Universidad de Alcalá

Escuela Politécnica Superior

Departamento de Teoría de la Señal y Comunicaciones

PhD Thesis

Speech Enhancement Algorithms for Audiological Applications

Author:

David Ayllón Álvarez

Supervisor:

Dr. Manuel Rosa Zurera
Dr. Roberto Gil Pita

October, 2013

Abstract

The improvement of speech intelligibility is a traditional problem which still remains open and unsolved. The recent boom of applications such as hands-free communications or automatic speech recognition systems and the ever-increasing demands of the hearing-impaired community have given a definitive impulse to the research in this area. This PhD thesis is focused on speech enhancement for audiological applications. Most of the research conducted in this thesis has been focused on the improvement of speech intelligibility in hearing aids, considering the variety of restrictions and limitations imposed by this type of devices. The combination of source separation techniques and spatial filtering with machine learning and evolutionary computation has originated novel and interesting algorithms which are included in this thesis. The thesis is divided in two main parts. The first one contains a preliminary study of the problem and a thorough review of the state of the art in this field, from which the goals of the thesis are defined. The second part contains a description of the research conducted to fulfill the objectives of this thesis, including the experimental work and the results obtained.

In a first stage, the speech enhancement problem is formally described and studied in the time-frequency domain. The particular engineering constraints and requirements demanded by hearing aids are also defined. Once the problem has been described, a review of the state of the art has been carried out. The review includes existing solutions to both the single-channel and multichannel speech enhancement problem, considering the noise reduction and the source separation approaches, as well as a review of the application of such algorithms in hearing aids.

The first problem addressed in this thesis is the sound source separation of undetermined mixtures in the time-frequency domain, without considering any type of computational restriction. The performance of the so-called DUET algorithm, which performs speech separation with only two microphones, has been evaluated in a variety of scenarios including linear and binaural anechoic mixtures, echoic mixtures, and mixtures of speech with other types of sources such as noise and music. The study reveals the lack of robustness of the original DUET algorithm, whose performance is notably decreased in echoic and binaural mixtures and when mixing speech with noise and music. In order to overcome this problem, a novel source separation algorithm that combines the mean shift clustering technique with the basis of DUET has been proposed. The clustering step in DUET, which is based on a weighted histogram, is replaced by a weighted-Gaussian kernel mean shift algorithm, which has been inferred for the problem at hand. The analysis of the results obtained demonstrates that the proposed algorithm clearly outperforms the original DUET and a modification thereof using k-means. Additionally, the proposed algorithm has been extended to the case of using any number of microphones and array geometry.

The automatic speech source enumeration problem, which is related to the source separation problem, has also been tackled. A novel algorithm based on information theoretic criteria and the estimation of the source delays between the signals received by two microphones has been proposed. The algorithm has obtained very good results and it has shown good robustness in the enumeration of anechoic mixtures up to 5 speech sources. Additionally, the potential of the algorithm to enumerate sources in echoic mixtures has been demonstrated.

The remaining of the thesis has been focused on hearing aids. The first problem related to hearing aids addressed in this thesis is the improvement of speech intelligibility in monaural hearing aids. First, a study of the computational resources available for signal processing in state-of-the-art commercial hearing aids has been carried out. The result of this study has been used to limit the computational cost of the speech enhancement algorithms for hearing aids proposed in this thesis. After that, a low-cost algorithm for single-channel speech enhancement has been proposed. The algorithm combines a generalized version of the LS estimator with a tailored feature selection algorithm based on evolutionary computation, with the purpose of estimating a time-frequency soft mask that maximizes the output PESQ value, which is a metric highly correlated with intelligibility. The mask is estimated using a novel set of features extracted from the STFT of the mixture. Excellent results are obtained even with low SNRs.

The next work approaches the speech enhancement problem in wireless-communicated binaural hearing aids. In this case, the two devices are connected with a wireless link, which increases the power consumption. The objective in this thesis is the design of low-cost speech enhancement algorithms that increase the energy efficiency of the wireless-communicated binaural hearing aids. First, an extremely low-cost binaural speech separation system that maximizes the WDO has been proposed. It is based on a quadratic discriminant that uses the ILD and ITD cues to classify each time-frequency point between speech or noise. The weights of the discriminant are calculated using a tailored evolutionary algorithm. The second low-cost algorithm uses the information from neighbor time-frequency points to estimate the IBM, using a generalized version of the LS-LDA, introducing a weighted MSE metric that allows estimating the IBM and maximizing the WDO factor at the same time. In both algorithms, a transmission schema to enhance the energy efficiency of the wireless system has been proposed. The schema quantizes the amplitude and phase values of each frequency band with a different number of bits. The bit distribution among frequency bands is optimized by evolutionary computation.

Finally, the last work included in this thesis concerns the design of beamformers for hearing aids fitted to a determined person. The beamforming filter coefficients can be easily fitted to a specific subject as long as the HRTF of that person is known. Unfortunately, this information is not available for every person that needs a new device, and the lack of this knowledge causes gain reduction and distortions. With this problem in mind, three different approaches to optimize the beamforming filter coefficients in case of unknown HRTF have been proposed. The three methods aim at maximizing the average array gain while minimizing the average speech distortions, using a design dataset. The experimental work has demonstrated that the proposed methods decrease significantly the gain reduction and distortions caused by computing the filter coefficients with unknown HRTF of the subject.

Resumen

La mejora de la calidad de la voz es un problema que, aunque ha sido abordado durante muchos años, aún sigue abierto. El creciente auge de aplicaciones tales como los sistemas manos libres o de reconocimiento de voz automático y las cada vez mayores exigencias de las personas con pérdidas auditivas han dado un impulso definitivo a este área de investigación. Esta tesis doctoral se centra en la mejora de la calidad de la voz en aplicaciones audiológicas. La mayoría del trabajo de investigación desarrollado en esta tesis está dirigido a la mejora de la inteligibilidad de la voz en audífonos digitales, teniendo en cuenta las limitaciones de este tipo de dispositivos. La combinación de técnicas de separación de fuentes y filtrado espacial con técnicas de aprendizaje automático y computación evolutiva ha originado novedosos e interesantes algoritmos que son incluidos en esta tesis. La tesis esta dividida en dos grandes bloques. El primer bloque contiene un estudio preliminar del problema y una exhaustiva revisión del estudio del arte sobre algoritmos de mejora de la calidad de la voz, que sirve para definir los objetivos de esta tesis. El segundo bloque contiene la descripción del trabajo de investigación realizado para cumplir los objetivos de la tesis, así como los experimentos y resultados obtenidos.

En primer lugar, el problema de mejora de la calidad de la voz es descrito formalmente en el dominio tiempo-frecuencia. Los principales requerimientos y restricciones de los audífonos digitales son definidas. Tras describir el problema, una amplia revisión del estudio del arte ha sido elaborada. La revisión incluye algoritmos de mejora de la calidad de la voz mono-canal y multi-canal, considerando técnicas de reducción de ruido y técnicas de separación de fuentes. Además, la aplicación de estos algoritmos en audífonos digitales es evaluada.

El primer problema abordado en la tesis es la separación de fuentes sonoras en mezclas infra-determinadas en el dominio tiempo-frecuencia, sin considerar ningún tipo de restricción computacional. El rendimiento del famoso algoritmo DUET, que consigue separar fuentes de voz con solo dos mezclas, ha sido evaluado en diversos escenarios, incluyendo mezclas lineales y binaurales no reverberantes, mezclas reverberantes, y mezclas de voz con otro tipo de fuentes tales como ruido y música. El estudio revela la falta de robustez del algoritmo DUET, cuyo rendimiento se ve seriamente disminuido en mezclas reverberantes, mezclas binaurales, y mezclas de voz con música y ruido. Con el objetivo de mejorar el rendimiento en estos casos, se presenta un novedoso algoritmo de separación de fuentes que combina la técnica de clustering mean shift con la base del algoritmo DUET. La etapa de clustering del algoritmo DUET, que esta basada en un histograma ponderado, es reemplazada por una modificación del algoritmo mean shift, introduciendo el uso de un kernel Gaussiano ponderado. El análisis de los resultados obtenidos muestran una clara mejora obtenida por el algoritmo propuesto en relación con el algoritmo DUET original y una modificación que usa k-means. Además, el algoritmo propuesto ha sido extendido para usar un array de micrófonos de cualquier tamaño y geometría.

A continuación se ha abordado el problema de la enumeración de fuentes de voz, que está relacionado con el problema de separación de fuentes. Se ha propuesto un novedoso algoritmo basado en un criterio de teoría de la información y en la estimación de los retardos relativos causados por las fuentes entre un par de micrófonos. El algoritmo ha obtenido excelente resultados y muestra robustez en la enumeración de mezclas no reverberantes de hasta 5 fuentes de voz. Además se demuestra la potencia del algoritmo para la enumeración de fuentes en mezclas reverberantes.

El resto de la tesis esta centrada en audífonos digitales. El primer problema tratado es el de la mejora de la inteligibilidad de la voz en audífonos monoaurales. En primer lugar, se realiza un estudio de los recursos computacionales disponibles en audífonos digitales de última generación. Los resultados de este estudio se han utilizado para limitar el coste computacional de los algoritmos de mejora de la calidad de la voz para audífonos propuestos en esta tesis. Para resolver este primer problema se propone un algoritmo mono-canal de mejora de la calidad de la voz de bajo coste computacional. El objetivo es la estimación de una mascara tiempo-frecuencia continua para obtener el mayor parámetro PESQ de salida. El algoritmo combina una versión generalizada del estimador de mínimos cuadrados con un algoritmo de selección de características a medida, utilizando un novedoso conjunto de características. El algoritmo ha obtenido resultados excelentes incluso con baja relación señal a ruido.

El siguiente problema abordado es el diseño de algoritmos de mejora de la calidad de la voz para audífonos binaurales comunicados de forma inalámbrica. Estos sistemas tienen un problema adicional, y es que la conexión inalámbrica aumenta el consumo de potencia. El objetivo en esta tesis es diseñar algoritmos de mejora de la calidad de la voz de bajo coste computacional que incrementen la eficiencia energética en audífonos binaurales comunicados de forma inalámbrica. Se han propuesto dos soluciones. La primera es un algoritmo de extremado bajo coste computacional que maximiza el parámetro WDO y esta basado en la estimación de una mascara binaria mediante un discriminante cuadrático que utiliza los valores ILD e ITD de cada punto tiempo-frecuencia para clasificarlo entre voz o ruido. El segundo algoritmo propuesto, también de bajo coste, utiliza además la información de puntos tiempo-frecuencia vecinos para estimar la IBM mediante una versión generalizada del LS-LDA. Además, se propone utilizar un MSE ponderado para estimar la IBM y maximizar el parámetro WDO al mismo tiempo. En ambos algoritmos se propone un esquema de transmisión eficiente energéticamente, que se basa en quantificar los valores de amplitud y fase de cada banda de frecuencia con un número distinto de bits. La distribución de bits entre frecuencias se optimiza mediante técnicas de computación evolutivas.

El último trabajo incluido en esta tesis trata del diseño de filtros espaciales para audífonos personalizados a una persona determinada. Los coeficientes del filtro pueden adaptarse a una persona siempre que se conozca su HRTF. Desafortunadamente, esta información no esta disponible cuando un paciente visita el audiólogo, lo que causa pérdidas de ganancia y distorsiones. Con este problema en mente, se han propuesto tres métodos para diseñar filtros espaciales que maximicen la ganancia y minimicen las distorsiones medias para un conjunto de HRTFs de diseño.

“Solo sé que no sé nada y,
al saber que no sé nada, algo sé;
porque sé que no sé nada.”

Sócrates

Acknowledgements

Some years ago I met a person that encouraged me to change the course of my life and start a research career. Thank you Manolo for giving me the opportunity and trusting me. And thank you Roberto for supporting me every day. You have been more than two good advisors during these years.

I am particularly grateful to Professor Harvey Silverman for advising me during the period I spent in Brown University. His great experience makes each of his advices become a valuable lesson. I would like to thank Dr. Fernando Seoane for his support during the two periods I spent in Boras University. And I would also like to thank the people from North Carolina State University for the time I spent there.

I want to express special gratitude to my colleagues from the Signal Theory and Communications department, specially the people from the 'S31' lab. We have shared many things together. And to my 'Swedish' friends from Boras Ruben, Javi and Juan Carlos.

I would specially like to thank my family and friends for their support. Although many of you do not understand the content of this book, you have helped to write it. And thank you Gema for your patience and support during these years.

This thesis has been funded by the Spanish Ministry of Science and Innovation, under projects TEC2009-14414-C03-03 and TEC2012-38142-C04-02, and the FPU program.

Finally, I want to thank you for reading these lines.

David Ayllón

Contents

List of Figures

List of Tables

List of Symbols

$x(t)$	Discrete-time signal
f_s	Sampling rate
N	Number of sources
M	Number of mixtures
$s_n(t)$	Sound signal produced by the n-th source
$x_m(t)$	Signal received by the m-th microphone
$n_m(t)$	Noise signal received by the m-th microphone
$h_{mn}(t)$	Acoustic impulse response between the n-th source and the m-microphone
k	Frequency index
K	Number of frequency bands
l	Time frame index
L	Number of time frames
$S_n(k,l)$	STFT of the signal produced by the n-th source
$X_n(k,l)$	STFT of the signal received by the m-th microphone
$N_m(k,l)$	STFT of the noise signal received by the m-th microphone
$M(k,l)$	Time-frequency mask
δ_K	Kronecker delta function
a_{mn}	Attenuation of the signal that travels from the n-th source to the m-th microphone
δ_{mn}	Time delay of the signal that travels from the n-th source to the m-th microphone
a_n	Relative amplitude difference between the n-th microphone and the reference sensor
δ_n	Relative time difference between the n-th microphone and the reference sensor
r	Reflection coefficient
$\mathbf{Q}$	Pattern matrix
$\mathbf{v}$	Weight vector
$\mathbf{t}$	Target vector
$\mathbf{w}_k$	Array weight vector for the k-th frequency
G_k	Array gain in the k-th frequency band for the steering direction
$H_{L/Rs}$	HRTF of the s-th subject
θ	Azimuth angle
ϕ	Elevation angle
$\mathbf{r}_m$	Position vector of the m-th microphone
$\mathbf{u}_{\theta\phi}$	Vector pointing to the DOA given by (θ, ϕ)
$\|\|x\|\|$	Euclidean norm
$(.)^T$	Matrix transpose operator
$(.)^H$	Matrix conjugate transpose operator
$(.)^*$	Complex conjugate operator

Acronyms

A

ADC	Analog-to-Digital Converter.
AM	Amplitude Modulation.
AR	Autoregressive Model.
ASR	Automatic Speech Recognition.
ATF	Acoustic Transfer Function.

B

BSS	Blind Source Separation.

C

CASA	Computational Auditory Scene Analysis.
CDB	Constant Directivity Beamformer.

D

DAC	Digital-to-Analog Converter.
DFT	Discrete Fourier Transform.
DOA	Direction Of Arrival.
DS	Delay and Sum.
DSP	Digital Signal Processor.
DUET	Degenerate Unmixing Estimation Technique.

E

EA	Evolutionary Algorithm.
EM	Expectation-Maximization.
EWMA	Exponentially-Weighted Moving Average.

F

FIR	Finite Impulse Response.
FM	Frequency Modulation.

G

GLSE	Generalized Least Squares Estimator.
GSC	Generalized Sidelobe Canceller.

H

HMM	Hidden Markov Model.
HRIR	Head-Related Impulse Response.
HRTF	Head-Related Transfer Function.

I

IBM	Ideal Binary Mask.
ICA	Independent Component Analysis.
IDFT	Inverse Discrete Fourier Transform.
ILD	Interaural Level Differences.
IPF	Instructions Per Frequency.
ISTFT	Inverse Short-Time Fourier Transform.
ITC	In-The-Canal.
ITD	Interaural Time Differences.

L

LCMV	Linearly Constrained Minimum Variance.
LMS	Least Mean Squares.
LPC	Linear Prediction Coefficients.
LS	Least Squares.
LS-GDA	Least Squares Generalized Discriminant Analysis.

M

MAC	Multiply ACcumulate.
MAP	Maximum A Posteriori.
MDL	Minimum Description Length.
ML	Maximum Likelihood.
MMSE	Minimum Mean-Square Error.
MSE	Mean Square Error.
MVDR	Minimum Variance Distortionless Response.
MWF	Multichannel Wiener Filter.

P

PDF	Probability Density Function.
PESQ	Perceptual Evaluation of Speech Quality.
PSD	Power Spectral Density.
PSR	Preserved-to-Signal Ratio.
PSS	Power Spectral Subtraction.

R

RIRG	Room Impulse Response Generator.

S

SDB	SuperDirective Beamformer.
SIR	Signal-to-Interference Ratio.
SNR	Signal-to-Noise Ratio.
SSS	Sound Source Separation.
STFT	Short-Time Fourier Transform.

T

TDOA	Time Difference Of Arrival.

W

W-LS-GDA	Weighted Least Squares Generalized Discriminant Analysis.
WDO	W-Disjoint Orthogonality.
WG-MS	Weighted Gaussian Mean Shift.
WMSE	Weighted Mean Square Error.

Part I

Preliminary work

Chapter 1

Introduction and motivation

1.1 Speech enhancement problem

When people is asked how they can listen to only one of the sounds in a mixture, they usually reply: "I just listen to one and try not to be distracted by the others" [Wang and Brown, 2006]. This answer presupposes a number of operations that the human brain has already done before the listener can only focus on the desired sound. The acoustic signal reaching our ears comprises sound waves originated in multiple sources and their reflections from surfaces in the environment. However, the person who is listening to a determined speaker does not bother to reject background noise or other cross-talking voices explicitly, the brain does it automatically. This problem was formally stated and named as 'Cocktail party problem' in [Cherry, 1953]:

> *"One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others...we may call it the cocktail party problem. No machine has yet been constructed to do just that".*

The design of machines that are able to listen to sounds in the same way than the humans do has been a very active line of research during many years. Nowadays, due to the rapid growth of digital systems in the last decades, it has become in one of the most interesting lines of research.

One of the main problems to solve in this research area is the enhancement of degraded speech signals. *Speech enhancement aims to improve the speech quality and intelligibility, introducing some kind of technology between the desired speech source and the human ear.* This enhancement is necessary due to the fact that the desired speech source is mixed with other sound sources transmitting energy at the same time, which can be either noise, music, or even different speech sources. In the case that those sources are in a closed space, reverberation also decreases the quality of the received signal. The aforementioned 'technology' is composed of a single or a set of microphones (microphone array), a system that enhances the signals gathered by these microphones, and a single or a set of loudspeakers that reproduce the enhanced signal to be listened to by the human ear. In the digital age in which we are currently living, the speech enhancement system is based on a digital signal processor (DSP) which allows running signal processing algorithms to deal with the problem at hand. Figure 1.1 shows an overview of the described speech enhancement system.

There are different applications where speech enhancement plays an important role. In some cases, we are interested in recovering only one source with good quality, removing all the remaining sources. On the other hand, there are cases in which we are interested in recovering all the different sources. Additionally, some applications require real-time processing, which

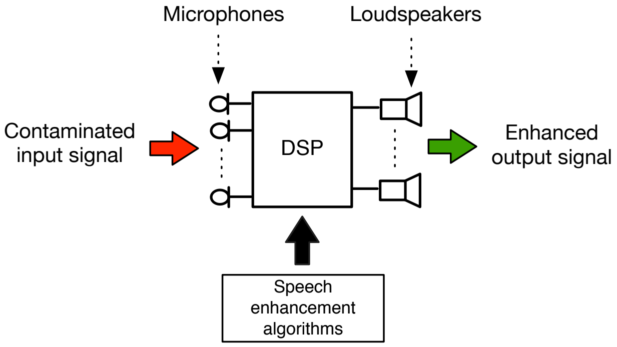


Figure 1.1: Speech enhancement system

increases the complexity of the problem. Some applications of speech enhancement that can be found in the daily life are:

- **Hearing aids**. Hearing loss affects an important percentage of people, and this figure is increasing due to the growing exposure to excessive noise in their daily lives. One of the main problems for hearing-impaired people is the reduction of speech intelligibility in noisy environments, which is mainly caused by the loss of temporal and spectral resolution in the auditory system of the impaired ear. The use of hearing aid devices that only provide amplification does not solve the problem, due to the fact that they amplify both speech and noise. Besides acoustic loss compensation, the DSPs of modern digital hearing aids include speech enhancement algorithms, as well as algorithms for echo cancellation and automatic sound classification.

- **Hands-free communication systems**. In recent years, the demand of hands-free communications for vehicles or teleconference systems has drastically increased the research and development of this kind of devices. The success of these systems relies on the quality of the acquired speech, which is contaminated by different types of noise and interferences. Consequently, the signals acquired by the microphones of the system are usually enhanced before being transmitted through the communication channel.

- **Automatic speech recognition (ASR)**. Much progress has been made in ASR in the last years. Smartphones, computers or smart TVs are only some examples of current technologies that include ASR. The probability of success in the recognition strongly depends on the quality of the acquired signal, and the performance of ASR systems rapidly degrades in the presence of noise. This fact makes a previous stage of speech enhancement necessary for ASR systems.

- **Recording systems.** Audio recordings have many applications such as security, automatic music transcription, audio information retrieval or electronic surveillance. One desirable operation to perform with these recordings is to recover the original sources with high quality, separating the different audio sources and removing background noise.

This thesis deals with the problems of sound source separation, noise reduction and speech source enumeration. The reduction of the computational complexity of speech enhancement algorithms will be also studied for the implementation in hearing aids. Systems with a single microphone and systems containing a microphone array are studied. The former case is more challenging due to the reduced information available in a single microphone. The latter is more interesting since the use of a microphone array includes spatial information, which gives rise to a wider range of algorithms. The aforementioned problems to solve have been a topic of study during many years, but they still remain open and unsolved due to their complexity. In a first approach, this thesis is focused on sound source separation algorithms and on the identification of the number of speakers, without considering computational restrictions. After that, the study is focused on low-cost algorithms for speech enhancement in hearing aid devices. These systems must work in real time but they have very low computational capacity due to their reduced battery life, which limits the power consumption. Hence, the computational cost of the signal processing algorithms developed for hearing aids must be low, implying that these algorithms must be relatively simple to be implemented in real time in this type of devices. An important part of this study is focused on binaural hearing aids, which is a recent topic of research. In binaural systems, the hearing-impaired person wears a different device in each ear, and these devices exchange information between them. Due to aesthetic reasons, it is desirable to connect them with a wireless link, which increases the power consumption. This wireless data link originates a new problem to solve: the reduction of the information exchanged between both devices without degrading excessively the performance of the binaural enhancement algorithm.

The remaining of this chapter contains a description of the problems to solve in digital hearing aids, a comprehensive review of the state of the art in this research field, and the main goals of this thesis. The chapter ends with a description of the structure of this thesis.

1.2 Speech enhancement in digital hearing aids

Hearing aids are electronic devices worn by hearing-impaired people ideally to improve the reduced intelligibility caused by hearing loss. Despite the fact that traditional devices do improve speech quality, their capability to improve speech intelligibility has been largely discussed. Simple devices often produce amplified noises when the user is in a multi-source environment (e.g. a crowded bar). Modern devices include some type of enhancement system to overcome this limitation, for instance, directional microphones or speech enhancement algorithms. However, in addition to the problems found by speech enhancement algorithms when improving intelligibility, their application in hearing aids entails three main additional problems: hearing-impaired listeners have greater susceptibility to the distortions introduced by signal processing algorithms, the small size of hearing devices limits the number of microphones assembled in the device, and the reduced life of the current batteries constrains the computational cost of the implemented algorithms.

1.2.1 Hearing impairment

The number of people with hearing loss is increasing at an alarming rate not only because of the aging of the world's population, but also because of the growing exposure to noise in daily life. Some figures confirming these facts are, for instance, that about one-third of Americans between the ages of 65 and 74, and about half the people who are 85 and older, have important hearing loss [NIA, 2005]. Or that about 16% of adult Europeans have hearing problems strong enough to adversely affect their daily life. The royal national institute for deaf people (RNID) has reported that there are 8.7 million deaf and hard of hearing people in the UK, and that just one in four hearing-impaired Britons owns a hearing aid [RNID, 2008]. All these facts compel scientists and engineers to enhance hearing aids in the effort of making them more accessible for people, especially the elderly.

Hearing loss is commonly represented by an audiogram, which shows the auditory threshold in logarithmic units (dB) for standardized frequencies measured by an audiometer. Hearing impairment implies larger thresholds than normal hearing but the level of loss among frequencies is not uniform and depends on each person. The degree of hearing loss is usually defined as the average hearing loss measured at a particular octave-band, and the level of loss is usually classified into mild (up to 40 dB), moderate (from 40 to 60 dB) and severe (over 60 dB). For hearing-impaired people suffering from mild to moderate hearing loss, a hearing aid is helpful, but in the case of severe hearing loss, the use of hearing aids is of little benefit, and some other solutions may be considered. Additionally, hearing loss can be unilateral, but in most cases it is bilateral, which means that both ears are affected with either the same or different degree of loss.

Hearing-impaired people face a variety of different auditory problems that reduce their ability of understanding. These problems are described below.

- **Decreased level of audibility**
  Depending on the level of hearing loss, a person will hear some sounds but miss some other sounds. In general, the high-frequency components of speech are weaker than the low-frequency components, and hearing loss of elderly people is higher at high frequencies. Consequently, hearing-impaired people tend to miss high-frequency information, basically consonants. This fact leads to miss essential parts of some phonemes reducing the intelligibility.

- **Reduced dynamic range**
  The dynamic range of the auditory system is defined as the level difference between the auditory threshold and the discomfort threshold (i.e. threshold of pain). For hearing-impaired people, the auditory threshold is increased in comparison to normal hearing people, hence the dynamic range is reduced. In order to avoid exceeding the discomfort threshold, hearing aids must amplify weak sounds more than intense sounds.

- **Reduced frequency resolution**
  Frequency resolution gradually decreases as the degree of hearing loss increases, and hearing-impaired people find difficult to distinguish between sounds of different frequencies simultaneously. This is due to the loss of sensitivity of the hair cells of the cochlea, which decreases the ability of discriminating frequencies.

- **Decreased temporal resolution**
  In general, weaker sounds are sometimes masked by intense sounds that immediately precede or follow them, which decreases the chances of intelligibility. In addition, the ability to hear weak sounds during short-time slots gradually decreases as the degree of hearing loss increases, and hearing-impaired people usually experience decreased temporal resolution, which involves that the speech intelligibility perceived by them is further decreased.

All the aforementioned problems combined together cause significant reduction in the speech intelligibility perceived by hearing-impaired people. The first two problems are commonly approached by a compression-expansion algorithm, which applies a frequency and signal level dependent gain customized for each person. The intelligibility decrease originated by the reduction in the temporal and frequency resolutions can be compensated by speech enhancement algorithms.

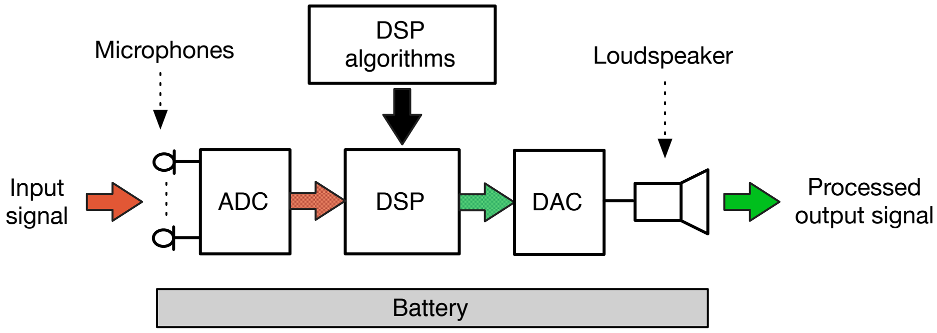


Figure 1.2: A simplified scheme of the typical structure of a digital hearing aid.

1.2.2 Signal processing in digital hearing aids

The introduction of the digital signal processor (DSP) in hearing aids opened a new era where these devices offer their users a greater flexibility to compensate for their hearing loss, providing a more natural sound quality than the previous analog hearing aids. The typical structure of a digital hearing aid is shown in figure 1.2. The device comprises the next elements:

- A single or multiple microphones that convert the acoustic signal into an electric signal.
- An analog-to-digital converter (ADC) to transform the continuos electric signal (analog) into a digital signal.
- A DSP, the main part of the device, that includes signal processing algorithms for different purposes.
- A digital-to-analog converter (DAC) to reconvert the digital processed signal into an analog signal.
- A tiny loudspeaker that produces the output acoustic signal from the processed analog electric signal.
- A small battery to supply power to the previous electronic devices.

The fact that hearing loss does not only result in sound attenuation, but also in distortions that lead to a reduction in speech intelligibility, motivates that modern digital hearing aids include a variety of signal processing algorithms for different purposes:

- A multi-band compression-expansion algorithm to compensate hearing loss and fit the output level into the dynamic range.
- Acoustic feedback cancellation to prevent the instability of the device due to the acoustic feedback that appears when part of the amplified output signal produced by the hearing aid returns through the external auditory canal and enters again the device, thus being again amplified.
- Automatic environmental classification in order to adapt the amplification or processing program to different listening conditions (e.g. a quiet room, a conference hall, a noisy street, etc.).

- Speech enhancement algorithms that aim to improve the speech intelligibility provided by hearing aids in different noisy environments.

All these algorithms must be implemented in the DSP embedded in the hearing aid. Unfortunately, the computational capability and the memory available in the DSP of such devices is highly restricted: the processor is forced to work at low-clock frequencies in order to minimize the power consumption and thus to maximize the battery life. The current batteries available for hearing aids and the expectation of a minimum battery life of one week entail that the DSPs found in such devices have on-chip processors with a selective clock speed that usually goes from 5.12 MHz down to 1.28 MHz, which is a relative low speed in comparison to the current state-of-art DSPs that can be used in other applications.

1.2.3 Speech enhancement algorithms for hearing aids

Imagine an elderly grandmother who wears a hearing aid in one or both ears. She is in a room where her family is celebrating her birthday. There are so many talks, music, the TV, and background noise mixing with each other that the old lady can not understand what her grandson is telling her. The solution to this problem would be that the hearing aids themselves were able to enhance only the voice of the grandson separately from the rest of the sounds without interest. The inclusion of speech enhancement algorithms in moderns devices aims to solve this problem. However, the design and implementation of this type of algorithms in digital hearing aids is strongly limited by two engineering constraints, which are not present in other speech enhancement applications such as hands-free devices or ASR systems. First, the limited computational resources restrict the complexity of the embedded algorithms, and second, the reduced dimensions of the device limit the number of microphones and their distance. As mentioned before, there are several signal processing algorithms running simultaneously in the DSP of modern digital hearing aids, trying to solve different problems: acoustic loss compensation, acoustic feedback cancellation, automatic environment classification or speech enhancement. These algorithms demand a significant part of the computational power of the device, and at the same time, electrical power. Bearing in mind the limited power of the processor, the computational cost of the algorithms used for speech enhancement must be very low, taking only a small part of the available computational resources.

Many hearing-impaired people have bilateral hearing loss and they are forced to wear two devices. Often, when hearing aids are worn at both ears, these devices operate independently. However, there is a new trend of binaural hearing aids that connects both devices in order to exchange information between them. Binaural hearing provides considerable benefits over using only one ear, due to the fact that the nature of the human auditory system is binaural. Humans are able to separate and selectively attend to individual sound sources in a cluttered acoustical environment taking advantages of the so-called spatial cues. Hence, it is fundamental that the hearing aid system preserves these cues, which notably increments the ability to localize sounds and consequently improves speech intelligibility. This obviously requires a communication link between both hearing devices. The simplest solution would be to connect them using a wire. However, most users do not like this approach because of the non-aesthetic aspect of the wire linking both hearing aids from one ear to the other. This enforces to use a wireless link between both devices, what unavoidably increases the power consumption and, consequently, reduces the battery life, one of the most important limiting factors for implementing signal processing algorithms on digital hearing aids. Roughly speaking, the current technology demands as much power to communicate both hearing aids as that required for the signal processing on a monaural device [Kates, 2008]. The reduction of the data rate helps cut down the power consumption, but it is done at the expenses of bringing down the performance of the enhancement algorithms.

1.3 State of the art of speech enhancement algorithms

Speech enhancement algorithms can be grouped according to different criteria:

- Number of channels: single-channel or multichannel, depending on the number of sensor that are considered.
- Algorithm output: whether the algorithm obtains all the sources in the mixture or only a cleaned version of the target speech signal.
- Mixture type: instantaneous, anechoic or echoic.
- Algorithm approach: statistical-based, source and channel models, human auditory model, spatial filter, time or frequency domain, etc.

In this review of the state of the art, the algorithms are divided into two different but not independent groups: noise reduction algorithms and sound source separation (SSS) algorithms. Both are further divided into single-channel and multichannel techniques. Noise reduction algorithms deal with the problem of estimating a speech signal from a corrupted version of itself with noise, assuming that all different sound sources are noise. SSS is related to the problem of estimating each original source contained in an audio mixture. This section contains a review of the most important approaches that have been proposed to solve both problems, as well as the suitability of these algorithms for speech enhancement in hearing aids.

1.3.1 Noise reduction algorithms

The goal of noise reduction algorithms is to provide a high quality speech signal and robustness against background noise, interfering sources and reverberation effects. Most single-channel algorithms operate in the time-frequency domain, whereas multichannel algorithms are based on spatial filtering.

1.3.1.1 Single-channel noise reduction

The problem faced by this type of algorithms is the estimation of a speech signal from a corrupted version of the signal composed of the desired speech and interfering noise. The complexity of the problem is relatively high, due to the limited information available in a single observation. Several approaches have been proposed for single-channel noise reduction during many years, the most relevant being discussed in this section.

The earliest approach for enhancing speech degraded by noise is the power spectral subtraction (PSS) method introduced in [Boll, 1979], which principle is to subtract the short-time spectral magnitude of noise from the noisy-speech magnitude, assuming uncorrelated and additive noise. The noise spectrum is usually estimated during speech pauses, and the phase is not processed assuming that phase distortion does not degrade speech intelligibility. Despite the good performance suppressing stationary background noise, the original algorithm has several drawbacks: the assumption of an accurate estimation of the noise spectrum, which is very difficult with low signal-to-noise ratio SNR; the speech distortions introduced as an annoying 'musical noise'; and the still noticeable remaining output noise. The algorithm presented in

[Berouti et al., 1979] aims to minimize the musical noise while further reducing the background noise. It consists in subtracting an overestimate of the noise power spectrum, as well as avoiding the spectral components to take values below a level. The algorithm obtains great noise reduction with very little effect on the intelligibility of the speech. This version has been traditionally considered as the basic implementation of the PSS algorithm. In [Virag, 1999], the PSS method is modified including a human hearing model based on the masking phenomenon commonly used in audio coding. The subtraction parameters are continuously adapted according to the noise masking threshold, obtaining a significant reduction of the musical and residual noise. Most implementations and variations of the PSS method perform subtraction over the entire speech spectrum. However, the spectrum of real-world noise is not flat, which implies that the noise signal does not affect the speech signal uniformly over the whole spectrum. Several implementations of non-linear spectral subtraction have been proposed, for instance, in [Kamath and Loizou, 2002], where a multi-band spectral subtraction technique for colored noise is proposed. The authors propose to split the frequency spectrum linearly into a number of non-overlapping bands. A traditional spectral subtractor [Berouti et al., 1979] with a different over-subtraction factor is applied to each band. They found that four is the optimal number of bands in terms of speech quality. The algorithm notably outperforms the original algorithm for different SNRs. Nevertheless, despite the speech distortion introduced by the PSS, it is perhaps the most popular algorithm for speech enhancement used today, thanks to its low complexity and high efficiency.

