPU load (%) can be reduced if downsampling is done ($D = 8$) at the expense of increasing the RMSE obtained (6.78) ms. On the other hand, the interpolated STLE-CC based alignment method has the advantage of using much lower bandwidth (125 bps) and CPU load (0.17 %) but achieving a RMSE equal to 6.51 ms. At this point it can be said that if computational resources and available bandwidth in the WASN are limited, the interpolated STLE-CC based alignment is the most appropriate, otherwise, the method based on the GCC-PHAT function can be chosen.

Besides RMSE, CPU load and bandwidth, the quality of the separated speech sources must be taken into account inasmuch as we focus on solving the cocktail party problem. In respect of the different versions of the 2-peak GCC-PHAT based alignment method, it can be highlighted that the two versions achieve roughly the same quality scores in terms of both SIR and Correct words (%), independently of the number of frame lengths used. Then, considering both quality measures, RMSE, CPU load (%) and required bandwidth, the downsampled version ($D = 8$) should be chosen since it obtains similar quality outcomes and requires less resources.

The interpolated STLE-CC based alignment method achieves slightly lower separation quality than the GCC-PHAT based method. Examples of these decreases in separation quality are the 3.32 dB of SIR using the MIX 7L method against a SIR=3.40 dB for the 2-peak GCC-PHAT based alignment method. The reader can note that regardless the synchronization solution, the MIX 7L proposal gets the highest quality at the expense of highest CPU load. And so, depending on the available computational resources in the wireless network, one of the three combinations of frame lengths should be chosen.

In view of the above results of CPU load and bandwidth required, the best separation solution for a WASN is to have a central node (sometimes also known as fusion center) that would make all the calculations, such as the STFTs, the calculation of the delays between speech mixtures, the mixture synchronization or the speech source separation. And so, the rest of nodes only transmit their captured speech mixtures to the central one. In this way, the amount of data for transmission is drastically reduced since it has been demonstrated (see Table 4.18) that downsampled speech mixtures can be sent to the central node obtaining quite good speech separation. Concretely, according to each alignment method used, the main task of each secondary node will be:

- If the 2-peak GCC-PHAT based alignment method is used, each secondary node sends the phase information to the central node, using in one case $N_b = 8$ and in other $N_b = 2$ bits per sample with a downsampling factor of $D = 1$ and $D = 8$, respectively.
- In the case of a alignment method based on the interpolated STLE function, the STLE of

each secondary node is transmitted to the central node using $N_b = 2$ bits per sample, but the synchronization data required for transmission to align the signals is reduced by the use of downsampling.

Only in cases in which the computational capability of nodes is limited, or in WASNs with high number of nodes and more bandwidth available, it would be more appropriate to distribute the processing.

4.8 Discussion

In this chapter, we present a novel study of the synchronization problem for WASNs from a signal processing point of view. First, we propose a theoretical framework that allows us to determine and quantize the effects of desynchronization over any STFT-based BSS algorithm. From this framework, the theoretical delay aimed at reducing the effective length of the analysis time frame is determined. This theoretical value lets us align speech mixtures and, consequently, minimize the effects of desynchronization due to acoustic propagation delays. Then, this alignment allows us to use lower analysis frame length with different speech separation algorithms. In the case of our GCC-based mixing matrix estimation stage with binary mask (Proposal 1 in Chapter 3), the most suitable frame length has demonstrated to be 128 ms. This value is very common in separation problems where microphones are close together, or in other words, in cases where propagation delays are not very significant. Additionally because an analysis frame length of 128 ms is used, a speech signal can be considered as nearly stationary, enabling BSS methods to perform better.

According to our proposed theoretical analysis, in the second part of this chapter, two novel alignment methods focused on determining these theoretical delays between speech mixtures are implemented. One of them is based on the GCC-PHAT function, and another one is based on the STLE of the mixtures as is exclusively useful for non-stationary sources. Both methods are tuned up to reduce the amount of synchronization information required for transmission and the computational cost.

From the experiments performed in this work, we conclude that the proposed GCC-PHAT based alignment method represents the best solution in terms of synchronization and separation performance, and in the particular case of non-stationary sources (for instance, speech), the interpolated STLE-CC achieves comparable results with an extremely reduced transmission bandwidth required for the synchronization data and lower CPU load.

In summary, two novel and efficient mixture alignment methods based on a new theoretical analysis have been proposed, allowing us to solve the synchronization problem of WASNs for BSS from a new perspective. These synchronization solutions are able to improve the quality of the separated signals in those cases with wide area WASNs, where sources are placed close to one of the nodes.

4.9 Summary of contributions

The main contributions of this chapter of the thesis are listed below:

- A new approach to solve the synchronization problem due to acoustic propagation delays in WASNs has been proposed. The desynchronization due to acoustic propagation delays is a factor that has apparently not yet been the subject of consideration in academic literature because the vast majority of works deal with the clock phase offset and frequency skew problems.

- Our solution of the synchronization problem is based on signal processing techniques and does not rely on synchronization protocols with message exchanges over wireless communications that consume resources, such as computation power in nodes or bandwidth.

- Our synchronization methodology relies on the analysis of the effects of desynchronization over the frame length used by STFT-based BSS algorithms. According to this study, the theoretical delay between speech mixtures that must be eliminated to synchronize them has been established. Synchronization of speech mixtures has been developed what is an important novelty since in the majority of synchronization methods only single speech signals are covered.

- According to the theoretical delay established in our analysis, two novel alignment methods have been implemented. The first one is a correct interpretation of the use of the GCC-PHAT function when delays between speech mixtures are calculated. The second method has the novelty of using the STLE function for synchronization. Additionally, considering the particularities of WASNs both methods have been designed to reduce the amount of information for transmission.

- Thanks to our synchronization stage prior to speech separation, classical BSS algorithms, like DUET with binary mask, can be used in sensor networks with distances between nodes greater than a few centimeters. This applicability can be observed in two aspects: a) the outcomes in terms of separation quality obtained by the classical separation methods increase considerably when our synchronization stage is included, and b) the frame length used is smaller when our alignment methods are used. Concretely, the values of frame length used are the usual ones when classical microphone networks are used.

- Our novel GCC-PHAT based mixing matrix estimation with binary mask has also been applied to WASNs, outperforming classical separation algorithms like DUET with binary mask.

Some of the contributions of this chapter have been published in the following contributions, [Llerena Aguilar et al., 2012, Llerena et al., 2013a, Llerena et al., 2013b] and two of them more recent are [Llerena-Aguilar et al., 2015, Llerena-Aguilar et al., 2016].

Chapter 5

Conclusions

In this chapter the main contributions of the thesis are presented and analyzed, paying special attention to the obtained results. These contributions will be summarized in Section 5.1. Considering the performed investigations, the main lines of research that remain open and future lines of research will be described in Section 5.2. Finally, a list of published work produced during this thesis is presented in Section 5.3.

5.1 Summary of conclusions

The overall aim of this thesis was to propose several solutions to the cocktail party problem since it can be considered an unresolved problem in certain circumstances, as for instance, in the presence of reverberation. Following this idea, new sound source separation methods have been developed aiming at overcoming some of these problems. Moreover, new useful tools are currently being used to carry out speech source separation, with WASNs being a representative example of it. This type of wireless sensor networks entails many advantages but also some problems that must be considered and solved. An important issue for classical BSS algorithms is the correct synchronization of nodes in wireless sensor networks. Synchronization problems degrade the results obtained by the separation algorithms and so, in this thesis, the synchronization problem in WASNs has been studied and analyzed in a novel way, what has given rise to new synchronization solutions.

Each of the next subsections illustrates the main contributions of this thesis according to the different objectives that have been achieved.

5.1.1 Introduction of new SSS algorithms to outperform methods based on sparsity

After doing an extensive literature review, it has been observed that one of the most known and successful type of SSS methods are those based on sparsity. As mentioned in previous chapters, the main assumption of these algorithms is that sparse representations of sound sources can be achieved in a certain domain, what facilitates the separation of sound sources. In respect of speech signals, the vast majority of separation algorithms work with time-frequency representations of speech because in this transformed domain they can be considered sparse. However, it has been observed that this assumption is not necessarily true since there is a degree of overlap between speech sources and so, these signals can only be considered almost sparse in this domain. Furthermore, there are some factors, as for instance, reverberation which causes delayed versions of the sources signals, increasing the degree of overlap and consequently, algorithms based on sparsity perform worse.

Despite of the aforementioned problems, methods based on sparsity have demonstrated to achieve very good results in many separation scenarios. One of the most important SSS algorithms based on sparsity is DUET, whose, perhaps, principal advantage is to develop source separation in underdetermined problems. For this and other reasons DUET will be used as reference method. While it is true that this method has some important limitations and in this sense, some contributions have been made. These contributions have been carried out considering the two parts in which it can be divided the algorithm: a) the mixing matrix estimation and b) the separation stages.

- First, the behavior of DUET has been thoroughly analyzed in different separation scenarios. In those scenarios, the level of reverberation, the size of the room, the number and locations of the talkers, and the type of microphone array have been varied aimed at evaluating the separation quality that DUET achieves. With respect to the type of microphone array, a two-microphone array with different microphone separations, mutual angles or type of microphones have been tested. It must be highlighted the inclusion of real MEMS microphone responses measured in an anechoic chamber. In addition, the use of different STFT frame lengths has been evaluated within these separation scenarios. From these experiments the most suitable frame lengths and two-microphone array configurations have been determined for the different separation scenarios.