Another common approach for single-channel noise reduction is the application of the Wiener filter [Wiener, 1949], which is an optimal estimator of the desired signal in the minimum mean-square error (MMSE) sense, in the stationary case. The PSS method can be also considered as a filter designed in the frequency domain, but the main difference resides in the fact that the Wiener filter estimates the speech waveform in the time domain rather than to estimate the speech spectrum. The solution of the optimal Wiener filter is obtained in [Trees, 1968], but is non-casual and depends on the clean signal power spectrum, therefore is not realizable. In practice, the Wiener filter can be estimated iteratively by assuming an all-pole model for speech production. The iterative Wiener filter (IWF) was originally formulated in [Lim and Oppenheim, 1979]. In this technique, the speech signal is modeled as the response of an all-pole system, and the approach is to solve for the maximum a posteriori (MAP) estimate of the speech signal given the noisy signal. Unfortunately, the convergence criteria, which is critical for the performance of the algorithm, is not specified, and allowing more iterations is not necessarily beneficial.

Although the Wiener filter achieves satisfactory noise reduction for some applications, it also introduces distortions in the speech signal that can be perceptually unacceptable when the SNR is very low. Many works have been focused on the reduction of these distortions. The design of the Wiener filter requires stationary speech and noise signals and assumes that the statistics of both signals are known a priori. In practice, these conditions are not met. The work in [Sambur, 1978] presents a least mean-square (LMS) adaptive filtering approach that takes advantage of the quasi-periodic nature of the speech waveform to supply a reference signal to the adaptive filter. The method has the advantage of requiring no a priori knowledge of the properties of the noise signal, and it obtains improvement in the quality of speech reducing the granular quantization noise. In [Chen et al., 2006], the relationship between noise reduction and speech distortion with the single-channel Wiener filter is formally studied. The authors demonstrated that the amount of noise attenuation is proportional to the amount of speech degradation. Depending on the final application of the enhanced speech, a tradeoff between the amount of noise reduction and speech distortion should be adopted. The authors propose three different approaches to control both parameters.

Statistical model-based algorithms rely on the MMSE estimator of the short-time spectral amplitude. The MMSE estimators have been very popular, partly because they have shown to be successful in eliminating musical noise even if the noise is only poorly stationary [Cappé, 1994]. The reason of the reduction of musical tones is the low variance estimate of the obtained spectra. In [Ephraim and Malah, 1984] the authors observe that the Wiener filter is optimal in the sense of MMSE signal spectral estimator, but is not an optimal spectral magnitude estimator under the Gaussian assumption. They derive an optimal MMSE short-time spectral amplitude estimator based on modeling speech and noise spectral components as statistically independent Gaussian random variables. The performance of the proposed approach is compared with the one of the Wiener filter, and it results in a significant reduction of the noise providing enhanced speech with colorless residual noise. The performance of the estimator is further increased minimizing the MSE of the log-spectra. The work in [Martin, 2005] presents a short-time spectral coefficients estimator in the MMSE sense, without assuming that the spectral coefficients of the noise and the clean speech signal follow a complex Gaussian probability density. The probability density function (PDF) of the clean speech spectral coefficients is modeled by a complex Laplacian or by a complex bilateral Gamma, and the PDF of the noise spectral coefficients is either modeled by a complex Gaussian or complex Laplacian. This estimator obtains higher noise reduction than the traditional MMSE estimator, and the residual noise is lower when the input noise has a Laplacian density.

Another linear filter that has been widely applied to the problem of speech enhancement is the so-called Kalman filter [Kalman and Emil, 1960], which is a casual filter applicable in cases where the desired and observed signals are non-stationary. For the application of the Kalman filter to speech enhancement, it is common to model the speech as a quasi-stationary autoregressive (AR) process, which is usually represented by linear predictive coefficients (LPC). These coefficients are unknown and should be estimated together with the signal. In [Goh et al., 1999] the authors propose a speech model originated from AR modeling that describes voiced and unvoiced speech as well as silence. The model is reformulated to be included in the Kalman filter. Furthermore, they introduce a mathematically equivalent algorithm which is computationally much more efficient, by exploiting the sparsity of the concerned matrices. Finally, a method based on expectation-maximization (EM) for estimating the model parameters is presented. The evaluation of the algorithm shows an improvement over existing methods as high as 4 dB of output SNR. On the other hand, the key step in the traditional Kalman filter is the minimization of the estimation error variance between the clean signal and its estimation. In [Ma et al., 2006] the authors propose to minimize that variance under the constraint that the energy of the estimation error is smaller than a masking threshold. This threshold is computed from both the time-domain forward masking and frequency-domain simultaneous masking properties of the human auditory system. Objective and subjective tests confirm that the performance obtained is better than the one obtained with the conventional methods.

A different approach for speech enhancement is the subspace approach, which arises to deal with the compromise between speech distortion and noise reduction. This approach is based on the decomposition of the noisy speech signal into two subspaces: the signal plus noise subspace and the noise subspace. The noise subspace is removed from the signal plus noise subspace and the speech signal is estimated from the remaining subspace. The space decomposition can be performed using either singular value decomposition (SVD) or eigenvalue decomposition (EVD). The SVD-based method in [Dendrinos et al., 1991] proposes that the signal information is contained in the eigenvectors corresponding to the largest singular values, and the noise information is contained in the those corresponding the smallest singular values. A different formulation to the subspace approach is given in [Ephraim and Van Trees, 1995] where the decomposition is done by applying the Karhunen-Loeve transform (KLT). The estimation of the speech signal aims to minimize a mathematically-derived speech distortion measure while keeping the energy of the residual noise in each spectral component below a threshold, in that way that the residual noise is masked by the speech signal. Listening tests demonstrate the improvement of this approach in comparison to the spectral subtraction approach. The main limitation of this method is that it was formulated under the assumption of white noise. The authors in [Hu and Loizou, 2003] derived a generalized subspace approach with built-in prewhitening, making no assumptions about the covariance matrix of the KLT-transformed noise vectors, hence obtaining an optimal estimator. The estimator obtained in [Ephraim and Van Trees, 1995] is a special case of this estimator when the noise is white.

Traditionally, more effort has been devoted to design speech enhancement algorithms capable of improving speech quality rather than speech intelligibility. Speech quality is highly subjective but intelligibility is related to the understanding of the underlying message. Unlike speech quality, which can be easily improved by removing the background noise (i.e. increasing the SNR), intelligibility is only improved by suppressing the background noise without distorting the target speech signal. The main factor that causes the absence of intelligibility improvement obtained with traditional algorithms is that methods such as the PSS or the Wiener filter optimize a cost function that correlates with speech quality but it does not necessarily correlate with speech intelligibility, for instance, the aforementioned MSE metric. A subjective evaluation of 13 different speech enhancement algorithms (including PSS, Wiener-type, and subspace algorithms) is performed in [Hu and Loizou, 2007], evaluating speech quality in terms of signal distortion, noise distortion and overall quality. Two important conclusions are drawn from this study: none of the studied algorithms provides significant benefit to the overall quality compared to the noisy (unprocessed) speech, and the algorithms that caused the lowest speech distortion were also the algorithms yielding the highest overall quality. The latter conclusion suggests that listeners are more influenced by speech distortions than by background noise. The recent work [Loizou and Kim, 2011] studies the different effects of the common distortions introduced by traditional speech enhancement algorithms and demonstrates that large gains in intelligibility can be achieved by controlling these distortions. Large amplification distortions (i.e over-estimation of the amplitude spectrum) were found to bear the most detrimental effect on speech intelligibility, and they can be avoided by imposing constraints on the estimated magnitude spectra.

Hearing-impaired people suffer an important reduction of speech intelligibility in noisy environments, which causes that algorithms that only improve speech quality have little benefit. Consequently, traditional algorithms for speech enhancement are not suitable for the objective of this thesis, and new approaches should be explored. An alternative to improve intelligibility is the design of enhancement algorithms that optimize objective measures correlated with intelligibility. The work in [Ma et al., 2009] examines different objective measures for predicting the intelligibility of speech in noisy conditions. A subjective quality assessment shows that the frequency-weighted segmental SNR (fwSNRseg) [Hu and Loizou, 2008] and the perceptual evaluation of speech quality (PESQ) [Recommendation, 2001] performed the best, obtaining a Pearson's correlation coefficient with intelligibility of 0.81 and 0.79, respectively. The modified coherence speech intelligibility index (CSII) and the normalized covariance metric (NCM) measures incorporating signal-specific weighting information, as described in [Ma et al., 2009], have been found to perform the best in terms of predicting speech intelligibility in noise. An example of an algorithm that aims to improve intelligibility is found in [Loizou, 2005]. It proposes the use of Bayesian estimators of the spectral magnitude of speech based on perceptually-motivated cost functions. These cost functions are variants of speech distortion measures, such as the Itakura-Saito and weighted likelihood-ratio distortion measures. The author proposes six different estimators and finds that the estimators that emphasize spectral valleys more than the spectral peaks performed the best in terms of having less residual noise and better speech quality.

In the last decade, and motivated in the auditory scene analysis (ASA), the computational auditory scene analysis (CASA) has become relevant as a new concept for speech enhancement and sound source separation with the aim to mimic the behavior of the human auditory system. The CASA approach proposes to design machine learning algorithms based on the mechanisms of the human auditory system to segregate mixtures of sound sources in the same way that human listeners do. The mechanism for the perceptual segregation of sounds is amply studied in [Bregman, 1994], and it is closely related to the cocktail party problem introduced in [Cherry, 1953]. The majority of the proposed CASA models ([Brown and Cooke, 1994]) are based on a time-frequency representation of the input signal with a cochleagram, which is obtained with a gammatone filterbank whose bandwidths are set according to the model of the human inner ear [Glasberg and Moore, 1990]. The time-frequency representation is sometimes obtained with a spectrogram based on the short-time Fourier transform (STFT) [Allen, 1977]. The separation of sound sources in CASA systems is normally achieved by identifying and grouping spectrotemporal regions in the mixture belonging to the same source, which originates time-frequency binary masks. The application of CASA to single-channel noise reduction consists in generating time-frequency masks to weight the different time-frequency regions, emphasizing regions dominated by the target speech and suppressing regions dominated by noise. The basic idea behind time-frequency masking is not new, since the traditional PSS and the Wiener filter can be also considered as real-valued time-frequency masks.

A different study that differs from the CASA approach but also leads to the use of time-frequency binary masks is made in [Yilmaz and Rickard, 2004]. The authors observe that the energy of a speech signal has a sparse distribution in time and frequency, which means that the most part of the energy of the signal is contained in small and isolated regions of the time-frequency representation. Due to this sparsity property, the overlap between different speech sources is small within a high resolution time-frequency representation, and this sparsity is quantitatively measured. If the speech sources were orthogonal (i.e. they do not overlap), they could be perfectly separated with time-frequency binary masks. Despite the existing amount of overlap, high quality separation can be obtained. The ideal binary mask (IBM) proposed in [Hu and Wang, 2004] is defined as the one that takes values of zero or one by comparing the local SNR in each time-frequency bin against a threshold, which is typically chosen as 0 dB. The IBM is formally proposed as a criterion for evaluating a CASA system in [Wang, 2005]. This is motivated by the auditory masking phenomenon where a louder sound masks a weaker sound within a critical band. Several psychoacoustic studies ([Brungart et al., 2006; Li and Loizou, 2008; Loizou and Kim, 2011]) demonstrate that *the application of the IBM to separate speech in noisy conditions entails an improvement in speech intelligibility* and improves notably the performance of ASR. Furthermore, the work in [Li and Wang, 2009] provides a formal study of the optimality of the IBM in terms of SNR. The IBM can be viewed as a quantified version of the Wiener filter where each time-frequency value is rounded to the closest binary value. However, this fact means that the IBM is only optimal in each local time-frequency unit. The study establishes the conditions for the IBM to be optimal at the global level, as well as evaluates the differences in the performance when these conditions are not met.

Unfortunately, the computation of the IBM needs to have access to the clean speech and noise signals, information that is not available in practice. Hence, the IBM should be estimated somehow from the corrupted signal, obtaining a binary mask that is just an approximation of the IBM. The study in [Li and Loizou, 2008] evaluates the impact on intelligibility caused by errors in the estimation of the IBM. Overall, there is a strong and negative correlation between the amount of error introduced in the binary mask and the obtained intelligibility scores. If the objective is to restore speech intelligibility, the error in the estimation should be lower than 10%, but if the objective is just to improve intelligibility, the error can be at most 20% or 30%, depending on the type of masker noise. The performance is affected mainly when time-frequency units dominated by noise are wrongly labeled as time-frequency units dominated by speech. The study also examines the effect of varying the local SNR threshold utilized to generate the IBM, finding a region that ranges from -20 to 5 dB where the intelligibility is almost invariable.

Two strategies have been followed for the estimation of the IBM. The first one is based on the CASA principles of grouping and segmentation, using features such as periodicity across frequency, common offsets and onsets, and common amplitude and frequency modulations. The main problem of this approach is the estimation of the fundamental frequency and the detection of onset/offset segments in noise. CASA techniques for noise reduction are inspired in speech source separation, and they will be reviewed in detail in the coming sections. *A conceptually and computationally simpler procedure to estimate the IBM to isolate speech from noise is the use of a classifier to identify time-frequency points as either speech-dominated or noise-dominated.* Some recent works based on this approach are:

- In [Ramírez et al., 2006] a support vector machine (SVM) is trained for speech/non-speech discrimination, using sub-band SNRs as input features. Two different alternatives for the calculation of the sub-band SNRs are proposed. The first alternative applies a noise reduction algorithm and passes the signal and the residual noise through a K-band filterbank to reduce the dimensionality. In the second alternative, a measure of the contextual deviation of the power spectrum from the background noise is defined. The measure is transformed to a wide K-band spectral representation where the sub-band SNRs are calculated. Results show that the second set of features obtain better results in terms of speech/non-speech classification, improving the performance of standard voice activity detectors (VADs).

- The work in [Kim et al., 2009] designs an accurate binary Bayesian classifier to estimate the IBM, capable of operating at negative SNR levels. Using amplitude modulation spectrograms (AMSs) as input features, Gaussian mixture models (GMMs) are trained to represent the distribution of each class. The target binary mask is generated comparing the local SNR to a frequency-variable threshold. The method is evaluated with low SNRs (0 and -5 dB) and with different types of noise (babble, factory and speech-shaped). Results indicate substantial improvements in intelligibility over the obtained by human listeners with the unprocessed signals.

- The algorithm in [Kim and Loizou, 2010] introduces a new binary mask based on the magnitude spectrum constraints proposed in [Loizou and Kim, 2011]. The algorithm combines the classification schema in [Kim et al., 2009] with the new mask for improving speech intelligibility. The algorithm is tested with low SNRs (0 and -5 dB) and with different types of noise (babble, airport and speech-shaped), obtaining higher PESQ scores than the previous algorithm.

1.3.1.2 Multi-channel noise reduction

Beamforming techniques achieve speech enhancement by using the principle of spatial filtering provided by a microphone array, normally composed of omnidirectional microphones, and assuming that the target source and the unwanted sources are physically separated in space. Spatial filtering aims to boost the signal coming from a determined direction, attenuating the interfering signals coming from different directions. In theory, a microphone array allows reducing the noise without distorting much the speech signal, in opposition to single-channel enhancement algorithms, which usually introduce distortions. Conventional beamforming techniques were initially developed for narrowband applications such as radar or communications, and later adapted to wideband signal processing. However, microphone arrays have some remarkable differences in comparison to the original beamforming applications: speech is a relative wideband signal, the assumption of far-field is not always valid, there is a high multi-path interference due to room reverberation in closed spaces, the signals and the environments are highly non-stationary, and the number of sensors is usually restricted. These differences have motivated the formulation of new techniques for microphone array applications. Beamforming techniques can be broadly grouped into data-independent (fixed) and data-dependent (adaptive). Data-independent techniques use fixed parameters during the processing of the input signal. On the other hand, data-dependent techniques update their parameters constantly depending on the input signal, adapting to changing noise conditions. Fixed techniques are simpler to implement, but they are limited when rejecting highly directive noise and changing noise sources, specially in the case of small arrays. Nevertheless, they are usually used in highly reverberant environments, applications where the position of the target source is known or assumed to be known (e.g. hearing aids or cars) and for creating multiple beams. An important factor to have into account when selecting a determined beamforming technique is the type of noise field in which the array is expected to work. The noise field is characterized by the degree of correlation between the noise signals at different spatial locations, and it is typically classified into coherent, incoherent or spatially white, and diffuse or spherically isotropic.

The simplest fixed beamforming technique is the delay and sum (DS) beamformer, in which delayed versions of the channels are equally combined at the output. The delay applied to each channel is an estimation of the time difference of arrival (TDOA) between each sensor and the reference one. The amplitude of the desired signal, which comes from the steering direction, is not modified, obtaining a distortionless response. The amount of noise attenuation increases as the number of microphones and the array length increase. Once the array geometry and the steering direction are established, the beam pattern of a DS beamformer is fixed, which is an important limitation when removing directional noise. That is the reason why a DS beamformer is usually combined with other techniques in real applications, for instance, in [Flanagan et al., 1985]. The DS beamformer belongs to a more general class of beamformers known as filter and sum, in which a different finite-impulse response FIR filter is applied to each channel. In fact, most types of beamformers belong to this class. A filter and sum beamformer allows applying a different complex weight to each channel, adjusting the beam pattern. The coefficients of these FIR filters can be computed according to different criteria, originating different types of beamformers. The most relevant are described next.

One of the first adaptations of beamforming techniques to microphone arrays was the constant directivity beamformer (CDB). The frequency response of a narrowband-designed array varies in a wide frequency range such as the speech range (i.e. 300 Hz to 3.4 kHz), which entails that the interfering signals are not equally attenuated in the entire band, resulting in a disturbing output speech. With the aim of overcoming this limitation, the CDB obtains an invariable response in a wide frequency band. An extended approach is the use of harmonically-nested subarrays, where each sensor may be used in more than one subarray. Each subarray is designed as a narrowband array and their outputs combined by bandpass filtering. A two-dimensional array based on a dual-beam analog system with scanning properties is described in [Flanagan et al., 1985]. The main drawback of the CDB approach is that the size of the array depends on the lowest frequency of operation (i.e. a several meters long array is required for 300 Hz.).

Hence, its implementation in microphone array systems is limited for many applications. This problem is solved using other design strategy such as superdirectivity.

The so-called superdirective beamformer (SDB) is inspired in the supergain that closely spaced endfire arrays (i.e. linear array whose direction of maximum radiation is along the axis of the array) show under diffuse noise conditions, when adjacent elements are separated by less than one-half wavelength [Hansen and Woodyard, 1938]. In this case, the filter coefficients are calculated to maximize the array gain, which is defined as the SNR improvement between the reference channel and the system output. The array gain is maximized by minimizing the power of the output signal, which is equivalent to reduce the noise variance, with the constraint that the signal in the desired direction is undistorted. The traditional solution of this optimization problem is the well-known minimum variance distortionless response (MVDR) beamformer, also known as Capon filter [Capon, 1969]. The Capon filter was originally designed for narrowband signals, but it was later adapted to wideband signals by splitting the signal into narrow frequency bands and applying a different filter to each band. The MVDR solution involves the knowledge of the noise covariance matrix. The classic SDB is defined for diffuse noise fields, which covariance matrix is constant (i.e. it only depends on the distance between microphones), and in this case the MVDR filter is considered a fixed beamformer. However, the solution is valid for any well-defined noise field introducing its coherence noise matrix, and that noise field can be non-stationary. In such a case, the noise covariance matrix can be recursively updated and the MVDR would be an adaptive beamformer. A large number of works based on SDB and the MVDR filter can be found in the literature. Some remarkable works are listed below.

- The initial solution of the SDB can lead to undesirable gain of incoherent noise due to electrical sensor noise, channel mismatch and errors in microphone spacing. The work in [Gilbert and Morgan, 1955] imposes a gain constraint to remove uncorrelated noise originated by self noise and phase error in the microphones. This idea is implemented in [Cox et al., 1986] to reduce uncorrelated white noise in a SDB, using a sensitivity constraint.

- The MVDR filter achieves perfect dereverberation when the acoustic transfer functions (ATFs) between the target source and the microphones are known. Unfortunately, the blind estimation of ATFs is not easy. A tradeoff between noise reduction and reverberation cancellation is observed in [Benesty et al., 2007]. The work in [Habets et al., 2010] provides a rigorous analysis of this tradeoff as well as analyzes the local and global behavior and derive novel forms of the MVDR filter. The study is conducted for different noise fields such as a mixture of coherent and non-coherent noise fields, entirely non-coherent noise field and diffuse noise field. The results show that the amount of noise reduction sacrificed for complete dereverberation depends on the direct-to-reverberation ratio (DRR) of the acoustic impulse response between the source and the reference microphone, and the desired response. The amount of noise reduction sacrificed decreases when the number of microphones increases.

- Additionally to the distortionless response constraint imposed by the MVDR, it is possible to design a SDB with multiple linear constraints, which is known as linearly constrained minimum variance (LCMV) beamformer [Er and Cantoni, 1983]. The MVDR beamformer is a particular case of the LCMV beamformer where only one constraint is applied. The work in [Habets et al., 2009] compares the noise reduction capability of the LCMV against the MVDR. The constraints are modified to include relative ATFs rather that ATFs to cancel coherent interferences. In a scenario composed by one desired source and one undesired source in spatially white noise, the MVDR performs significantly better noise reduction compared to the LCMV beamformer.

Adaptive beamforming techniques are able to adapt to changing acoustic environments, obtaining higher noise reduction than fixed techniques, but being less robust due to their higher sensitivity to errors in the steering vector. Most adaptive techniques rely on the minimization of the MSE between a reference signal highly correlated to the desired signal, and the output signal. However, the application of the classical LMS algorithm to minimize the MSE introduces distortions in the target speech signal. This limitation is solved by the famous Frost's algorithm presented in [Frost, 1972], where the optimization of the filter coefficients is converted into a constrained LMS minimization problem, where the constraint is given by a determined transfer function for the target signal. This constraint is usually applied to speech signals ensuring constant gain and linear phase. The Frost's algorithm is an adaptive version of the MVDR beamformer. An alternative to this adaptation is the generalized sidelobe canceller (GSC) [Griffiths and Jim, 1982]. This structure creates a double path for the signal: a standard fixed beamforming path (normally a MVDR filter), and an adaptive path composed of a blocking matrix, which removes the desired signal from this path, and a set of adaptive filters that minimize the output noise power. The filter coefficients are adapted with a LMS algorithm and the output signal is obtained by subtracting the signals at the end of both paths. The GSC structure is generalized to include multiple constraints in [Buckley, 1986], implementing a LCMV broadband beamformer. A proof of the equivalence between the GSC and the LCMV beamformer based on complementary subspaces is presented in [Breed and Strauss, 2002].

A problem associated with the GSC is the assumption that the received signals are simple delayed versions of the source signal. However, real environments have arbitrary ATFs and the performance of this method is seriously affected in reverberating rooms. In [Gannot et al., 2001] the authors adapt the GSC to deal with any ATF, estimating ATF ratios instead of ATFs exploiting the non-stationarity of speech. The algorithm is derived in the frequency domain and can be also implemented in the time domain, but the latter increases the computational burden. The method is experimentally evaluated using speech and noise signals recorded in a real reverberating room. Nevertheless, the hard constraints imposed in LCMV techniques cause a lack of freedom in the choice of the filters, limiting the achieved level of noise reduction. A different approach for adaptive beamforming is the use of a soft constraint, allowing to introduce some distortions in the target signal but controlling that these distortions are not perceived by the human ear. The most important technique is the adaptive microphone array system for noise reduction (AMNOR) introduced in [Kaneda and Ohga, 1986]. The method proposes to control the tradeoff between minimizing the output noise power and reducing the signal degradation by introducing a fictitious desired signal only during noise periods. The optimization criterion proposed (AMNOR criterion) maintains the degradation of the frequency response of the desired signal below a determined value. Subjective tests confirm the superiority of the AMNOR criterion over conventional criteria for noise reduction. The main limitations of this technique are the need for an accurate speech/noise detection and the knowledge of the ATFs between the source and the microphones.

In practice, the filter and sum beamformer rarely exhibits the performance ideally expected, mainly due to reverberations that arrive from nearly all directions caused by multipath room reflections. The performance can be further improved introducing a single-channel noise suppression filter at the output of the array. The use of a time-varying post-filter allows incorporating frequency filtering using the knowledge provided by the spatial filtering previously performed. However, the non-uniform modification of frequency components can lead to signal distortions, in the same sense that the ones described in single-channel speech enhancement algorithms. The solution provided by the MVDR beamformer is optimum in the maximum likelihood (ML) sense and it maximizes the SNR for a narrowband signal. The multichannel Wiener filter (MWF) provides the optimum broadband filter in the MMSE sense, and its solution can be factorized into a MVDR beamformer followed by a single-channel Wiener post-filter [Edelblute et al., 1967]. The Wiener post-filter technique uses the cross-spectral densities or coherence function between the signals of the different channels to calculate the filter coefficients. The filter is derived under the assumption of spatially uncorrelated noise and reverberation, but it also performs reasonably well in case of diffuse noise field. The approach was first applied in [Allen et al., 1977] with the purpose of reducing reverberation. The algorithm applies the post-filter at the output of a simple DS beamformer with only two microphones. The work in [Zelinski, 1990] extends the algorithm to a higher number of microphones and tests it in a typical office room with a two-dimensional 4-microphone array. The performance of the filter is drastically incremented, with barely noticeable residual noise. Post-filtering techniques based on [Zelinski, 1990] assumes zero cross-correlation between the noise on different sensors, which is inaccurate, especially at low frequencies and for small arrays. The work in [McCowan and Bourlard, 2003] presents a generalization of the post-filter technique, replacing the assumption of incoherent noise with the assumption of a known noise field coherence function. The method is tested in a real reverberant office environment, improving the performance of the original postfilter technique.

In recent years, speech models based on CASA have been successfully applied to single-channel speech enhancement. The combination of spatial information with CASA techniques allows improving the performance of traditional beamformers as well as performing SSS by filtering the signals from different directions. A common approach based on the time-frequency domain is to apply a binary time-frequency mask to the beamformer output. An useful example is the algorithm presented in [Levi and Silverman, 2010] which applies a CASA-inspired time-frequency binary mask to the output of a DS beamformer steered to the target direction. Each time-frequency point of the beamformer output is labeled as target-dominated or noise-dominated using a discriminator based on the steering response power with phase transform (SRP-PHAT). The algorithm is successfully implemented in real-time in [Ayllón et al., 2011], and the performance is improved in [Do and Silverman, 2011], using a null-steering MVDR beamformer with phase transform (MVDR-PHAT). Special attention is paid to binaural speech enhancement and separation systems, which are multichannel systems inspired in CASA that takes advantages of the spatial resolution provided by two microphones placed at both sides of the human head. In this case, the signals that arrive to the microphones are modified by the so-called head-related transfer function (HRTF). This fact causes physical cues that the human brain uses to localize sound sources coming from different spatial locations. The most important cues are the interaural time difference (ITD) and the interaural level difference (ILD) [Rayleigh, 1907]. The works in [Hawley et al., 2004] evaluates the advantages of using binaural hearing in detriment of monaural hearing to discriminate sources. In [Roman et al., 2001], time-frequency binary masks are generated using spatial cues to separate sources, and the algorithm in [Roman et al., 2006] estimates de IBM comparing each time-frequency point of the mixture signal with the output of an adaptive beamformer that cancels the target source. The main application of binaural beamforming is speech enhancement in binaural hearing aids. Some examples will be discussed in detail in the following sections.

1.3.2 Sound source separation

The SSS problem consists in estimating the original source signals from the mixture signals. The complete estimation process involves also dereverberation, which is a complex problem that is not considered in this thesis. SSS methods can be grouped according to different criteria:

1. Number of channels: multichannel or single-channel.
2. Type of mixture: instantaneous, anechoic or echoic.
3. Relative number of mixtures and sources: overdetermined, determined or underdetermined.
4. Amount of information about the sources or the mixing process: blind or semi-blind source separation (BSS), source-based models, CASA techniques.

In this review of the state of art, the different SSS techniques have been divided into single-channel and multichannel. Multichannel techniques are generally based on statistical assumptions. The use of several observations makes possible to estimate several sources with very little information about them. However, in single-channel mixtures, blind estimation of the original sources from a single mixture is very difficult, and some a priori knowledge is necessary. Single-channel mixtures are naturally underdetermined, and most techniques for SSS are based on CASA, whereas multichannel techniques involves statistically-based BSS methods, as well as beamforming techniques and binaural CASA models to exploit spatial information. The most significative contributions in both cases are reviewed in this section.

A related problem that requires special attention is source enumeration. Most SSS algorithms assume to know the number of sources in advance, assumption that is not usually met in real situations. Hence, the problem of estimating the number of speech sources automatically from a mixture is studied separately and a review of algorithms that aim to solve this problem is also provided.

1.3.2.1 Single-channel source separation

CASA methods are psychoacoustically motivated techniques based on auditory perception that try to separate sound sources in the same way that the human auditory system does. CASA basis have been already described, and its application to single-channel SSS consists in identifying and grouping those time-frequency regions belonging to each source to generate a time-frequency mask for each original speech source. Two types of grouping can be distinguished: simultaneous grouping that aims to group sounds that overlap in time (i.e. frequency components), and sequential grouping that aims to put together successive speech sections from the same speaker that are separated in time. CASA single-channel techniques can be roughly divided into feature-based and model-based.

Feature-based methods make use of intrinsic sound properties such as proximity in frequency and time, periodicity (harmonicity), amplitude modulation (AM) and frequency modulation (FM), temporal continuity and onset/offset events. Many relatively simple algorithms have been proposed for the extraction of these features from a single speech signal. Unfortunately, the extraction of these features from a mixture of speech signals is more complex, and new advanced algorithms for feature extraction have been proposed. The most relevant approaches for feature extraction and speech separation algorithms that involves one of several of these features are described below.

- **Pitch estimation:** A big amount of algorithms that estimate the pitch of isolated speech signals can be found in the literature. However, the estimation of multiple pitches in speech mixtures is a harder task, mainly due to the mutual overlap between voices that weakens the pitch cues. Algorithms that use pitch estimations to separate speech sources need to estimate multiple pitches, but algorithms that estimate multiple pitches need to separate the speech sources first. Single pitch estimation methods exist in both the time domain and the spectral domain. There are many variants of both approaches, but most rely on the same ideas. Most of spectral techniques are based on pattern matching, finding periodicity in the sequence of peaks in the power spectrum, hence they are highly dependent on frequency resolution. This idea was first applied in [Schroeder, 1968], and many variants have been proposed, for instance, [Duifhuis et al., 1982]. Time domain methods try to find periodic patterns in the time signal normally using the autocorrelation function. This idea was introduced in [Rabiner, 1977] and variations are found, for instance, in [Klapuri, 2005].

  In the case of multiple pitch estimation, there are two strategies. The first alternative is to estimate a single pitch from the mixture, removing the speech source corresponding to that pitch, and estimating again the pitch from the remaining mixture, repeating the procedure until all sources are extracted. The previous estimations can be refined in further iterations. The second and more elegant alternative is to jointly estimate all the pitches at the same time. Again, methods exist in the spectral and temporal domain, but the former are more common for multiple estimations. Spectral approaches are based on the work in [Parsons, 1976] that seeks the harmonic series that best match the spectrum. Another work based on the previous one is [Vincent et al., 2010]. Some works in the time domain are [De Cheveigné, 1993; De Cheveigné and Kawahara, 1999]. The main limitation of spectral domain methods is frequency resolution, which is limited by the size of the analysis window, and it affects directly the accuracy of the pitch estimation. On the other hand, time domain methods are limited by the sampling resolution and their computational efficiency is lower than spectral techniques.

  Multiple pitch estimations are useful to perform simultaneous grouping of voiced speech segments, grouping harmonics that are scattered in the frequency spectrum. In [Parsons, 1976], voiced segments are separated using a comb filter with large responses at the fundamental frequency and its harmonics. This is a common solution adopted in many following works, for instance, in the model proposed in [Brown and Cooke, 1994] which combines pitch, FM and onset/offset detection. Early CASA models perform relatively well in low frequencies where the harmonics are resolved, but their performance is reduced in high frequencies where the harmonics are unresolved. The work in [Hu and Wang, 2004] groups resolved and unresolved harmonics differently. Resolved harmonics are grouped according their periodicity, and unresolved harmonics are grouped according to AM rates.

- **Onset and Offset detection:** The identification of sudden intensity changes (increase or decrease) is easily achieved by finding the peaks and valleys of the first-order derivative of the intensity function with respect to time. However, many peaks and valleys can be originated by background noise, hence the intensity function should be previously smoothed applying a low-pass filter. An example of speech segmentation based on onset and offset analysis is [Hu and Wang, 2007]. Onset/offset detection is usually combined with pitch estimation to separate voiced segments, but they are very useful to segregate unvoiced speech that lacks periodicity. A significative example is [Hu and Wang, 2008].

- **Amplitude modulation extraction:** AM detection is a common problem in signal processing which corresponds with the extraction of the envelope of the signal, which variations are assumed to be much slower than the carrier frequency (i.e. the fundamental frequency in the case of speech). Common methods are the Hilbert transform method and the half-wave rectification following to low-pass filtering. The amplitude modulation spectrum (AMS), which is the spectral representation of the signal envelope, is a very useful feature for speech separation. Some examples of algorithms that use AMS for separation are [Kim and Loizou, 2010; Kim et al., 2009].

- **Frequency modulation extraction:** FM corresponds with variations of the carrier frequency, which occurs at rates much slower than the carrier frequency itself. The FM feature used in CASA refers to a change in frequency of a sound component, and it can be detected either from a two-dimensional cochleagram [Brown and Cooke, 1994] or from the response of a band-pass filter [Kumaresan and Rao, 1999].

Model-based methods understand the problem of source separation as inference, where some constraints should be included to be able to recover an approximation of the original signals. The CASA approach uses a parametric model of the sources which parameters are estimates from the mixture. The most common approach is the use of hidden Markov models (HMM) for the sources. The constraints included in the model represent the prior knowledge about the expected sources, and they can be either explicit or implicit. Explicit signal models typically use a dictionary containing the possible signals, for instance, in [Roweis, 2001], or they consider that the signals are contained in a subspace [Jang and Lee, 2002]. Feature-based models based on periodicity can be considered implicit signal models.

Unsupervised learning algorithms have been also applied to single-channel source separation. These methods usually apply a simple non-parametric model and use less prior information of the sources, learning the information directly from the data. One of the most popular approaches is based on non-negative matrix factorization (NMF). The work in [Virtanen, 2007] combines NMF with a sparseness constraint for single-channel SSS based on minimizing a cost function which is a weighted sum of three terms: a reconstruction error term, a temporal continuity term, and a sparseness term.

1.3.2.2 Multichannel source separation

Multichannel SSS can be divided into two main approaches: one basically inspired in the independent component analysis (ICA), and the other relies on sparse representations of speech in which only a small number of the source components differs significantly from zero.

The first BSS techniques applied to SSS were based on ICA [Comon, 1994]. The ICA main assumption is that the sources are statistically independent and non-Gaussian, and the separation problem is formulated as a mixing matrix estimation problem. Further assumptions about the number of microphones and the mixing process are required. ICA tries to find the independent components of the mixture by maximizing the statistical independence of the estimated components either minimizing the mutual information or maximizing the non-Gaussianity. The main limitations of ICA are: the original formulation is not valid for underdetermined mixtures, the mixing matrix needs to be stationary during a period of time (i.e. the sources can not move), the sources should come from different spatial directions, and the number of sources must be known in advance. The algorithms based on ICA work very well when the signals are mixed instantaneously, but they do not perform so well in a reverberant environment. Many efforts have been carried out to adapt the original ICA to undetermined and reverberant mixtures.

The work in [Hyvärinen and Oja, 1997] describes a fast implementation of the ICA algorithm, which is denominated FastICA. The algorithm finds, one at a time, all non-Gaussian independent components, regardless of their probability distributions. The convergence of the algorithm is guaranteed, and the algorithm is 10 to 100 times faster than gradient-based ICA algorithms. The work in [Parra and Spence, 2000] exploits the non-stationarity of speech to estimate the multiple channels of echoic speech mixtures. The multi-path channels are identified using a LS optimization to estimate a forward model. An efficient FastICA (EFICA) algorithm is described in [Koldovsky et al., 2006], where the accuracy given by the residual error variance attains the Cramer-Rao lower bound. The algorithm assumes that the PDFs of the independent signals are generalized Gaussian distributions. The computation time is only three times higher than the standard FastICA.