- Once the best configurations for DUET have been established, its separation stage has been studied. This stage of the DUET algorithm uses time-frequency binary masking for separating speech sources. The use of BM entails some problems when sound sources are separated. Bearing it in mind, a masking technique based on l_1-norm minimization has been evaluated. Carrying out the same experiments as in the previous point, both separation stages have been compared as it can be observed in Table 3.7. From these results the outperformance of the separation stage using BM has been demonstrated in all the separation scenarios.

- The mixing matrix estimation stage in DUET is based on clustering time and level differences between two microphones. This clustering process is developed by means of a weighted two-dimensional histogram whose peaks are used to calculate the mixing parameters. The success of this clustering stage depends on how well the clusters are defined. In some situations, these clusters are not well defined and the DUET algorithm does not perform correctly. Aiming at overcoming these limitations a novel sound source separation algorithm is introduced in this thesis. This algorithm contains a new mixing matrix estimation method based on the geometric analysis of the separation scenario. The procedure of this method consists in estimating time differences and from them, level differences are calculated using a geometric relationship between them. Once level and time differences are estimated, the mixing matrix will be calculated. A comparative study between DUET and this algorithm has been developed, demonstrating that our proposal outperforms DUET in all the experiments.

- The correct estimation of time differences plays a key role in our proposed source separation algorithm. In this sense, an important study of classical TDE methods has been made. Those TDE methods have been tested in the different separation scenarios, that is, with different number of speech sources in mixtures, level of reverberation, etc. This comparative study has concluded that GCC-PHAT is the most appropriate for being applied in cocktail party problems. The combination of using the GCC-PHAT method and our mixing matrix

estimation method based on a geometric analysis has proven to obtain highest separation quality than DUET as it can be observed in Table 3.12.

- Evaluating different microphone separations, it has been established that the accuracy of the GCC-PHAT method decreases for distances between microphones of a few centimeters. In order to overcome this limitation, a new TDE algorithm that works in the frequency domain has been implemented. Several experiments have shown that it gets higher accuracy than the GCC-PHAT method for these short distances (see Figures 3.10 and 3.12). This higher accuracy has also been observed when this TDE algorithm is combined with our speech separation method since it reaches slightly better outcomes in terms of speech quality of the separated sources. Table 3.15 illustrates results in terms of speech quality and intelligibility when GCC-PHAT and our new TDE method are used for a microphone separation equal to 5 cm.

- Our novel SSS algorithm has a classical separation stage that is performed by means of time-frequency binary masking. This stage has been chosen since it enables the separation of speech sources in underdetermined problems but it has an important disadvantage since it causes the emergence of musical noise artifacts. A novel musical noise reduction algorithm has been introduced in this thesis. The basic idea of this method is to combine the different time-frequency representations of the separated speech sources when different window sizes are used by the separation algorithm. The new methodology introduces slight improvements on the quality of the separated speech sources (in terms of SIR) regardless the number of speech sources, the level of reverberation, the microphone separation or the separation algorithm used. Nevertheless, these improvements are more evident in terms of speech intelligibility.

To summarize, a new separation method based on a geometric analysis of the separation scenario has been introduced. Looking at the different tables and figures in Chapter 3, it is clear that this proposal outperforms the DUET algorithm in practically all the separation cases, that is, for two and three talkers in rooms, for different microphone array configurations, levels of reverberation, etc. It must be mentioned the clear improvements for reverberant environments because in these scenarios, most of the separation algorithms in the literature do not separate sound sources correctly. Our separation method requires good estimation of time differences and in this sense, a classical method, GCC-PHAT, has been established as the most suitable one for developing it. However, it was observed that GCC-PHAT performs worse when microphone separations are very short, what led us to the development of a new TDE algorithm. Finally, in order to improve the classical separation stage of our separation method which introduces musical noise artifacts, a new methodology that combines the use of several window sizes has been introduced.

5.1.2 Speech mixture synchronization in WASNs to improve the performance of sound source separation methods

The use of WASNs involves a large number of benefits, such as wider areas are covered, larger number of microphones can be used or power and size constraints are avoided. Nevertheless, due to the characteristics of these networks, an important issue is detected: the synchronization problem. This synchronization problem considerably affects many type of signal processing algorithms designed to work in wired networks, BSS algorithms being no exception. It has been analyzed that desynchronization is mainly caused by two factors: a) the clock problem and b)

the different distances between sources and nodes. Making an extensive literature review, it has been observed that the vast majority of synchronization solutions only pay their attention to the first cause. Many synchronization protocols have been designed obtaining very high accuracy when the clock problem is presented. The objective of this thesis is to deal with the second cause, that is, the differences in propagation delay due to different distances between nodes and sources.

Considering the aforementioned problem, the most important contributions of this thesis in this regard are:

- It has been studied how the use of WASNs as the one described in Section 4.3 affects the performance of the DUET algorithm and of our separation method based on the geometric analysis of the separation scenario. The main outcomes of this study are included from Table 4.2 to Table 4.5. In these tables we can see that the separation algorithms need larger frame lengths and achieve low separation quality. $Kf_s = 128$ ms and 256 ms have demonstrated to be the most suitable frame lengths.

- A novel theoretical study of the effects of desynchronization over short-time based algorithms has been introduced. From this study, the requirements of the effective length of the time frame in BSS systems have been established. Specifically, a lower bound of the time frame was determined.

- A new synchronization solution based on signal processing techniques has been proposed. This synchronization stage must be applied before the BSS algorithm. It consists in calculating delays between the speech mixtures captured by the different microphones in WASNs and aligning them according to the delays obtained. This solution has the advantage of not requiring synchronization protocols with message exchanges over wireless communications what uses the available bandwidth.

- The theoretical delay between speech mixtures that helps BSS algorithms to obtain the highest separation quality has been defined. In the vast majority of applications delays are calculated between single signals (only one source signal) with at most noise components and so, there is not a universal definition of the delay between mixture signals.

- The effects of aligning speech mixtures using our theoretical delay have been observed when the DUET algorithm and our separation method are used. It was demonstrated that higher separation quality in terms of both speech quality and speech intelligibility is obtained. Furthermore, the separation algorithms achieve better results with shorter frame lengths when the synchronization stage is included.

- Aiming at aligning speech mixtures, a comparative study of classical TDE algorithms has been made. This study concludes that the most robust method for our separation scenarios (reverberant rooms) is the GCC-PHAT one. However, it has been checked that the delay that GCC-PHAT obtains does not correspond with our theoretical delay. While our theoretical delay is the difference between the average value of the propagation delays of all the sources at each microphone, the GCC-PHAT method obtains the difference of the propagation delays of only the dominant speech source.

- Two new TDE algorithms according to the definition of our theoretical delay have been implemented. On the one hand a reinterpretation of the GCC-PHAT function, that is, instead of using only one peak (the predominant one), a number of peaks equal to the

number of speech sources in mixtures is searched in the GCC-PHAT function. On the other hand a novel method based on the STLE function. Using both TDE methods, the separation algorithms get quite good separation quality in the different separation scenarios. Note that both methods have also been designed to reduce their computational load.

- In order to minimize the amount of synchronization information for transmission, a synchronization schema is proposed in our WASNs. It consists of one reference node that develops all the tasks (synchronization and separation stages) and the rest of them are secondary nodes, with the only function of transmitting the signal captured by their microphones to the reference node.

- Additionally, to further reduce the amount of synchronization information for transmission, optimized versions of the two TDE methods have been implemented. These versions represent a trade-off between the reduction of the amount of information to be transmitted and the quality of the separated sources. The TDE method based on the STLE function has demonstrated to be more efficient, obtaining slightly worse results than the GCC-PHAT based one in terms of speech quality and speech intelligibility as it can be observed in Table 4.18. For instance, the Interpolated STLE-CC method needs a bandwidth equal to 125 bps, while one optimize version of the 2-peak GCC-PHAT based one uses 4 Kbps when downsampling has been applied and the number of bits to code the phase is minimum ($N_b = 2$). From Table 4.18 we can observed that both separation solutions (TDE method + our separation algorithm) can be implemented in real-time.

To conclude, the application of traditional BSS algorithms in WASNs requires to solve the synchronization problem due to the differences in propagation delays. In this sense, a new mathematical development is introduced to determine the effects of desynchronization on short-time based algorithms and to establish the theoretical delay between mixtures. From it, two novel TDE methods have been implemented paying special attention to the amount of information for transmission that they use. Furthermore, considering the computational load associated to them, these methods have been optimized. Very good separation quality has been obtained with the different solutions.

5.2 Future research lines

Many challenges remain open in the field of blind source separation. During this doctoral thesis, the objective was to solve some important problems in this field. Interesting solutions have been proposed, but it must be said that from the work of this thesis, different promising research lines could be open. Considering the research carried out in this thesis, the following future research lines are listed below:

- The algorithms proposed during this thesis have been evaluated in simulated rooms with different dimensions and important levels of reverberation, obtaining quite good results. The following step must be to test these algorithms in different noisy environments, varying the level and type of noise source signals. Finally, our proposals should be used in real rooms.

- The new speech separation method proposed in Chapter 3 has been developed for two-microphone arrays. The results are very promising since with the simplest microphone array our method achieves quite good separation quality. The following step must be to

generalize this method to perform separation with microphone arrays with more than two microphones. It seems intuitive that if the number of microphones increases, and so the available information, higher separation quality will be obtained.