A more recent approach for SSS is based on the assumption that the sources are sparse and the data do not overlap in the time-frequency domain. The sparsity based approach solves the underdetermined separation problem and the algorithms can be further divided into two categories: the first type of algorithms are based on MAP estimation of the sources, usually performed by l_1-norm minimization, after estimating the mixing matrix either by clustering or by using the ML criterion; the second type of algorithms are based on extracting the signals by means of time-frequency masking, which can be calculated using different criteria. A relevant example of the first type of algorithms is the one described in [Bofill and Zibulevsky, 2001]. The algorithm exploits the sparsity of speech and music signals when they are represented in the STFT domain. The authors propose the use of a clustering algorithm to estimate the mixing matrix from only two sensors, and a shortest path separation procedure based on the l_1-norm to recover the most sparse original signals from the mixtures. The algorithm also identifies the number of sources in the mixture. Tests with speech and music mixtures show good separation performance even in the case of separating 6 sources from only 2 mixtures. Another interesting algorithm is the line orientation separation technique (LOST) algorithm described in [O'Grady and Pearlmutter, 2008]. The algorithm considers that the problem of audio source separation is equivalent to the separation of linear subspaces in a mixture of oriented lines and separates any number of sources from any number of instantaneous mixtures by identifying lines in a scatter plot. The orientation of each line is estimated using an EM procedure. The demixing procedure in case of undetermined mixtures is performed using l_1-norm minimization.

The best-known algorithm for SSS based on sparsity and time-frequency masking is the degenerate unmixing estimation technique (DUET) [Rickard and Yilmaz, 2002; Yilmaz and Rickard, 2004]. In these works, the authors introduce the concept of approximate W-disjoint orthogonality (WDO) to measure the orthogonality of speech signals in the STFT domain. The experiments carried out demonstrate that there exists a time-frequency binary mask that allows separating each speech source from the mixtures, similar to the one inspired in CASA, but the problem still remains in the estimation of the IBM from the observations. Unlike traditional CASA approaches that use a single mixture, the DUET algorithm uses two mixtures to estimate the IBM. The algorithm proposes to construct a weighted two-dimensional histogram from estimations of the delay and level differences between the two microphones. The weighted histogram shows peaks corresponding to each source. Unsupervised clustering is applied to identify these peaks from the smoothed histogram, and these peaks are used to estimate the mixing parameters of each source. The demixing procedure is performed via time-frequency masking, generating binary masks based on a proximity criteria. Listening experiments show that the WDO measure is fairly correlated with subjective separation performance. Hence, the WDO measure is proposed as a good indicator of the separation performance for this type of SSS methods. There are three main limitations of the DUET algorithm: the number of sources must be known in advance, its performance is reduced in echoic mixtures, and the use of time-frequency binary masks introduces residual musical noise. Many methods based on DUET have been proposed in the last decade, and some of the most relevant are listed below.

- A multichannel DUET algorithm is described in [Melia and Rickard, 2006], combining the sparse assumption with the estimation of signal parameters via rotational invariance technique (ESPRIT). The method, denominated DESPRIT, is limited to linear arrays.
- A new algorithm for SSS, denominated time-frequency ratio of mixtures (TIFROM), is presented in [Abrard and Deville, 2005]. Using two microphones, it allows separating speech sources from instantaneous linear mixtures even if the original signals almost fully overlap in the time-frequency domain. The only condition required is the existence of slight differences in the time-frequency distributions of the original signals, i.e. each source only needs to occur alone in a small time-frequency area. The algorithm calculates time-frequency ratios of the mixed signals to identify those small time-frequency areas and estimates the mixing matrix. This approach is much less restrictive than ICA and sparsity-based approaches.
- Music signals do not meet so well the WDO assumption as pure speech signals do, due to their harmonic structure. The DUET algorithm is combined with CASA techniques to perform stereo music source separation in [Woodruff and Pardo, 2006]. The algorithm has three steps: a cross-channel histogram is performed using spatial cues (i.e. similar to the DUET algorithm), the pitches of the original signals are estimated from the previous histogram to generate harmonic masks, and the harmonic amplitude envelopes are obtained from the pitch estimations.
- The method in [Araki et al., 2007] proposes a generalized multichannel DUET algorithm that is valid for any number of sensors and geometry. The method performs k-means clustering using normalized amplitude and time differences between sensors. The phase differences are weighted to obtain a variance comparable to the one of the level differences.

The two main approaches, ICA and sparseness, have been also combined in order to overcome their individual limitations. In [Araki et al., 2004] the authors combine both approaches with the aim of reducing the distortions associated to time-frequency binary masking. The algorithm estimates the time-frequency points where only one source is active, removes that source from the observations and applies ICA to the remaining mixtures. The time-frequency source estimation is inspired in DUET. Furthermore, the authors propose to reduce the distortions associated to binary masking using a directivity pattern based continuous mask instead. The mask is generated with a null beamformer. The use of a soft mask reduces distortions even in reverberant rooms. On the other hand, ICA and time-frequency masking can be also combined the other way around: a time-frequency mask can be applied to the ICA outputs, as a post-processing technique. For instance, in [Kolossa and Orglmeister, 2004] the time-frequency masking is applied to the output of two frequency domain ICA methods. The time-frequency masks are determined from the ratio of the demixed signal energies. The approach notably increases the output SNR.

1.3.2.3 Automatic speech source enumeration

Up to this point, no concern has been given to find the number of speech sources present within a mixture. The original ICA and DUET algorithms, which are two of the main algorithms for multichannel SSS, assume to know the number of sources in advance, which is an important limitation.

There are many different approaches to signal enumeration, and those based on information theoretic criteria have largely been used in array signal processing [Krim and Viberg, 1996]. These algorithms are based on the estimation of a parametric model order for the observed process. Two such criteria for order estimation of an observed process are the Akaike information criterion (AIC) [Akaike, 1974] and the Rissanen's minimum description length (MDL) principle [Rissanen, 1978], which have inspired many algorithms to solve the aforementioned problem, for instance, [Cheng et al., 2012; Krim and Cozzens, 1994; Valaee and Kabal, 2004; Wax and Ziskind, 1989]. Unfortunately, most of those algorithms have been applied to problems where the relative bandwidth of the signals is low, such as radar, sonar or mobile communications. The wideband nature of speech requires a different approach. Furthermore, the information theoretic approach is normally reduced to the over-determined case. In the last years, several algorithms for speech enumeration have been proposed, some of them included as an initial step in multichannel SSS and source localization algorithms. They are usually based on TDOA and pitch estimations. Some relevant examples are listed below.

- The authors in [Luengo et al., 2003] extend the MDL information theoretic criterion to estimate the number of sources in undetermined speech mixtures. The algorithm exploits the sparsity of the sources to construct an autocorrelation matrix from the angles of the observations to which the MDL criterion is applied.

- The method in [Klapuri, 2003] is based on an iterative search, where the fundamental frequency of the most prominent sound is estimated, the sound is subtracted from the mixture, and the process is repeated for the residual signal. The method stops when the weight associated to a candidate fundamental frequency falls below a threshold.

- The algorithm in [Katmeoka et al., 2004] uses a mixture of tied Gaussian mixtures to model the multiple harmonic structure of a speech mixture. The harmonic structure of each source and the number of sources are estimated from that model combining an EM algorithm with an information criterion.

- The method in [Gilbert and Payton, 2009] estimates the number of sources in instantaneous and non-instantaneous linear mixtures containing additive white Gaussian noise, using two sensors. The estimation combines TDOA and pitch estimates using a harmonic windowing function.

- In [Arberet et al., 2010] the authors propose a method that counts (and locates) the number of speech sources in underdetermined multichannel mixtures. The method is based on a local confidence measure which detects the time-frequency regions where robust information is available, and uses a technique similar to the generalized cross-correlation with phase transform (GCC-PHAT) to estimate the delays associated to each source. The method is valid for instantaneous and anechoic mixtures.

1.3.3 Speech enhancement algorithms for hearing aids

Directional microphones have been used in hearing aids for over 25 years and have proved to significantly increase speech intelligibility in various noisy environments [Hawkins and Yacullo, 1984]. However, they are usually not applicable to small in-the-canal devices for reasons of size and the assumption of a free sound field which is not met inside the ear canal. Nevertheless, directional microphones are not under the scope of this thesis. A comprehensive review is given in [Chung, 2004].

Besides directional microphones, modern hearing aids include one or several omnidirectional microphones combined with speech enhancement algorithms to improve intelligibility. The implementation of algorithms for speech enhancement in hearing aids presents particular challenges:

- The requirement or real-time processing limits the processing delay to few milliseconds, which in turns limits the algorithmic complexity.
- The reduced battery life limits the clock speed of the processor, which also limits the computational capability of the device.
- The number of microphones in multichannel systems is reduced due to the dimensions of the device.
- The number of frequency bands used for the analysis of the input signal is relatively small.
- Hearing-impaired listeners have greater susceptibility to interference from background noise than normal listeners. They typically require a signal-to-interference ratio (SIR) that is 5-10 dB higher than a normal hearing person in order to achieve the same level of speech understanding [Plomp, 1986].

Bearing in mind the above limitations, this section discusses the suitability of the different speech enhancement approaches for their implementation in hearing aids.

1.3.3.1 Single-channel algorithms

Single-channel noise reduction in hearing aids is even more challenging than in the general case. As it was previously described, single-channel noise reduction algorithms tend to reduce noise introducing distortions in the signal. Impaired listeners are more sensitive to speech distortions than normal listeners. Consequently, the effect that these distortions have on intelligibility can be minimized for normal listeners but it is magnified for hearing-impaired listeners. Among the single-channel noise reduction algorithms previously described, those based on the Wiener filter and the MMSE estimator have been traditionally implemented in hearing aids [Hamacher et al., 2005]. Unfortunately, these methods may improve the SNR, but they could not yet prove to enhance the speech intelligibility. Despite their limitations, single-channel noise reduction systems are still implemented in modern hearing aids.

Concerning single-channel algorithms for speech separation inspired in CASA, they usually are either too complex or the performance is too limited to be directly applicable to practical hearing systems. These algorithms typically involve complex operations for feature extraction, segregation and grouping, which makes a real-time implementation difficult. In addition, the performance of such algorithms is not good enough for the implementation in a hearing aid [Wang, 2008]. Nevertheless, the application of time-frequency masking is a promising approach, as long as the mask computation is relatively simple.

1.3.3.2 Multichannel algorithms

Recently, high-end hearing aids including multiple microphones have demonstrated to provide reasonable improvements in intelligibility and listening comfort. Multichannel SSS algorithms, such as those based on ICA or clustering, have reduced application in hearing aids due to their complexity. Hence, multichannel speech enhancement in hearing aids has been dominated by beamforming techniques. The design approach varies: some systems enhance the signal coming from the target direction (usually straight-ahead), whilst other systems suppress the noise from a specific direction (usually coming from the back). Both fixed and adaptive beamformers have been successfully implemented in modern hearing aids. The required processing time of fixed beamformers is relatively low, as long as the filter coefficients that satisfy the design constraints can be previously computed and easily included as constant values in the embedded algorithm. An additional advantage is that fixed beamformers are more robust than adaptive beamformers to minor steering errors and reflections correlated with the desired signal. However, their performance is reduced when rejecting directional interferences. The are several works that analyze the effects of the array geometry and the number of microphones for several types of fixed beamforming techniques, evaluating the intelligibility improvement introduced for hearing-impaired subjects. Some remarkable works are [Kates and Weiss, 1996; Liu and Sideman, 1996; Saunders and Kates, 1997; Stadler and Rabinowitz, 1993]. A common and affordable approach is the use of independent small endfire arrays, often integrated into behind-the-ear devices, with low microphone distances of around 1-2 cm [Peterson and Zurek, 1987]. The use of external larger arrays have been also proposed, for instance, with microphones placed in eyeglasses [Zwicker and Beckenbauer, 1988], but this solution is not comfortable for hearing aid users. Adaptive beamforming requires higher computational capability and it is more sensitive to steering direction errors, but it has better performance rejecting interferences. However, the evaluation of the performance is highly influenced by the acoustic environment, which makes the measurement of the benefit obtained over fixed beamforming difficult. An example that uses the MVDR filter is found in [Spriet et al., 2005] where the filter is used to implement a MWF. One promising approach is the application of a GSC structure. Some GSC-based algorithms for hearing aids are [Berghe and Wouters, 1998; Greenberg, 1998; Greenberg and Zurek, 1994; Hoffman et al., 1994].

Hearing loss usually affects to both ears and the hearing-impaired person is forced to wear a hearing device in each side. Bilateral systems perform independent processing in the left and right hearing aids, which originates that the spatial cues are distorted, decreasing the localization ability of the user. A recent trend motivated by the availability of wireless data links between the right and left hearing aids is the design of binaural beamformers. In such a case, the speech is enhanced by combining the information from both ears. Binaural systems work with dual-channel input-output signal, although more than one microphone could be placed in each device. The main advantage of binaural processing is the availability of spatial cues (ILD and ITD) that can be used to separate sounds. However, these cues must be preserved in the binaural output in order to maintain the original spatial information. A simple example of binaural noise reduction is found in [Wittkop et al., 1997], where the ILD and ITD estimates are compared with a reference value for the frontal direction. Binaural fixed beamformers have low computational complexity, but they only preserve the spatial cues of speech (i.e. the target signal). The work in [Lotter and Vary, 2006] designs a dual-channel superdirective beamformer and obtains the binaural output signal by applying adaptive spectral weights to the beamformer input channels. The spectral weights are computed from the monaural output of the beamformer. The desired signal is passed unfiltered. The performance is further increased applying a MWF.

Some examples of binaural adaptive beamforming based on the GSC structure are [Campbell and Shields, 2003; Welker et al., 1997], which use a two-microphone sub-band adaptive GSC-like structure to adaptively cancel out interfering sources. Beamforming combined with CASA techniques allows preserving the binaural cues of speech and noise, but the use of time-frequency masking introduces some distortions. In [Roman et al., 2006], a binaural adaptive beamformer is trained to form a null in the front direction. A single time-frequency mask is then calculated comparing the responses from the front cardioid and the back cardioid. The binary masking algorithm is very simple, making feasible its implementation in a hearing aid. The system is designed using real measurements obtained from a KEMAR manikin. In [Rutledge, 2009] an adaptation of the MVDR beamformer is combined with monaural CASA attributes. The simultaneous and temporal grouping steps are performed with clustering and Kalman filtering. Finally, the analysis of the robustness of different binaural speech enhancement systems in hearing aids is carried out in [Rohdenburg et al., 2007], using objective perceptual quality measures.

Binaural hearing aids require the exchange of information between the left and the right devices. Due to aesthetic reasons, the best solution is the use of a wireless link for data transmission, which notably increases the power consumption, one of the main limitations in these devices. This fact opens a new area of research: how to reduce the amount of information transmitted (bit-rate) without altering the performance of the enhancement system. One of the first related works is [Roy and Vetterli, 2006] which evaluates the gain provided by collaborating hearing aids as a function of the communication rate, using an information theoretic approach. In [Doclo et al., 2007] the authors evaluate the decrement of noise reduction achieved by a binaural MWF when reducing the bandwidth of the transmission link. The work in [Srinivasan and Den Brinker, 2009] proposes two approaches to reduce data transmission. The first approach is to transmit only an estimation of the undesired signal at a determined bit rate, and the second approach is to transmit the complete received signal at the determined bit rate. The second schema transmits more information, but it requires higher transmission rate. Furthermore, the authors evaluate the transmission of only the low-frequency components.

1.4 Scope of the thesis

Speech enhancement is an extensive and active field of research, with applicability in a variety of trending applications. The previous sections contain a thorough review of the state of the art in this field, and it gives an idea of the broad range of solutions that have been proposed during many years. However, due to the complexity of the problem and the higher requirements demanded by new applications, the speech enhancement problem remains largely open and unsolved. In this broad framework, the following objective is set as principal in this thesis:

> *To design speech enhancement algorithms based on source separation and spatial filtering suited to audiological applications, with special emphasis on state-of-the-art hearing aid devices.*

The main objective establishes some constraints to the general speech enhancement problem. *In the case of audiological applications, it is primordial to increase the speech intelligibility rather than the speech quality*, due to the fact that people with hearing disorders suffer a lack of understanding of speech in noise. Additionally, in the case of hearing aid devices, the signal processing algorithms must work in real time. This fact, together with the low computational resources available in hearing aids, force speech enhancement algorithms implemented in such devices to be relatively simple.

The next particular problems, which arise from the main objective, are addressed in this thesis:

- To improve the robustness of time-frequency SSS methods based on clustering, increasing their performance in the separation of different types of sources and mixtures. Specifically, speech, noise and music sources will be considered, as well as anechoic, echoic and binaural mixtures.

- To formulate solutions for the problem of automatic speech enumeration, which is necessary as an initial step in most speech source separation algorithms.

- To design low-cost single-channel speech enhancement algorithms for monaural hearing aids. The algorithms should increase the speech intelligibility and be feasible to be implemented in a state-of-the-art commercial hearing aid.

- To design low-cost speech enhancement systems that increase the energy efficiency of wireless-communicated binaural hearing aids. In addition to the requirements of monaural devices, binaural systems should optimize the data transmission in order to reduce the power consumption associated to the wireless link.

- To generalize the design of customized microphone arrays for speech enhancement in monaural and binaural hearing aids. In order to maximize the enhancement provided by a microphone array, the HRTF of the hearing aid user should be considered in the computation of the beamformer filter coefficients. Unfortunately, the availability of this information is limited in practice, and the array can not be fitted to a specific user. The design can be generalized to optimize the output speech quality of the array for any unknown subject.

1.5 Structure of the thesis

This thesis has been divided in two main blocks, which are organized and presented as follows:

- The first block contains the preliminary study of the problem. It contains two chapters. The first chapter, which is the current one, contains a global description of the problem, a comprehensive review of the state of the art of the problem addressed in this thesis, as well as a description of the main objectives of this thesis. The second chapter establishes the theoretical basis necessary for the understanding of speech enhancement algorithms, as well as the material used in the experiments carried out for the evaluation of these algorithms.

- The second block, which is the main block, contains a description of the research conducted to fulfill the objectives of this thesis, as well as it describes the experimental work and results obtained. The chapters of this block correspond with each of the goals of the thesis, and they are the next:

  - In chapter 3, the performance of the so-called DUET algorithm is evaluated in a variety of scenarios, demonstrating the need for more advanced clustering techniques in such situations. A novel source separation algorithm that combines the mean shift clustering technique with the basis of DUET is proposed in this chapter. Additionally, an algorithm for automatic speech source enumeration is presented in this chapter.
  - Chapter 4 tackles the problem of single-channel speech enhancement and its application to monaural hearing aids, considering that the main goal is to improve the intelligibility of speech in noise, rather than to improve the speech quality. A novel algorithm that increases intelligibility is proposed. The algorithm is feasible to be implemented in a state-of-the-art hearing device.

- In chapter 5, different approaches to design low-cost speech enhancement algorithms that increase the energy efficiency of the wireless-communicated binaural hearing aid are proposed. The algorithms are feasible to be implemented in state-of-the-art hearing aids.
- Chapter 6 deals with the design of superdirective beamformers for monaural and binaural hearing aids considering the head shadow effect but assuming unavailable head measurements of the subjects.
- Chapter 7 summarizes the main results obtained during this research and the main contributions of the thesis. The chapter also contains a description of future research lines that have been opened or broaden with the realization of this thesis. Finally, a list of publications derived from this thesis is also included.
- The last section of the thesis includes the bibliography used in this thesis.

Chapter 2

Background and materials

2.1 Introduction

This chapter pretends to formally describe the speech enhancement problem in the time-frequency domain, as well as to provide the mechanisms to evaluate the quality of the algorithms proposed in this thesis. First, the main properties of speech signals and their differences with other types of sounds are established. The signals from different types of sound sources are usually mixed together, existing different mixing models to describe the way the sources are mixed. The different mixing models are formally presented in this chapter, providing a formal description of the problem to solve. This thesis approaches the speech enhancement problem in the time-frequency domain. Hence, the motivation for the use of such domain is presented, and the mathematical basis used to transform the sound signals into the time-frequency domain are also described. Finally, the last section of this chapter describes the material used to evaluate the algorithms proposed in this thesis: objective measurements, sound databases and procedures to generate different types of mixtures.

2.2 Sound sources and mixtures

2.2.1 Types of sound sources

Different types of sound sources can be mixed together: speech, music and noise from different kind of sources (babble noise, ambient noise, narrow-band noise, etc.). According to the goal of this thesis (i.e. speech enhancement), it is considered that the sound mixtures contain at least one speech source which is the desired or target source. The temporal and spectral differences of the different types of sources are essential for many speech enhancement algorithms. The basic structure and main characteristics of speech, music and noise sources are briefly described.

2.2.1.1 Speech sources

Speech represents a sequence of sounds produced by the human vocal apparatus system to convey a determined message. The nature of speech signals makes their analysis to differ from the analysis of other types of signals. The main characteristics of speech are the following:

- **Non-stationarity**
  Speech signals can be modeled as non-stationary stochastic processes. Fortunately, speech is considered to be quasi-stationary over short periods of time (of the order of 20 ms).

This fact enables the analysis of short-time signal segments as stationary processes, but it involves a windowing process in the analysis and synthesis stages. The short-time Fourier transform (STFT) is a powerful mathematical tool commonly applied to perform this analysis. It will be described in more detail in the coming sections.

- **Wideband signal**
  Speech signals have a bandwidth of approximately 7 kHz, but most of the information is in the band up to 4 kHz. This is a relatively wide bandwidth compared to the sampling frequency (usually 8 or 16 kHz). Hence, speech signals are considered wideband signals. It is very important to consider this property in the design of speech enhancement algorithms, specially in those based on filters in the frequency domain.

- **Non-Gaussianity**
  Speech signals are highly non-Gaussian but they are closer to have a super-Gaussian PDF. This property should be considered by statistical-based algorithms and can be advantageous to perform speech enhancement.

- **Speech production model**
  The commonly assumed speech production model decomposes speech into an excitation signal and a vocal tract related filter. The excitation signal corresponds with the glottal flow and it is represented by an impulse train generator with the same period as the speech signal (the inverse of the fundamental frequency) for voiced segments, and a random noise generator for unvoiced segments. The vocal tract filter can be approximated by an AR model. A schema of this speech production model is represented in figure 2.1.

- **Pitch**
  The pitch is the subjective perception of the fundamental frequency and it is a characteristic property of each human voice. The pitch of voiced segments varies over time, but stays within a range of about 40 Hz centered around an average of 140 Hz for male voices and 200 Hz for female voices. The pitch is an interesting property for speech source separation due to the fact that it allows discriminating between different speakers.

2.2.1.2 Music sources

Music signals usually contain speech signals but also musical instruments that produce differences between the speech and music spectra. Speech tends to have a well-defined spectrum with well established and predictable perceptual characteristics. In contrast, musical spectra

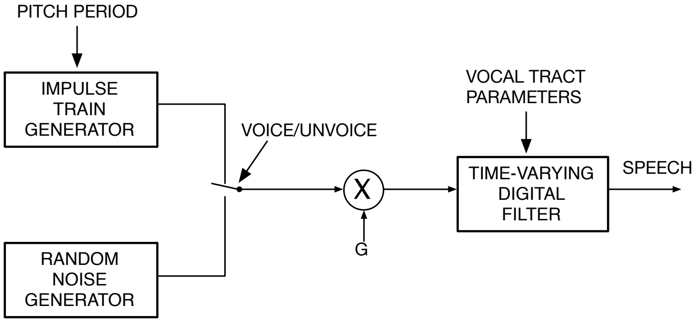
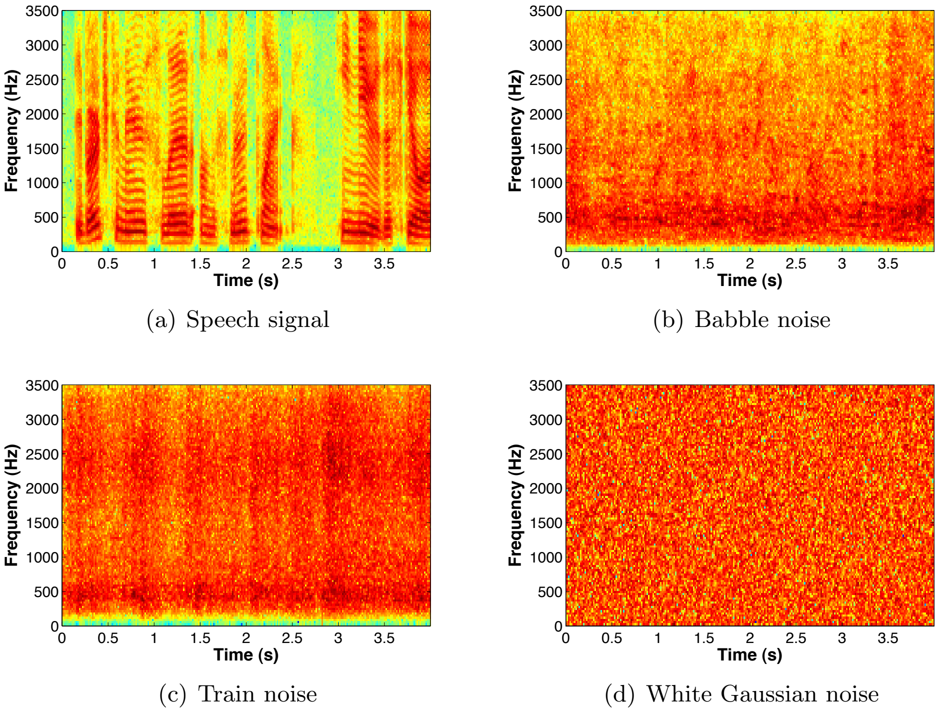


Figure 2.1: Speech production model.



(a) Speech signal

(b) Babble noise

(c) Train noise

(d) White Gaussian noise

Figure 2.2: Spectrograms of clean speech (a) and different types of noise: babble noise (b), train noise (c) and white Gaussian noise (d).

are highly harmonic, and the spectral characteristics strongly depend on the instrument being played. Furthermore, the fundamental frequencies in music vary in an wide range that goes from 30 Hz to 4 kHz. Nevertheless, music signals will be considered as interfering signals in this thesis.

2.2.1.3 Noise sources

There is a wide variety of noise sources that usually interfere with the desired speech: aircraft, bus, cafe, car, kindergarten, living room, nature, school, shops, sports, traffic, train, train station, etc. Noise sources can be divided into stationary or non-stationary. Stationary noise is produced in homogeneous noisy environments, for instance, the aircraft cabin noise or a factory noise. Non-stationary noise refers to other non-homogeneous noises, for example, children shouting in a kindergarten or babble noise. The spectra of the different types of noise differ, thus affecting in a different way to speech intelligibility: non-stationary noise affects more the speech intelligibility than stationary noise as well as it is more difficult to remove. The spectral differences of different types of noise are clearly appreciated in figure 2.2, which represents the spectrograms of a clean speech signal (a), a babble noise (b), a noise recorded in a train car (c), and white Gaussian noise (d).

2.2.2 Mixing models

There exist different mixing scenarios where the sound signals described in the previous section are mixed together. A mixing model is described by a mathematical expression of the observations generated by the mixing process, which are the mixture of signals received by the microphones of the system. Prior to describe the different mixing models, it is worth clarifying the notation used in this thesis. In signal processing, it is common to assume that the variable (t) represents continuos time and the variable $[n]$ discrete time (i.e. $[n] = (nT_s)$ where T_s is the sampling period). *In this thesis, all the signals are defined in discrete time, and the variable (t) is adopted to represent the discrete time signals.* Henceforth, $x(t)$ represents a discrete time signal, where $t = 0, ..., T-1$ are T observations of the signal.

Let us consider a set of M microphones that receive the signals coming from N different sources, $s_n(t)$, $n \in \{1, \cdots, N\}$, to generate M mixtures, $x_m(t)$, $m \in \{1, \cdots, M\}$. The general expression for the additive mixing model is given by

$$x_m(t) = \sum_{n=1}^{N} s_n(t) * h_{mn}(t), \quad m = 1, ..., M, \tag{2.1}$$

where $h_{mn}(t)$ is the impulse response of a linear time-invariant (LTI) filter that describes the acoustic channel between the n-th source and the m-th microphone, and the operator $*$ represents linear convolution. The filter $h_{mn}(t)$ is commonly denominated acoustic impulse response, and its transformation into the frequency domain is known as acoustic transfer function (ATF). The signals received by the microphones are the result of the convolution between the original sources $s_n(t)$ and the filter $h_{mn}(t)$ (i.e. $s_n(t) * h_{mn}(t)$). The type of the mixing model depends on the assumptions made about the ATF.

2.2.2.1 Instantaneous mixing model

The instantaneous or linear mixing model is the simplest model and it assumes that the signals received by the microphones are just a scaled version of the original signals, and it is expressed as

$$x_m(t) = \sum_{n=1}^{N} a_{mn} \cdot s_n(t), \quad m = 1, ..., M, \tag{2.2}$$

where a_{mn} are the scaling factors. In this case, $h_{mn}(t) = a_{mn} \cdot \delta_K(t)$, where $\delta_K(t)$ is the Kronecker delta function.

2.2.2.2 Anechoic mixing model

The anechoic or delayed mixing model introduces into the previous model different delays between the sources and the microphones, and it is given by

$$x_m(t) = \sum_{n=1}^{N} a_{mn} \cdot s_n(t - \delta_{mn}), \quad m = 1, ..., M, \tag{2.3}$$

where a_{mn} and δ_{mn} represent the attenuation and delay respectively, introduced by the channel in the signal that travels from the n-th source to the m-th microphone. The impulse response of the acoustic channel filter is $h_{mn}(t) = a_{mn} \cdot \delta_K(t - \delta_{mn})$. The sampling frequency is assumed to be high enough to allow delays δ_{mn} lower than one sample.

2.2.2.3 Echoic mixing model

The echoic or convolutive mixing model also considers the reflections produced by the environment, that is, the microphones receive several delayed and attenuated versions of the same source signal. The process is described by

$$x_m(t) = \sum_{n=1}^{N} \sum_{p=1}^{N_p} a_{mnp} \cdot s_n(t - \delta_{mnp}), \quad m = 1, ..., M, \tag{2.4}$$

where N_p is the number of different paths that the signals take from the sources to the microphones, and a_{mnp} and δ_{mnp} are the attenuations and delays introduced in the p-th path. In this case, the acoustic impulse response is given by $h_{mn}(t) = \sum_{p=1}^{N_p} a_{mnp} \cdot \delta_K(t - \delta_{mnp})$.

2.2.2.4 Noisy model

The observations of the three mixing models previously described are noise free, but they can easily include an additive noise term that reflects uncorrelated noise. This term contemplates acoustic noise (e.g. isotropic noise), and inaccuracies in the measurements performed by the sensors. The general expression of the mixing model with additive noise is given by

$$x_m(t) = \sum_{n=1}^{N} h_{mn}(t) * s_n(t) + n_m(t), \quad m = 1, ..., M, \tag{2.5}$$

where $n_m(t)$ represents the noise at the m-th sensor.

2.2.2.5 Matrix notation

The mixing models are usually expressed in matrix notation to simplify their formulation. Let us define $\mathbf{x} = [x_1(t), ..., x_M(t)]^T$ as an $M \times 1$ vector of mixtures and $\mathbf{s} = [s_1(t), ..., s_N(t)]^T$ as an $N \times 1$ vector of sources, where the operator $(.)^T$ denotes matrix transposition. The mixing matrix is defined according to

$$\mathbf{A} = \begin{bmatrix} h_{11}(t) & ... & h_{1N}(t) \\ \vdots & \ddots & \vdots \\ h_{M1}(t) & ... & h_{MN}(t) \end{bmatrix}, \tag{2.6}$$

and the general mixing model is given by $\mathbf{x} = \mathbf{A} * \mathbf{s} + \mathbf{n}$, where $\mathbf{n} = [n_1(t), ..., n_M(t)]^T$ is the $M \times 1$ noise vector and $*$ denotes the element-wise convolution operation (note that in case of instantaneous mixtures the operation becomes a simple matrix product).

2.3 Time-frequency representation of speech

The non-stationary nature of speech motivates the analysis of speech signals in both the time and frequency domains simultaneously. The classical Fourier analysis represents the frequency content of a signal, but it do not provide information about the time of appearance of frequency components or sudden changes of energy. The statistics of speech signals vary with time, and their frequency content can only be considered stationary in short-time segments around 20 ms. In a time-frequency representation, the frequency domain only reflects the behavior of a short-time segment of the signal. The most used time-frequency representations of speech signals are the short-time Fourier transform (STFT) [Allen, 1977] and the discrete wavelet transform (DWT) [Akansu and Haddad, 2000]. The algorithms described in this thesis are only based on the STFT, which is described in depth in this section.

The super-Gaussian PDF associated to speech signals causes the sparse representation of speech in certain domains. Sparsity refers to the property by which most of the sample values of a signal are zero or close to zero. Speech signals show a certain level of sparsity when they are represented in the time or in the frequency domain, but this property is increased when they are represented in the time-frequency domain. This property is very useful for speech source separation due to the fact that the probability of two of more sources being simultaneously active is low in sparse representations. Sparsity is formally described in this section.

2.3.1 Discrete short-time Fourier transform

The discrete STFT is a time-localized spectral transformation based on the discrete Fourier transform (DFT). The DFT is an orthogonal transformation which uses complex exponentials as basis functions. The DFT coefficients $X(k)$ of a discrete time signal $x(t)$ composed of T samples are calculated according to

$$X(k) = \sum_{t=0}^{T-1} x(t)e^{-j\frac{2\pi}{T}kt}, \quad k = 0, ..., T-1, \tag{2.7}$$

where the variable k represents frequency. Note that the coefficients $X(k)$ are complex values that comprise the magnitude spectrum $|X(k)|$ and the phase spectrum $\angle X(k)$. The inverse discrete Fourier transform (IDFT) is given by

$$x(t) = \frac{1}{T}\sum_{k=0}^{T-1} X(k)e^{j\frac{2\pi}{T}kt}, \quad t = 0, ..., T-1. \tag{2.8}$$

The DFT is a frequency localized transformation, where the analog frequencies equivalent to the normalized frequency of the basis functions are fixed and given by $f_k = \frac{kf_s}{T}$, with $k = 0, ..., T-1$, where f_s is the sampling frequency. The samples of speech signals are real numbers, which causes that the DFT is symmetric. Due to this fact, only $K = \lfloor \frac{T}{2} + 1 \rfloor$ frequency bands are considered. The DFT provides a spectral analysis of the signal, but it lacks of the temporal information required for speech analysis.

The STFT can be viewed as a two-dimensional transformation (i.e. frequency and time) which is calculated by splitting the input signal into segments using a sliding time-limited window and then calculating the DFT of each of the segments. The complex DFT coefficients of each frame are stored as a column in a matrix. The segments (frames) usually overlap with each other to avoid artifacts at the boundaries. Considering a discrete time input signal $x(t)$, it is segmented into frames according to

$$x_l(r) = w(r)x(r + lD), \quad r = 0, ..., R-1, \tag{2.9}$$

where $x_l(r)$ is the windowed l-th frame of the signal, r is a local time index, R is the window length, and D is the hop size which represents the number of samples that the sliding window moves between two consecutive frames. The STFT is obtained by calculating the DFT of each windowed segment of the signal, and it is given by the expression

$$X(k, l) = \sum_{r=0}^{R-1} w(r)x(r + lD)e^{-j\frac{2\pi}{T}kr}, \quad k = 0, ..., T-1, \tag{2.10}$$

where $X(k, l)$ is the STFT point corresponding to the k-th frequency bin of the l-th frame. The squared magnitude of the STFT, usually in dBs, yields the spectrogram of the signal,

$SPG = 20\log_{10}|X(k,l)|$, which is a visual representation of the variation of the energy of the spectral components with time (see figure 2.2 for some examples). The STFT is invertible under some conditions, which means that the original signal could be perfectly reconstructed from its transform by applying the inverse STFT (ISTFT) if these conditions are fulfilled. The synthesis process is basically the opposite of the analysis: compute the IDFT of each spectral segment and perform an overlap-add method using a synthesis window. The choice of the window is important to obtain perfect reconstruction. Some widely used windows are the Hamming and Hanning windows [Harris, 1978; Nuttall, 1981], which have into account frequency resolution and sidelobe behavior.

The frequency resolution provided by the STFT is the same that the one provided by the DFT, but it also provides a fixed time resolution of $t_m = \frac{D}{f_s}$ seconds. Both time and frequency resolutions depend on the window length R: when the width of the window increases the frequency resolution increases, but it implies a decrement in the temporal resolution; on the other side, if the window length is smaller, the time resolution is higher, but the frequency resolution is poorer. In the analysis of speech signals, the window length should guarantee stationarity (i.e around 20 ms.), and that value depends on the sampling rate. For instance, if the sampling rate is 8000 Hz, a window of 128 samples provides a time resolution of 16 ms. Additionally, the maximum processing delay allowed in real-time systems such as hearing aids restricts the window length.

2.3.2 Sparsity of speech signals in the time-frequency domain

Let $s_j(t) \in \mathbf{s}$, $j \in \{1, \cdots, N\}$, to be the equivalent discrete-time version of a set of square integrable bandlimited functions and suppose that there exist a linear transformation U that maps the set $\mathbf{s}$ into another family of signals $\mathbf{S}$, $U : s_j \to S_j$, with the following properties:

1. U is invertible on $\mathbf{s}$, i.e. $U^{-1}(U(s_j)) = s_j, \;\; \forall s_j \in \mathbf{s}$.

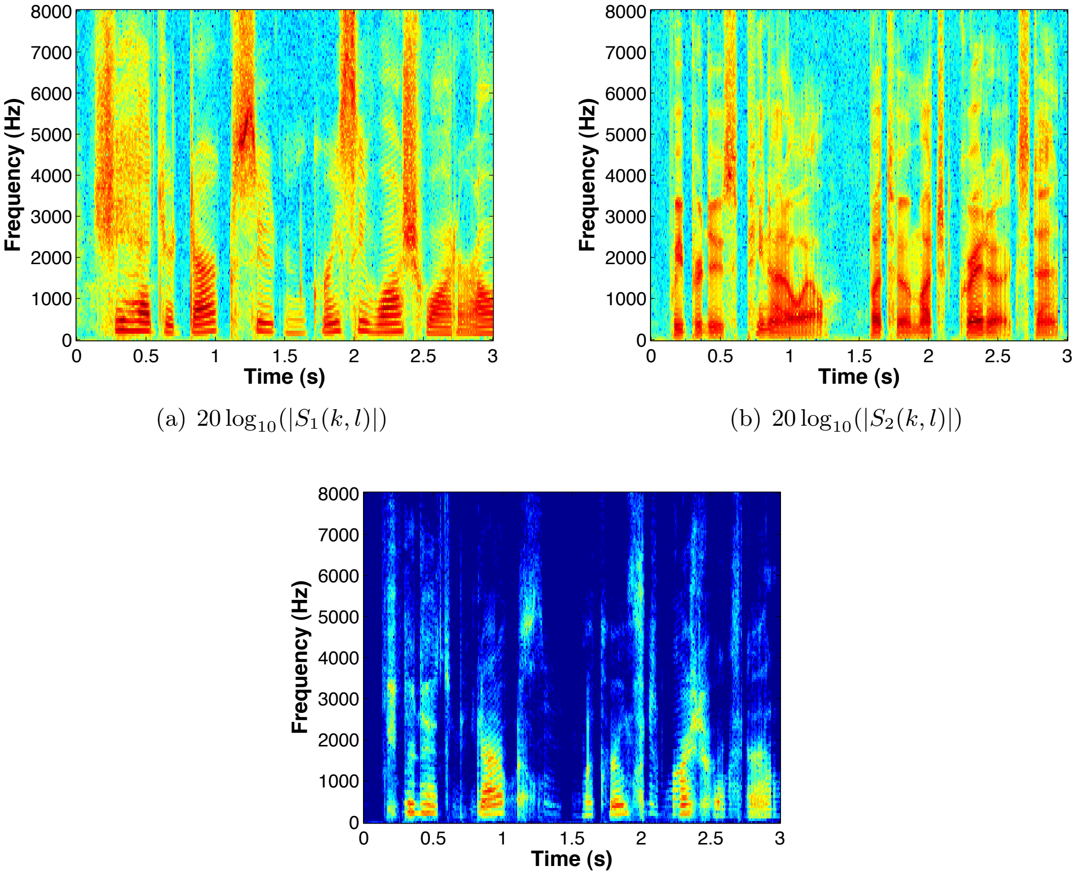
2. $\Lambda_j \cap \Lambda_k = \emptyset$ for $j \neq k$, where Λ_j represents the set of non-zero values of S_j.

If the two previous conditions are met, the transformation projects the signals to a domain on which they have disjoint representation (i.e. they do not overlap), and consequently the signals can be perfectly separated in that domain. In practice, the second condition is difficult to be fulfilled, and can only be satisfied in some approximate sense. Transforms that result in sparse representations of the signals of interest, which are representations where a small percentage of the signal coefficients capture a large percentage of the signal energy, can lead to satisfy the second condition approximately.

The discrete STFT has proved to be an approximate sparse representation of speech signals. Let consider $S_1(k,l)$ and $S_2(k,l)$ to be the STFT of two speech signals. The signals would be completely disjoint in the transformed domain only if

$$S_1(k,l)S_2(k,l) = 0, \; \forall \; k, m. \tag{2.11}$$

Unfortunately, condition (2.11) will not be satisfied for simultaneous speech signals because the time-frequency representation of active speech is rarely zero. However, speech is sparse in the time-frequency domain in the sense that a small percentage of the time-frequency coefficients of the STFT contain a large percentage of the overall energy. This fact implies that the magnitude of most of the coefficients is small, and it is unlikely that the large magnitude coefficients of different sources coincide, which leads the signals to be disjoint in an approximate sense. Figure 2.3 shows the spectrogram of two speech signals, $20\log_{10}(|S_1(k,l)|)$ and $20\log_{10}(|S_2(k,l)|)$, and



(a) $20\log_{10}(|S_1(k,l)|)$

(b) $20\log_{10}(|S_2(k,l)|)$

(c) $20\log_{10}(|S_1(k,l)S_2(k,l)|)$

Figure 2.3: Spectrogram of two speech signals, $20\log_{10}(|S_1(k,l)|)$ and $20\log_{10}(|S_2(k,l)|)$, and the spectrogram of their product in the time-frequency domain, $20\log_{10}(|S_1(k,l)S_2(k,l)|)$. The signals are sampled at 16 kHz and a Hamming window of 256 points with 50% of overlap is used to calculate the STFT.

the spectrogram of their product in the time-frequency domain, $20\log_{10}(|S_1(k,l)S_2(k,l)|)$. The signals are sampled at 16 kHz and a Hamming window of 256 points with 50% of overlap is used to calculate the STFT. It is clear that the product signal (c) contains less energy than the original speech signals in (a) and (b).

2.4 Time-frequency masking methods for speech enhancement

The application of SSS methods for speech enhancement is straightforward: the desired speech source can be separated from the remaining sources in the mixture (speech, music or noise), which are considered interfering sources, obtaining a cleaned version of the desired source. The SSS methods explored in this thesis are based on time-frequency masking, which exploits the sparsity property of speech when it is represented in the time-frequency domain. In this section, the mechanisms to perform sound separation via time-frequency masking are formally described.

2.4.1 W-disjoint orthogonality

Two signals are considered to be W-disjoint orthogonal (W-DO) if their STFT transformations do not overlap [Yilmaz and Rickard, 2004]. Assuming that this property is strictly met, in a mixture $X_m(k,l)$ of N speech sources at most one of them will be non-zero for a given time-frequency point. In such a case, it is possible to perfectly demix the signals by identifying the active source in each time-frequency point. Defining a time-frequency binary mask, $M_n(k,l)$, for the separation of the n-th source, $S_n(k,l)$, according to

$$M_n(k,l) := \begin{cases} 1, & S_n(k,l) \neq 0 \\ 0, & \text{otherwise} \end{cases}, \tag{2.12}$$

$S_n(k,l)$ can be demixed according to

$$\hat{S}_n(k,l) = M_n(k,l)X_m(k,l), \tag{2.13}$$

where $\hat{S}_n(k,l)$ is the estimated n-th source, which is expected to be a perfect estimation of the original source $S_n(k,l)$, in case of complete sparsity. Clearly, the hard restriction for W-DO sources is not strictly met by speech signals. However, if the restriction is relaxed to the assumption that the probability of two sources having large energy in the same time-frequency point is low, the speech sources are considered to be approximate W-DO sources. In spite of this assumption, good separation of speech sources can be achieved using time-frequency masking.

It is established in [Yilmaz and Rickard, 2004] that the performance of a given time-frequency mask depends on two properties: the amount of preserved target source and the amount of suppressed interfering sources. These two conditions are measured by the preserved-to-signal ratio (PSR) and the signal-to-interference ratio (SIR), respectively. The PSR indicates the amount of the energy of the target source preserved by the mask after separation. For the n-th source of the mixture, it is calculated as

$$PSR_n = \frac{||M_n(k,l) \cdot S_n(k,l)||^2}{||S_n(k,l)||^2}, \tag{2.14}$$

where $M_n(k,l)$ is the time-frequency mask computed for the separation of the n-th source $S_n(k,l)$. If the sources are W-DO, the mask defined in (2.12) preserves all the energy of the desired signal, obtaining the maximum value $PSR_n = 1$.

On the other hand, the SIR is an indicator of how the mask suppresses the interfering signals. For the n-th source it is given by

$$SIR_n = \frac{||M_n(k,l) \cdot S_n(k,l)||^2}{||M_n(k,l) \cdot Y_n(k,l)||^2}, \tag{2.15}$$

where $Y_n(k,l)$ is the STFT of the signal interfering with the n-th source, which is composed of the addition of all signals of the mixture except the desired signal:

$$Y_n(k,l) = \sum_{\substack{j=1 \\ j \neq n}}^{N} S_j(k,l). \tag{2.16}$$

When the sources are W-DO, the mask in (2.12) completely suppresses the energy of the interfering signals, and then the SIR is infinite, $SIR_n = \infty$. Both the PSR and the SIR are combined into the WDO factor to measure the approximate W-DO associated with a time-frequency mask. The WDO factor for the n-th source is calculated using the following expression

$$WDO_n = \frac{||M_n(k,l) \cdot S_n(k,l)||^2 - ||M_n(k,l) \cdot Y_n(k,l)||^2}{||S_n(k,l)||^2} = PSR_n - \frac{PSR_n}{SIR_n}. \tag{2.17}$$

It is clear that W-DO sources perfectly separated with the mask defined in (2.12) have a value of WDO=1, which is the maximum value. However, this value is only achievable by perfect W-DO sources and it obviously decreases (i.e. $WDO \leq 1$) with approximate W-DO sources, due to the fact that a small part of the source signals overlap, which implies that the mask is not able neither to preserve all the energy of the desired signal nor to reject all the energy of the interfering signals. Therefore, the WDO factor is a good indicator of the quality of the separation achieved by a time-frequency binary mask for approximate W-DO sources.

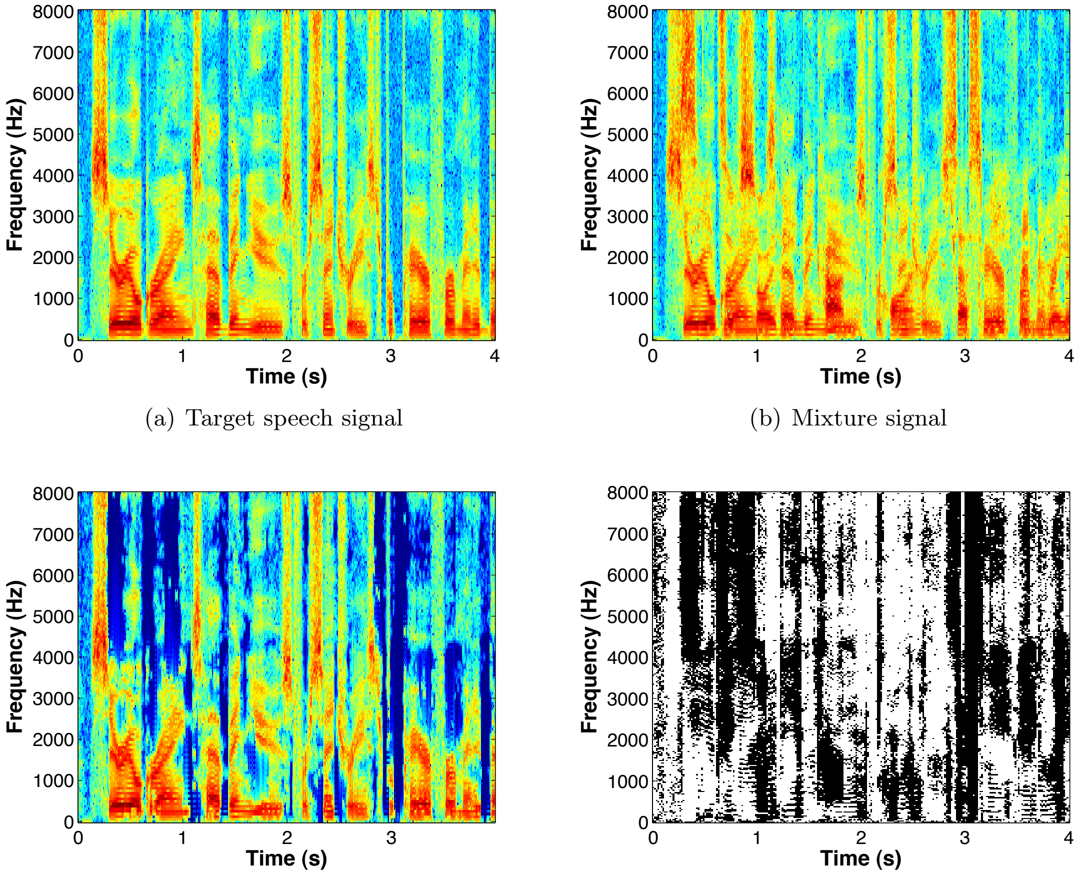
2.4.2 Ideal binary mask for approximate W-DO sources

The binary mask in (2.12) is defined for strictly W-DO sources, but it is not valid for approximate W-DO sources due to their mutual overlap. Nevertheless, the relaxed assumption that the probability of two sources having large energy in the same time-frequency point is low, motivates the application of time-frequency binary masks that assign each time-frequency bin to the predominant source. According to this, the binary mask associated to the n-th source is defined as

$$M_n(k,l) := \begin{cases} 1, & 20\log(\frac{|S_n(k,l)|}{|Y_n(k,l)|}) \geq x \\ 0, & \text{otherwise} \end{cases}, \tag{2.18}$$

which means that a time-frequency bin is associated to the source that has x dB more energy than its interfering sources. The effect of varying the energy threshold x is analyzed in [Yilmaz and Rickard, 2004], finding that a value of 0 dB maximizes the WDO. Hence, the 0 dB binary mask defined in (2.18) represents an upper bound of WDO for any other mask, and it will be useful to measure and compare the quality of the separation obtained by methods based on time-frequency masking.

The majority of CASA algorithms have also applied a time-frequency binary mask for sound separation. The ideal binary mask (IBM) defined in [Hu and Wang, 2001, 2004] is the same that the one defined in (2.18) for 0 dB, and it has been widely applied to the separation of not only speech sources, but also speech from noise and music. It has been proven in [Loizou and Kim, 2011] that *the IBM maximizes the articulation index (AI), a metric known to highly correlate with speech intelligibility*. Additionally, the work in [Li and Loizou, 2008] demonstrates that speech intelligibility is almost invariable when the energy threshold varies in the range that goes from -20 to 5 dB. The IBM defined in (2.18) is the optimal time-frequency mask in terms of WDO and it has also been established as a goal for CASA systems. Unfortunately, the computation of the IBM requires access to the desired speech and interfering signals, but this information is not available in practice. Hence, one of the main tasks to be performed by time-frequency masking SSS algorithms is the estimation of the IBM from the mixtures to separate the sound sources with enough quality. Figure 2.4 shows an example of the application of the 0 dB IBM in a mixture of two speech signals. The figure contains the spectrograms of the target speech signal (a), the mixture of the target signal with other speech signal with the same power (b), and the estimated target signal (c) using the IBM shown in (d). The spectrogram of the original target signal (a) and the estimated signal (c) are closely similar, which means that the desired signal is separated with high quality.



(a) Target speech signal

(b) Mixture signal

(c) Estimated target signal

(d) Ideal binary mask

Figure 2.4: Application of the 0 dB IBM in a mixture of two speech signals added with the same power. The target speech signal (a) is mixed with other speech signal with the same power (b), and estimated (c) using the IBM in (d). The spectrogram of the original target signal (a) and the estimated signal (c) are closely similar.

Finally, despite the good performance achieved by the application of binary masking for source separation, it has an important drawback that affects the quality of the separated signals: the introduction of a nonlinear distortion called musical noise. The assignation of the energy of each time-frequency bin to a single source causes spurious peaks in the processed spectrum. When the enhanced signal is reconstructed in the time domain, these peaks result in short sinusoidals whose frequencies vary from frame to frame. This type of noise is specially annoying during speech pauses and in cases of low SNR, where the noise is not masked by the speech signal. The musical noise can be reduced either performing temporal smoothing by using relatively smaller frame shifts or using soft masks instead of binary masks.

2.5 Evaluation of speech enhancement algorithms

The material used to evaluate the speech enhancement algorithms proposed in this thesis is described in this section. First, different objective measurements of speech quality are defined. The measurement of intelligibility is quite subjective but there are some indicators highly correlated with it. In order to generalize the obtained results, the algorithms must be evaluated against an extended set of mixtures containing a variety of different sources. These mixtures are generated in a controlled way, using the sound sources available in some public databases. Those databases used in this thesis are described in this section. Finally, the procedure to simulate different types of mixtures is also described.

2.5.1 Objective measurements

2.5.1.1 Signal-to-noise ratio (SNR)

The well-known SNR is a measure that compares the power of the desired signal with the power of background noise, and it is usually expressed in logarithmic scale (dB). Its application to evaluate speech enhancement algorithms is straightforward in situations where speech is corrupted by stationary background noise. In such a case, it is calculated according to

$$SNR(dB) = 10\log_{10}(\frac{P_{signal}}{P_{noise}}), \tag{2.19}$$

where P_{signal} is the power of the enhanced speech signal and P_{noise} is the power of the acoustic background noise.

2.5.1.2 Signal-to-interference ratio (SIR)

The SIR is also a widely used measure that represents the ratio between the power of the desired signal and the power of a set of interference signals. The SIR is applicable to measure the ability of speech enhancement algorithms to reject other spatially localized sound sources that interfere with the desired speech source. The SIR in dB is expressed as

$$SIR(dB) = 10\log_{10}(\frac{P_{signal}}{P_{interf}}), \tag{2.20}$$

where P_{interf} is the addition of the power of all interference sources. The SIR is very useful to evaluate the quality of source separation algorithms, and its application to time-frequency SSS algorithms has been already described in section 2.4.1.

2.5.1.3 WDO factor for time-frequency masking

The WDO factor defined in expression (2.17) is useful to measure the quality of the separation achieved by time-frequency masking SSS algorithms. A high SIR value means that most of the energy of the interference signals has been rejected, but it says nothing about the power of the desired signal, which may have been also supressed, resulting in a non-intelligible separated signal. This fact is considered by the WDO factor, which measures both the amount of interference signal rejected and the amount of desired signal preserved. Subjective listening tests performed in [Yilmaz and Rickard, 2004] demonstrated that there is a fairly relationship between the WDO measure and the subjective intelligibility of the separated sources: $WDO > 0.8$ leads to perfect intelligibility, $0.6 < WDO < 0.8$ implies minor artifacts, $0.4 < WDO < 0.6$ means distorted but intelligible, $0.2 < WDO < 0.4$ corresponds with very distorted and barely intelligible signals, and $WDO < 0.2$ results in signals that are not intelligible at all.

2.5.1.4 Perceptual evaluation of speech quality (PESQ)

The PESQ is a standard measure recommended by [ITU-T, 2001] to evaluate the speech quality of handset telephony and narrowband speech codecs, although it has been also adapted for perceptual evaluation of voice quality in speech enhancement algorithms. The PESQ algorithm compares the enhanced signal with the clean signal, producing a score between 1.0 and 4.5, with high values indicating better quality. Several works report high correlation between PESQ and subjective listening tests, [Hu and Loizou, 2008; Ma et al., 2009; Rix et al., 2001], which demonstrates that *the PESQ score is also a good indicator of speech intelligibility.*

2.5.1.5 Frequency-weighted segmental SNR (fwSNRseg)

The frequency-weighted segmental SNR (fwSNRseg) proposed in [Hu and Loizou, 2008] is a measure highly correlated with speech intelligibility and defined as

$$fwSNRseg = \frac{10}{L}\sum_{l=0}^{L-1}\frac{\sum_{k=1}^{K} W(k) log_{10}\frac{|S(k,l)|^2}{(|S(k,l)|-|\hat{S}(k,l)|)^2}}{\sum_{k=1}^{K} W(k)}, \tag{2.21}$$

where $S(k,l)$ and $\hat{S}(k,l)$ are the clean signal spectrum and the estimated signal spectrum respectively, $W(k)$ is a weight applied to the k-th frequency band, K is the total number of frequency bands and L is the total number of time frames. The weighted magnitude spectra $S(k,l)$ and $\hat{S}(k,l)$ is obtained by multiplying the DFT magnitude spectra by 25 overlapping Gaussian-shaped windows spaced according to the critical bands. The weights $W(k)$ are taken from table B.1 of the [ANSI, 1997] standard to compute the articulation index.

2.5.2 Public databases of sound sources

2.5.2.1 TIMIT database

The TIMIT database [Fisher et al., 1986] is a corpus of read speech designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of ASR systems. The database contains a total of 630 recordings from American English speakers of different sexes and dialects (a single speaker in each file). The speech signals are sampled at 16 kHz and saved in 16 bits PCM format. The signals of this database will be used to generate different types of mixtures according to the mixing models defined in section 2.2.2, in order to evaluate the algorithms proposed in this thesis.

2.5.2.2 NOIZEUS database

The NOIZEUS database [Hu and Loizou, 2007] is a noisy speech corpus developed to facilitate comparison of speech enhancement algorithms among research groups. The database contains 30 IEEE sentences [Rothauser et al., 1969] produced by three male and three female speakers, corrupted by 8 different real-world noises at different SNRs (0dB, 5dB, 10dB, and 15dB). The noise signals belong to the AURORA database [Hirsch and Pearce, 2000] and include suburban train noise, babble, car, exhibition hall, restaurant, street, airport and train-station noise. The signals are sampled at 8 kHz and saved in wave format (16 bits PCM).

2.5.3 Generation of synthetic mixtures

2.5.3.1 Anechoic mixtures

This type of mixtures only requires to add an attenuated and delayed version of each original source, according to the mixing parameters (see expression (2.3)). The attenuations and delays of the mixing model will be generally expressed with respect to the reference sensor, i.e. the attenuation in the reference sensor is 1 and the delay 0, the successive being calculated with respect to this sensor.

2.5.3.2 Echoic mixtures

Echoic mixtures, which are described by expression (2.4), are simulated using a room impulse response generator (RIRG). The acoustic impulse response between two points inside a defined room is calculated by using the simple image method described in [Allen and Berkley, 1979]. This method considers the dimensions of the room, the reflection coefficient and the number of virtual sources, to calculate the impulse response of the acoustic channel associated to each pair of source and microphone, both located in any point inside the room. The original method has been completed to consider also the attenuation due to distance and the directivity pattern of the microphones. Henceforth, this method is labeled as RIRG.

2.5.3.3 Binaural mixtures

Binaural mixtures are two-channel mixtures in which the microphones are situated in both human ear canals. It is well known that any kind of sound is modified by the anatomy of the body before entering the eardrum. Such modifications consist of a series of reflections, attenuations and time delays originated by the shape of the outer ear (pinna), the shape of the head and of the body (torso and shoulders) and some spatial characteristics. These effects, which are highly dependent on the DOA, introduce variations in the amplitude and phase of the original signal, originating monaural and binaural spatial cues that allow the human auditory system to localize sounds in the space. The amplitude variations are due to the so-called head shadow effect. The time variations are caused by the different paths that signals coming from different directions travel around the head until they reach the eardrum. Such modifications introduced in the signals that arrive at the eardrum are characterized by the head-related impulse response (HRIR), which can be obtained either by mathematical models or experimental measurements. Its transformation into the frequency domain is known as head-related transfer function (HRTF). It is important to clarify that the HRTF varies significantly from person to person, because it is strongly related to anatomy. Therefore, there will be a different HRTF for each subject, ear, and DOA. In this thesis, the HRTF's are labeled as H_{Ls} and H_{Rs}, for the left (L) and right (R) ear of the s-th subject, respectively.

Several mathematical models for simulating the head effects can be found in the literature. A 3D-model considering three independent models for the head, the pinna and the room is proposed in [Brown and Duda, 1997]. The head is assumed to be spheric, introducing time delays according to the DOA and shadowing by a one-pole/one-zero filter. The pinna model is a 5-th order FIR filter bank and the reflections due to the room are modeled by introducing delays and weights to the input signal. The model is improved in [Duda et al., 1999] with a more realistic model for the time differences based on an ellipsoid head model. However, none of the existing models is very accurate due to the use of physical approximations. In addition, they need anthropometric body measurements in any case. Moreover, experimental databases exist, measuring the HRIR with a microphone placed in the eardrum of different subjects or dummies, and varying the position of the source covering all possible directions.

In this thesis, binaural mixtures are generated using the CIPIC database [Algazi et al., 2001], which comprises recordings of the HRIR with in-the-canal microphones in 43 different human subjects and 2 KEMAR mannequins. The recordings are performed for different spatial directions, splitting the space in 50 angles corresponding to elevation and 25 angles corresponding to azimuth, having a total of 1250 different source directions along the sphere. The database also provides a set of 27 anthropometric measurements of the head, torso and pinna of the subjects. The generation of binaural mixtures can be performed either in the time domain, filtering the discrete time source signal with the corresponding HRIR, or in the frequency domain, multiplying each frame of the STFT of the source signal by the corresponding HRTF.

Part II

Proposed methods and results

Chapter 3

Time-frequency sound source separation for general purpose applications

3.1 Introduction

One of the most successful algorithms for source separation based on time-frequency masking is the so-called DUET algorithm. This algorithm performs good separation in case of anechoic mixtures of speech sources using only two microphones, but its performance drops with other types of mixtures and sound sources. The DUET algorithm is based on clustering estimates of the mixing parameters under the assumption of an anechoic mixing model. However, in some situations, for instance, echoic or binaural mixtures, these clusters are not well defined and the identification of the mixing parameters fails. In this chapter, a novel time-frequency masking algorithm for SSS that combines a generalization of the mean shift clustering technique with the DUET algorithm is presented. The proposed method aims to overcome the limitations of DUET and it is further generalized for any number of sensors, which is practical for many applications based on microphone arrays where more than two microphones are available, for instance, in multi-conference systems, police surveillance, vehicles or aircrafts, room monitoring system, etc. In such applications the separation can be improved using the information collected by more than two microphones.

A different problem associated to SSS algorithms is the automatic enumeration of speech sources in a mixture. Most SSS algorithms assume to know the number of sources in advance. A novel algorithm for source enumeration of speech sources based on information theoretic criteria is also presented in this chapter.

3.2 The DUET algorithm for SSS

The degenerate unmixing estimation technique (DUET) [Rickard and Yilmaz, 2002; Yilmaz and Rickard, 2004] is a SSS algorithm that allows recovering any number of sources from only two mixtures, thus solving the undetermined problem. The algorithm assumes an anechoic mixing model (i.e. attenuated and delayed sources) and exploits the approximate W-DO property of speech sources when they are represented in the time-frequency domain. The mixing parameters are estimated by clustering the relative attenuation-delay pairs between the two microphones. The estimation of the mixing parameters is used to generate time-frequency binary masks. The details of the algorithm are explained in this section.

3.2.1 Local estimates of the mixing parameters

The anechoic mixing model given by expression (2.3) can be rewritten in the time-frequency domain. In this section, $X_1(k,l)$ and $X_2(k,l)$ are the STFT of the signals received by the two microphones, and $S_n(k,l)$, with $n = 1, ..., N$, are the attenuated and delayed version of the original sources received by the first microphone, which is assumed to be the reference sensor. According to this, the mixing model can be expressed by

$$\begin{bmatrix} X_1(k,l) \\ X_2(k,l) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 & \cdots & 1 \\ a_1 e^{-i\omega\delta_1} & \cdots & a_n e^{-i\omega\delta_n} & \cdots & a_N e^{-i\omega\delta_N} \end{bmatrix} \begin{bmatrix} S_1(k,l) \\ \cdots \\ S_n(k,l) \\ \cdots \\ S_N(k,l) \end{bmatrix}, \tag{3.1}$$

where a_n and δ_n are the level and time differences between both microphones for the n-th source, respectively, and $\omega = \frac{\pi k}{K-1}$ with $k = 0, \cdots, K-1$. With the further assumption of W-DO sources (i.e. only one source is active at every (k,l) point), the previous mixing process becomes into

$$\begin{bmatrix} X_1(k,l) \\ X_2(k,l) \end{bmatrix} = \begin{bmatrix} 1 \\ a_j e^{-i\omega\delta_j} \end{bmatrix} S_j(k,l), \tag{3.2}$$

where $S_j(k,l)$ is the active source at each (k,l) point. According to this, the ratio of the STFT of the two mixtures does not depend on the sources but only on the mixing parameters related to the active source $S_j(k,l)$:

$$R(k,l) = \frac{X_2(k,l)}{X_1(k,l)} = a_j e^{-i\omega\delta_j}. \tag{3.3}$$

According to this, the local mixing parameters for each time-frequency point are easily estimated from

$$\hat{a}(k,l) = |R(k,l)| \tag{3.4}$$

$$\hat{\delta}(k,l) = -\frac{1}{\omega}\angle(R(k,l)). \tag{3.5}$$

When the sources are strictly W-DO, the previous estimators can only take the values of the mixing parameters. In practice, the W-DO assumption is only approximated, and the local mixing parameter estimates are not exactly the mixing parameters, but they will cluster around the mixing parameters. When these clusters are well defined and sufficiently separated, they can be identified, and the true mixing parameters can be estimated from their centers.

Spatial aliasing can affect the delay estimator in (3.5) when the distance between microphones is relatively large. The phase is unique only when $|\omega\delta_j| < \pi$, which yields the condition for the distance between microphones to avoid spatial aliasing $d < \frac{c}{2f_{max}}$. For example, considering that the highest frequency of interest is 4 kHz and the speed of sound is 340 m/s, the maximum distance between microphones to avoid aliasing is 4.25 cm, which can be a problem in some applications such as binaural mixtures. This limitation is overcome by analyzing the phase difference between adjacent time-frequency points, according to

$$R'(k,l) = \frac{X_2(k,l)}{X_1(k,l)} \left(\frac{X_2(k+\Delta k,l)}{X_1(k+\Delta k,l)} \right)^* = a_j^2 e^{i\Delta\omega\delta_j}, \tag{3.6}$$

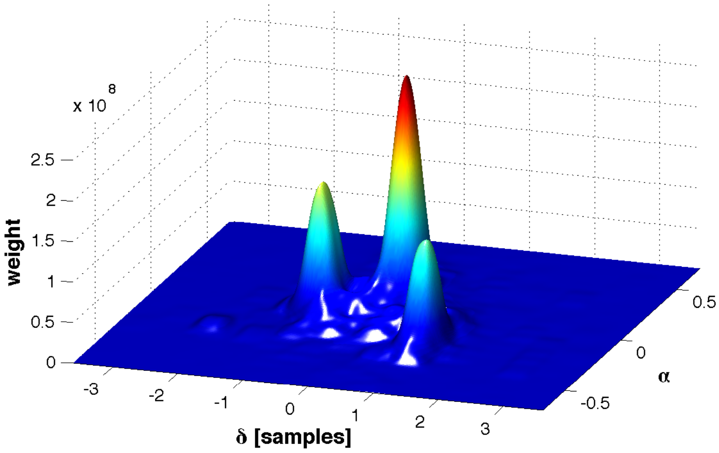


Figure 3.1: DUET two-dimensional weighted histogram of the symmetric attenuation ($\hat{\alpha}$) and the delay ($\hat{\delta}$) estimators, for a mixture of three speech sources. The peaks of the histogram have been smoothed using a two-dimensional FIR filter.

where the operator $(.)^*$ represents complex conjugation. The constraint to avoid phase ambiguity in (3.6) has been relaxed to $|\Delta\omega\delta_j| < \pi$, and the value of $\Delta\omega$, which is controlled by Δk, can be made arbitrarily small by oversampling along the frequency bands. Hence, the delay estimator for large microphone distances becomes

$$\hat{\delta}'(k,l) = -\frac{1}{\Delta\omega}\angle(R'(k,l)). \tag{3.7}$$

3.2.2 Clustering the local mixing parameter estimates

The clustering process is performed by a two-dimensional smoothed weighted histogram motivated by the form of the ML estimators deduced in [Yilmaz and Rickard, 2004]. First, instead of using the attenuation estimator $\hat{a}(k,l)$ directly, a symmetric attenuation estimator $\hat{\alpha}(k,l)$ is proposed to avoid problems in case that the microphone signals are swapped:

$$\hat{\alpha}(k,l) = \hat{a}(k,l) - \frac{1}{\hat{a}(k,l)}. \tag{3.8}$$

The local symmetric attenuation and delay estimators $(\hat{\alpha}(k,l), \hat{\delta}(k,l))$ are weighted by a time-frequency dependent weight, usually being $|X_1(k,l)X_2(k,l)|$, to construct a two-dimensional weighted histogram. Different clusters centered on the actual mixing parameters will appear in the histogram. Assuming that these clusters are reasonably separated, the histogram can be further smoothed. The number of clusters corresponds with the number of sources in the mixture, and the centers (peaks) of the clusters are the mixing parameters associated to each source. Figure 3.1 represents the two-dimensional weighted histogram of the local symmetric attenuation and the delay estimators, for a mixture of three speech sources. The figure has been generated according to the description of the DUET algorithm, using a two-dimensional FIR filter to smooth the histogram. The three clusters are clearly identified and their centers represent the mixing parameters associated to each source.