- The computational cost associated with our new separation method has been analyzed, concluding that it can be applied to real-time solutions since it only entails the calculation of time differences, and level differences are directly calculated using the theoretical relationship. It must be mentioned that in the problems addressed in this thesis (2-3 microphones, 2-3 speech sources), this method method can work with only one processor, but for more complex separation problems (higher number of microphones and sources), it will be interesting to parallelize processes using tools as graphics processor units.

- The novel TDE methods introduced in Chapter 4 work properly for two and three speech sources. Their performance for higher number of speech sources in mixtures must be studied.

- Continuing the research of BSS methods based on sparsity, one goal may be to achieve more sparse representations of speech. One way of obtaining these sparse representations could be the use of non-spatial cues, as for instance pitch, which are very common in single-channel separation problems.

5.3 List of publications

The following paragraphs present a list of published work produced during the course of candidature for the degree. The works accepted and published in journals and conference proceedings that support the main contributions of the thesis are listed below:

- **International journals**

  - Llerena-Aguilar, C., Gil-Pita, R., and Rosa-Zurera, M. (2016). A novel speech separation method based on the geometric analysis of the problem. *IEEE Transactions on Audio, Speech and Language Processing.* (Working on it)
  - Llerena-Aguilar, C., Gil-Pita, R., Rosa-Zurera, M., and Utrilla-Manso, M. (2016). Efficient speech mixture synchronization in wireless acoustic sensor networks oriented to blind source separation problems. *Signal Processing.* (Working on it)
  - Llerena-Aguilar, C., Gil-Pita, R., Rosa-Zurera M., Ayllón, D., Utrilla-Manso, M., and Llerena, F. (2016). Synchronization based on mixture alignment for sound source separation in wireless acoustic sensor networks. *Signal Processing*, 118:177-187.
  - Llerena, C., Gil-Pita, R., Álvarez, L., and Rosa-Zurera, M. (2013). Synchronizing speech mixtures in speech separation problems under reverberant conditions. *Artificial Intelligence and Soft Computing*, 7894(1):568-579.

- **International conferences**

  - Llerena-Aguilar, C., Ramos-Auñón, G., Llerena-Agular, F.J., Sánchez-Hevia, H.A., and Rosa-Zurera, M. (2015). Improving speech mixture synchronization in blind source separation problems. In *Audio Engineering Society Convention 138.* Audio Engineering Society.

- Llerena, C., Álvarez, L., Gil-Pita, R., and Rosa-Zurera, M. (2013). Introducing synchronization of speech mixtures in blind sparse separation problems. In *Audio Engineering Society Convention 134*. Audio Engineering Society.
- Llerena, C., Gil, R., Álvarez, L., Cuadra, L., and Ayllón, D. (2012). Comparing two methods based on time-frequency analysis to estimate the instantaneous mixing matrix in blind audio source separation. In *Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*. Audio Engineering Society.

- **National conferences**

  - Llerena-Aguilar, C., Gil-Pita, R., Rosa-Zurera, M., Ayllón, D., and Sánchez-Hevia, H.A. (2014). Introducción de algoritmos clásicos de separación de mezclas de voz sincronizadas en redes wireless. In *XXIX Simposio nacional de la Unión Científica Internacional de Radio (URSI)*.
  - Llerena-Aguilar, C., Álvarez-Pérez, L., Ayllón-Álvarez, D., Gil-Pita, R., and Rosa-Zurera, M. (2011). Comparativa de dos métodos para la estimación de la matriz de mezclas en separación ciega de fuentes. In *XXVI Simposio nacional de la Unión Científica Internacional de Radio (URSI)*.

Some other conference papers and journals also related to the content of this thesis are the next:

- **International journals**

  - Gil-Pita, R., Ayllon, D., Ranilla, J., LLerena-Aguilar, C., and Diaz, I. (2015). A computationally efficient sound environment classifier for hearing aids. *Biomedical Engineering, IEEE Transactions on*, 62(10).
  - Seoane, F., Mohino-Herranz, I., Ferreira, J., Alvarez, L., Buendia, R., Ayllón, D., Llerena, C., and Gil-Pita, R. (2014). Wearable biomedical measurement systems for assessment of mental stress of combatants in real time. *Sensors*, 14(4):7120-7141.
  - Álvarez-Pérez, L., Alexandre-Cortizo, E., Cuadra-Rodríguez, L., Gil-Pita, R., and Llerena-Aguilar, C. (2013). Speech enhancement in noisy environments in hearing aids driven by a tailored gain function based on a Gaussian mixture model. *Artificial Intelligence and Soft Computing*, 7894(1):503-514.
  - Seoane, F., Ferreira, J., Alvarez, L., Buendia, R., Ayllón, D., Llerena, C., and Gil-Pita, R. (2013). Sensorized garments and textrode-enabled measurement instrumentation for ambulatory assessment of the autonomic nervous system response in the atrec project. *Sensors*, 13(7):8997-9015.

- **International conferences**

  - Sánchez-Hevia, H. A., Llerena-Aguilar, C., Ramos-Auñón, G., and Gil-Pita, R. (2015). Automatic vocal percussion transcription aimed at mobile music production. In *Audio Engineering Society Convention 138*. Audio Engineering Society.
  - Ramos-Aunón, G., Mohino-Herranz, M.I., Sánchez-Hevia, H.A., Llerena-Aguilar, C., and Ayllón-Álvarez, D. (2015). Two-sensor EEG-based stress detection system. In *6th IASTED International Conference on Computational Intelligence*.

- Álvarez, L., Alexandre, E., Llerena, C., Gil-Pita, R., and Rosa-Zurera, M. (2013). Combination of growing and pruning algorithms for multilayer perceptrons for speech/music/noise classification in digital hearing aids. In *Audio Engineering Society Convention 134*. Audio Engineering Society.
- Ferreira, J., Álvarez, L., Buendía, R., Ayllón, D., Llerena, C., Gil-Pita, R., and Seoane, F. (2013). Bioimpedance-based wearable measurement instrumentation for studying the autonomic nerve system response to stressful working conditions. In *Journal of Physics: Conference Series*, volume 434, page 012-015. IOP Publishing.
- Mohino-Herranz, M.I., Goni-Ibaceta, M., Álvarez-Pérez, L., Llerena-Aguilar, C., and Gil-Pita, R. (2013). Detection of emotions and stress through speech analysis. In *10th IASTED International Conference on Signal Processing, Pattern Recognition and Applications*.
- Goni-Ibaceta, M., Mohino-Herranz, M.I., Llerena-Aguilar, C., Gil-Pita, R., and Rosa-Zurera, M. (2013). Feature selection in mental stress analysis using multiple biological signals. In *10th IASTED International Conference on Signal Processing, Pattern Recognition and Applications*.
- Ayllón, D., Benito-Olivares, V., Llerena-Aguilar, C., Gil Pita, R., and Rosa Zurera, M. (2012). Three-dimensional microphone array for speech enhancement in hands-free systems for cars. In *Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*. Audio Engineering Society.
- Ayllón, D., Gil-Pita, R., Jarabo-Amores, P., Rosa-Zurera, M., and Llerena-Aguilar, C. (2011). Energy-weighted mean shift algorithm for speech source separation. In *Statistical Signal Processing Workshop (SSP), 2011 IEEE*, pages 785-788. IEEE.
- Álvarez, L., Llerena, C., Alexandre, E., Gil-Pita, R., and Rosa-Zurera, M. (2011). Selection of approximated activation functions in neural network-based sound classifiers for digital hearing aids. In *Audio Engineering Society Convention 130*. Audio Engineering Society.

- **National conferences**

  - Ayllón-Álvarez, D., Gil-Pita, R., Rosa-Zurera, M., Llerena-Aguilar, C., and Sánchez, H. (2014). Design of energy-efficient speech enhancement algorithms for binaural hearing aids. In *XXIX Simposio nacional de la Unión Científica Internacional de Radio (URSI)*.
  - Ayllón-Álvarez, D., Llerena-Aguilar, C., and Gil-Pita, R. (2011). Separación de fuentes de voz mediante mean shifts. In *XXVI Simposio nacional de la Unión Científica Internacional de Radio (URSI)*.
  - Álvarez-Pérez, L., Llerena-Aguilar, C., and Alexandre-Cortizo, E. (2011). Selección de funciones de activación aproximadas para clasificadores basados en redes neuronales en audífonos digitales. In *XXVI Simposio nacional de la Unión Científica Internacional de Radio (URSI)*.

Bibliography

[Abrard and Deville, 2005] Abrard, F. and Deville, Y. (2005). A time–frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. *Signal Processing*, 85(7):1389–1403.

[Akyildiz et al., 2002] Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., and Cayirci, E. (2002). Wireless sensor networks: a survey. *Computer networks*, 38(4):393–422.

[Allen and Berkley, 1979] Allen, J. B. and Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950.

[Amari, 1998] Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276.

[Anemüller and Kollmeier, 2003] Anemüller, J. and Kollmeier, B. (2003). Adaptive separation of acoustic sources for anechoic conditions: A constrained frequency domain approach. *Speech Communication*, 39(1):79–95.

[Anzalone et al., 2006] Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (2006). Determination of the potential benefit of time-frequency gain manipulation. *Ear and hearing*, 27(5):480.