3.2.3 Separation of the sources

The sources are separated via time-frequency masking, generating time-frequency binary masks from the estimated mixing parameters. The mixing parameters for each source are obtained by locating the centers of the clusters in the histogram. The DUET algorithm does not propose any automatic peak identification method, and it performs this task manually assuming to know the number of sources in advance. Once these peaks have been identified, the generation of the binary masks is straightforward: each time-frequency point of the mixture is assigned to the peak (i.e. source) which is closest to the local mixing parameter estimate of that point. The proposed measure of closeness is based on the likelihood function [Yilmaz and Rickard, 2004]. Let $(\hat{\alpha}_n, \hat{\delta}_n)$ to be the estimated center of the n-th cluster, hence $\hat{a}_n$ and $\hat{\delta}_n$ are the n-th source mixing parameter estimates where $\hat{a}_n$ is obtained through

$$\hat{a}_n = \frac{\hat{\alpha}_n + \sqrt{\hat{\alpha}_n^2 + 4}}{2}. \tag{3.9}$$

Each time-frequency point is assigned to a source according to

$$J(k,l) := \underset{\mathbf{n}}{argmin} \frac{|\hat{a}_n e^{-i\omega\hat{\delta}_n} X_1(k,l) - X_2(k,l)|^2}{1 + \hat{a}_n^2}, \tag{3.10}$$

and the time-frequency binary mask for the n-th source is generated according to

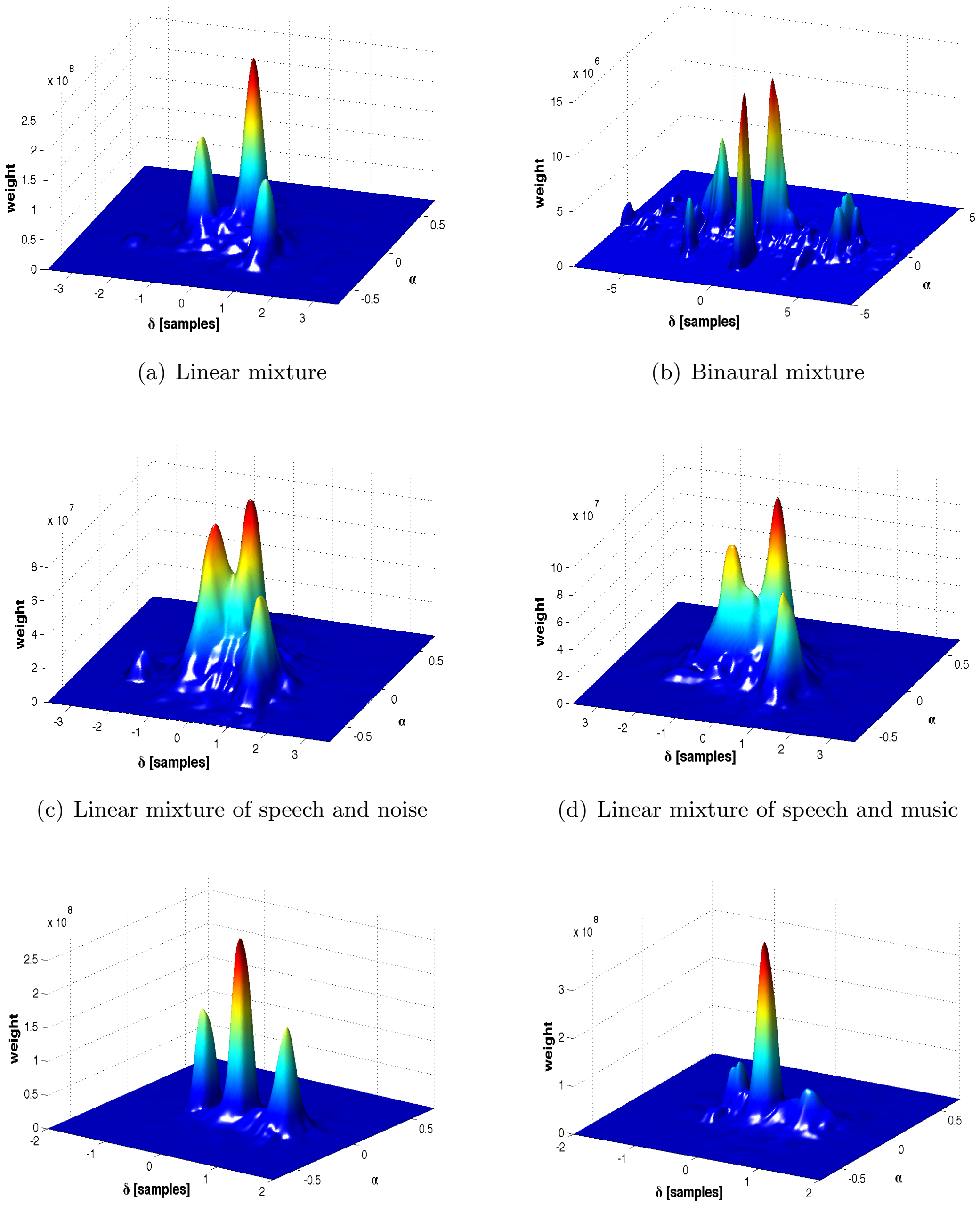
$$M_n(k,l) := \begin{cases} 1 & J(k,l) = n \\ 0 & otherwise. \end{cases} \tag{3.11}$$

Finally, the original sources are demixed combining the binary masks in (3.11) and the ML source estimators deduced in [Yilmaz and Rickard, 2004]:

$$\hat{S}_n(k,l) = M_n(k,l) \left(\frac{X_1(k,l) + \hat{a}_n e^{-i\omega\hat{\delta}_n} X_2(k,l)}{1 + \hat{a}_n^2} \right). \tag{3.12}$$

3.3 The weighted-Gaussian kernel mean shift (WG-MS) algorithm for SSS

The key step of the DUET algorithm resides in the clustering stage, which is performed by a two-dimensional weighted histogram, and allows estimating the mixing parameters by identifying the peaks of the clusters. The performance of the algorithm directly depends on the correct estimation of the mixing parameters, which is conditioned to have well defined and separated clusters. This situation happens in the example shown in figure 3.1, which represents an anechoic mixture of speech sources of the same power. However, there are different situations where the clusters are not so well defined and the mixing parameters are not estimated with good accuracy, for instance, in echoic and binaural mixtures or in mixtures of speech with different types of sources such as music and noise. In these cases, the use of more sophisticated clustering techniques may improve the mixing parameter estimation and thus the performance of the separation algorithm. By way of illustration, figure 3.2 represents the densities obtained by the two-dimensional weighted histogram of DUET in the case of an anechoic linear mixture (a), anechoic binaural mixture (b), speech-noise mixture (c), speech-music mixture (d), echoic mixture with reflection coefficient of 0.1 (e) and echoic mixture with reflection coefficient of 0.5 (f) of three sources. In (a), (b), (e) and (f) the three sources are speech, in (c) the mixture



(a) Linear mixture

(b) Binaural mixture

(c) Linear mixture of speech and noise

(d) Linear mixture of speech and music

(e) Echoic mixture with r=0.1

(f) Echoic mixture with r=0.5

Figure 3.2: Clusters obtained by the two-dimensional weighted histogram of DUET in the case of an anechoic linear mixture (a), anechoic binaural mixture (b), speech-noise mixture (c), speech-music mixture (d), echoic mixture with reflection coefficient of 0.1 (e) and echoic mixture with reflection coefficient of 0.5 (f) of three sources. In (a), (b), (e) and (f) the three sources are speech, in (c) the mixture contains two speech sources and one noise source and in (d) the mixture contains one speech source, one vocal music source and one instrumental music source. The echoic mixtures in (e) and (f) have been generated with the RIRG. The peaks of the histogram have been smoothed using a two-dimensional FIR filter.

contains two speech sources and one noise source, and in (d) the mixture contains one speech source, one vocal music source and one instrumental music source. The binaural mixture in (b) has been generated with the HRTFs of the CIPIC database, and the echoic mixtures in (e) and (f) have been generated with the RIRG. The peaks of the histogram have been smoothed using a two-dimensional FIR filter. The well defined three peaks that appear in the case of the anechoic linear mixture in (a) are not so easily identifiable in the case of the binaural mixture in (b), were many peaks of appreciable amplitude appear. Additionally, in the case of mixing speech with noise and music, which are sources that fulfill in a less degree the W-DO condition, the peaks are closer, and hence their identification becomes more difficult. The effect of reverberation is noticeable comparing subfigures (e) and (f). When reverberation increases, the mixture contains reflected signals that are copies of the originals but with different amplitudes and delays, which originates a considerable number of small peaks that can interfere with the clusters of the original signals if they are relatively close (see subfigure (f)).

In this section, the mean shift method for clustering is modified to be combined with the DUET algorithm to improve its performance. The first and last stages of the DUET algorithm, which are the extraction of the local estimates of the mixing parameters (see section 3.2.1) and the separation of the sources via time-frequency masking (see section 3.2.3), are unaltered. The clustering step for the estimation of the mixing parameters (see section 3.2.2) is replaced by a modified version of the mean shift algorithm that is deduced next.

3.3.1 The mean shift algorithm for clustering and mode seeking

Mean shift is a non-parametric clustering and mode seeking technique of an unknown probability density of a multidimensional feature space without calculating the probability density itself [Cheng, 1995]. The method is based on kernel density estimation (also known as Parzen window technique [Parzen, 1962; Rosenblatt, 1956]) and its main advantage is that it performs independently of the number of modes and shape of the clusters.

The mean shift algorithm was first introduced in [Fukunaga and Hostetler, 1975]. Considering a finite set of Q data points $\mathbf{x} = [\mathbf{c}_1, \cdots \mathbf{c}_i, \cdots, \mathbf{c}_Q]$, on a N-dimensional euclidean space, $\mathbf{c}_i \in \Re^N$, the mean shift vector $\mathbf{m}(\mathbf{x})$ is formulated as

$$\mathbf{m}(\mathbf{x}) = \frac{\sum_{i=1}^{Q} \mathbf{c}_i k(\|\frac{\mathbf{x}-\mathbf{c}_i}{h}\|^2)}{\sum_{i=1}^{Q} k(\|\frac{\mathbf{x}-\mathbf{c}_i}{h}\|^2)} - \mathbf{x}, \tag{3.13}$$

where $k(x)$ is a kernel function (i.e a symmetric but not necessarily positive function that integrates to one) and $h > 0$ is a smoothing parameter called bandwidth. The mean shift vector represents the difference between a weighted mean, using the kernels for weights, and $\mathbf{x}$, which is the center of the kernel (window). This fact yields the mean shift vector is an estimate of the ascendent gradient of data density, which means that the vector always points towards the direction of maximum increase in density [Comaniciu and Meer, 2002]. Hence, the mean shift vector define a path leading to a maximum of the estimated density, and the modes of the density are such maxima. The mean shift algorithm evaluates iteratively the mean shift vector $\mathbf{m}(\mathbf{x})$ shifting the kernel along its direction, $\mathbf{x}_{p+1} \leftarrow \mathbf{m}(\mathbf{x}_p)$, where p is the iteration counter. Iterations halt when the shift is smaller than a threshold, considering that a stationary point has been reached. Since the shift direction is an estimate of the ascendent gradient, the result is that the point approximates to the closest mode (i.e. local maxima) of the distribution. When this procedure is applied to all data points simultaneously, they are grouped together forming clusters around the modes.

The bandwidth parameter h is used by the kernel function to control the radius of candidate points to be neighbors in space of the evaluated point. This parameter plays an important role in the convergence of the algorithm: a large bandwidth considers more points, hence the convergence is slower, and, in the other hand, a small bandwidth considers fewer points and the convergence is faster. However, the use of a small bandwidth increases the risk of prematurely converging to a false local extreme. The bandwidth selection problem is analyzed in [Comaniciu et al., 2001].

3.3.2 Mean shift vector with Gaussian kernel

The original expression of the mean shift vector in (3.13) is first reformulated introducing a Gaussian kernel, which shape adapts appropriately with the modes of the actual distribution. The multivariate Gaussian kernel probability density estimator is given by [Parzen, 1962; Rosenblatt, 1956])

$$\hat{f}(\mathbf{x}) = \frac{1}{Q(2\pi)^{N/2}h^N}\sum_{i=1}^{Q} e^{-\frac{(\mathbf{x}-\mathbf{c}_i)^T(\mathbf{x}-\mathbf{c}_i)}{2h^2}}, \tag{3.14}$$

where $\mathbf{c}_i$ corresponds with the centers of the kernels. In order to determine the mean shift vector, it is necessary to evaluate the gradient of the probability density estimator. The n-th dimension of the gradient of the density function evaluated in $\mathbf{x}$ can be expressed as [Fukunaga and Hostetler, 1975]

$$\frac{\partial \hat{f}(\mathbf{x})}{\partial x_n} = \frac{\hat{f}(\mathbf{x})}{h^2}\left(\frac{\sum_{i=1}^{Q} c_{ni} e^{-\frac{(\mathbf{x}-\mathbf{c}_i)^T(\mathbf{x}-\mathbf{c}_i)}{2h^2}}}{\sum_{i=1}^{Q} e^{-\frac{(\mathbf{x}-\mathbf{c}_i)^T(\mathbf{x}-\mathbf{c}_i)}{2h^2}}} - x_n\right). \tag{3.15}$$

The n-th component of the mean shift vector $\mathbf{m}(\mathbf{x})$ evaluated in the point $\mathbf{x}$, i.e. $m_n(\mathbf{x})$, is given by

$$m_n(\mathbf{x}) = \frac{\sum_{i=1}^{Q} c_{ni} e^{-\frac{(\mathbf{x}-\mathbf{c}_i)^T(\mathbf{x}-\mathbf{c}_i)}{2h^2}}}{\sum_{i=1}^{Q} e^{-\frac{(\mathbf{x}-\mathbf{c}_i)^T(\mathbf{x}-\mathbf{c}_i)}{2h^2}}} - x_n. \tag{3.16}$$

3.3.3 Definition of the input feature space

The DUET algorithm clusterizes the two-dimensional feature space composed of the local symmetric attenuation estimator $\hat{\alpha}(k,l)$ and the local delay estimator $\hat{\delta}(k,l)$, which are defined in expressions (3.8) and (3.5) (or (3.7) for large microphone distances) respectively. They represent the estimates of the level and time differences between microphones for the k-th frequency bin and the l-th time frame. The two time-frequency matrixes ($\hat{\alpha}(k,l)$ and $\hat{\delta}(k,l)$) originates the Q input data samples, where $Q = K \cdot L$ (K is the number of frequency bins and L the number of time frames). The bidimensional input data set $\mathbf{x}$ is given by

$$\mathbf{x} = \begin{bmatrix} \alpha(1,1) & \cdots & \alpha(1,L) & \alpha(2,1) & \cdots & \alpha(K,L) \\ \delta(1,1) & \cdots & \delta(1,L) & \delta(2,1) & \cdots & \delta(K,L) \end{bmatrix}. \tag{3.17}$$

3.3.4 Mean shift vector with weighted-Gaussian kernel

The fact that the energy of each time-frequency point of one mixture of sound sources varies from point to point must be considered when separating the signals in the time-frequency domain:

the information provided by high-energy time-frequency points is more relevant that the one provided by low-energy time-frequency points. However, the expression of the mean shift vector with Gaussian kernel deduced in (3.16) does not allow introducing such weighting. In order to adapt the mean shift algorithm to the SSS problem, a generalized expression of the mean shift vector is inferred, in which the smoothing parameter varies in function of the dimension and the kernel, and a weighted average of the kernels is used instead of the standard averaging.

Let us define a weighted version of the generalized multivariate Gaussian kernel as

$$\hat{f}(\mathbf{x}) = \frac{1}{Q(2\pi)^{N/2}\sum_{i=1}^{Q} p_i}\sum_{i=1}^{Q} p_i|\mathbf{H}_i|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{x}-\mathbf{c}_i)^T\mathbf{H}_i^{-1}(\mathbf{x}-\mathbf{c}_i)}, \tag{3.18}$$

where p_i is a weighting factor that is applied to each kernel and $\mathbf{H}_i = diag([h_{1i}^2, ..., h_{Ni}^2])$, where h_{ni} is the smoothing parameter of the i-th kernel in the n-th dimension. The n-th dimension of the gradient of the density estimator given by (3.18) results in

$$\frac{\partial \hat{f}(\mathbf{x})}{\partial x_n} = \frac{1}{Q(2\pi)^{N/2}\sum_{i=1}^{Q} p_i}\sum_{i=1}^{Q} p_i\frac{c_{ni}-x_n}{h_{ni}^2}|\mathbf{H}_i|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{x}-\mathbf{c}_i)^T\mathbf{H}_i^{-1}(\mathbf{x}-\mathbf{c}_i)}. \tag{3.19}$$

Considering equation (3.18), equation (3.19) can be rewritten obtaining

$$\frac{\partial \hat{f}(\mathbf{x})}{\partial x_n} = \hat{f}(\mathbf{x})\frac{\sum_{i=1}^{Q} p_i\frac{c_{ni}-x_n}{h_{ni}^2}|\mathbf{H}_i|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{x}-\mathbf{c}_i)^T\mathbf{H}_i^{-1}(\mathbf{x}-\mathbf{c}_i)}}{\sum_{i=1}^{Q} p_i|\mathbf{H}_i|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{x}-\mathbf{c}_i)^T\mathbf{H}_i^{-1}(\mathbf{x}-\mathbf{c}_i)}}. \tag{3.20}$$

This expression can only lead to a mean shift like expression if the bandwidth parameter h_{ni} does not depend on the current kernel, and therefore $\mathbf{H}_i = \mathbf{H}$. Using this consideration and rearranging terms, expression (3.20) becomes into

$$\frac{\partial \hat{f}(\mathbf{x})}{\partial x_n} = \frac{\hat{f}(\mathbf{x})}{h_n^2}\left(\frac{\sum_{i=1}^{Q} c_{ni}p_i e^{-\frac{1}{2}(\mathbf{x}-\mathbf{c}_i)^T\mathbf{H}^{-1}(\mathbf{x}-\mathbf{c}_i)}}{\sum_{i=1}^{Q} p_i e^{-\frac{1}{2}(\mathbf{x}-\mathbf{c}_i)^T\mathbf{H}^{-1}(\mathbf{x}-\mathbf{c}_i)}} - \mathbf{x}_n\right) = \frac{\hat{f}(\mathbf{x})}{h_n^2}m_n^W(\mathbf{x}), \tag{3.21}$$

where $m_n^W(\mathbf{x})$ is the new expression for the n-th dimension of the weighted Gaussian kernel mean shift vector. This expression allows implementing a weighted generalized version of the mean shift algorithm, in which each kernel influences the result in a different way, depending on the corresponding value of p_i. Furthermore, the shape of each kernel can be varied along each dimension using different values of h_n. Note that in the special case in which the bandwidth parameter h_n is the same in all dimensions, ($h_n = h$), and the weighting factor p_i is the same for all the kernels, equation (3.21) becomes into the standard mean shift expression shown in equation (3.16).

The proposed algorithm for sound separation replaces the clustering step of the original DUET algorithm by the mean shift iterative algorithm introducing the weighted-Gaussian mean shift vector deduced in expression (3.21), and using different smoothing parameters for each of the two dimensions of the input feature space. The best values for the smoothing parameters were obtained experimentally: $h_1 = 0.5$ for the attenuation dimension and $h_2 = 0.8$ for the delay dimension. According to the purpose of the weighting values p_i, which is to give more importance to high-energy points and less to low-energy points, their values must be directly related to the energy of every time-frequency point. The weighting function proposed in this

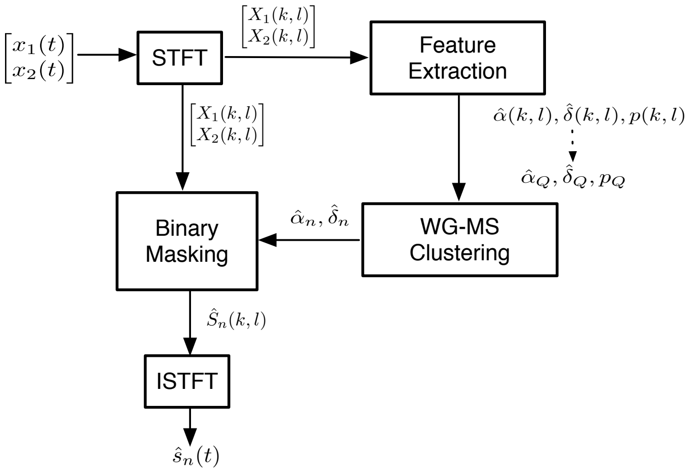

$\begin{bmatrix}x_1(t)\\x_2(t)\end{bmatrix}$
STFT
$\begin{bmatrix}X_1(k,l)\\X_2(k,l)\end{bmatrix}$
Feature Extraction
$\hat{\alpha}(k,l), \hat{\delta}(k,l), p(k,l)$
$\hat{\alpha}_Q, \hat{\delta}_Q, p_Q$
$\begin{bmatrix}X_1(k,l)\\X_2(k,l)\end{bmatrix}$
Binary Masking
$\hat{\alpha}_n, \hat{\delta}_n$
WG-MS Clustering
$\hat{S}_n(k,l)$
ISTFT
$\hat{s}_n(t)$

Figure 3.3: Algorithm overview.

algorithm is the squared geometric average of the absolute values of the STFT of the signals at both microphones, which is calculated from the two mixtures as

$$p(k,l) = |X_1(k,l)||X_2(k,l)|. \tag{3.22}$$

The time-frequency matrix $p(k,l)$ is merged into a vector of length $Q = K \cdot L$, resulting in the weighting vector p_Q. This weighting allows including energy related information in the mean shift clustering algorithm. Experimental results show that weighing the different kernels that compose the mean shift vector improves the performance of the algorithm when estimating the mixing parameters. The centroids of the clusters originated by the mean shift algorithm are obtained by a simple ad-hoc algorithm that searches iteratively for maximum values in the clustered feature space, only considering local maxima that are separated a minimum distance from each other. The centroids provide an estimation of the mixing parameters of each source. The separation step is carried out according to the described in the DUET algorithm.

Figure 3.3 contains an overview of the whole algorithm, which is labeled as WG-MS algorithm. The signals $x_1(t)$ and $x_2(t)$ are the two mixtures containing N sources $s_n(t)$. The local symmetric attenuation and delay estimates as well as the weighting factor are extracted from the STFT decomposition of both mixtures, and the KxL matrixes are merged to compose the input data set according to (3.17). The mean shift algorithm obtains clusters using the previous estimates and weighting values, and it estimates the mixing parameters from the centers of the clusters. The estimates of the mixing parameters are used to generate binary masks that are applied to the mixtures to estimate the original sources, $\hat{S}_n(k,l)$. The ISTFT is applied to convert the estimated signals into the time domain $\hat{s}_n(t)$. In addition, figure 3.4 shows the probability density function estimated from an anechoic mixture of three speech sources using the weighted-Gaussian kernel estimator in (3.18). The three modes corresponding to each source are clearly identified.

3.3.5 Experimental work

Different tests have been carried out for the assessment of the proposed WG-MS algorithm. Since the algorithm performs time-frequency masking, the quality of the separated sources is measured by means of the SIR defined in expression (2.15) and the WDO factor defined in

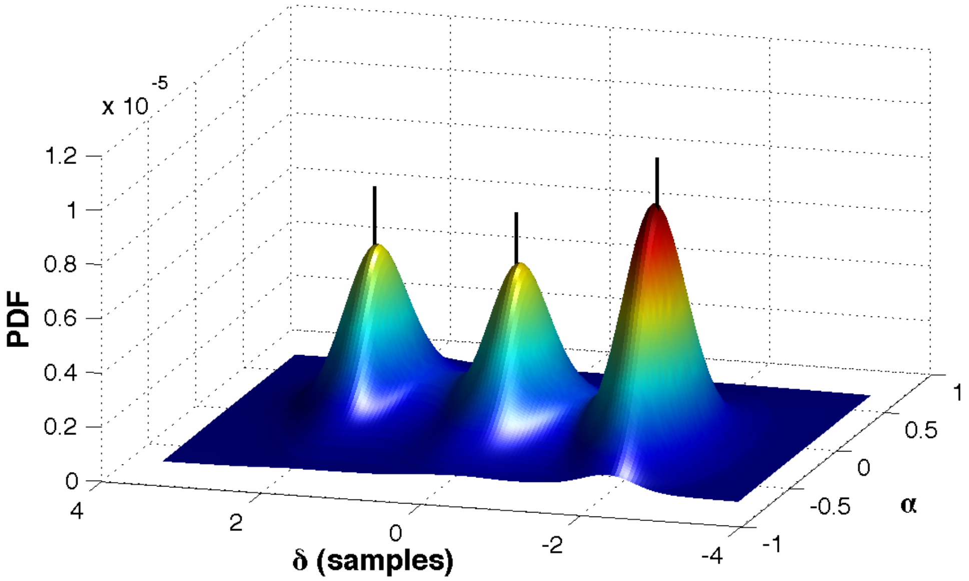


Figure 3.4: Probability density function estimated from an anechoic mixture of three speech sources by the weighted-Gaussian kernel estimator.

expression (2.17). The performance achieved by the proposed algorithm (labeled as WG-MS) is compared with the one obtained by an implementation of the original DUET algorithm described in [Rickard, 2007] (labeled as DUET) and a modification thereof replacing the clustering step by the so-called k-means technique [MacQueen, 1967] (labeled as DUET-KM). The time-frequency decomposition is performed by a STFT with a 256-DFT Hamming window and 50% of overlap. All sound signals are sampled at 16 kHz and normalized and mixed with the same power.

In order to generalize the results, the algorithm is evaluated with several types of mixtures:

- Linear anechoic mixtures of 2, 3 and 4 speech sources. The time and level differences introduced in the mixtures for the three cases are summarized in table 3.1.

- Binaural anechoic mixtures of 2, 3 and 4 speech sources. The binaural signals are generated in the time domain, filtering the original signals with the HRTF's from the CIPIC database. The DOA of each source, which is described by the azimuth and elevation angles, is randomly selected among the available directions in the database for each mixture.

- Linear mixtures of 2 sources mixing speech with noise and speech with music. The position of the sources is the one shown in table 3.1 for $N = 2$.

- Echoic mixtures of 2, 3 and 4 speech sources, generated with the RIRG, varying the reflection coefficient from 0 to 0.5. The microphones are placed in the center of the room and the sources are randomly located around the microphones.

All the speech signals have been randomly selected from the TIMIT database. The noise and music signals have been randomly selected from a database that contains a wide variety of different types of noise and both vocal and instrumental music signals.

Table 3.2 contains a comparison in terms of SIR and table 3.3 in terms of WDO between the three separation methods for linear and binaural mixtures of 2, 3 and 4 speech sources. The SIR

Table 3.1: Level differences (LD) and time differences (TD) between microphones introduced in linear mixtures of 2, 3 and 4 sources.

Sources	LD	TD
$N = 2$	[1.1, 0.9]	[-2, 2]
$N = 3$	[1.1, 1, 0.9]	[-2, 0, 2]
$N = 4$	[1.1, 0.9, 1, 1.1]	[-2, -1, 0, 2]

and WDO values have been averaged over 400 mixtures in the case of 2 sources, 200 mixtures in the case of 3 sources, and 100 mixtures in the case of 4 sources. Additionally, both tables show the average values of the mean SIR and WDO for all sources. Considering linear mixtures, the proposed WG-MS method obtains slightly lower SIR than DUET but slightly higher than DUET-KM, on average. Nevertheless, the WDO value is increased by 1.91% on average when compared to the DUET algorithm, and by 3.78% on average when compared to the DUET-KM algorithm. Results shown are quite similar between the DUET and the WG-MS methods in the case of 2 sources, where the WDO values are very high, which means that both algorithms separate 2 sources very well. In the case of 3 and 4 sources, the WG-MS algorithm obtains better results than the other two, getting an average increase of the WDO of 3.63% and 1.98% respectively, when compared to DUET, and an average increase of the WDO of 2.1% and 6.66% respectively, when compared to DUET-KM.

Concerning binaural mixtures, the SIR and WDO values obtained are notably lower than in the linear case, in general, due to the higher complexity entailed by the use of binaural mixtures. The proposed WG-MS algorithm increases the SIR by 20.9% on average compared to the DUET algorithm, and by 7.29% on average when compared to the DUET-KM algorithm. In addition, the WG-MS obtains an average increase of the WDO of 6.16%, 12.74% and 25.43%, for the 2, 3 and 4 sources cases respectively, when compared to DUET algorithm, and an average increase of the WDO of 3.25%, 10.95% and 16.93% respectively when compared to the DUET-KM algorithm.

Moreover, table 3.4 and table 3.5 contain the SIR and WDO values, respectively, obtained by the three methods in the case of mixtures of speech and noise and of speech and music. In both cases the SIR and WDO values have been averaged over 500 mixtures. Furthermore, the tables show the average values of the two sources. In the case of mixing speech with noise, the proposed WG-MS method improves the SIR by 11.16% on average when compared to DUET, and by 0.37% when compared to DUET-KM, as well as improves the WDO by 7.81% on average average when compared to DUET, and by 7.17% when compared to DUET-KM. In the case of speech and music mixtures, the WG-MS method improves the SIR by 14.21% and 5.45%, on average, when compared to DUET and DUET-KM respectively, and the WDO by 7.43% and 2.97%, on average, respectively.

Finally, table 3.6 shows the SIR and the WDO values obtained in the separation of echoic mixtures of 2, 3 and 4 speech sources with the DUET, DUET-KM and WG-MS algorithms, for reflection coefficient values of 0, 0.1, 0.3 and 0.5. The values shown are the average of the N sources of the mixture. The WG-MS algorithm increases the SIR obtained by the DUET algorithm in 40%, 24.7%, and 329% on average, for the 2, 3 and 4 sources cases respectively. Comparing the SIR obtained by the WG-MS and DUET-KM algorithms, the former increases in 39%, 22.7% and 93% on average, for 2, 3 and 4 sources respectively, the values of the DUET-KM. The SIR increments are so large in this case due to the extremely low SIR values obtained by the DUET and the DUET-KM algorithms in the 4 sources case. Concerning the WDO values, the WG-MS algorithm obtains an average increase of 11.7%, 10.2% and 19.7% for the 2, 3 and

Table 3.2: Averaged SIR (dB) values obtained in the separation of linear and binaural mixtures of 2, 3 and 4 speech sources with the DUET, DUET-KM and WG-MS algorithms.

Sources		Linear mixtures			Binaural mixtures		
		DUET	DUET-KM	WG-MS	DUET	DUET-KM	WG-MS
$N = 2$	S_1	14.77	12.49	14.12	8.87	9.03	10.32
	S_2	12.59	12.18	12.68	9.10	9.19	9.61
	Average	13.68	12.34	13.40	8.98	9.11	9.97
$N = 3$	S_1	10.17	8.20	9.15	5.26	5.39	5.04
	S_2	10.03	9.08	10.36	4.49	4.64	4.93
	S_3	8.52	8.47	8.70	5.79	5.23	6.04
	Average	9.57	8.58	9.40	5.18	5.09	5.34
$N = 4$	S_1	7.50	5.78	6.43	3.18	6.03	6.21
	S_2	7.05	6.96	7.32	4.15	6.02	5.94
	S_3	6.53	7.69	6.62	3.37	4.23	4.95
	S_4	6.15	6.50	5.06	3.94	3.97	4.65
	Average	6.81	6.73	6.36	3.66	5.06	5.44

Table 3.3: Averaged WDO values obtained in the separation of linear and binaural mixtures of 2, 3 and 4 speech sources with the DUET, DUET-KM and WG-MS algorithms.

Sources		Linear mixtures			Binaural mixtures		
		DUET	DUET-KM	WG-MS	DUET	DUET-KM	WG-MS
$N = 2$	S_1	0.939	0.896	0.933	0.776	0.803	0.824
	S_2	0.895	0.894	0.903	0.782	0.799	0.830
	Average	0.917	0.895	0.918	0.779	0.801	0.827
$N = 3$	S_1	0.837	0.797	0.840	0.612	0.609	0.649
	S_2	0.820	0.833	0.845	0.611	0.626	0.697
	S_3	0.737	0.801	0.795	0.636	0.656	0.751
	Average	0.798	0.810	0.827	0.620	0.630	0.699
$N = 4$	S_1	0.740	0.671	0.746	0.509	0.583	0.672
	S_2	0.724	0.671	0.738	0.536	0.583	0.679
	S_3	0.689	0.702	0.705	0.535	0.529	0.633
	S_4	0.675	0.658	0.693	0.514	0.550	0.640
	Average	0.707	0.676	0.721	0.523	0.561	0.656

4 sources cases respectively, in comparison to the DUET algorithm, and an average increase of 2.3%, 6.7% and 14.8% for the 2, 3 and 4 sources cases respectively, in comparison to the DUET-KM algorithm.

3.3.6 Discussion

The analysis of the results obtained in this section demonstrates that the WG-MS algorithm clearly outperforms the original DUET and its modification using k-means, specially when separating more than two sources. This improvement is reduced in the case of linear mixtures of speech sources, where the clusters generated by the DUET histogram are well defined, and the DUET algorithm already obtains almost perfect separation. On the other hand, the improvement is amply noticeable in case of binaural and echoic mixtures and when speech is mixed with non-speech sources. The reason is that the binaural and echoic mixing models as well as the nature of noise and music sources originate local mixing parameter estimates much more

Table 3.4: Averaged SIR (dB) values for the separation of linear mixtures of speech-noise and speech-music with the DUET, DUET-KM and WG-MS algorithms. S_1 is the speech source and S_2 the noise/music source.

Sources	Speech-Noise			Speech-Music		
	DUET	DUET-KM	WG-MS	DUET	DUET-KM	WG-MS
S_1	24.28	21.65	26.99	14.82	12.25	15.81
S_2	14.62	21.51	16.32	8.54	13.05	10.86
Average	19.45	21.58	21.66	11.68	12.65	13.34

Table 3.5: Averaged WDO values for the separation of linear mixtures of speech-noise and speech-music with the DUET, DUET-KM and WG-MS algorithms. S_1 is the speech source and S_2 the noise/music source.

Sources	Speech-Noise			Speech-Music		
	DUET	DUET-KM	WG-MS	DUET	DUET-KM	WG-MS
S_1	0.940	0.839	0.966	0.884	0.837	0.922
S_2	0.724	0.835	0.828	0.731	0.850	0.814
Average	0.832	0.837	0.897	0.808	0.843	0.868

Table 3.6: Averaged SIR (dB) and WDO values obtained in the separation of echoic mixtures of 2, 3 and 4 speech sources with the DUET, DUET-KM and WG-MS algorithms, for reflection coefficient values of 0, 0.1, 0.3 and 0.5. The values shown are the average of the N sources of the mixture.

Sources		SIR (dB)			WDO		
		DUET	DUET-KM	WG-MS	DUET	DUET-KM	WG-MS
$N = 2$	$r = 0.0$	12.19	12.16	15.60	0.787	0.823	0.844
	$r = 0.1$	11.66	11.78	15.54	0.759	0.819	0.840
	$r = 0.3$	9.57	9.75	14.02	0.703	0.781	0.807
	$r = 0.5$	7.56	7.59	11.59	0.659	0.745	0.752
$N = 3$	$r = 0.0$	4.42	4.08	5.92	0.582	0.600	0.674
	$r = 0.1$	3.97	4.23	5.45	0.579	0.591	0.657
	$r = 0.3$	3.55	3.79	4.41	0.564	0.589	0.610
	$r = 0.5$	3.18	3.27	3.29	0.557	0.575	0.575
$N = 4$	$r = 0.0$	1.46	1.68	2.77	0.481	0.499	0.535
	$r = 0.1$	1.32	1.44	2.19	0.461	0.492	0.524
	$r = 0.3$	0.77	1.21	1.96	0.409	0.419	0.511
	$r = 0.5$	0.15	0.56	1.66	0.381	0.398	0.491

scattered in the feature space than in the case of linear mixtures of speech sources, causing that the clusters obtained by the DUET histogram are not as well defined as in the linear case. The replacement of the DUET histogram by a simple clustering technique such as k-means clearly improves the results, which are largely outperformed by the proposed WG-MS algorithm. Nevertheless, despite the improvement obtained by the WG-MS algorithm in echoic mixtures in comparison to DUET and DUET-KM, the algorithm does not show a solid robustness against reverberations, and its performance drops when reverberation increases. The use of microphone arrays with more than 2 microphones may improve the results.

3.4 Generalization of the WG-MS algorithm for any number of sensors

In this section, the WG-MS algorithm for source separation proposed in the previous section, originally using only 2 microphones, is generalized to consider the signals recorded by a microphone array of any number of sensors and geometry. In such a case, different level and time differences can be estimated from each pair of sensors, thus increasing the dimension of the input feature space for clustering. This fact does not require to derive a new expression of the mean shift vector, since the weighted-Gaussian mean shift vector proposed in the previous section is valid for any dimension of the input data. However, in this case the mean shift algorithm provides a different estimation of the source mixing parameters from each pair of microphones. Consequently, the decision rule to decide whether a time-frequency point belongs to a source or to another (expression (3.10)) should be modified to consider the different source mixing parameter estimations and the different mixtures. The demixing step (expression (3.12)) should be also adapted to consider all the mixtures.

3.4.1 Mixing model

Let consider $X_m(k,l)$, $m = 1, \cdots, M$, to be the STFT of M mixtures of sound signals (M being now any number) and $S_n(k,l)$, $n = 1, \cdots, N$ the STFT of the N original signals. Assuming the first element as reference signal, the anechoic mixing model in the time-frequency domain is given by

$$\begin{bmatrix} X_1(k,l) \\ \cdots \\ X_m(k,l) \\ \cdots \\ X_M(k,l) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 & \cdots & 1 \\ & & \cdots & & \\ A_{m1} & \cdots & A_{mn} & \cdots & A_{mN} \\ & & \cdots & & \\ A_{M1} & \cdots & A_{Mn} & \cdots & A_{MN} \end{bmatrix} \begin{bmatrix} S_1(k,l) \\ \cdots \\ S_n(k,l) \\ \cdots \\ S_N(k,l) \end{bmatrix}, \tag{3.23}$$

where $A_{mn} = a_{mn} e^{-i\omega\delta_{mn}}$ is now defined as the relative complex difference of the m-th microphone with respect to the reference one, for the n-th source, a_{mn} and δ_{mn} being the level and time differences respectively (i.e. the mixing parameters). Assuming W-DO sources, the mixing model becomes into

$$\begin{bmatrix} X_1(k,l) \\ \dots \\ X_m(k,l) \\ \dots \\ X_M(k,l) \end{bmatrix} = \begin{bmatrix} 1 \\ \dots \\ A_{mj} \\ \dots \\ A_{Mj} \end{bmatrix} S_j(k,l), \tag{3.24}$$

where $S_j(k,l)$ is the active source at each (k,l) point. Let define now $B_{mn} = a^b_{mn} e^{-i\omega\delta^b_{mn}}$ as the relative complex difference between the m-th and the $(m-1)$-th microphones, for the n-th source, with a^b_{mn} and δ^b_{mn} being the time and level differences respectively. Note that A_{mn} can be expressed as $A_{mn} = \prod_{i=2}^{m} B_{in}$. Hence, the signal received by the m-th microphone of the array can be expressed as a function of signal received by the $(m-1)$-th microphone according to

$$X_m(k,l) = B_{mn} \cdot X_{m-1}(k,l). \tag{3.25}$$

3.4.2 Estimation of the sources from multiple mixtures

Considering an array of M elements, there exist $M-1$ pairs of consecutive microphones. According to expression (3.25), the signal at each microphone can be expressed as the signal at the previous microphone multiplied by the relative complex differences between them (B_{mn}). Consequently, there now exist $M-1$ different estimates of the local mixing parameters, one from each pair of microphones. This fact makes that the input feature space for the mean shift algorithm has a dimension of $2(M-1)$ in this case. The mean shift vector derived in expression (3.21) is valid for any dimension of the input space, but the algorithm provides $M-1$ different estimates of the source mixing parameters, $\hat{A}_{mn} = \hat{a}_{mn} e^{i\omega\hat{\delta}_{mn}}$, with $m = 1, \cdots, M-1$. Consequently, expressions (3.10) and (3.12) are not further valid, and a new expression for the ML source estimator is inferred below.

Let us consider a source s_j. In the time-frequency points where that source is dominant, the mixtures can be modeled as $X_m(k,l) = A_{mj}S_j(k,l) + N_m(k,l)$, $m = 1, \cdots, M$, where the variables $N_m(k,l)$ are independent and identically distributed (i.i.d) white complex Gaussian noise signals with zero mean and variance σ^2. These noise signals represent the contributions of other sources at the time-frequency points where s_j is the dominant source. Assuming that each time-frequency point of the observations $X_m(k,l)$, $m = 1, \cdots, M$, is an independent and Gaussian variable, the likelihood function L_j for the source s_j, obtained from the observations $X_m(k,l)$, is given by

$$L_j = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}\sum_{m=1}^{M} |X_m(k,l) - A_{mj}S_j(k,l)|^2}. \tag{3.26}$$

Maximizing expression (3.26) is equivalent to minimizing

$$L'_j = \sum_{m=1}^{M} |X_m(k,l) - A_{mj}S_j(k,l)|^2. \tag{3.27}$$

To minimize expression (3.27), the system of equations $\partial L'_j / \partial S_i = 0$, with $i = 1, \cdots, N$, must be solved, having

$$\frac{\partial L'_j}{\partial S_i(k,l)} = -\sum_{m=1}^{M} (X_m(k,l) - A_{mi}S_i(k,l))^* A_{mi}. \tag{3.28}$$

Solving this system of equations leads to the ML estimator of $S_j(k,l)$ given by

$$\hat{S}_j^{ML}(k,l) = \frac{\sum_{m=1}^{M} X_m(k,l) A_{mj}^*}{\sum_{m=1}^{M} |A_{mj}|^2}. \tag{3.29}$$

The new measurement of closeness is obtained by replacing $S_j(k,l)$ in (3.27) by its estimation in (3.29). According to this, each time-frequency point is assigned to a source using the closeness function defined by

$$J(k,l) := \underset{\mathbf{n}}{argmin} \sum_{m=1}^{M} \left| X_m(k,l) - \hat{A}_{mn} \frac{\sum_{m=1}^{M} X_m(k,l)\hat{A}_{mn}^*}{\sum_{m=1}^{M} |\hat{A}_{mn}|^2} \right|^2, \tag{3.30}$$

where $\hat{A}_{mn}$ are the $M-1$ estimates of the source mixing parameters obtained by the mean shift algorithm. The time-frequency binary masks $M_n(k,l)$ are obtained with expression (3.11), and

Table 3.7: Averaged SIR (dB) and WDO values obtained in the separation of echoic mixtures of 2, 3 and 4 speech sources with the generalized WG-MS algorithm with 2, 3 and 4 microphones, and for reflection coefficient values of 0, 0.1, 0.3 and 0.5. The values shown are the average of the N sources of the mixture.

Sources		SIR (dB)			WDO		
		M=2	M=3	M=4	M=2	M=3	M=4
$N=2$	$r=0.0$	15.60	20.51	21.63	0.844	0.856	0.860
	$r=0.1$	15.54	20.53	21.82	0.840	0.858	0.849
	$r=0.3$	14.02	19.41	17.67	0.807	0.841	0.837
	$r=0.5$	11.59	15.90	15.55	0.752	0.789	0.786
$N=3$	$r=0.0$	5.92	10.51	9.51	0.674	0.699	0.697
	$r=0.1$	5.45	10.33	9.48	0.657	0.698	0.686
	$r=0.3$	4.41	9.18	8.32	0.610	0.687	0.671
	$r=0.5$	3.29	5.41	4.97	0.575	0.640	0.634
$N=4$	$r=0.0$	2.77	3.50	3.71	0.535	0.567	0.575
	$r=0.1$	2.19	3.79	3.70	0.524	0.569	0.571
	$r=0.3$	1.96	2.84	3.14	0.511	0.551	0.564
	$r=0.5$	1.66	2.52	2.78	0.491	0.534	0.541

the original sources are demixed by multiplying the binary masks with the ML source estimator in (3.29), resulting in

$$\hat{S}_n(k,l) = M_n(k,l)\left(\frac{\sum_{m=1}^{M} X_m(k,l)\hat{A}_{mn}^*}{\sum_{m=1}^{M} |\hat{A}_{mn}|^2}\right). \tag{3.31}$$

3.4.3 Experimental work

The generalized version of the WG-MS algorithm has been tested with the same echoic mixtures evaluated by the original WG-MS algorithm, increasing the number of microphones to 3 and 4. That means that the algorithm has been evaluated when separating echoic mixtures of 2, 3 and 4 speech sources, generated with the RIRG, varying the reflection coefficient from 0 to 0.5. The microphones are placed in the center of the room and the are sources randomly located around the microphones.

Table 3.7 contains the averaged SIR (dB) and WDO values obtained in the separation of echoic mixtures of 2, 3 and 4 speech sources with the generalized WG-MS algorithm with 2, 3 and 4 microphones, and for reflection coefficient values of 0, 0.1, 0.3 and 0.5. The values shown are the average of the N sources of the mixture. The values corresponding to 2 microphones are the same that the ones contained in table 3.6 for the WG-MS algorithm. The increment in the number of microphones from 2 to 3 always leads to higher SIR and WDO values, the difference being larger when the reflection coefficient is higher. Thus, the WDO value is incremented by 3.7% in the case of $r = 0$ and by 8.3 % in the case of $r = 0.5$, on average, for the different number of sources. An additional increment of the number of microphones from 3 to 4 does not always bring better results in terms of SIR and WDO. The WDO values in the case of separating 2 and 3 sources are quite similar, and the use of 4 microphones only increments the WDO values obtained in the case of using 3 microphones when separating 4 sources.

3.4.4 Discussion

This section describes a generalization of the proposed WG-MS algorithm for speech separation, using a microphone array of any number of elements. In this case, the dimension of the input feature space is increased to $2(M-1)$, but the WG-MS clustering proposed in the previous section is valid for any dimension. However, the ML source estimator derived in case of two mixtures is not further valid, and a new one has been inferred.

The performance of the method has been evaluated in echoic mixtures generated with the RIRG, varying the reflection coefficient from 0 to 0.5. The microphones have been placed in a fixed position in the center of the room, and the sources have been placed around the microphones at random positions. Results show that the increment in the number of elements of the array usually entails an increment in the performance of the separation, in terms of WDO. Finally, although the results support the robustness of the proposed method against reverberations, more experiments should be carried out in this direction.

3.5 Source enumeration of speech sources based on information theoretic criteria

Speech source enumeration is a problem that remains largely open and unsolved. Determining the number of speech sources is a critical first step for SSS algorithms, such as the DUET algorithm and the proposed WG-MS algorithm, which assumes to know the number of sources in advance. In this section, a novel method for speech source enumeration based on information theoretic criteria is proposed.

The proposed algorithm considers a microphone array composed of two sensors and assumes a different DOA for the speech sources, which implies that each source causes a different relative delay between the two microphones. Under the assumption of approximate W-DO sources, a local delay estimator is computed. The PDF of this estimator shows peaks associated to each of the sources. Consequently, the problem of speech source enumeration is equivalent to count the number of peaks in this PDF. The proposed method fits an autoregressive (AR) model to the local delay estimator in order to estimate its PDF, and then applies the Rissanen's minimum description length (MDL) principle [Rissanen, 1978] to minimize the length of the code required to describe the data, finding the model that best fits the data with the lowest possible order. Unlike other algorithms that apply the MDL principle to the observed data, the proposed method applies it to the prediction error of the AR model, as proposed in [Krim and Cozzens, 1994].

3.5.1 Consistent local delay estimator

Let us consider the mixing model described in expression (3.1) and the local delay estimator given in (3.5). The fact that the delay estimator changes with the position of the sources together with the assumption of approximate W-DO sources, causes that the PDF of the delay estimator presents peaks associated to each of the sources. Figure 3.5 shows a histogram-based PDF estimate of the PDF of the local delay estimator $\hat{\delta}(k,l)$ in the case of a linear mixture of three speech sources causing a delay of [-1, 0, 1] samples between microphones. The correlation between the number of modes of the PDF and the number of sources is clear, hence the problem of speech source enumeration is equivalent to count the number of modes in that PDF.

It was mentioned before that the local delay estimator given by expression (3.5) can be ambiguous due to the periodicity of the phase, and it is reliable only if $|\omega\delta_j| < \pi$, condition that is guaranteed when $\omega_{max}\delta_{jmax} < \pi$, which means that for relative delays between microphones

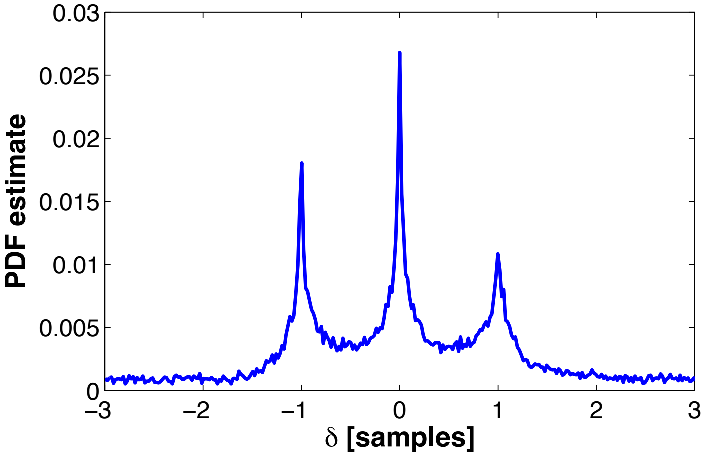


Figure 3.5: Histogram-based PDF estimate of $\hat{\delta}(k, l)$ in the case of a linear mixture of 3 speech sources mixed with microphone time differences of [-1, 0, 1] samples.

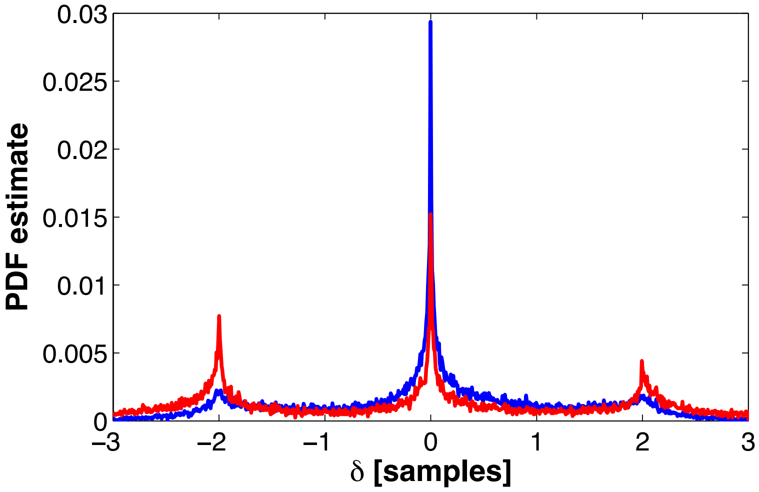


Figure 3.6: Histogram-based PDF estimate of $\hat{\delta}(k, l)$ with unresolved phase ambiguity (blue line) and with the phase unwrapped (red line), in the case of a linear mixture of 3 speech sources mixed with microphone time differences of [-2, 0, 2] samples.

larger than one sample, the estimated phase will be inaccurate. In order to overcome this limitation, the phase $\omega\delta_j$ is unwrapped by changing absolute phase jumps greater or equal to π to their 2π complement. This operation is performed along the frequencies of each frame. Once the angular frequency term is removed from the unwrapped phase, the local delay estimator $\hat{\delta}(k, l)$ reflects the true values even if they are larger than one sample. An example of this situation is found in figure 3.6, where the source delays in a linear mixture of three speech sources are set to two samples. In the PDF estimate represented with a blue line, the phase ambiguity has not been resolved, and the peaks corresponding to sources with delays of two samples are barely perceived, whereas the peak located at 0 has notably larger amplitude due to phase wrapping. However, in the PDF estimate represented with a red line, where the phase has been unwrapped, the three peaks are clearly identifiable and have similar amplitude.

Finally, the local delay estimations that have been made from time-frequency points with low

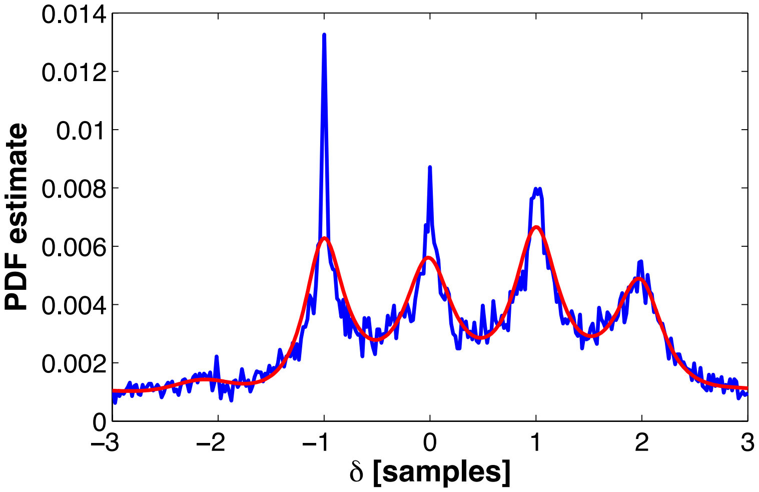


Figure 3.7: Histogram-based PDF estimate (blue line) and 4^{th}-order AR-model estimate (red line) of $\hat{\delta}(k,l)$ in the case of an anechoic mixture of 4 speech sources mixed with microphone time differences of [-1, 0, 1, 2] samples.

energy are not consistent, and they are consequently removed. The energy of a time-frequency point is measured here by the geometric mean of the energy of the signals in both microphones

$$E(k,l) = 10\log_{10}(|X_1(k,l)||X_2(k,l)|), \tag{3.32}$$

rejecting those delay estimations made from points where $E(k,l) < Th$, where Th is an ad-hoc threshold. The number of delays rejected is R.

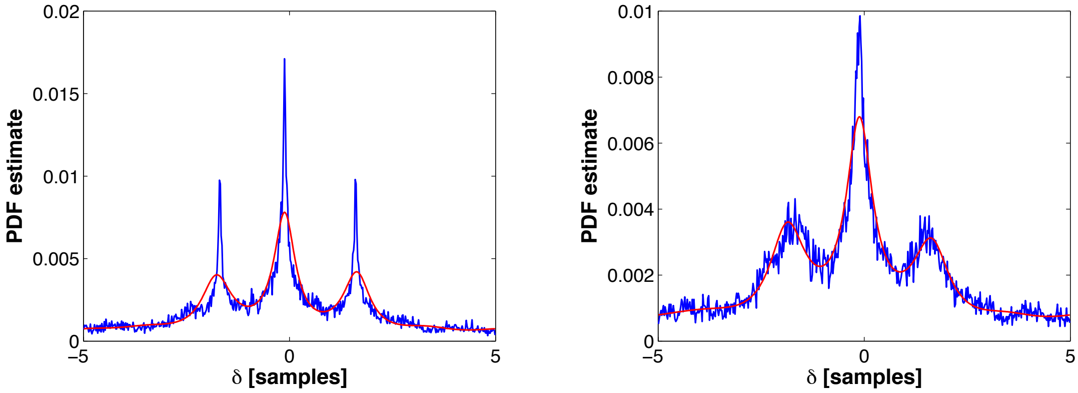
3.5.2 Parametric model-based PDF estimation

Let us define the random variable $\delta = [\delta_1, \cdots, \delta_Q]$ containing the local delay estimations corresponding to every valid time-frequency point (i.e. the matrix $\hat{\delta}(k,l)$ is merged into a vector), where $Q = K \cdot L - R$ is the number of data samples. The information regarding the number of sources is contained in the PDF of the random variable δ, which is denoted by $f(\delta)$. Due to the fact that the PDF of a random variable has similar properties to a power spectral density (PSD), the PDF can be estimated using parametric spectral density estimation techniques [Kay, 1998]. Using an autoregressive (AR) model, the PDF is estimated fitting the model

$$\hat{f}(\delta) = \frac{G}{\left|1 - \sum_{p=1}^{P} \gamma_p e^{-i2\pi p\delta}\right|^2} \tag{3.33}$$

to the available observations, where γ_p are the AR model coefficients, P is the order of the model, and G is a constant related to the minimum square prediction error. In the case of PSD estimation, the Yule-Walker equations allow obtaining the AR coefficients from an estimation of the autocorrelation of the original data. However, the PSD and the autocorrelation function form a Fourier transform pair. This fact helps to solve the problem at hand, in which, under the assumption that $f(\delta) = 0$ if $|\delta| > \pi$, the role of the autocorrelation function is played by the sequence $\phi_\delta[m]$, defined as

$$\phi_\delta[m] = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\delta m} f(\delta)d\delta = \frac{1}{2\pi}E\{e^{i\delta m}\}. \tag{3.34}$$



(a) r=0

(b) r=0.4

Figure 3.8: Histogram-based PDF estimate (blue line) and 3^{th}-order AR-model estimate (red line) of $\hat{\delta}(k,l)$ in the case of an echoic mixture of 3 speech sources. The left image corresponds with a mixture without reverberation and the right image corresponds with a mixture with a reflection coefficient of 0.4.

Using the sample mean as an estimator of the probabilistic expectation, the sequence $\phi_\delta[m]$ can be estimated as

$$\hat{\phi}_\delta[m] = \frac{1}{2\pi Q}\sum_{q=1}^{Q} e^{i\delta_q m}, \quad m = 0, 1, ...P. \tag{3.35}$$

When this sequence is known, the AR model coefficients are estimated using the Yule-Walker equations. If the previous assumption ($f(\delta) = 0$ if $|\delta| > \pi$) is not fulfilled, the delays should be normalized before applying this method.

Figure 3.7 shows an example of this method estimating the PDF of the local delay estimators in a linear mixture of 4 speech sources, introducing source delays of [-1, 0, 1, 2] samples. The blue line represents the histogram-based PDF estimate and the red line represents the AR model estimation with $P = 4$, which clearly fits the four peaks contained in the PDF. The fact that the AR model provides an smoothed estimation of the PDF is very useful to apply the method in echoic mixtures. By way of illustration, figure 3.8 contains the histogram-based PDF estimate (blue line) and 3^{th}-order AR model estimate (red line) of $\hat{\delta}(k,l)$ in the case of an echoic mixture of 3 speech sources obtained with the RIRG. The left figure corresponds with a mixture without reverberation and the right figure corresponds with a mixture with a reflection coefficient of 0.4. In case of reverberation, the peaks of the histogram-based PDF are clearly worse identifiable in comparison to the case without reverberation. Nevertheless, the AR model performs a perfect estimation in both cases with the same model order.

3.5.3 Application of the MDL principle for enumeration

The prediction error related to fitting the AR model to the data is a monotonically decreasing function of the order model P. However, from a certain value of P, the AR model fits with enough accuracy the true PDF. In this context, the MDL principle suggests choosing the model that provides the shortest description of the data, and it considers that the order of that model is an estimation of the number of sources. Additionally, encoding the prediction error is equivalent to encode the best representation of the data [Krim and Cozzens, 1994].

In the AR model described in (3.33), the coefficients can be calculated using the Yule-Walker equations, obtaining the estimation

$$\hat{\phi}_\delta[m] = \sum_{p=1}^{P} \gamma_p \hat{\phi}_\delta[m-p] + \sigma_\epsilon^2 \delta_K[m], \quad 0 \leq m \leq P, \tag{3.36}$$

where $\hat{\phi}_\delta[m]$ represents the values estimated with expression (3.35), σ_ϵ^2 is the variance of the model input random process ($\epsilon(n)$), and δ_K is the Kronecker delta. For $m = 0$, the value of σ_ϵ^2 can be easily obtained. The model input random process is assumed to be a Gaussian random process with zero mean and variance σ_ϵ^2:

$$f(\epsilon) = \frac{1}{\sigma_\epsilon \sqrt{2\pi}} e^{-\frac{\epsilon^2}{2\sigma_\epsilon^2}}. \tag{3.37}$$

Rissanen in [Rissanen, 1978] proposed the minimal code length required to describe the observed data and the free parameters (model parameters) as a general criterion for model order determination. The number of bits needed to encode the data judges the model selected. He chooses the estimator that achieves a short total code length, and the MDL criterion is formulated as

$$MDL(p) = -log(f(\mathbf{y}|\mathbf{a}) + \frac{p}{2} log(Q), \tag{3.38}$$

where $\mathbf{y}$ is the vector composed of the data with the autocorrelation sequence (related to the delays). Considering the fact that $f(\mathbf{y}|\mathbf{a}) = f(\epsilon(n))$, introducing (3.37) into (3.38), and after some manipulations, the MDL criterion is expressed as

$$MDL(p) = Qlog(\pi) + Qlog(\frac{1}{Q}||\epsilon_p||^2) + \frac{p}{2} log(Q), \tag{3.39}$$

where ϵ_p is the prediction error associate to the AR model of order p. Finally, the number of speech sources is given by

$$\hat{N} = \underset{p \in \{1,\dots,P\}}{min} MDL(p). \tag{3.40}$$

In summary, the steps of the algorithm for source enumeration are the following:

1. Compute the STFT of the signals recorded at the two microphones, $X_1(k,l)$ and $X_2(k,l)$.

2. Unwrap the phase of the ratio of the two mixtures and calculate the local delay estimator for each time-frequency point.

3. Remove the local delay estimations that have been made from time-frequency points with low energy.

4. Construct the sequence $\hat{\phi}_\delta[m]$ according to expression (3.35).

5. Estimate the AR model parameters γ_p using the Yule-Walker equations and evaluate the prediction error ϵ_p, varying the order of the model P.

6. Evaluate the MDL for each model order, using expression (3.39). The minimum value of the MDL is the estimation of the number of sources.

Finally, figure 3.9 contains a graphical description of the algorithm.

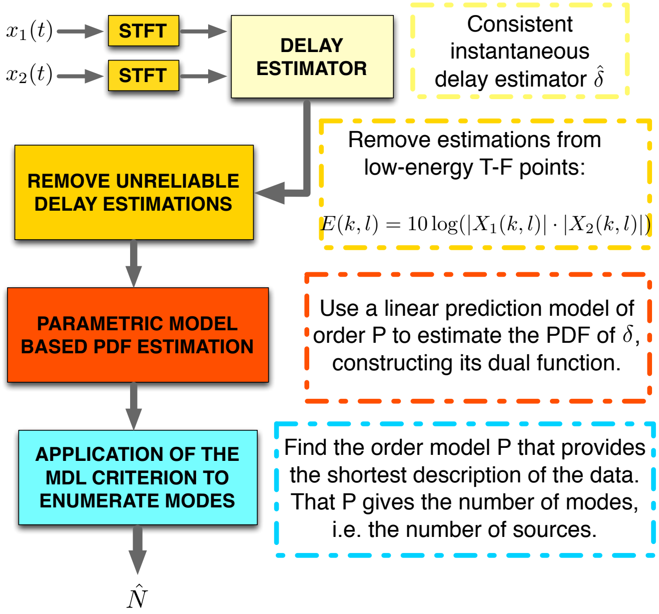

$x_1(t)$
$x_2(t)$
STFT
STFT
DELAY ESTIMATOR
Consistent instantaneous delay estimator $\hat{\delta}$
REMOVE UNRELIABLE DELAY ESTIMATIONS
Remove estimations from low-energy T-F points:
$E(k,l) = 10\log(|X_1(k,l)| \cdot |X_2(k,l)|)$
PARAMETRIC MODEL BASED PDF ESTIMATION
Use a linear prediction model of order P to estimate the PDF of δ, constructing its dual function.
APPLICATION OF THE MDL CRITERION TO ENUMERATE MODES
Find the order model P that provides the shortest description of the data. That P gives the number of modes, i.e. the number of sources.
$\hat{N}$

Figure 3.9: Source enumeration algorithm overview.

Table 3.8: Source delays for linear mixtures

N	Delays (samples)
2	[-1, 0]
3	[-1, 0, 1]
4	[-1, 0, 1, -2]
5	[-1, 0, 1, -2, 2]

3.5.4 Experiments

The proposed source enumeration algorithm is evaluated for anechoic mixtures of 2, 3, 4 and 5 speech sources. A total of 50 different mixtures are generated for each number of sources, selecting the speech sources randomly from the TIMIT database. The time-frequency decomposition is performed by a STFT with frames of 256 samples and 50% of overlap, using a Hamming window. The sampling rate is 16 kHz. All signals have been normalized and mixed with the same power, and the threshold value to remove low-energy time-frequency points is set to 0 dB. The source delays introduced in the mixtures are summarized in table 3.8.

Figure 3.10 plots the enumeration accuracy rate averaged over 50 mixtures, for each number of sources evaluated. The enumeration in the cases of 2 and 3 sources is almost perfectly performed, but when the number of sources increases, the error in the estimation also increases, as it was expected. Nevertheless, the accuracy rate in the 5 sources case is still 80%, which is a noticeable good value for speech enumeration.

3.5.5 Discussion

This section describes a novel method to solve the problem of speech source enumeration, based on AR modeling of the PDF of the local delay estimations. It is demonstrated that the number of sources is equivalent to the order of the AR model when it is selected with the MDL criterion. The proposed algorithm obtains excellent results in the enumeration of sources in anechoic

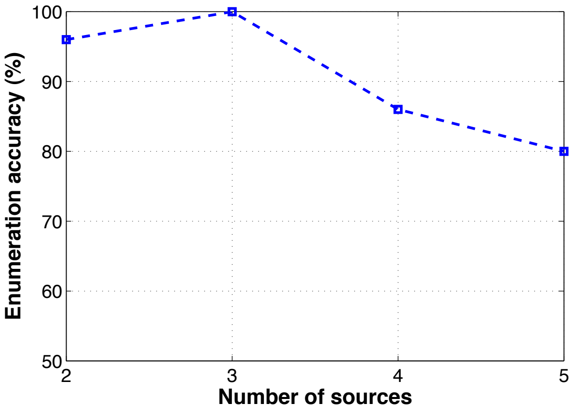


Figure 3.10: Averaged accuracy rate (%) obtained in the estimation of the number of sources in anechoic mixtures of 2, 3, 4 and 5 sources.

speech mixtures. Additionally, it has been shown that the proposed AR model also estimates with accuracy the PDF of the source relative delays in the case of echoic mixtures. In this type of mixtures, due to reverberation, the width of the peaks that appear in the PDF is larger, hence the peaks are not so well defined. Fortunately, the smooth estimation of the PDF performed by the AR model overcomes this problem. Further investigations should be carried out in this direction.

Finally, is is worth highlighting that this technique is specially useful in source separation applications, where the number of sources is an input parameter for the algorithms. The presented results are promising, but further research must be done to check the robustness in noisy environments, the dependence with the relative positions of the speech sources and the energy of the sources.

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 of scenarios including linear and binaural anechoic mixtures, echoic mixtures, and mixtures of speech with other types of sources such as noise and music. It has been demonstrated the need for more advanced clustering techniques in such situations.

- A novel source separation algorithm that combines the mean shift clustering technique with the basis of DUET has been proposed. The clustering step in DUET, which is based on a weighted histogram, is replaced by a generalized version of the mean shift algorithm. A weighted-Gaussian kernel mean shift vector has been inferred and included in an iterative process to clusterize the bidimensional feature input space composed of the level and time differences between the two microphones. The proposed WG-MS algorithm has been tested in different scenarios: linear and binaural anechoic speech mixtures, echoic speech mixtures with different reverberation coefficients, and anechoic mixtures of speech with noise and speech with music. The WG-MS algorithm has been compared to the original DUET algorithm and a modification thereof which introduces the k-means algorithm in the clustering step. The analysis of the results obtained demonstrates that the WG-MS algorithm clearly outperforms the original DUET and its modification using k-means.

- The WG-MS algorithm, which was originally proposed for two microphones, has been extended to the case of any number of microphones and array geometry. The mean shift algorithm allows clustering a feature space of any dimension. A new ML source estimator that considers any number of microphones has been inferred. Several experiments varying the number of microphones support the suitability of the method, which shows a special robustness in the case of echoic mixtures.

- A novel speech source enumeration algorithm has been proposed. The algorithm is based on information theoretic criteria and the estimation of the source delays between the signals received by two microphones. The algorithm has obtained very good results and it has shown good robustness in the enumeration of anechoic mixtures up to 5 speech sources. Additionally, the potential of the algorithm to enumerate sources in echoic mixtures has been demonstrated.

The contributions obtained in this chapter have originated the publications [Ayllón et al., 2010], [Ayllón et al., 2011a], [Ayllón et al., 2012a], [Ayllón et al., 2012b] and [Ayllón et al., 2013d].

Chapter 4

Single-channel speech enhancement for monaural hearing aids

4.1 Introduction

This chapter tackles the problem of single-channel speech enhancement and its application to monaural hearing aids, considering that the main goal is to improve the intelligibility of speech in noise. Single-channel speech enhancement can be performed from two different approaches: noise reduction and source separation. A comprehensive review of single-channel speech enhancement algorithms has been carried out in sections 1.3.1.1 and 1.3.2.1. Nevertheless, single-channel source separation algorithms inspired in CASA are either too complex or the performance is too limited to be applicable in hearing aids. These algorithms typically involve complex operations for feature extraction, segregation and grouping, which makes difficult a real-time implementation. Nevertheless, the time-frequency masking approach inspired in CASA can be useful in hearing aids, as long as the mask computation is relatively simple.

The main problem associated to single-channel noise reduction algorithms resides in the fact that they are commonly designed to improve the speech quality rather than to improve the speech intelligibility, which is the final purpose for hearing impaired people. The correct approach is to design the algorithms to optimize an objective measure correlated with speech intelligibility instead of correlated with speech quality. It has been demonstrated in [Ma et al., 2009] that the fwSNRseg and the PESQ are two good objective measures highly correlated with speech intelligibility. The other alternative, originated in the field of CASA, is time-frequency masking. This approach is based on the application of a gain function or mask to the time-frequency representation of a corrupted speech signal, removing portions of the signal that are considered noise and allowing the remaining signal to pass through unaltered. The mask may be either a binary mask (i.e. takes values of 0 and 1) or a soft mask (i.e. takes continuous values between 0 and 1). The ideal binary mask (IBM) commonly defined in CASA systems [Hu and Wang, 2001, 2004] is the one that takes values of zero or one by comparing the local SNR in each time-frequency point against a threshold, which is usually set to 0 dB. It is demonstrated in [Loizou and Kim, 2011] that the IBM maximizes the articulation index (AI), a metric highly correlated with speech intelligibility [Kryter, 1962]. Consequently, the use of the IBM for noise reduction also entails an improvement in speech intelligibility. Unfortunately, the computation of the IBM needs to have access to the clean speech and noise signals, information that is not available in practice.

The design of a speech enhancement algorithm based on time-frequency masking consists in estimating the IBM from the corrupted observations of the signal. The CASA approach performs this estimation using features inspired in the human auditory system (pitch, amplitude and frequency modulation, onset/offset, etc.). However, it is conceptually and computationally simpler to use machine learning techniques to identify each time-frequency point as speech-dominated or noise-dominated.

In this chapter, a time-frequency masking algorithm is proposed for single-channel speech enhancement in monaural hearing aids. The algorithm is designed bearing in mind the reduced computational resources available in state-of-the-art commercial hearing aids. The system uses a soft mask and is designed to maximize the output PESQ, which is an objective measure correlated with intelligibility.

4.2 Evaluation of different time-frequency masks in terms of PESQ

Several works have evaluated the intelligibility gain associated to the IBM inspired in CASA. However, the use of a different mask may lead to a higher intelligibility improvement. In this section, three time-frequency masks (one binary mask and other two soft masks) are proposed and compared to the IBM. The comparison is made in terms of PESQ, which maximization is the final objective of the algorithm proposed in this chapter. Additionally, the influence of varying the frequency resolution of the time-frequency decomposition is examined. The best mask will be used as starting point in the speech enhancement algorithm.

4.2.1 Definition of the time-frequency masks

Let us consider $X(k,l) = S(k,l)+N(k,l)$ to be the STFT of a speech signal $S(k,l)$ contaminated by other noise and interfering signals denoted as $N(k,l)$. The magnitude of the complex values $S(k,l)$ and $N(k,l)$ are used to determine whether a time-frequency point is dominated by speech or noise, thus generating a time-frequency mask $M(k,l)$. The clean speech signal is estimated by applying the mask to the mixture, i.e. $\hat{S}(k,l) = M(k,l) \cdot X(k,l)$. The criterion to assign a time-frequency point to speech or noise may vary, originating different types of masks. In this section, four time-frequency masks are evaluated: the CASA IBM, a binary mask that maximizes the $fwSNRseg$, another soft mask that also maximizes the $fwSNRseg$, and a Wiener soft mask.

4.2.1.1 CASA IBM

The IBM proposed in CASA systems has been already discussed, and in the case of a threshold of 0 dB, it is defined as

$$M(k,l) := \begin{cases} 1, & \text{if } |S(k,l)| > |N(k,l)| \\ 0, & \text{otherwise} \end{cases} . \tag{4.1}$$

This binary mask assigns all the energy of a determined time-frequency point to the signal with higher energy (speech or noise).