[Araki et al., 2004] Araki, S., Makino, S., Blin, A., Mukai, R., and Sawada, H. (2004). Underdetermined blind separation for speech in real environments with sparseness and ICA. In *Acoustics, Speech, and Signal Processing. Proceedings (ICASSP'04). IEEE International Conference on*, volume 3, pages 881–884. IEEE.

[Araki et al., 2006] Araki, S., Sawada, H., Mukai, R., and Makino, S. (2006). Blind sparse source separation with spatially smoothed time-frequency masking. In *Proc. 2006 Int. Workshop on Acoustic Echo and Noise Control*.

[Araki et al., 2007] Araki, S., Sawada, H., Mukai, R., and Makino, S. (2007). Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. *Signal Processing*, 87(8):1833–1847.

[Arvind, 1994] Arvind, K. (1994). Probabilistic clock synchronization in distributed systems. *Parallel and Distributed Systems, IEEE Transactions on*, 5(5):474–487.

[Ayllón et al., 2013] Ayllón, D., Gil-Pita, R., and Rosa-Zurera, M. (2013). Design of microphone arrays for hearing aids optimized to unknown subjects. *Signal Processing*, 93(11):3239–3250.

[Bahari et al., 2015] Bahari, M. H., Bertrand, A., and Moonen, M. (2015). Blind sampling rate offset estimation based on coherence drift in wireless acoustic sensor networks. In *Proc. of the European Signal Processing Conference*, pages 2336–2340.

[Bechler et al., 2003] Bechler, D., Kroschel, K., et al. (2003). Considering the second peak in the GCC function for multi-source TDOA estimation with a microphone array. In *Proc. of IWAENC*, pages 315–318. Citeseer.

[Bell and Sejnowski, 1995] Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159.

[Benesty et al., 2004] Benesty, J., Chen, J., and Huang, Y. (2004). Time-delay estimation via linear interpolation and cross correlation. *Speech and Audio Processing, IEEE Transactions on*, 12(5):509–519.

[Benesty et al., 2008] Benesty, J., Chen, J., and Huang, Y. (2008). *Microphone array signal processing*, volume 1. Springer Science & Business Media.

[Benesty et al., 2007] Benesty, J., Chen, J., Huang, Y., and Dmochowski, J. (2007). On microphone-array beamforming from a MIMO acoustic signal processing perspective. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(3):1053–1065.

[Bertrand, 2011] Bertrand, A. (2011). Applications and trends in wireless acoustic sensor networks: a signal processing perspective. In *Communications and Vehicular Technology in the Benelux (SCVT), 18th IEEE Symposium on*, pages 1–6. IEEE.

[Bofill and Zibulevsky, 2001] Bofill, P. and Zibulevsky, M. (2001). Underdetermined blind source separation using sparse representations. *Signal processing*, 81(11):2353–2362.

[Boyd et al., 2006] Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. (2006). Randomized gossip algorithms. *IEEE/ACM Transactions on Networking (TON)*, 14(SI):2508–2530.

[Brandstein and Ward, 2001] Brandstein, M. and Ward, D. (2001). *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media.

[Bregman, 1990] Bregman, A. S. (1990). Auditory scene analysis: The perceptual organization of sound. 1990.

[Broadbent, 2013] Broadbent, D. E. (2013). *Perception and communication*. Elsevier.

[Brown and Cooke, 1994] Brown, G. J. and Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech and Language*, 8(4):297–336.

[Buckley, 1986] Buckley, K. M. (1986). Broad-band beamforming and the generalized sidelobe canceller. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(5):1322–1323.

[Cao and Liu, 1996] Cao, X.-R. and Liu, R.-W. (1996). General approach to blind source separation. *Signal Processing, IEEE Transactions on*, 44(3):562–571.

[Capon, 1969] Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418.

[Cappé, 1994] Cappé, O. (1994). Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE transactions on Speech and Audio Processing*, 2(2):345–349.

[Cardoso, 1998] Cardoso, J.-F. C. (1998). Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025.

[Carter, 1987] Carter, G. C. (1987). Coherence and time delay estimation. *Proceedings of the IEEE*, 75(2):236–255.

[Carter, 1993] Carter, G. C. (1993). *Coherence and time delay estimation: an applied tutorial for research, development, test, and evaluation engineers*. IEEE.

[Carter et al., 1973] Carter, G. C., Nuttall, A. H., and Cable, P. G. (1973). The smoothed coherence transform. *Proceedings of the IEEE*, 61(10):1497–1498.

[Chen et al., 2003] Chen, J., Benesty, J., and Huang, Y. (2003). Robust time delay estimation exploiting redundancy among multiple microphones. *Speech and Audio Processing, IEEE Transactions on*, 11(6):549–557.

[Chen et al., 1998] Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61.

[Cherry, 1953] 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.

[Ching and Chan, 1988] Ching, P. and Chan, Y. (1988). Adaptive time delay estimation with constraints. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(4):599–602.

[Chowdhury et al., 2013] Chowdhury, R. H., Reaz, M. B., Ali, M., Bakar, A. A., Chellappan, K., and Chang, T. G. (2013). Surface electromyography signal processing and classification techniques. *Sensors*, 13(9):12431–12466.

[Cichocki and Amari, 2002] Cichocki, A. and Amari, S.-i. (2002). *Adaptive blind signal and image processing: learning algorithms and applications*, volume 1. John Wiley & Sons.

[Cobos and Lopez, 2010] Cobos, M. and Lopez, J. J. (2010). Two-microphone separation of speech mixtures based on interclass variance maximization. *The Journal of the Acoustical Society of America*, 127(3):1661–1672.

[Comon, 1994] Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.

[Cooke, 1991] Cooke, M. (1991). *Modelling auditory processing and organisation*. PhD thesis, Department of Computer Science, University of Shelfield.

[Cooke, 2005] Cooke, M. (2005). *Modelling auditory processing and organisation*, volume 7. Cambridge University Press.

[Cooley and Tukey, 1965] Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of computation*, 19(90):297–301.

[Couvreur and Couvreur, 2004] Couvreur, L. and Couvreur, C. (2004). Blind model selection for automatic speech recognition in reverberant environments. In *Real World Speech Processing*, pages 115–129. Springer.

[Couvreur et al., 2001] Couvreur, L., Ris, C., and Couvreur, C. (2001). Model-based blind estimation of reverberation time: application to robust ASR in reverberant environments. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, volume 1, pages 2631–2634.

[Coviello and Sibul, 2004] Coviello, C. and Sibul, L. (2004). Blind source separation and beamforming: algebraic technique analysis. *Aerospace and Electronic Systems, IEEE Transactions on*, 40(1):221–235.

[Cox et al., 1986] Cox, H., Zeskind, R. M., and Kooij, T. (1986). Practical supergain. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(3):393–398.

[Darwin and Hukin, 1999] Darwin, C. and Hukin, R. (1999). Auditory objects of attention: the role of interaural time differences. *Journal of Experimental Psychology: Human perception and performance*, 25(3):617.

[Dau et al., 1996] Dau, T., Püschel, D., and Kohlrausch, A. (1996). A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements. *The Journal of the Acoustical Society of America*, 99(6):3623–3631.

[Daubechies et al., 1992] Daubechies, I. et al. (1992). *Ten lectures on wavelets*, volume 61. SIAM.

[Daudet and Sandler, 2004] Daudet, L. and Sandler, M. (2004). MDCT analysis of sinusoids: exact results and applications to coding artifacts reduction. *Speech and Audio Processing, IEEE Transactions on*, 12(3):302–312.

[Daudet and Torrésani, 2002] Daudet, L. and Torrésani, B. (2002). Hybrid representations for audiophonic signal encoding. *Signal Processing*, 82(11):1595–1617.

[Davies and Mitianoudis, 2004] Davies, M. and Mitianoudis, N. (2004). Simple mixture model for sparse overcomplete ICA. *IEE Proceedings-Vision, Image and Signal Processing*, 151(1):35–43.

[De Fréin and Rickard, 2011] De Fréin, R. and Rickard, S. T. (2011). The synchronized short-time-Fourier-transform: properties and definitions for multichannel source separation. *Signal Processing, IEEE Transactions on*, 59(1):91–103.

[de Veth et al., 2001] de Veth, J., Mauuary, L., Noe, B., de Wet, F., Sienel, J., Boves, L., and Jouvet, D. (2001). Feature vector selection to improve ASR robustness in noisy conditions. In *INTERSPEECH*, pages 201–204.

[Deller Jr et al., 1993] Deller Jr, J. R., Proakis, J. G., and Hansen, J. H. (1993). *Discrete time processing of speech signals*. Prentice Hall PTR.

[Deng and Geisler, 1987] Deng, L. and Geisler, C. D. (1987). A composite auditory model for processing speech sounds. *The Journal of the Acoustical Society of America*, 82(6):2001–2012.

[Doclo and Moonen, 2003] Doclo, S. and Moonen, M. (2003). Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments. *EURASIP Journal on Applied Signal Processing*, pages 1110–1124.

[Donoho and Elad, 2002] Donoho, D. L. and Elad, M. (2002). Maximal sparsity representation via l_1 minimization. *IEEE Trans. Inf. Theory*.

[Donoho and Elad, 2003] Donoho, D. L. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202.

[Donohue et al., 2007] Donohue, K. D., Agrinsoni, A., and Hannemann, J. (2007). Audio signal delay estimation using partial whitening. In *SoutheastCon, 2007. Proceedings. IEEE*, pages 466–471. IEEE.