4.2.1.2 Maximum fwSNRseg binary mask

The time-frequency mask proposed here is a binary mask that tries to maximize the $fwSNRseg$ value. The local signal-to-residual spectrum ratio (SNR_{ESI}) is defined in [Loizou and Kim, 2011] as

$$SNR_{ESI}(k,l) = \frac{|S(k,l)|^2}{(|S(k,l)| - |\hat{S}(k,l)|)^2}, \tag{4.2}$$

and, for each frequency band, it can be calculated by

$$SNR_{ESI}(k) = \frac{1}{L}\sum_{l=0}^{L-1} \frac{|S(k,l)|^2}{(|S(k,l)| - |\hat{S}(k,l)|)^2}. \tag{4.3}$$

According to this, the $fwSNRseg$ defined in expression (2.21) can be expressed as

$$fwSNRseg = \frac{\sum_{k=1}^{K} W(k) SNR_{ESI}(k)}{\sum_{k=1}^{K} W(k)}. \tag{4.4}$$

As $W(k) \in \mathbb{R}^+$, the maximization of the SNR_{ESI} implies the maximization of the $fwSNRseg$. Consequently, the SNR_{ESI} also correlates with speech intelligibility. The binary mask proposed here maximizes the local SNR_{ESI}. Considering that the estimated amplitude spectrum is given by $|\hat{S}(k,l)| = M(k,l) \cdot |S(k,l) + N(k,l)|$, the local SNR_{ESI} (expression (4.2)) is obtained by

$$SNR_{ESI}(k,l) = \begin{cases} \frac{|S(k,l)|^2}{(|S(k,l)| - |S(k,l)+N(k,l)|)^2} & \text{if } M(k,l) = 1, \\ 1 & \text{if } M(k,l) = 0. \end{cases} \tag{4.5}$$

In order to maximize somehow the SNR_{ESI} (and consequently the $fwSNRseg$), its value needs to be higher than 1 ($SNR_{ESI} > 0$ dB) in the case of $M(k,l) = 1$. Otherwise, if the value SNR_{ESI} is lower than 1, the mask should be $M(k,l) = 0$. This is equivalent to the next inequality:

$$\frac{|S(k,l)|^2}{(|S(k,l)| - |S(k,l)+N(k,l)|)^2} \underset{M(k,l)=0}{\overset{M(k,l)=1}{\gtrless}} 1. \tag{4.6}$$

Solving (4.6) in the case of $M(k,l) = 1$ leads the next relationship:

$$|S(k,l)| > \frac{|S(k,l) + N(k,l)|}{2}. \tag{4.7}$$

Finally, the binary mask that maximizes the SNR_{ESI} is defined as

$$M(k,l) := \begin{cases} 1, & \text{if } |S(k,l)| > \frac{|S(k,l)+N(k,l)|}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{4.8}$$

4.2.1.3 Maximum fwSNRseg soft mask

Let us consider now a soft time-frequency mask that can take any value between 0 and 1. In such a case, the SNR_{ESI} is given by

$$SNR_{ESI}(k,l) = \frac{|S(k,l)|^2}{(|S(k,l)| - M(k,l) \cdot |X(k,l)|)^2}. \tag{4.9}$$

In order to maximize the previous expression, the denominator should be equal to 0, that is, $|S(k,l)| - M(k,l) \cdot |X(k,l)| = 0$, which yields the soft mask defined by

$$M(k,l) := \frac{|S(k,l)|}{|S(k,l) + N(k,l)|}. \tag{4.10}$$

To avoid amplification distortions, the values are bounded between 0 and 1.

4.2.1.4 Wiener soft mask

The transfer function of the non-causal Wiener filter is given by [Wiener, 1949]

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_n(\omega)}, \tag{4.11}$$

where $P_s(\omega)$ is the PSD of the clean speech signal and $P_n(\omega)$ the PSD of the noise signal. According to this, the next time-frequency soft mask is defined

$$M(k,l) := \frac{|S(k,l)|^2}{|S(k,l)|^2 + |N(k,l)|^2}, \tag{4.12}$$

which takes continuous values between 0 and 1.

4.2.2 Evaluation of the time-frequency masks

The four time-frequency masks defined in the previous section are now evaluated in terms of PESQ. The masks are generated assuming that the original speech and noise sources are known. Additionally, the masks are evaluated using different numbers of frequency bands for the computation of the STFT.

4.2.2.1 Experimental setup

Speech and noise mixtures of different SNRs are generated, using the speech and noise signals contained in the NOIZEUS database. The procedure to generate the mixtures is the next:

1. The 30 clean speech signals are normalized and linked together, one after the other, obtaining a speech segment of 80 seconds length.

2. The 30 noise signals of each type of noise are also normalized and linked together, one after the other, obtaining 8 different segments of 80 seconds length, one for each noise.

3. The clean speech segment is repeated 8 times, generating a speech segment of 640 seconds length.

4. The 8 different noise signals are linked together, generating a noise segment of 640 seconds length.

5. The clean speech and noise segments are normalized and then mixed with the desired SNR.

Figure 4.1 shows graphically the database setup procedure.

The four proposed masks are generated from the original clean speech and noise signals and applied to the mixture to estimate the clean signal. The PESQ of the estimated speech signal is evaluated, using the original clean speech signal as reference. The frequency resolution of the time-frequency decomposition is modified varying the number of frequency bands of the STFT, with values of 32, 64, 128 and 256 (the sampling rate is 8 kHz, so higher values correspond to very long analysis windows). A Hamming window with 50% of overlap is used.

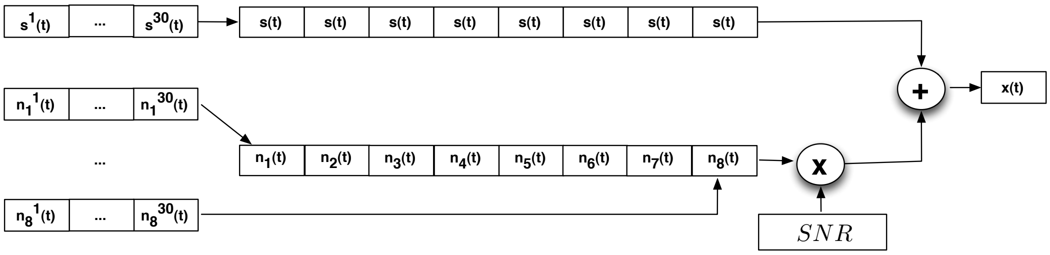


Figure 4.1: Database setup procedure.

4.2.2.2 Results and discussion

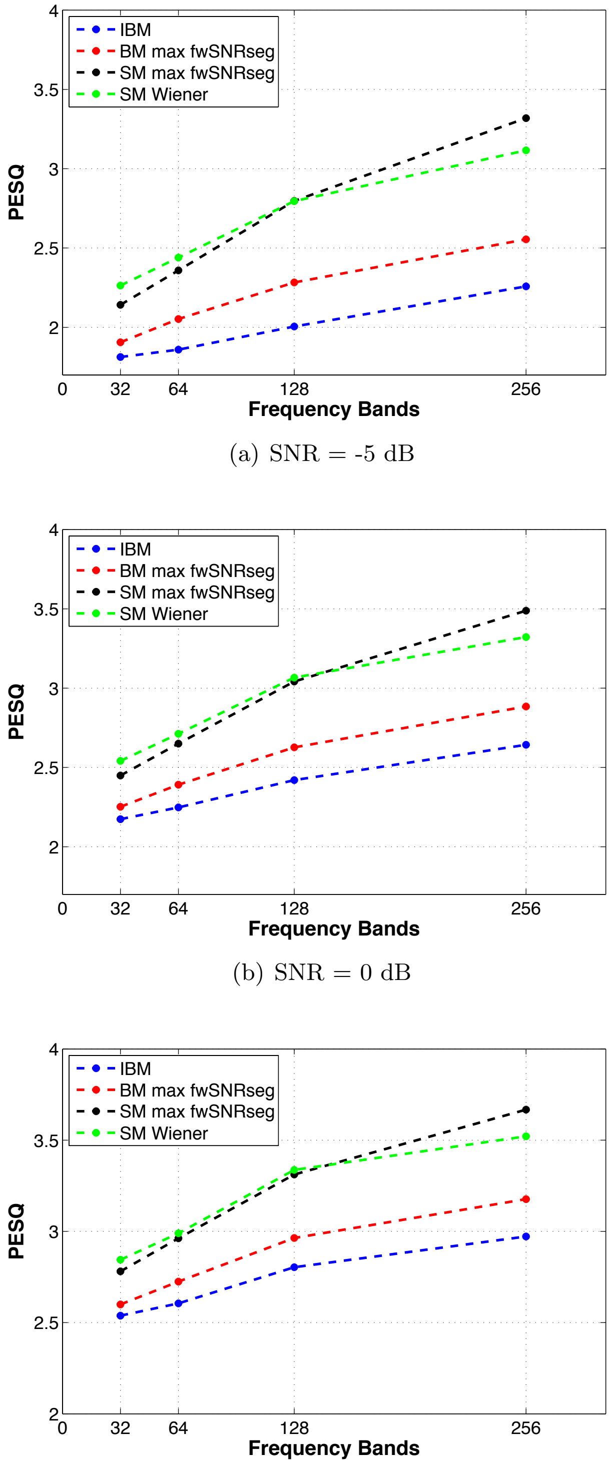
Figure 4.2 represents the PESQ obtained by the different time-frequency masks when varying the number of frequency bands of the STFT. The speech and noise mixtures are generated with a SNR of -5 dB in (a), 0 dB in (b), and 5 dB in (c). The three graphs confirm that the PESQ obtained by time-frequency masking depends on the frequency resolution. As it was expected, an increment of the number of frequency bands implies a notable increment of the PESQ values (whenever the length of the time window guarantees stationarity). Comparing the different time-frequency masks, the two soft masks clearly improve the PESQ values obtained by the binary masks. The Wiener mask obtains the best results for 32, 64 and 128 frequency bands and the soft mask that maximizes the fwSNRseg obtains the best results for 256 frequency bands. Concerning the two binary masks, the one that maximizes the fwSNRseg obtains higher PESQ values than the IBM in any case.

The DSP embedded in digital hearing aids works with adjustable sampling rates that usually vary between 8 and 16 kHz, with short analysis windows of 128 or 64 points to ensure low latency. Hence, the DFT-based time-frequency analysis only contains 64 or 32 frequency bands. According to this, the Wiener soft mask seems to be the most suitable mask among the evaluated to improve intelligibility in hearing aids.

4.3 Computational resources available for speech enhancement in hearing aids

The implementation of speech enhancement algorithms in hearing aids is strongly restricted by the reduced computational capability and the memory available in the DSP embedded in such devices. The processor is forced to work at-low clock frequencies in order to minimize the power consumption and thus to maximize the battery life. The time-frequency analysis is based on a DFT filterbank and usually implemented in a specific processor, hence it does not imply any extra consumption of computational resources. Nevertheless, a considerable part of the computational capabilities of the DSP is already dedicated to the algorithms that aim to compensate hearing losses (i.e. the multi-band compression-expansion algorithm). Therefore, the design and implementation of speech enhancement algorithms in hearing aids are constrained by the remaining resources of the DSP, resources that other signal processing algorithms also demand to perform other tasks such as acoustic feedback cancellation or automatic sound classification. Consequently, an efficient speech enhancement algorithm should only use a small part of the available computational resources, allowing to run other types of algorithms simultaneously.

Let us now restrict the computational cost available for speech enhancement quantitatively,



(a) SNR = -5 dB

(b) SNR = 0 dB

(c) SNR = 5 dB

Figure 4.2: PESQ obtained with the different time-frequency mask when varying the number of frequency bands of the STFT. The speech and noise mixtures are generated with a SNR of -5 dB in (a), 0 dB in (b), and 5 dB in (c).

considering the characteristics of a state-of-the-art commercial device. Common DSPs embedded in hearing aids have a processor with a selective clock speed that usually goes from 5.12 MHz down to 1.28 MHz. They have a Harvard architecture containing a multiplier-accumulator (MAC) with a set of instructions completed in a clock cycle, hence the number of mega instructions per second (MIPS) is the clock speed value. The sampling rate f_s is usually adjustable but limited by the output frequency range of the loudspeaker, 16 kHz normally being the allowed maximum sampling rate. The analysis and synthesis windows have a length of L_{WIN} samples working with 50% of overlap, and the DFT-based frequency analysis contains K frequency bands. According to this, the number of instructions available to process each frequency band (IPF) of a frame is calculated using the next expression:

$$IPF = \frac{MIPS}{K} \cdot \frac{L_{WIN}/2}{f_s}. \tag{4.13}$$

In the special case of a processor with a clock speed of 5.12 MHz (5 MIPS), working with a sampling rate of 8 kHz, an analysis window of 64 samples, and 32 frequency bands, the number of instructions available to process each frequency band of a frame is 625. These instructions are shared between the multi-band compression-expansion algorithm, which is an indispensable algorithm, and the algorithms dedicated to feedback cancellation, automatic sound classification, and speech enhancement. Henceforth, the IPF calculated in this section is used as reference value and the speech enhancement algorithm proposed in this chapter will be constrained to use only a part of the total number of available instructions.

4.4 Proposed algorithm to improve the intelligibility of speech in noise for monaural hearing aids

Figure 4.3 shows a block diagram of the enhancement algorithm proposed in this chapter. It is divided into two parts, the training stage (top) and the enhancement stage (bottom). The frequency analysis is performed by a 64-points DFT for each time frame, using a Hamming window with an overlap of 50%.

The main goal is to obtain a low-cost system that maximizes the output PESQ score. This is obtained in different steps. The first one is the estimation of the Wiener soft mask defined in expression (4.12) with a generalized linear least squares estimator (GLSE), minimizing the MSE between the estimated mask and the Wiener mask. The weights obtained with the GLSE are labeled as $\mathbf{v_{MSE}}$. The estimator combines a set of simple features, included in matrix $\mathbf{Q}$, which are extracted from the spectrum of the mixture. Note that the Wiener soft mask does not maximize the PESQ score and the estimated mask with the GLSE will obtain a lower PESQ score that the one obtained by the Wiener mask due to estimation errors. This fact motivates the next step of the algorithm. In the second step, the weights $\mathbf{v_{MSE}}$ are improved to increase the output PESQ value, obtaining $\mathbf{v_{opt}}$. The third step is to reduce the computational cost of the system using a feature selection algorithm that selects the subset of features that best approximates the mask estimated in the case of using all features. In the last step, the weights selected $\mathbf{v_{sel}}$ are finally re-improved to increase the output PESQ value, obtaining $\mathbf{v_{selopt}}$.

In the enhancement stage, the weights $\mathbf{v_{selopt}}$ calculated in the training stage are used to generate the mask from the features extracted from the mixture signal. The clean speech signal is estimated by applying the estimated mask $\hat{M}(k,l)$ to the spectrum of the mixture. The different parts of the algorithm are detailed in the next subsections.

Finally, it is worth clarifying that all the algorithms included in the training stage are carried

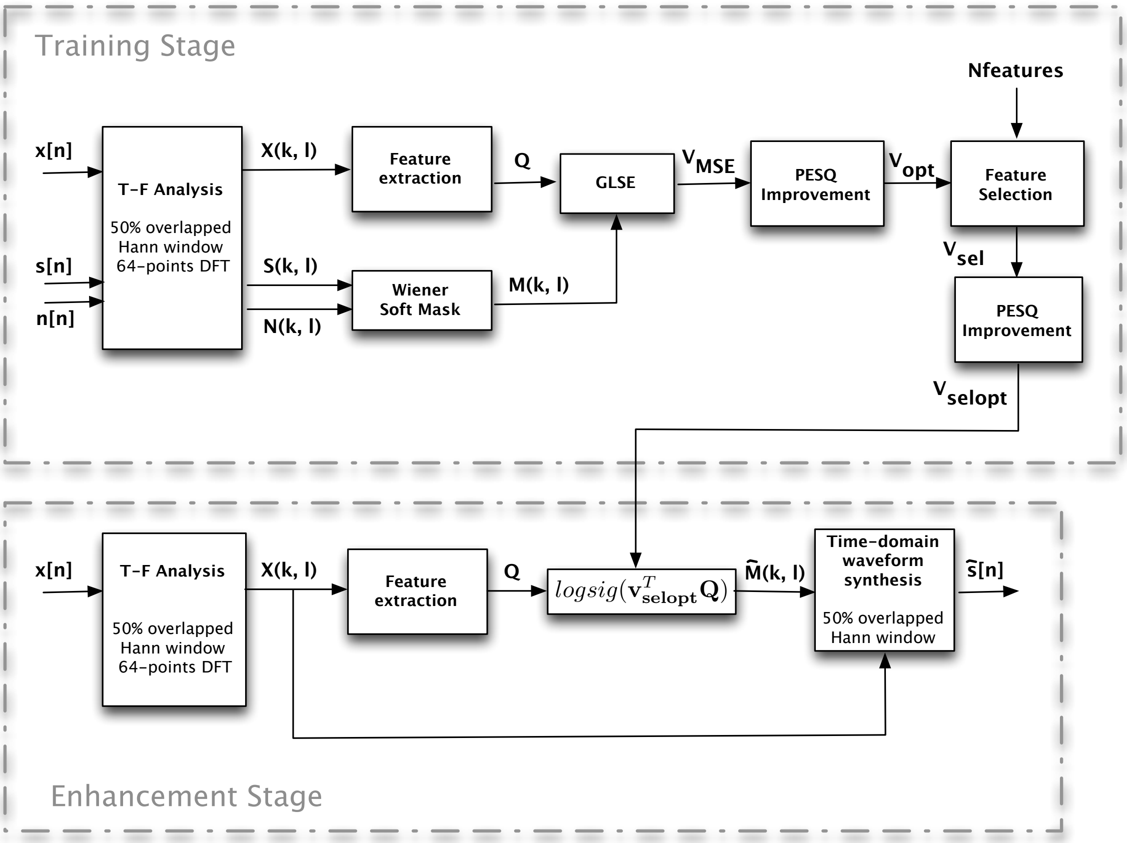


Figure 4.3: Algorithm overview. Training stage (top) and speech enhancement stage (bottom)

out offline on a computer. Only when the design has been completed, the optimum solution is then implemented on the digital hearing aid.

4.4.1 Generalized least squares estimator (GLSE)

The least squares estimation (LSE) is an approach that fits a parametrized mathematical model to the observed data by minimizing the MSE between the observed data and their expected values. In the case that the model combines linearly the unknown parameters, the method is known as linear least squares.

Let us consider the pattern vector $\mathbf{x_i}$ (i.e. the observations of the model) containing P input features, $\mathbf{x_i} = [x_1, x_2, \ldots, x_P]^T$, which are extracted from the mixture signal in the problem at hand. The pattern matrix $\mathbf{P}$ of size PxL is defined as a matrix that contains the patterns $\mathbf{x_i}$ of a set of L data samples, $\mathbf{P} = [\mathbf{x_1}, ..., \mathbf{x_L}]$, and the matrix $\mathbf{Q}$ is defined as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{1} \\ \mathbf{P} \end{bmatrix}, \tag{4.14}$$

where $\mathbf{1}$ is a row vector of length L. The output of the linear estimator is obtained as a linear combination of the input features (i.e. observations) according to

$$\mathbf{y} = \mathbf{v}^T\mathbf{Q}, \tag{4.15}$$

where the vector $\mathbf{v} = [v_0, v_1, v_2, \ldots, v_P]^T$ contains the bias v_0 and the weights applied to each of the P input features and $\mathbf{y}$ is a vector of size 1xL containing the output of the LSE for the L input patterns. In the problem at hand, there are a different vector $\mathbf{v}$ and matrix $\mathbf{Q}$ for each frequency band, which are denominated $\mathbf{v}_k$ and $\mathbf{Q}_k$ respectively, and the L data samples correspond with each of the time frames. Hence, the vector $\mathbf{y}_k$ is a vector of size 1xL containing the output of the LSE for the L input patterns in the k-th frequency band, $\mathbf{y}_k = [y(k,1), ..., y(k,l), ..., y(k,L)]^T$, which is calculated as $\mathbf{y_k} = \mathbf{v_k}^T \mathbf{Q_k}$. The estimation of the binary mask corresponds with the output of the LSE, i.e. $\hat{M}(k,l) = y(k,l)$.

The design of the estimator consists in finding the best value of $\mathbf{v_k}$ in order to minimize the error between the obtained output and its desired value. In the case of supervised learning, the desired output values are available and used for training. The desired output values for each data sample are contained in the target vector defined as $\mathbf{t_k} = [t(k,1), t(k,2), \cdots, t(k,L)]^T$. In the proposed enhancement algorithm, the target values are those corresponding to the Wiener soft mask previously calculated. The estimation error is defined as the difference between the output value of the estimator (4.15) and the desired value

$$\mathbf{e_k} = \mathbf{y_k} - \mathbf{t_k} = \mathbf{v_k}^T \mathbf{Q_k} - \mathbf{t_k}, \tag{4.16}$$

and the MSE for the k-th frequency band is computed according to

$$MSE_k = \frac{1}{L} \left\| \mathbf{y}_k - \mathbf{t}_k \right\|^2 = \frac{1}{L} \left\| \mathbf{v_k}^T \mathbf{Q_k} - \mathbf{t}_k \right\|^2 . \tag{4.17}$$

In the least squares approach, the weights are adjusted in order to minimize the MSE. The minimization of the MSE is obtained by deriving the expression (4.17) with respect to every weight of the linear combination and setting the result equal to zero, which yields the next expression

$$\mathbf{v_k} = \mathbf{t_k} \mathbf{Q_k}^T \left(\mathbf{Q_k} \mathbf{Q_k}^T \right)^{-1} . \tag{4.18}$$

The performance of the linear least squares estimator can be improved by introducing non-linear transformations of the input features, which are still linearly combined, unlike the non-linear least squares approach. The simplest example is the use of a quadratic transformation of the input features, which leads to the quadratic least squares estimator. In the general case, the matrix $\mathbf{Q}$ can be defined as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{1} \\ f_1(\mathbf{P}) \\ \ldots \\ f_{N_T}(\mathbf{P}) \end{bmatrix}, \tag{4.19}$$

where $f_1, \ldots, f_{N_T}$ are N_T linear or non-linear transformations performed over the original input features contained in $\mathbf{P}$. The weight vector is then defined as $\mathbf{v} = [v_0, v_1, \ldots, v_{N_T \cdot P}]^T$, and it can be also obtained using expression (4.18). Henceforth, this is denominated generalized least squares estimator (GLSE).

The estimation of each time-frequency point of the Wiener soft mask is obtained by expression (4.15), using the features proposed in the next section.

4.4.2 Proposed features for estimation

The DFT-based analysis filterbank included in the hearing aid provides the input to calculate the energy of each frequency band of the mixture signal for each time frame, $|X(k,l)|^2$.

This information can be used as input feature, but, according to the aforementioned GLSE, further transformations of this feature can be included as input features. Specifically, the logarithm and the square logarithm of the energy of the time-frequency point, $\log(|X(k,l)|^2)$ and $\log^2(|X(k,l)|^2)$, have been experimentally found to provide the most meaningful information to the GLSE. Additionally, the information of neighbor time-frequency points is also included as input features. The logarithm and the squared logarithm of all frequency bands of the current time frame are included as input features. The information about the previous time frames is also included, considering two different alternatives:

1. The logarithm and the squared logarithm of the energy of the $(D-1)$ previous time frames of all frequency bands are directly included as input features.

2. An exponentially-weighted moving average (EWMA) of the logarithm and the squared logarithm of the energy of the previous time frames is calculated for each frequency band. The EWMA is a type of infinite impulse response filter that applies weighting factors which decrease exponentially. In this work, the values are calculated according to

$$pm(k,l) = (1-2^{-\alpha}) \cdot pm(k,l-1) + 2^{-\alpha} \cdot x(k,l), \quad \alpha \in \mathbb{Z}^+, \tag{4.20}$$

   where $pm(k,l)$ is the EWMA for the k-th frequency band in the l-th time frame, $x(k,l)$ here represents the input value (i.e. the logarithm or the squared logarithm of the energy), and α is a smoothing factor that controls the degree of weighting decrease. A lower value of α discounts older observations faster. The EWMA is calculated with $(D-1)$ different values of α, then having $(D-1)$ different averages for each frequency band, which are all included as input features. From the computational point of view, the use of exponential values $(2^{-\alpha})$ as filter coefficients is equivalent to shift a value α bits in memory, which reduces the computational cost associated to the computation of the EWMA.

In summary, each time-frequency point has a total of $P = 2KD$ input features for its classification, considering two options, the use of instantaneous values of the previous time frames (INST) or the use of EWMA values of the previous time frames (EWMA).

4.4.3 Optimization algorithm to increase the output PESQ score

The weights $\mathbf{v_{MSE}}$ obtained by the GLSE are calculated to minimize the MSE between the estimated mask and the Wiener soft mask. Nevertheless, the estimated Wiener soft mask is not the one that maximizes the output PESQ score and it is only used as a first approximation. The fact that the time-frequency mask that maximizes the PESQ score is unknown motivates the use of a heuristic optimization method to approximate the unknown global solution. Stochastic optimization (SO) methods are iterative optimization methods that generate and use random changes in the parameters of the optimization process [Spall, 2005]. According to this approach, and considering the weight vector $\mathbf{v_{MSE}}$ as starting point, the proposed optimization method consists in modifying iteratively the weight vector using a random factor, using the PESQ score as fitness (cost) function. The vector is either scaled or shifted alternatively in each iteration. The vector is scaled according to $\mathbf{v}_{i+1} = \mu \cdot \mathbf{v}_i$, where i is the iteration counter. The vector is shifted according to $\mathbf{v}_{i+1} = \mathbf{v}_i + \mu \cdot \mathbf{S}$, where $\mathbf{S}$ is a zero-mean unit-variance Gaussian random vector with the same size of $\mathbf{v}$, which represents the shift direction. The parameter μ is a value that controls the modification (scale or shift) introduced into the weighting vector, and its best value is searched iteration by iteration. Additionally, the output is smoothed by using a log-sigmoid function, i.e., $\mathbf{y} = logsig(\mathbf{v}^T\mathbf{Q})$. The log-sigmoid function is defined as $logsig(x) \equiv \frac{1}{1+e^{-x}}$.

The steps of the proposed iterative optimization algorithm are the next:

1. The weights $\mathbf{v_{MSE}}$ calculated with the GLSE by expression (4.18) are taken as initial values.

2. A set of candidate μ values is generated. The algorithm scales the vector every third iteration, shifting the vector the other two intermediate iterations:

    - When the vector is scaled (i.e. $\mathbf{v}_{i+1} = \mu \cdot \mathbf{v}_i$) the set of candidate scaling values is generated according to $\mu = 10^{x/i}$, where x contains linearly equally spaced points between -2 and 2. The use of logarithmic values allows exploring a wider range using a lower number of points. Dividing the values of x by the iteration counter refines the search.
    - When the vector is shifted (i.e. $\mathbf{v}_{i+1} = \mathbf{v}_i + \mu \cdot \mathbf{S}$) the set of μ values is generated considering logarithmically distributed values between -10 and 10. In this case, the μ value varies the variance of the Gaussian matrix $\mathbf{S}$, which controls the size of the shift.

3. Evaluate the output PESQ with each of the μ candidate values of the generated set. The output is obtained according to $\mathbf{y} = logsig((\mu\mathbf{v})^T\mathbf{Q})$ in case of scaling and $\mathbf{y} = logsig((\mathbf{v} + \mu \cdot \mathbf{S})^T\mathbf{Q})$ in case of shifting. The μ value that achieves higher PESQ score is selected as the best candidate.

4. The weights are then updated (scaled or shifted) according to the best μ value.

5. Repeat steps 2 to 4 a total of 100 times.

The number of iterations and the number of candidate μ points have been found to obtain a good tradeoff between design time and performance. The optimized weights obtained in this step are labeled as $\mathbf{v_{opt}}$.

4.4.4 Feature selection algorithm

In the enhancement stage, the estimation of the time-frequency soft mask only involves the operation $logsig(\mathbf{v}^T\mathbf{Q})$ using the weights calculated in the training stage. The implementation of the proposed estimator is relatively simple, its computational cost being directly related to the number of features. Considering that the MAC operation is executed in a single instruction, the number of instructions required by the estimator can be reduced to $2P$, P being the number of input features included in $\mathbf{Q}$ (i.e. $P = 2KD$). Assuming that the output of the time-frequency analysis is $|X(k,l)|^2$, and according to the standard assembler language used in this type of DSPs, the number of instructions required for the computation of the input features is $14+4(D-1)$, as shown in figure 4.4. According to this, the number of instructions necessary to process each frequency band is $IPF = 4KD + 14 + 4(D-1)$. Considering $K = 32$ and $D = 5$, the IPF is 670, value that exceeds the reference value calculated in section 4.3.

The computational cost of the algorithm (i.e. the IPF) can be reduced by decreasing the number of features used for the estimation of the mask. For this purpose, a feature selection algorithm is proposed. Considering K=32 and D=5, the number of features available to estimate the mask of each time-frequency point is $P = 2KD = 320$, which leads to a huge number of possible combinations. Consequently, to perform an exhaustive search is not affordable, and a feature selection algorithm based on evolutionary computation is proposed.

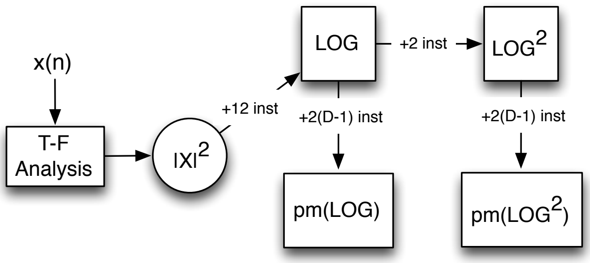


Figure 4.4: Number of instructions associated to the computation of the proposed features. The output of the time-frequency analysis block is the squared modulus (energy) of the STFT.

Evolutionary algorithms (EA) exhibit a great potential to solve certain problems that otherwise would be intractable. This type of algorithms are inspired in natural evolution laws such as selection, mutation and crossover, to iteratively search for the optimum solution from the solutions obtained in previous iterations [Haupt and Haupt, 2004]. EA are commonly used in engineering to solve optimization and search problems in a wide range of applications, for instance, automatic speech/music discrimination [Ruiz-Reyes et al., 2010], antenna array design in [Chabuk et al., 2012], mobile robot localization [Kwok et al., 2006], or feature selection for sound classification in hearing aids [Alexandre et al., 2007]. The three main parts of an evolutionary algorithm are the generation of the candidate solutions of the population, the evaluation of a fitness function, and the evolution of the population [Alexandre et al., 2007]. The candidate solutions are defined for each specific problem, and they are composed of a set of elements that may be binary or continuos values. The fitness function is defined as the cost function to optimize, and it also depends on the specific problem to solve. The evolution is based on the crossover and mutation operators which characteristics can be also adapted to each specific problem. The common steps of the evolutionary algorithms implemented in this thesis are listed below.

1. Generation of an initial population of candidate solutions. The size of the population N_P is a crucial issue for the EA performance. On the one hand, a large population could cause more genetic diversity (and thus, a higher search space) and consequently suffer from slower convergence. On the other hand, with a very small population, only a reduced part of the search space could be explored, thus increasing the risk of prematurely converging to a local extreme. In each specific problem, the population size should be chosen as a tradeoff between computational complexity and performance. The candidate solutions of the initial population can be either generated at random or initialized with a determined set of values, for instance, to accelerate the convergence of the algorithm.

2. In some cases, it is necessary to validate the new population in order to check if all candidate solution fulfill some constraints applied to the possible solutions. Those candidates that do not fulfill the constraints are iteratively regenerated until they are valid.

3. Evaluation of the fitness (cost) function for each candidate in the population.

4. A selection process is performed, using the results of the evaluation of the fitness function as ranking. It consists in selecting a subpopulation of N_{SP} candidate individuals that best fit the fitness function. These elite individuals are those that will survive to the next

generation. In some cases, not all of the worst individuals in the current generation are replaced, in other cases they are. The first replacement strategy, called the “steady state approach”, prevents the algorithm from prematurely converging to a local minima.

5. Breed the new generation by recombining the parents by using a crossover operator. A number of $N_P - N_{SP}$ novel candidates for the next generation are generated by random crossover of the previously selected N_{SP} best candidates. The probability that the crossover operator is applied to each individual is called crossover probability, $P_C < 1$. The crossover operator implemented in the EAs included in this thesis is an uniform crossover (UX) operator with $P_C = 0.5$. The offspring has approximately half of the elements from the first parent and the other half from the second parent, and these elements are randomly selected.

6. To mutate or randomly change the offspring. The main purpose of a mutation operator is to maintain diversity within the population and inhibit premature convergence to local extreme. Not all the offspring elements are mutated: the probability that the mutation operator is applied is called mutation probability, $P_M < 1$, and its value is usually found empirically. The mutation operator depends on the characteristics of each candidate solution.

7. To iterate steps $2 - 6$ until a maximum number of generations is reached or if the best value of the fitness function remains unchanged for a given number of iterations.

The values of the parameters of the EA (population size, crossover rate, mutation probability, mutation scheme, and number of generations) should be chosen in each specific problem to obtain a good tradeoff between design time and performance.

The goal of the EA proposed in this section is the selection of determined number of features (Nfeatures) among the whole set with the aim of obtaining the maximum output PESQ value, trying to approximate the value obtained by the system when the classifier uses all features. Each candidate solution of the EA contains the indexes of the selected input features. The ideal feature selection algorithm would first compute the weights with the GLSE ($\mathbf{v_{MSE}}$) to estimate the Wiener mask using the selected subset of features (i.e. the candidate solution), optimize that weights to maximize the output PESQ value ($\mathbf{v_{opt}}$), and use that value as cost function to select the best candidate solution. This process would have to be repeated for each candidate solution and iteration. Unfortunately, the evaluation of the PESQ function is computationally expensive. Using a powerful computer with a 2 x 2.96 GHz 6-core processor and 32 GB of RAM, the evaluation of the PESQ function for the design set takes around 19 s. Assuming a population of 100 candidate solutions and 1000 iterations, the time required by the EA is approximately 528 hours. This implementation is not affordable and an efficient alternative is proposed.

The proposed feature selection algorithm, rather than searching for the subset of features that maximizes the output PESQ, it searches for the subset of features that best approximates the time-frequency mask estimated in the case of using all the input features, i.e, $\mathbf{y} = logsig(\mathbf{v_{opt}}^T\mathbf{Q})$. In this case, the weights $\mathbf{v_{opt}}$ are taken as input values and they are not recalculated for each candidate solution. Instead, only the weights corresponding to the selected features of each candidate solution are considered, setting the remaining weights to zero. The selected weight vector is labeled as $\mathbf{v_{sel}}$, and each candidate solution is a vector of the same size of $\mathbf{v}$ containing values of one in the positions of the selected features and zero in the remaining positions. The cost function considered by the feature selection algorithm is the MSE between the mask estimated using all features, $\mathbf{y} = logsig(\mathbf{v}_{\mathbf{opt}}^T\mathbf{Q})$, and the mask obtained using only the weights of the selected features, $\mathbf{y} = logsig(\mathbf{v}_{\mathbf{sel}}^T \mathbf{Q})$. Hence, dropping the *logsig* function, the cost function is $MSE = \mathbf{v}_{\mathbf{opt}}^T \mathbf{Q} - \mathbf{v}_{\mathbf{sel}}^T \mathbf{Q}$.

The complete steps of the feature selection algorithm are the next:

1. An initial population of 100 candidate solutions is generated. Each candidate solution contains P random binary values (0 or 1).

2. The candidates of the population are validated to fulfill the constraint of total number of features. If a candidate solution exceeds the maximum number of features (Nfeatures), random positions are decreased by one (avoiding negative values). The process iterates until the candidate solution fulfills the requirement.

3. The cost function is evaluated for each candidate solution of the population. It consists of the computation of the MSE between the mask estimated using all features and the mask obtained using the weights of the features selected by each candidate solution.

4. A selection process is applied, using the MSE of each solution as ranking. It consists in selecting the best 10% of the solutions of the population, removing the remaining solutions.