[Duda et al., 2000] Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification (Pt. 2)*. Wiley-Interscience.

[Ehrenberg et al., 1978] Ehrenberg, J., Ewart, T., and Morris, R. (1978). Signal-processing techniques for resolving individual pulses in a multipath signal. *The Journal of the Acoustical Society of America*, 63(6):1861–1865.

[Elad and Bruckstein, 2002] Elad, M. and Bruckstein, A. M. (2002). A generalized uncertainty principle and sparse representation in pairs of bases. *Information Theory, IEEE Transactions on*, 48(9):2558–2567.

[Ellis, 1996] Ellis, D. P. (1996). *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachusetts Institute of Technology.

[Elson et al., 2002] Elson, J., Girod, L., and Estrin, D. (2002). Fine-grained network time synchronization using reference broadcasts. *ACM SIGOPS Operating Systems Review*, 36(SI):147–163.

[Elson and Römer, 2003] Elson, J. and Römer, K. (2003). Wireless sensor networks: A new regime for time synchronization. *ACM SIGCOMM Computer Communication Review*, 33(1):149–154.

[Ephraim et al., 2003] Ephraim, Y., Lev-Ari, H., and Roberts, W. J. (2003). A brief survey of speech enhancement 1. *The Electrical Engineering Handbook*, pages 12–15.

[Er and Cantoni, 1983] Er, M. H. and Cantoni, A. (1983). Derivative constraints for broadband element space antenna array processors. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 31(6):1378–1393.

[Esquef et al., 2002] Esquef, P. A., Välimäki, V., and Karjalainen, M. (2002). Restoration and enhancement of solo guitar recordings based on sound source modeling. *Journal of the Audio Engineering Society*, 50(4):227–236.

[Etter and Stearns, 1981] Etter, D. M. and Stearns, S. D. (1981). Adaptive estimation of time delays in sampled data systems. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 29(3):582–587.

[Fabrizio and Farina, 2011] Fabrizio, G. A. and Farina, A. (2011). Exploiting multipath for blind source separation with sensor arrays. In *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pages 2536–2539. IEEE.

[Feder and Weinstein, 1988] Feder, M. and Weinstein, E. (1988). Parameter estimation of superimposed signals using the EM algorithm. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(4):477–489.

[Févotte and Doncarli, 2004] Févotte, C. and Doncarli, C. (2004). Two contributions to blind source separation using time-frequency distributions. *Signal Processing Letters, IEEE*, 11(3):386–389.

[Févotte and Godsill, 2005] Févotte, C. and Godsill, S. J. (2005). A bayesian approach to time-frequency based blind source separation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA*. Citeseer.

[Flanagan et al., 1985] Flanagan, J., Johnston, J., Zahn, R., and Elko, G. (1985). Computer-steered microphone arrays for sound transduction in large rooms. *The Journal of the Acoustical Society of America*, 78(5):1508–1518.

[Frost III, 1972] Frost III, O. L. (1972). An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8):926–935.

[Fuchs, 1999] Fuchs, J.-J. (1999). Multipath time-delay detection and estimation. *Signal Processing, IEEE Transactions on*, 47(1):237–243.

[Gannot et al., 2001] Gannot, S., Burshtein, D., and Weinstein, E. (2001). Signal enhancement using beamforming and nonstationarity with applications to speech. *Signal Processing, IEEE Transactions on*, 49(8):1614–1626.

[Gil-Pita et al., 2015] Gil-Pita, R., Ayllon, D., Ranilla, J., LLerena-Aguilar, C., and Diaz, I. (2015). A computationally efficient sound environment classifier for hearing aids. *Biomedical Engineering, IEEE Transactions on*, 62(10).

[Gilbert and Morgan, 1955] Gilbert, E. and Morgan, S. (1955). Optimum design of directive antenna arrays subject to random variations. *Bell System Technical Journal*, 34(3):637–663.

[Goh et al., 1998] Goh, Z., Tan, K.-C., and Tan, B. (1998). Postprocessing method for suppressing musical noise generated by spectral subtraction. *Speech and Audio Processing, IEEE Transactions on*, 6(3):287–292.

[Griffin et al., 2015] Griffin, A., Alexandridis, A., Pavlidi, D., Mastorakis, Y., and Mouchtaris, A. (2015). Localizing multiple audio sources in a wireless acoustic sensor network. *Signal Processing*, 107:54–67.

[Griffiths and Jim, 1982] Griffiths, L. J. and Jim, C. W. (1982). An alternative approach to linearly constrained adaptive beamforming. *Antennas and Propagation, IEEE Transactions on*, 30(1):27–34.

[Grünbaum, 2003] Grünbaum, B. (2003). *Neighborly Polytopes*. Springer.

[Günel et al., 2008] Günel, B., Hacihabiboğlu, H., and Kondoz, A. M. (2008). Acoustic source separation of convolutive mixtures based on intensity vector statistics. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(4):748–756.

[Habets et al., 2010] Habets, E., Benesty, J., Cohen, I., Gannot, S., and Dmochowski, J. (2010). New insights into the MVDR beamformer in room acoustics. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(1):158–170.

[Habets et al., 2009] Habets, E. A., Benesty, J., Gannot, S., Naylor, P., Cohen, I., et al. (2009). On the application of the LCMV beamformer to speech enhancement. In *Applications of Signal Processing to Audio and Acoustics (WASPAA'09), IEEE Workshop on*, pages 141–144. IEEE.

[Herault and Jutten, 1986] Herault, J. and Jutten, C. (1986). Space or time adaptive signal processing by neural network models. In *Neural networks for computing*, pages 206–211. AIP Publishing.

[Hu and Wang, 2001] Hu, G. and Wang, D. (2001). Speech segregation based on pitch tracking and amplitude modulation. In *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on the*, pages 79–82. IEEE.

[Hu and Wang, 2002a] Hu, G. and Wang, D. (2002a). Monaural speech separation. In *Advances in neural information processing systems*, pages 1221–1228.

[Hu and Wang, 2002b] Hu, G. and Wang, D. (2002b). On amplitude modulation for monaural speech segregation. In *Proceedings of the 2002 International Joint Conference on Neural Networks, IJCNN*, volume 2, pages 12–17.

[Hu and Wang, 2004] Hu, G. and Wang, D. (2004). Monaural speech segregation based on pitch tracking and amplitude modulation. *Neural Networks, IEEE Transactions on*, 15(5):1135–1150.

[Hu and Wang, 2006] Hu, G. and Wang, D. (2006). An auditory scene analysis approach to monaural speech segregation. *Topics in acoustic echo and noise control*, pages 485–515.

[Hu and Wang, 2007] Hu, G. and Wang, D. (2007). Auditory segmentation based on onset and offset analysis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(2):396–405.

[Hu and Wang, 2010] Hu, G. and Wang, D. (2010). A tandem algorithm for pitch estimation and voiced speech segregation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(8):2067–2079.

[Hu and Loizou, 2007] Hu, Y. and Loizou, P. C. (2007). Subjective comparison and evaluation of speech enhancement algorithms. *Speech communication*, 49(7):588–601.

[Hu and Loizou, 2008] Hu, Y. and Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1):229–238.

[Huang et al., 2001] Huang, X., Acero, A., Hon, H.-W., and Foreword By-Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice Hall PTR.

[Huang and Benesty, 2003] Huang, Y. and Benesty, J. (2003). Adaptive multichannel time delay estimation based on blind system identification for acoustic source localization. In *Adaptive Signal Processing*, pages 227–247. Springer.

[Huang et al., 1999] Huang, Y., Benesty, J., and Elko, G. W. (1999). Adaptive eigenvalue decomposition algorithm for real time acoustic source localization system. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 2, pages 937–940. IEEE.

[Hurley and Rickard, 2009] Hurley, N. and Rickard, S. (2009). Comparing measures of sparsity. *Information Theory, IEEE Transactions on*, 55(10):4723–4741.

[Hyvärinen, 1999] Hyvärinen, A. (1999). Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural computation*, 11(7):1739–1768.

[Hyvärinen and Oja, 1997] Hyvärinen, A. and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492.

[Jacovitti and Scarano, 1993] Jacovitti, G. and Scarano, G. (1993). Discrete time techniques for time delay estimation. *Signal Processing, IEEE Transactions on*, 41(2):525–533.

[Jiang et al., 2014] Jiang, Y., Wang, D., Liu, R., and Feng, Z. (2014). Binaural classification for reverberant speech segregation using deep neural networks. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(12):2112–2121.

[Jordan and Bach, 2005] Jordan, M. I. and Bach, F. R. (2005). Blind one-microphone speech separation: A spectral learning approach. In *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, volume 17, page 65. MIT Press.

[Jutten et al., 1985] Jutten, C., Herault, J., and Ans, B. (1985). Detection de grandeurs primitives dans un message composite par une architecture de calcul neuromimetrique en apprentissage non supervise. In *Proc. Gretsi*.

[Kaneda and Ohga, 1986] Kaneda, Y. and Ohga, J. (1986). Adaptive microphone-array system for noise reduction. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(6):1391–1400.

[Kates and Arehart, 2005] Kates, J. M. and Arehart, K. H. (2005). Coherence and the speech intelligibility index. *The Journal of the Acoustical Society of America*, 117(4):2224–2237.

[Kidd et al., 2015] Kidd, G., Mason, C. R., Best, V., and Swaminathan, J. (2015). Benefits of acoustic beamforming for solving the cocktail party problem. *Trends in hearing*, 19.

[Kim and Loizou, 2010] Kim, G. and Loizou, P. C. (2010). Improving speech intelligibility in noise using a binary mask that is based on magnitude spectrum constraints. *Signal Processing Letters, IEEE*, 17(12):1010–1013.

[Kirlin et al., 1981] Kirlin, R. L., Moore, D. F., and Kubichek, R. F. (1981). Improvement of delay measurements from sonar arrays via sequential state estimation. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 29(3):514–519.

[Kleiner et al., 1993] Kleiner, M., Dalenbäck, B.-I., and Svensson, P. (1993). Auralization-an overview. *Journal of the Audio Engineering Society*, 41(11):861–875.

[Knapp and Carter, 1976] Knapp, C. H. and Carter, G. C. (1976). The generalized correlation method for estimation of time delay. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(4):320–327.

[Koldovsky et al., 2006] Koldovsky, Z., Tichavsky, P., and Oja, E. (2006). Efficient variant of algorithm fastICA for independent component analysis attaining the Cramer-Rao lower bound. *Neural Networks, IEEE Transactions on*, 17(5):1265–1277.

[Kollmeier and Koch, 1994] Kollmeier, B. and Koch, R. (1994). Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *The Journal of the Acoustical Society of America*, 95(3):1593–1602.

[Kollmeier et al., 1993] Kollmeier, B., Peissig, J., and Hohmann, V. (1993). Real-time multiband dynamic compression and noise reduction for binaural hearing aids. *Journal of rehabilitation research and development*, 30:82–82.

[Kolossa and Orglmeister, 2004] Kolossa, D. and Orglmeister, R. (2004). Nonlinear postprocessing for blind speech separation. In *Independent Component Analysis and Blind Signal Separation*, pages 832–839. Springer.

[Krim and Viberg, 1996] Krim, H. and Viberg, M. (1996). Two decades of array signal processing research: the parametric approach. *Signal Processing Magazine, IEEE*, 13(4):67–94.

[Kulowski, 1985] Kulowski, A. (1985). Algorithmic representation of the ray tracing technique. *Applied Acoustics*, 18(6):449–469.

[Lee et al., 2000] Lee, T.-W., Girolami, M., Bell, A. J., and Sejnowski, T. J. (2000). A unifying information-theoretic framework for independent component analysis. *Computers and Mathematics with Applications*, 39(11):1–21.

[Lee et al., 1999] Lee, T.-W., Lewicki, M. S., Girolami, M., and Sejnowski, T. J. (1999). Blind source separation of more sources than mixtures using overcomplete representations. *Signal Processing Letters, IEEE*, 6(4):87–90.

[Lee et al., 1998] Lee, T.-W., Ziehe, A., Orglmeister, R., and Sejnowski, T. (1998). Combining time-delayed decorrelation and ICA: Towards solving the cocktail party problem. In *Acoustics, Speech and Signal Processing. Proceedings of the 1998 IEEE International Conference on*, volume 2, pages 1249–1252. IEEE.

[Lemmon et al., 2000] Lemmon, M. D., Ganguly, J., and Xia, L. (2000). Model-based clock synchronization in networks with drifting clocks. In *Proceedings of 2000 Pacific Rim International Symposium on Dependable Computing*, page 177. IEEE.

[Lewicki and Sejnowski, 2000] Lewicki, M. S. and Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural computation*, 12(2):337–365.

[Li et al., 2001] Li, M., McAllister, H. G., Black, N. D., Pérez, D., and Adrián, T. (2001). Perceptual time-frequency subtraction algorithm for noise reduction in hearing aids. *Biomedical Engineering, IEEE Transactions on*, 48(9):979–988.

[Li and Loizou, 2008] Li, N. and Loizou, P. C. (2008). Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *The Journal of the Acoustical Society of America*, 123(3):1673–1682.

[Li et al., 2010] Li, P., Guan, Y., Wang, S., Xu, B., and Liu, W. (2010). Monaural speech separation based on MAXVQ and CASA for robust speech recognition. *Computer Speech and Language*, 24(1):30–44.

[Li et al., 2004] Li, Y., Cichocki, A., and Amari, S.-i. (2004). Analysis of sparse representation and blind source separation. *Neural computation*, 16(6):1193–1234.

[Linsker, 1989] Linsker, R. (1989). An application of the principle of maximum information preservation to linear systems. In *Advances in neural information processing systems*, pages 186–194.

[Liu et al., 2001] Liu, C., Wheeler, B. C., O' Brien Jr, W. D., Lansing, C. R., Bilger, R. C., Jones, D. L., and Feng, A. S. (2001). A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers. *J. Acoust. Soc. Am*, 110(6):3218–3231.

[Llerena et al., 2013a] Llerena, C., Álvarez, L., Gil-Pita, R., and Rosa-Zurera, M. (2013a). Introducing synchronization of speech mixtures in blind sparse separation problems. In *Audio Engineering Society Convention 134*. Audio Engineering Society.

[Llerena et al., 2012] Llerena, C., Gil, R., Álvarez, L., Cuadra, L., and Ayllón, D. (2012). Comparing two methods based on time-frequency analysis to estimate the instantaneous mixing matrix in blind audio source separation. In *Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*. Audio Engineering Society.

[Llerena et al., 2013b] Llerena, C., Gil-Pita, R., Álvarez, L., and Rosa-Zurera, M. (2013b). Synchronizing speech mixtures in speech separation problems under reverberant conditions. In *Artificial Intelligence and Soft Computing*, pages 568–579. Springer.

[Llerena-Aguilar et al., 2016] Llerena-Aguilar, C., Gil-Pita, R., Rosa-Zurera, M., Ayllón, D., Utrilla-Manso, M., and Llerena, F. (2016). Synchronization based on mixture alignment for sound source separation in wireless acoustic sensor networks. *Signal Processing*, 118:177–187.

[Llerena Aguilar et al., 2012] Llerena Aguilar, C., Mohino, I., and Perez Álvarez, L. (2012). A comparative study of time-delay estimation techniques for convolutive speech mixtures. In *Advances in Computer Science*, pages 291–296.

[Llerena-Aguilar et al., 2015] Llerena-Aguilar, C., Ramos-Auñón, G., Llerena-Aguilar, F. J., Sánchez-Hevia, H. A., and Rosa-Zurera, M. (2015). Improving speech mixture synchronization in blind source separation problems. In *Audio Engineering Society Convention 138*. Audio Engineering Society.

[Lyon, 1982] Lyon, R. F. (1982). A computational model of filtering, detection, and compression in the cochlea. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, volume 7, pages 1282–1285. IEEE.

[Lyon, 1983] Lyon, R. F. (1983). A computational model of binaural localization and separation. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, volume 8, pages 1148–1151. IEEE.

[Madhu et al., 2008] Madhu, N., Breithaupt, C., and Martin, R. (2008). Temporal smoothing of spectral masks in the cepstral domain for speech separation. In *Acoustics, Speech and Signal Processing, IEEE International Conference on ICASSP'08*, pages 45–48. IEEE.

[Makino et al., 2005] Makino, S. A. S., Sawada, H., and Mukai, R. (2005). Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask. *Proceedings of the ICASSP 2005*, 3:81–84.

[Mandel, 2010] Mandel, M. I. (2010). *Binaural model-based source separation and localization.* PhD thesis, Columbia University.

[Markovich-Golan et al., 2012] Markovich-Golan, S., Gannot, S., and Cohen, I. (2012). Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming. In *Proc. of the International Workshop on Acoustic Signal Enhancement (IWAENC), Aachen*, pages 1–4.

[Martin, 2001] Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *Speech and Audio Processing, IEEE Transactions on*, 9(5):504–512.

[Master, 2006] Master, A. S. (2006). *Stereo music source separation via Bayesian modeling*. PhD thesis, Citeseer.

[McCowan, 2001] McCowan, I. (2001). Microphone arrays: A tutorial. *Queensland University, Australia*, pages 1–38.

[Melia and Rickard, 2007] Melia, T. and Rickard, S. (2007). Underdetermined blind source separation in echoic environments using DESPRIT. *EURASIP Journal on Applied Signal Processing*, 2007(1):90–90.

[Mills, 1991] Mills, D. L. (1991). Internet time synchronization: the network time protocol. *Communications, IEEE Transactions on*, 39(10):1482–1493.

[Mirzaei et al., 2014] Mirzaei, S., Norouzi, Y., et al. (2014). Blind speech source localization, counting and separation for 2-channel convolutive mixtures in a reverberant environment. *Proceedings Interspeech 2014*.

[Mirzaei et al., 2015] Mirzaei, S., Norouzi, Y., et al. (2015). Blind audio source counting and separation of anechoic mixtures using the multichannel complex NMF framework. *Signal Processing*, 115:27–37.

[Mitianoudis and Davies, 2003] Mitianoudis, N. and Davies, M. E. (2003). Using beamforming in the audio source separation problem. In *Signal Processing and Its Applications. Proceedings. Seventh International Symposium on*, volume 2, pages 89–92. IEEE.

[Mitianoudis and Davies, 2004] Mitianoudis, N. and Davies, M. E. (2004). Audio source separation: Solutions and problems. *International Journal of Adaptive Control and Signal Processing*, 18(3):299–314.

[Mitianoudis and Stathaki, 2005] Mitianoudis, N. and Stathaki, T. (2005). Overcomplete source separation using Laplacian mixture models. *IEEE Signal Processing Letters*, 12(4):277–280.

[Miyabe et al., 2013a] Miyabe, S., Ono, N., and Makino, S. (2013a). Blind compensation of interchannel sampling frequency mismatch with maximum likelihood estimation in STFT domain. In *Acoustics, Speech and Signal Processing, IEEE International Conference on ICASSP'13*, pages 674–678. IEEE.

[Miyabe et al., 2013b] Miyabe, S., Ono, N., and Makino, S. (2013b). Optimizing frame analysis with non-integrer shift for sampling mismatch compensation of long recording. In *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on WASPAA'13*, pages 1–4. IEEE.

[Mock et al., 2000] Mock, M., Frings, R., Nett, E., and Trikaliotis, S. (2000). Continuous clock synchronization in wireless real-time applications. In *Reliable Distributed Systems. SRDS-2000. Proceedings The 19th IEEE Symposium on*, pages 125–132. IEEE.

[Murata et al., 2001] Murata, N., Ikeda, S., and Ziehe, A. (2001). An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1):1–24.

[Nakatani and Okuno, 1999] Nakatani, T. and Okuno, H. G. (1999). Harmonic sound stream segregation using localization and its application to speech stream segregation. *Speech Communication*, 27(3):209–222.

[Naqvi et al., 2012] Naqvi, S. M., Wang, W., Khan, M. S., Barnard, M., and Chambers, J. (2012). Multimodal (audio–visual) source separation exploiting multi-speaker tracking, robust beamforming and time–frequency masking. *IET signal processing*, 6(5):466–477.

[Naqvi et al., 2010] Naqvi, S. M., Yu, M., Chambers, J., et al. (2010). A multimodal approach to blind source separation of moving sources. *Selected Topics in Signal Processing, IEEE Journal of*, 4(5):895–910.

[Nishiura et al., 2000] Nishiura, T., Yamada, T., Nakamura, S., and Shikano, K. (2000). Localization of multiple sound sources based on a CSP analysis with a microphone array. In *Acoustics, Speech, and Signal Processing. Proceedings of the IEEE International Conference on ICASSP'00*, volume 2, pages II1053–II1056. IEEE.

[O'grady, 2007] O'grady, P. D. (2007). *Sparse separation of under-determined speech mixtures.* PhD thesis, National University of Ireland, Maynooth.

[Omlor and Giese, 2011] Omlor, L. and Giese, M. A. (2011). Anechoic blind source separation using wigner marginals. *The Journal of Machine Learning Research*, 12:1111–1148.

[Ono et al., 2013] Ono, N., Koldovsky, Z., Miyabe, S., and Ito, N. (2013). The 2013 signal separation evaluation campaign. In *Machine Learning for Signal Processing (MLSP), IEEE International Workshop on*, pages 1–6. IEEE.

[Pahm et al., 1992] Pahm, D., Garrat, P., and Jutten, C. (1992). Separation of a mixture of independent sources through a ML approach. In *Proc. European Signal Processing Conf*, pages 771–774.

[Palmer, 1999] Palmer, S. E. (1999). *Vision science: Photons to phenomenology*, volume 1. MIT press Cambridge, MA.

[Parra and Spence, 2000] Parra, L. and Spence, C. (2000). Convolutive blind separation of non-stationary sources. *Speech and Audio Processing, IEEE Transactions on*, 8(3):320–327.

[Parra and Alvino, 2002] Parra, L. C. and Alvino, C. V. (2002). Geometric source separation: Merging convolutive source separation with geometric beamforming. *Speech and Audio Processing, IEEE Transactions on*, 10(6):352–362.

[Patterson et al., 1995] Patterson, R. D., Allerhand, M. H., and Giguere, C. (1995). Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *The Journal of the Acoustical Society of America*, 98(4):1890–1894.

[Pawig et al., 2010] Pawig, M., Enzner, G., and Vary, P. (2010). Adaptive sampling rate correction for acoustic echo control in voice-over-IP. *Signal Processing, IEEE Transactions on*, 58(1):189–199.

[Pearlmutter and Parra, 1996] Pearlmutter, B. A. and Parra, L. C. (1996). A context-sensitive generalization of ICA. *Advances in neural information processing systems*, 151.

[Pedersen et al., 2005] Pedersen, M. S., Wang, D., Larsen, J., and Kjems, U. (2005). Overcomplete blind source separation by combining ICA and binary time-frequency masking. In *2005 IEEE International Workshop on Machine Learning for Signal Processing*, pages 15–20.

[Pedersen et al., 2008] Pedersen, M. S., Wang, D., Larsen, J., and Kjems, U. (2008). Two-microphone separation of speech mixtures. *Neural Networks, IEEE Transactions on*, 19(3):475–492.

[Pierce et al., 1991] Pierce, A. D. et al. (1991). *Acoustics: an introduction to its physical principles and applications*. Acoustical Society of America Melville, NY.

[Pietrzyk, 1998] Pietrzyk, A. (1998). Computer modeling of the sound field in small rooms. In *Audio Engineering Society Conference: 15th International Conference: Audio, Acoustics and Small Spaces*. Audio Engineering Society.

[Proakis John and Manolakis Dimitris, 1996] Proakis John, G. and Manolakis Dimitris, G. (1996). Digital signal processing, principles, algorithms, and applications. *Pentice Hall*.

[Rabiner and Schafer, 1978] Rabiner, L. R. and Schafer, R. W. (1978). *Digital processing of speech signals*. Prentice Hall.

[Raine et al., 2003] Raine, D., Langley, P., Murray, A., Dunuwille, A., and Bourke, J. (2003). P3223 surface 12-lead electrocardiogram waveform analysis in patients with atrial fibrillation: a tool for evaluating the effects of intervention? *European Heart Journal*, 5(24):605.

[Reed et al., 1981] Reed, F., Feintuch, P. L., Bershad, N. J., et al. (1981). Time delay estimation using the LMS adaptive filter–static behavior. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 29(3):561–571.

[Rennie et al., 2003] Rennie, S., Aarabi, P., Kristjansson, T., Frey, B. J., and Achan, K. (2003). Robust variational speech separation using fewer microphones than speakers. In *Acoustics, Speech, and Signal Processing. Proceedings of the IEEE Conference on*, volume 1, pages 88–91. IEEE.

[Rice, 2006] Rice, J. (2006). *Mathematical statistics and data analysis*. Cengage Learning.

[Rickard, 2007] Rickard, S. (2007). The DUET blind source separation algorithm. In *Blind Speech Separation*, pages 217–241. Springer.

[Rickard et al., 2001] Rickard, S., Balan, R., and Rosca, J. (2001). Real-time time-frequency based blind source separation. *Proc. Workshop on Independent Component Analysis and Blind Signal Separation*, pages 651–656.

[Rickard and Dietrich, 2000] Rickard, S. and Dietrich, F. (2000). Doa estimation of many W-disjoint orthogonal sources from two mixtures using DUET. In *Statistical Signal and Array Processing, Proceedings of the Tenth IEEE Workshop on*, pages 311–314. IEEE.

[Rickard and Yilmaz, 2002] Rickard, S. and Yilmaz, Ö. (2002). On the approximate W-disjoint orthogonality of speech. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'02*, volume 1, pages I–529. IEEE.

[Ristaniemi et al., 2002] Ristaniemi, T., Raju, K., and Karhunen, J. (2002). Jammer mitigation in DS-CDMA array system using independent component analysis. In *Communications, (ICC'02), IEEE International Conference on*, volume 1, pages 232–236. IEEE.

[Roman and Wang, 2008] Roman, N. and Wang, D. L. (2008). Binaural speech segregation. In *Speech and Audio Processing in Adverse Environments*, pages 525–549. Springer.

[Römer, 2001] Römer, K. (2001). Time synchronization in ad hoc networks. In *Proceedings of the 2nd ACM international symposium on Mobile ad hoc networking and computing*, pages 173–182. ACM.

[Roth, 1971] Roth, P. R. (1971). Effective measurements using digital signal analysis. *IEEE spectrum*, 4(8):62–70.

[Rothauser et al., 1969] Rothauser, E., Chapman, W., Guttman, N., Nordby, K., Silbiger, H., Urbanek, G., and Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust*, 17(3):225–246.

[Saarnisaari, 1996] Saarnisaari, H. (1996). ML time delay estimation in a multipath channel. In *Spread Spectrum Techniques and Applications, Proceedings of the IEEE 4th International Symposium on*, volume 3, pages 1007–1011. IEEE.

[Saito, 2004] Saito, N. (2004). The generalized spike process, sparsity, and statistical independence. *Modern signal processing*, 46:317–340.

[Saruwatari et al., 2003] Saruwatari, H., Kurita, S., Takeda, K., Itakura, F., Nishikawa, T., and Shikano, K. (2003). Blind source separation combining independent component analysis and beamforming. *EURASIP Journal on Applied Signal Processing*, pages 1135–1146.

[Saruwatari et al., 2005] Saruwatari, H., Mori, Y., Takatani, T., Ukai, S., Shikano, K., Hiekata, T., and Morita, T. (2005). Two-stage blind source separation based on ICA and binary masking for real-time robot audition system. In *Intelligent Robots and Systems (IROS'05), IEEE/RSJ International Conference on*, pages 2303–2308. IEEE.

[Sawada et al., 2011] Sawada, H., Araki, S., and Makino, S. (2011). Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(3):516–527.

[Schmalenstroeer and Haeb-Umbach, 2013] Schmalenstroeer, J. and Haeb-Umbach, R. (2013). Sampling rate synchronisation in acoustic sensor networks with a pre-trained clock skew error model. In *Signal Processing Conference (EUSIPCO'13), Proceedings of the 21st European*, pages 1–5. IEEE.

[Schmalenstroeer et al., 2015] Schmalenstroeer, J., Jebramcik, P., and Haeb-Umbach, R. (2015). A combined hardware–software approach for acoustic sensor network synchronization. *Signal Processing*, 107:171–184.

[Seoane et al., 2013] Seoane, F., Ferreira, J., Álvarez, L., Buendia, R., Ayllón, D., Llerena, C., and Gil-Pita, R. (2013). Sensorized garments and textrode-enabled measurement instrumentation for ambulatory assessment of the autonomic nervous system response in the atrec project. *Sensors*, 13(7):8997–9015.

[Seoane et al., 2014] Seoane, F., Mohino-Herranz, I., Ferreira, J., Álvarez, L., Buendia, R., Ayllón, D., Llerena, C., and Gil-Pita, R. (2014). Wearable biomedical measurement systems for assessment of mental stress of combatants in real time. *Sensors*, 14(4):7120–7141.

[Smaragdis, 1998] Smaragdis, P. (1998). Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22(1):21–34.

[Spieth et al., 1954] Spieth, W., Curtis, J. F., and Webster, J. C. (1954). Responding to one of two simultaneous messages. *The Journal of the Acoustical Society of America*, 26(3):391–396.

[Sur et al., 2014] Sur, S., Wei, T., and Zhang, X. (2014). Autodirective audio capturing through a synchronized smartphone array. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, pages 28–41. ACM.

[Taal et al., 2011] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2125–2136.

[Takigawa et al., 2004] Takigawa, I., Kudo, M., and Toyama, J. (2004). Performance analysis of minimum l_1-norm solutions for underdetermined source separation. *Signal Processing, IEEE Transactions on*, 52(3):582–591.

[Tatlas et al., 2015] Tatlas, N.-A., Potirakis, S. M., Mitilineos, S. A., and Rangoussi, M. (2015). On the effect of compression on the complexity characteristics of wireless acoustic sensor network signals. *Signal Processing*, 107:153–163.

[Torkkola, 1996] Torkkola, K. (1996). Blind separation of delayed sources based on information maximization. In *Acoustics, Speech, and Signal Processing, Proceedings of the IEEE International Conference on ICASSP'96*, volume 6, pages 3509–3512. IEEE.

[Torkkola, 1999] Torkkola, K. (1999). Blind separation for audio signals-are we there yet? In *Proc. Workshop on Independent Component Analysis and Blind Signal Separation*, pages 11–15.

[Tremblay et al., 1987] Tremblay, R., Carter, G. C., and Lytle, D. W. (1987). A practical approach to the estimation of amplitude and time-delay parameters of a composite signal. *Oceanic Engineering, IEEE Journal of*, 12(1):273–278.

[Vaidyanathan and Hoang, 1988] Vaidyanathan, P. and Hoang, P.-Q. (1988). Lattice structures for optimal design and robust implementation of two-channel perfect-reconstruction QMF banks. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(1):81–94.

[van Der Kouwe et al., 2001] van Der Kouwe, A. J., Wang, D., and Brown, G. J. (2001). A comparison of auditory and blind separation techniques for speech segregation. *Speech and Audio Processing, IEEE Transactions on*, 9(3):189–195.

[Van Veen and Buckley, 1988] Van Veen, B. D. and Buckley, K. M. (1988). Beamforming: A versatile approach to spatial filtering. *IEEE assp magazine*, 5(2):4–24.

[Vielva et al., 2001] Vielva, L., Erdogmus, D., and Principe, J. C. (2001). Underdetermined blind source separation using a probabilistic source sparsity model. In *Proc. ICA*, pages 675–679.

[Vincent et al., 2012] Vincent, E., Araki, S., Theis, F., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, V., Lutter, D., and Duong, N. Q. (2012). The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges. *Signal Processing*, 92(8):1928–1936.

[Vincent et al., 2006] Vincent, E., Gribonval, R., and Févotte, C. (2006). Performance measurement in blind audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1462–1469.

[Vincent et al., 2005] Vincent, E., Jafari, M. G., Abdallah, S. A., Plumbley, M. D., and Davies, M. E. (2005). Blind audio source separation. *Centre for Digital Music, Queen Mary University of London, Technical Report C4DM-TR-05-01.*

[Virag, 1999] Virag, N. (1999). Single channel speech enhancement based on masking properties of the human auditory system. *Speech and Audio Processing, IEEE Transactions on*, 7(2):126–137.

[Wang, 2008] Wang, D. (2008). Time-frequency masking for speech separation and its potential for hearing aid design. *Trends in amplification.*

[Wang and Brown, 2006a] Wang, D. and Brown, G. (2006a). Fundamentals of computational auditory scene analysis. *Computational Auditory Scene Analysis: Principles, Algorithms and Applications.*

[Wang and Brown, 2006b] Wang, D. and Brown, G. J. (2006b). *Computational auditory scene analysis: Principles, algorithms, and applications.* Wiley-IEEE Press.

[Wang et al., 2009] Wang, D., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (2009). Speech intelligibility in background noise with ideal binary time-frequency masking. *The Journal of the Acoustical Society of America*, 125(4):2336–2347.

[Wang et al., 2010] Wang, L., Ding, H., and Yin, F. (2010). Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals. *EURASIP Journal on Audio, Speech, and Music Processing*, pages 1–13.

[Wang et al., 2011] Wang, L., Ding, H., and Yin, F. (2011). A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(3):549–557.

[Webster and Thompson, 1954] Webster, J. and Thompson, P. (1954). Responding to both of two overlapping messages. *The Journal of the Acoustical Society of America*, 26(3):396–402.

[Wehr et al., 2004] Wehr, S., Kozintsev, I., Lienhart, R., and Kellermann, W. (2004). Synchronization of acoustic sensors for distributed ad-hoc audio networks and its use for blind source separation. In *Multimedia Software Engineering, Proceedings of the IEEE Sixth International Symposium on*, pages 18–25. IEEE.

[Weintraub, 1985] Weintraub, M. (1985). *A theory and computational model of monaural auditory sound separation.* PhD thesis, Stanford University.

[Winter et al., 2007] Winter, S., Kellermann, W., Sawada, H., and Makino, S. (2007). Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l_1-norm minimization. *EURASIP Journal on Applied Signal Processing*.

[Winter et al., 2005] Winter, S., Sawada, H., and Makino, S. (2005). On real and complex valued l_1-norm minimization for overcomplete blind source separation. In *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on*, pages 86–89. IEEE.

[Woodruff and Pardo, 2007] Woodruff, J. and Pardo, B. (2007). Using pitch, amplitude modulation, and spatial cues for separation of harmonic instruments from stereo music recordings. *EURASIP Journal on Applied Signal Processing*.

[Woodruff et al., 2010] Woodruff, J., Prabhavalkar, R., Fosler-Lussier, E., and Wang, D. (2010). Combining monaural and binaural evidence for reverberant speech segregation. In *INTERSPEECH*, pages 406–409. Citeseer.

[Wu, 2002] Wu, Y. (2002). Time delay estimation of non-gaussian signal in unknown Gaussian noises using third-order cumulants. *Electronics Letters*, 38(16):930–931.

[Wu et al., 2011] Wu, Y.-C., Chaudhari, Q., and Serpedin, E. (2011). Clock synchronization of wireless sensor networks. *Signal Processing Magazine, IEEE*, 28(1):124–138.

[Yilmaz and Rickard, 2004] Yilmaz, O. and Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *Signal Processing, IEEE transactions on*, 52(7):1830–1847.

[Yu et al., 2013] Yu, W., Jiajun, L., Ning, C., and Wenhao, Y. (2013). Improved monaural speech segregation based on computational auditory scene analysis. *EURASIP Journal on Audio, Speech, and Music Processing*, 2(1):1–15.

[Zarzoso et al., 1997] Zarzoso, V., Nandi, A., and Bacharakis, E. (1997). Maternal and foetal ECG separation using blind source separation methods. *Mathematical Medicine and Biology*, 14(3):207–225.

[Zeremdini et al., 2015] Zeremdini, J., Messaoud, M. A. B., and Bouzid, A. (2015). A comparison of several computational auditory scene analysis (CASA) techniques for monaural speech segregation. *Brain Informatics*, 2(3):155–166.

[Zhang et al., 2008] Zhang, C., Florêncio, D., and Zhang, Z. (2008). Why does PHAT work well in lownoise, reverberative environments? In *Acoustics, Speech and Signal Processing, IEEE International Conference on ICASSP'08*, pages 2565–2568. IEEE.

[Zibulevsky et al., 2001] Zibulevsky, M., Pearlmutter, B., et al. (2001). Blind source separation by sparse decomposition in a signal dictionary. *Neural computation*, 13(4):863–882.

[Ziegler, 1995] Ziegler, G. M. (1995). *Lectures on polytopes*, volume 152. Springer Science & Business Media.