5. The remaining 90% solutions of the new generation are then generated by uniform crossover of the best candidates.

6. Mutations are applied to the candidate solutions of the new population that are repeated, excluding the best solution. Mutations consist of changing the values of random positions of the candidate solution.

7. The process is repeated from step 2 to 6 until 1000 generations are evaluated.

The features contained in the best solution obtained in the last iteration are considered to be the optimized solution. The weights corresponding to that features are labeled as $\mathbf{v}_{\mathbf{sel}}$. The values of the parameters of the evolutionary algorithm (population size, crossover rate, mutation scheme and number of generations) have been found to obtain a quite good tradeoff between design time and performance for the experiments carried out.

Finally, the weights obtained in this step $\mathbf{v}_{\mathbf{sel}}$ are optimized again to increase the output PESQ, using the algorithm described in the previous section. The optimized selected weights are labeled as $\mathbf{v}_{\mathbf{selopt}}$, and they are the output weights of the training stage.

4.5 Experimental work and results

4.5.1 Database setup

Speech and noise mixtures of different SNRs are generated, using the speech and noise signals contained in the NOIZEUS database described in section 2.5.2.2. The procedure to generate the mixtures is the next:

1. The 30 clean speech signals are normalized and linked together, one after the other, obtaining a speech segment of 80 seconds lenght.

2. The 30 noise signals of each type of noise are also normalized and linked together, one after the other, obtaining 8 different segments of 80 seconds length, one for each noise.

3. The clean and noise signals are split into two different parts, one for training and another for testing. The training set consists of the 60% of the signals and the test set consists of the 40% of the signals. Hence, the training speech and noise signals are 56 seconds length and the test speech and noise signals are 24 seconds length.

4. The training and test clean speech segments are repeated 8 times, generating a signal of 448 seconds length in the case of training and 192 seconds length in the case of test.

5. The 8 different noise segments are linked together, for the training and test sets separately.

6. The clean and noise signals of both sets are normalized and mixed with the desired SNR.

Figure 4.5 shows graphically the database setup procedure.

4.5.2 Experiments and results

The training stage has been executed 6 times, varying the maximum number of features selected for all frequency bands (Nfeatures) in the feature selection algorithm, among the values 1024, 512, 256, 128, 64 and 32. These values correspond with an average of 32, 16, 8, 4, 2 and 1 features selected for each frequency band. The best value of D, which represents the previous time frames in the case of instantaneous features and the number different averages in the case of EWMA features, has been obtained experimentally, finding that $D = 5$ represents a good tradeoff between speech improvement and computational cost. The α values used to calculated the EWMA are -8, -4, -2, -1 and 0. With this value of D, the IPF values associated are 94, 62, 46, 38, 34 and 32, for each value of Nfeatures, respectively. These IPF values represent only a reduced part of the total computational resources available for signal processing, specifically, a 15%, 9.9%, 7.4%, 6.1%, 5.4% and 5.1%, respectively.

Figure 4.6 represents the averaged PESQ values obtained by the proposed algorithm, as a function of the number of features (Nfeatures), using instantaneous features with $D = 5$ (blue color) and using EWMA features with $D = 5$ (red color). The solid line represents the values obtained in the training set and the dashed line represents the values obtained in the test set. The SNR is -5 dB in (a), 0 dB in (b), and 5 dB in (c). Comparing the PESQ values obtained in the design set, the EWMA option shows higher PESQ values for any SNR and number of features. The PESQ values decrease as the number of features decreases, as it is expected. Considering the values obtained in the test set, the EWMA option obtains higher PESQ values

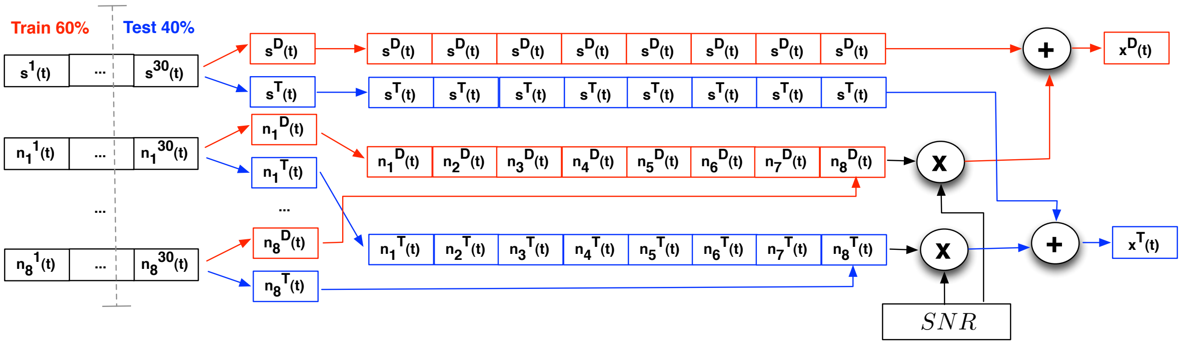
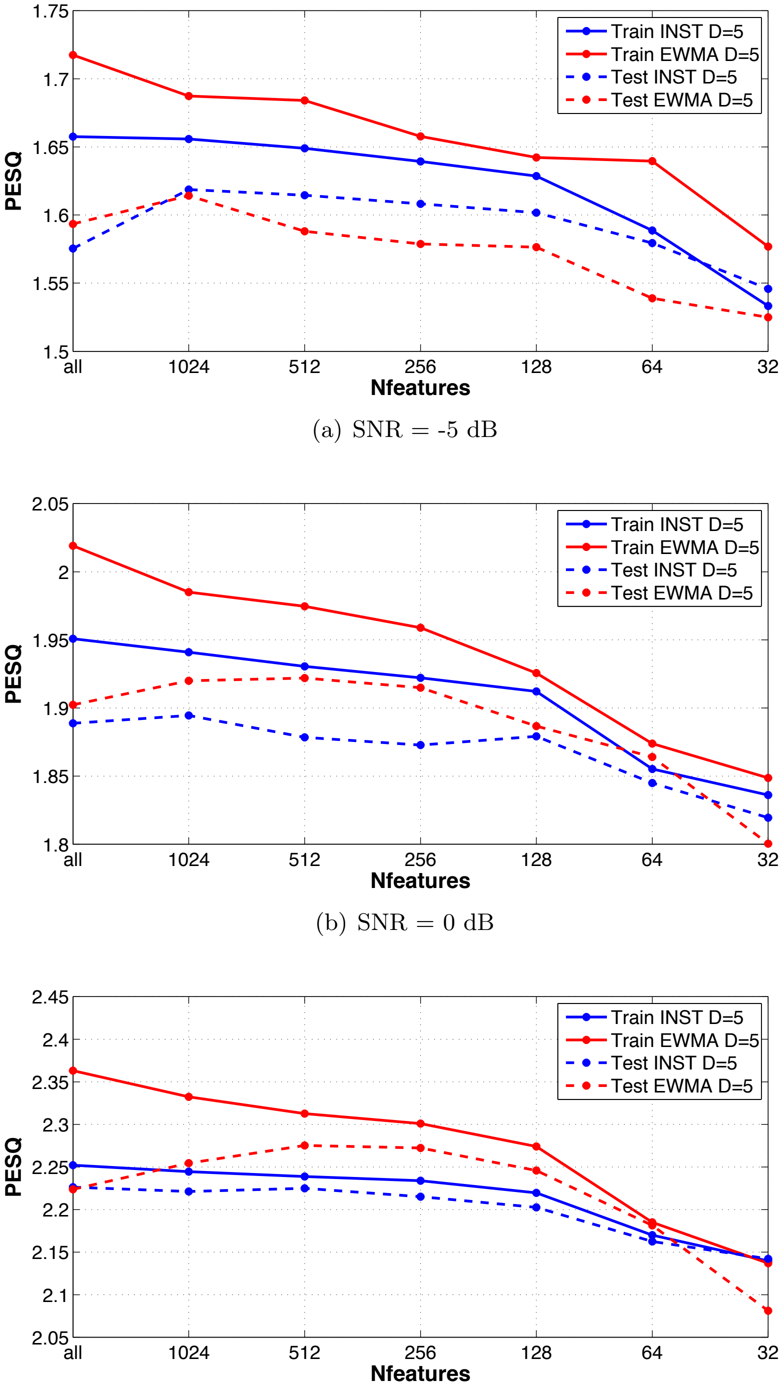

Train 60%
Test 40%
s1(t)
...
s30(t)
n11(t)
n130(t)
n81(t)
n830(t)
sD(t)
sT(t)
n1D(t)
n1T(t)
n8D(t)
n8T(t)
n1D(t) n2D(t) n3D(t) n4D(t) n5D(t) n6D(t) n7D(t) n8D(t)
n1T(t) n2T(t) n3T(t) n4T(t) n5T(t) n6T(t) n7T(t) n8T(t)
x
+
xD(t)
xT(t)
SNR

Figure 4.5: Database setup procedure.



(a) SNR = -5 dB

(b) SNR = 0 dB

(c) SNR = 5 dB

Figure 4.6: Averaged PESQ values obtained by the proposed algorithm, as a function of the number of features (Nfeatures), using instantaneous features with $D = 5$ (blue color) and using EWMA features with $D = 5$ (red color). The solid line represents the values obtained in the training set and the dashed line represents the values obtained in the test set. The SNR is -5 dB in (a), 0 dB in (b), and 5 dB in (c).

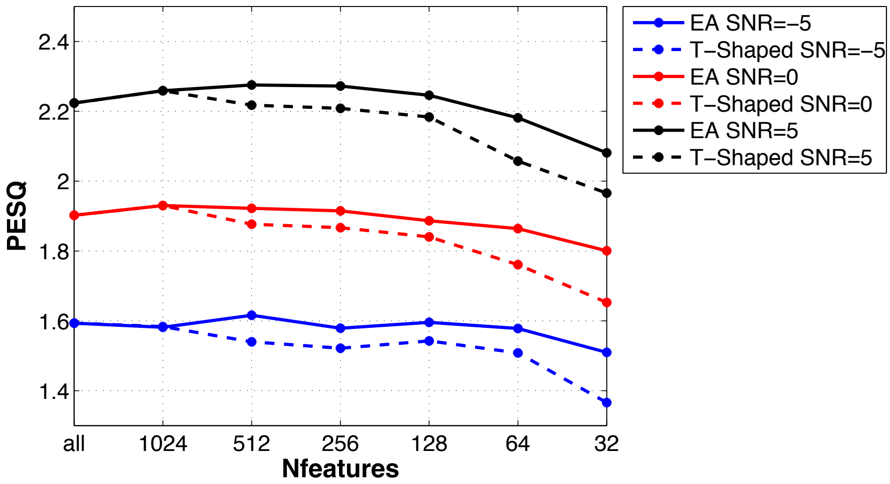


Figure 4.7: PESQ values averaged over the test set, with the different number of features (Nfeatures), in the case of EWMA features and $D = 5$, using the proposed feature selection algorithm (solid line) and the T-shaped footprint (dashed line). The different colors represent different SNRs.

in the cases of 0 and 5 dB of SNR for any number of features, except for the lowest value (Nfeatures=32), in which case the INST option slightly improves the PESQ values obtained by the EWMA option. In the case of a SNR of -5 dB, the INST option outperforms the EWMA option for any number of features, except in the case of 1024, where the values obtained by the two options are practically the same. A noticeable observation deduced from these results is the fact that the PESQ values obtained in the test set do not always decrease with the number of features. In the case of EWMA, the PESQ score obtained when all features are used for classification is lower than the one obtained when only 1024 features are selected and, in the cases of 0 and 5 dB of SNR, it is even lower than the values obtained when only 512 and 256 features are selected. This fact suggests that the generalization of the trained estimator is worse when a excessively high number of features is considered.

In order to assess the validity of the feature selection algorithm, the results are compared to the case of selecting a constant T-shaped pattern of input features. For each time-frequency point, the T-shaped footprint selects the P_T upper adjacent frequencies, P_T lower adjacent frequencies, and the P_T previous time frames, which corresponds with a total number of selected features of $P = 3P_T + 1$. Figure 4.7 contains the PESQ values averaged over the test set with the different number of features (Nfeatures), in the case of EWMA features and $D = 5$, using the proposed feature selection algorithm (solid line) and the T-shaped footprint (dashed line). The different colors represent different SNRs. From this plot it is deduced that the proposed feature selection algorithm outperforms the feature selection based on a T-shaped footprint, and the difference between both solutions increases when the number of features decreases. Hence, the PESQ values obtained by both methods in the case of selecting 1024 features is the same for all SNRs, but in the case of selecting only 32 features, the difference between the PESQ values obtained by both methods is approximately a 10%.

With the goal of providing more meaning to the results of the experiments obtained in this chapter, which are represented in terms of PESQ scores, a relationship between PESQ and SNR is established, calculating the PESQ scores associated to the current database generated with different SNR values. The mixtures are generated in the same way that the one described in

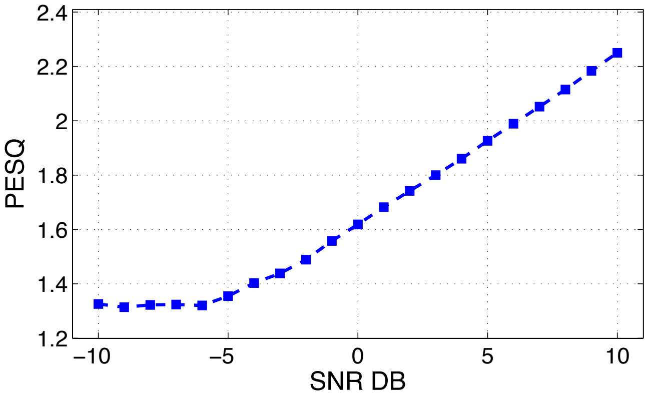


Figure 4.8: Relationship between the PESQ score and the input SNR.

figure 4.1, with SNR values that go from -10 to 10 in steps of 1. The PESQ is calculated comparing the mixture signal with the original clean speech signal. Figure 4.8 shows the relationship between the PESQ score and the input SNR. From this graph it is deduced that an improvement of 0.2 in the PESQ score corresponds with an increment of approximately 5 dB in the SNR.

Finally, table 4.1 contains the PESQ values corresponding to the unprocessed mixture (UN) and the obtained by the proposed algorithm in the test set using different number of features (Nfeatures), with the best option (EWMA or INST). The increment in the PESQ scores obtained by the proposed algorithm is clearly demonstrated for all input SNRs, even in the case of using only 32 features. The results show an average increment in the PESQ value of 17% in the case of selecting all features, 19% in the case of selecting 1024 and 256 features, and 12% in the case of selecting only 32 features, with respect to the values obtained in the unprocessed signal. In the case of using 256 features, the absolute increment in the PESQ score is 0.31 on average, which corresponds to an increment in the SNR of approximately 6 dB. And, in the case of using the lowest number of features, i.e. 32, the absolute increment in the PESQ score is still 0.19 on average, which corresponds to an increment in the SNR of approximately 4.5 dB.

4.6 Discussion

This chapter deals with the problem of single-channel speech enhancement for monaural hearing aids, using a time-frequency masking approach. The proposed algorithm has been designed to work in low SNRs and bearing in mind the reduced computational resources available in hearing aids. The three time-frequency masks proposed in this work outperforms the results obtained by the CASA IBM, in terms PESQ score, which has been demonstrated to correlate with speech

Table 4.1: PESQ values corresponding to the unprocessed mixture (UN) and the obtained by the proposed algorithm in the test set using different number of features (Nfeatures), with the best option (EWMA or INST).

SNR	UN	Nfeatures			
		All	1024	256	32
- 5 dB	1.35	1.59	1.62	1.61	1.55
0 dB	1.61	1.90	1.92	1.92	1.82
5 dB	1.92	2.23	2.26	2.27	2.14

intelligibility. The benefits associated to the use of soft masks instead of binary masks for single channel noise reduction are clearly demonstrated. The speech enhancement algorithm has been designed to maximize the output PESQ score, estimating the proposed Wiener soft mask with a low-cost linear estimator and re-optimizing its weights to increase the PESQ score. In order to reduce the computational cost, a feature selection algorithm has been proposed. According to the results obtained with the experiments carried out in this work, the proposed system clearly improves the output speech quality, even for a low SNR of -5 dB. An average improvement of 0.31 in the output PESQ score is obtained using only 7.4 % of the computational resources available for signal processing, which is equivalent to an increment of 6 dB in the SNR. The computational cost can be further reduced to only a 5.1% obtaining an average improvement of 0.19 in the output PESQ score, which is still a good value equivalent to 4.5 dB. Regarding the two alternatives to incorporate the information of the previous time frames, the EWMA option is recommended for normal and low SNRs (down to 0 dB) but the INST option is more suitable for extremely low SNRs (-5 dB).

In summary, the speech enhancement system presented in this chapter represents a real feasible solution to be implemented in monaural hearing aids, obtaining good noise reduction levels even for low SNRs, and consuming a minimum part of the computational resources available for signal processing.

4.7 Summary of contributions

The main contributions included in this chapter are listed below.

- Three different time-frequency masks have been proposed and compared to the CASA IBM for single-channel noise reduction. One of them is also a binary mask and the other two are soft masks. The experiments carried out have demonstrated that the use of soft masks instead of binary masks is beneficial for single-channel speech enhancement.

- It has been proved that the performance of the time-frequency masks depends on the frequency resolution of the STFT used to compute the time-frequency representation of the signals.

- A study of the computational resources available for signal processing in state-of-the-art commercial hearing aids has been carried out. The result of this study has been used to limit the computational cost of the proposed algorithm.

- A generalization of the least squares estimator, the GLSE, is introduced. The estimator is adapted to use any transformation of the input features.

- A novel set of features to estimate the mask associated to each time-frequency point has been proposed. The main novelty resides in the fact that the information of neighbor time-frequency points is included as input features. Additionally, two different alternatives to introduce the information regarding to the previous time frames have been proposed: the use of instantaneous values and the use of different EWMA.

- A low-cost algorithm for single-channel speech enhancement in monaural hearing aids has been proposed. The algorithm aims at maximizing the output PESQ score with a tailored optimization algorithm that uses a previous estimation of the proposed Wiener soft mask, which is estimated with the GLSE. In order to reduce the computational cost, a feature selection algorithm based on evolutionary computation has been also proposed. The results obtained with the algorithm have shown good intelligibility increase using a small part of the available computational resources.

- A relationship between the PESQ score and the SNR has been obtained, using the NOIZEUS database.

The contributions of this chapter have originated the publication [Ayllón et al., 2013b].

Chapter 5

Design of speech enhancement algorithms for binaural hearing aids

5.1 Introduction

Many hearing-impaired people, basically the elderly, have bilateral hearing loss and they are forced to wear two devices. When hearing aids are worn at both ears, these devices usually operate independently. However, there is a new trend of binaural hearing aids that connects both devices in order to exchange information between them. Binaural hearing provides considerable benefits over using a single ear, incrementing the ability of localizing sounds and consequently improving the speech intelligibility. Binaural listening takes advantage of the use of the so-called spatial cues. The interaural time differences (ITDs) and interaural level differences (ILDs) are two of the most important spatial cues for the estimation of the source azimuth angle, that is the location of the interesting sound source in the horizontal plane, which is the main priority for hearing aid users. ITD basically refers to the difference in arrival time of a sound between two ears. For instance, if the sound source is nearer to the right ear than to the left one, then the signal entering the right ear will arrive sooner than the one entering the left ear. Similarly, ILD is related to the different attenuation that the signals at both ears suffer from, depending on the location of the sound source. Binaural hearing aids obviously require a communication link between both hearing devices. The simplest solution would be to connect them by using a wire. However, most users do not like this approach because of the non-aesthetic aspect of the wire linking both hearing aids from one ear to the other. This enforces to use a wireless link between both devices, what unavoidably increases the power consumption and, consequently, reduces the battery life, one of the most important limiting factors for implementing signal processing algorithms on digital hearing aids. Quantizing the parameters to be transmitted will cut down the number of bits used to represent such parameters to be transmitted (which are related in some sense to the power consumption), but at the expense of degrading the ability of the system to enhance the desired speech. This fact opens a new problem: how to reduce the bit rate transmitted between both devices without decreasing the performance of the speech enhancement algorithm. In this respect, data coding can be applied to the signals before transmission in order to reduce the data rate, which includes a number of additional design tradeoffs such as the transmission latency that can be tolerated, and the acceptable decrease in battery life in the hearing aid.

The goal in this chapter is the design of low-cost speech enhancement algorithms that increase the energy efficiency of the wireless-communicated binaural hearing aids, improving the performance in comparison to monaural systems. In this context, there are two problems to solve. The first problem is the design of low-cost speech enhancement algorithms for binaural hearing aids, which are designed combining time-frequency masking and supervised machine learning techniques. The second problem is to increase the energy efficiency of the wireless-communicated binaural hearing aids for speech enhancement. The proposed approach is based on quantizing some of the parameters to be transmitted (and avoiding the transmission of others found to be unnecessary), the number of quantization bits being computed by means of evolutionary computation techniques aiming at finding a balance between low bit rate and good speech enhancement.

5.2 Designing speech-centered separation systems with quantized transmission for speech enhancement in binaural hearing aids

The assumed acoustic scenario and the binaural speech enhancement system proposed in this chapter are represented in figure 5.1. In this scenario, the hearing aid user who wears two hearing aids wants to understand the speech produced by an interlocutor. Assuming that the user is looking at the desired interlocutor, the sound arriving at both devices is a mixture of the desired source coming from the straight ahead direction (the green circle) and a combination of undesired sound sources coming from other directions (the gray cloud). The origin of the undesired sources may vary: different speakers, babble, music, traffic noise, TV, etc. Hence, the signals entering the left and the right hearing aids, $x_L(k,l)$ and $x_R(k,l)$, can be expressed as

$$\begin{aligned} x_L(t) &= s_L(t) + n_L(t) \\ x_R(t) &= s_R(t) + n_R(t), \end{aligned} \tag{5.1}$$

where $s_L(t)$ and $s_R(t)$ are the signals coming from the target source that arrive at the left and right hearing aids respectively, and $n_L(t)$ and $n_R(t)$ represent the combination of undesired sources (noises coming from other directions) entering the left and right hearing aids respectively. The filterbank of each device computes the STFT of each frame, obtaining $X_L(k,l)$ and $X_R(k,l)$ for the left and right ear, respectively. The amplitude (in dB) of the STFT is represented by $A_L(k,l)$ and $A_R(k,l)$ for the left and right hearing aids respectively, and it is calculated according to

$$\begin{aligned} A_L(k,l) &= 20\log_{10}|X_L(k,l)| \\ A_R(k,l) &= 20\log_{10}|X_R(k,l)|. \end{aligned} \tag{5.2}$$

The use of the logarithmic transformation of the squared amplitude provides more meaningful information from the human hearing point of view. The phase of the STFT is represented by $\phi_L(k,l)$ and $\phi_R(k,l)$ for the left and right hearing aids respectively.

The speech enhancement algorithms proposed in this chapter are based on the computation of an optimized binary mask using the information of the two binaural mixtures. It has already been discussed that the computation of the binary mask can be performed attending to different criteria, basically divided in two: the maximization of metrics correlated with intelligibility, or the estimation of the IBM defined in CASA, which has been proved to correlate with intelligibility [Brungart et al., 2006; Li and Loizou, 2008; Loizou and Kim, 2011]. In order to preserve the binaural cues, the mask must be the same for the left and the right devices. The mask is only calculated in one of the devices and transmitted to the other one, thus reducing the

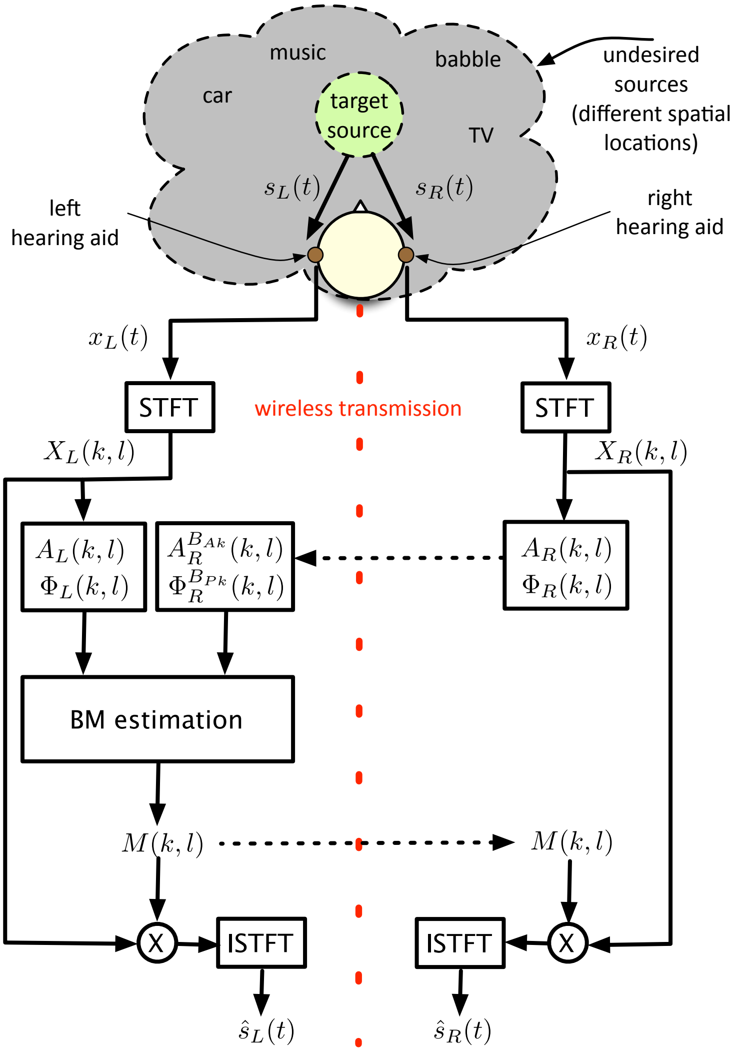


Figure 5.1: Binaural speech enhancement system overview.

computational load in one of the devices. In the schema shown in figure 5.1, the right device transmits the amplitude and phase of the STFT of its received signal, $A_R(k,l)$ and $\phi_R(k,l)$, to the left device, which calculates de binary mask $M(k,l)$ and transmits it to the right device. Once both devices have the same mask, they apply it to the STFT of their received signals and compute the ISTFT to obtain a clean version of the original target source, which is directly played in the loudspeakers of the hearing devices. The number of bits transmitted can be reduced by transmitting a quantized low bit version of $A_R(k,l)$ and $\phi_R(k,l)$, instead of their values themselves. The transmitted quantized version of $A_R(k,l)$ and $\phi_R(k,l)$ are labeled as $A_R^{B_{Ak}}(k,l)$ and $\phi_R^{B_{Pk}}(k,l)$, where B_{Ak} and B_{Pk} are the number of bits used to quantize the k-th frequency band of the amplitude and phase values, respectively, and $B_k = B_{Ak} + B_{Pk}$ is the total number of bits. The quantized values from the right device and those directly computed by the left device, $A_L(k,l)$ and $\phi_L(k,l)$, are used by the left device to calculate the binary mask $M(k,l)$. Due to the fact that the binary mask, which is transmitted from the left to the right device, only contains values of 0 and 1, it is coded with only 1 bit, hence K bits are transmitted for each frame. The key point of the system proposed in this chapter is that the values $A_R(k,l)$ and $\phi_R(k,l)$ of each frequency band are quantized with a different number of bits B_{Ak} and B_{Pk}, limiting the total number of bits transmitted for each frame. The assignation of the number of bits to the different frequency bands is carried out by optimizing the performance of the speech enhancement system, avoiding to transmit unnecessary information.

The proposed transmission schema only makes sense when the latency of the system allows a delay higher than the transmission time plus the processing time. The system can also be implemented symmetrically, for instance, transmitting the information of half of the frequency bands from the left to the right device and the other half from the right to the left device. In this case, each device would calculate half of the mask and it would transmit it to the other device. For the sake of simplicity, the schema in figure 5.1 is adopted in this work, considering that the proposed algorithms are also valid for the symmetric schema. Finally, it is worth clarifying that the data transmission is not continuous: first, the amplitude and phase information is transmitted from the right to the left device, and after the processing time, the mask is transmitted from the left to the right device. This fact allows transmitting at the maximum bit rate available in the device (around 300 kbps in commercial devices) but only during a small part of the processing time of each frame.

The algorithms proposed in this chapter are divided into two approaches, in both cases considering the limited computational resources of the hearing aids. The first approach is inspired in the DUET algorithm and it is based on the computation of a time-frequency binary mask that maximizes the WDO factor obtained for the separation of the desired source. It has been previously described how the WDO factor is correlated with the intelligibility of the sources separated via binary masking (see section 2.5.1.3). The estimation of each time-frequency point of the binary mask is performed using the quantized version of the ILD and ITD of the current time-frequency point. The use of the information of a single point allows maximizing the WDO factor independently for each frequency band. Once the separation system is designed for each frequency band, the bit rate is constrained and the number of bits transmitted for each frequency band are optimized. The second approach is based on the estimation of a time-frequency mask using the information of neighbor time-frequency points. This fact forces to design the binary mask for all frequencies at once. Due to computational reasons, the procedure followed in this second approach differs from the previous one. The IBM defined in CASA is estimated with a low-cost linear classifier that uses an extended set of features considering the information of neighbor time-frequency points. The optimization is performed in terms of MSE and a weighted version of the MSE. The proposed procedure allows calculating the mask for all frequencies at once at the same time than optimizing the transmission schema.

Finally, it is worth mentioning that all the design methods described in this chapter are carried out offline on a computer. Only when the design has been completed, the optimum solution is then implemented on the digital hearing aid.

5.3 Case 1: Designing a binaural speech separation system using the information of a single time-frequency point

In this first approach, the source separation algorithm is designed to maximize the WDO factor, and, as it will be shown in the following, the optimization can be performed independently for each frequency band. The binary mask is estimated using a quadratic discriminant that uses the ILD and ITD values of the current time-frequency point, which are labeled as $L(k,l) = A_L(k,l) - A_R(k,l)$ and $P(k,l) = \phi_L(k,l) - \phi_R(k,l)$, to decide whether that point belongs either to the speech source or to any noise source. The weights of the quadratic discriminant are computed by a tailored evolutionary algorithm. Once the separation algorithm is designed, a second EA optimizes the bit distribution among frequency bands. The method is described next.

5.3.1 WDO quality factor: measuring the disjointness in a two-channel problem

The WDO factor described in section 2.4.1 is a good indicator of the quality of the separation achieved by a time-frequency binary mask for approximate W-DO sources, and for each source in the mixture, it is given by expression (2.17). In the problem addressed in this chapter, two different sources are considered, the target speech and the undesired noise which comprises all the interfering sound sources the user is not interested in. The speech is binaurally separated by

$$\begin{aligned}\hat{S}_L(k,l) &= M(k,l)X_L(k,l)\\ \hat{S}_R(k,l) &= M(k,l)X_R(k,l),\end{aligned} \tag{5.3}$$

and the noise can be separated by

$$\begin{aligned}\hat{N}_L(k,l) &= (1-M(k,l))X_L(k,l)\\ \hat{N}_R(k,l) &= (1-M(k,l))X_R(k,l).\end{aligned} \tag{5.4}$$

Let us define $p_s(k,l) = S_L(k,l)^2 + S_R(k,l)^2$ and $p_n(k,l) = N_L(k,l)^2 + N_R(k,l)^2$. According to expression (2.17), the WDO factor associated to the separation of the speech source can be expressed as

$$WDO = \frac{\sum\limits_{(k,l)} M(k,l)(p_s(k,l) - p_n(k,l))}{\sum\limits_{(k,l)} p_s(k,l)}. \tag{5.5}$$

This expression can be rewritten as

$$WDO = \sum_{(k,l)} M(k,l)K(k,l) \tag{5.6}$$

where $K(k,l) = (p_s(k,l) - p_n(k,l))/P_S$ and $P_S = \sum\limits_{(k,l)} p_s(k,l)$ is a constant values for a given mixture.

The WDO quality factor deduced in expression (5.6) is very intuitive and plays an important conceptual role in this work, due to the fact the the algorithm described in the next section aims to maximize its value. The following key conclusions are derived from expression (5.6):

1. The value of WDO in (5.6) for a given mixture can be evaluated using the IBM, that is, $M(k,l) = 1$ when the term $K(k,l)$ is positive ($p_s(k,l) > p_n(k,l)$), and 0 in other case. This value provides a valuable mathematical insight since it is not possible for any feasible, implementable quantized-based WDO approach to be higher than this limiting WDO. This value is known as WDO 0-dB in the literature [Yilmaz and Rickard, 2004].

2. The proposed separation system can be designed and implemented in an independent way for each frequency band, then the problem of maximizing its corresponding WDO

function can be easily decomposed in so many problems as frequency bands, making easier the optimization of the separation system.

3. For practical implementation, the elements taking part in the WDO function can be quantized, and, by properly selecting the number of bits in this quantization, reducing the bit rate. Depending on the method to compute the proper number of quantization bits it will be possible to reach different implementable realizations.

The details of the algorithms to design an energy efficient wireless-communicated binaural hearing aid separation system that maximizes the WDO are described in the next sections.

5.3.2 Quadratic discriminant based separation system with constrained transmission bit rate

The proposed acoustic scenario assumes that the target speech source is located in the front direction and the undesired noise sources can be placed on any other different location. As explained and represented in figure 5.1, this is a realistic hypothesis since the hearing-impaired person wearing the two hearing aids will be looking at his/her interlocutor in the aim of enhancing speech understanding. This is the reason why those STFT components $X_L(k,l)$ and $X_R(k,l)$ that are mostly influenced by the speech source will lead to values of $L(k,l)$ and $P(k,l)$ close to zero, while those components basically caused by the other sounds coming from the other directions will lead to $L(k,l)$ and $P(k,l)$ values far from zero. With this scenario in mind, it is proposed here to implement a separation system that can be seen as a non-linear spatial filter of the input signals. Taking into account that the optimization of the WDO function can be done independently for each of such frequency bands, and that the values of the $L(k,l)$ and $P(k,l)$ will have different behaviors for each band, the optimization problem can be solved independently for each band.

Bearing in mind that the optimal boundary will have the shape of a closed space centered in zero, using the measurements $L(k,l)$ and $P(k,l)$, the objective is to determine whether a given sample of the STFT belongs to either the speech source or to the other noise sources. For this purpose, the next quadratic discriminant is proposed

$$y(k,l) = v_{k0} + v_{k1}L(k,l) + v_{k2}L(k,l)^2 + v_{k3}P(k,l) + v_{k4}P(k,l)^2, \tag{5.7}$$

where $y(k,l)$ is the output of the discriminant, $L(k,l)$, $L(k,l)^2$, $P(k,l)$ and $P(k,l)^2$ label its inputs, and $\mathbf{v}_k = [v_{k0}, v_{k1}, v_{k2}, v_{k3}, v_{k4}]$ are the frequency-dependent coefficients whose value has to be computed. If $y(k,l) > 0$ then $M(k,l) = 1$, and the corresponding STFT samples $X_L(k,l)$ and $X_R(k,l)$ are mostly influenced by the speech source; if this is not the case, then $M(k,l) = 0$, and the corresponding STFT samples are assumed to be mostly affected by the noise source so that they should be attenuated (in the effort of speech enhancement).

The computational cost associated to the current classifier is very low, since it only needs to perform 5 MAC operations. Considering that the selected data is consecutively stored in memory, and the processor performs the MAC operation in a single instruction, the number of instructions necessary to process each frequency band is approximately IPF= 5, excluding the computation of the input features.

5.3.2.1 Computation of the quadratic discriminant coefficients

Designing the quadratic discriminant separation system demands to compute the adequate values for $\mathbf{v}_k$ so that the system properly discriminates both sources. In this work, a tailored evolutionary algorithm is proposed for this purpose. In the problem at hand, for each frequency band and each proposed combination of B_{Ak} and B_{Pk} (the number of quantization bits for amplitude and phase values, respectively), a single EA is used to compute those values of the coefficients $\mathbf{v}_k$ that maximize the quality function WDO (the objective function here) obtained in the separation process. Any candidate solution is thus a coefficient vector with the structure $C_q = [v_{q0}, v_{q1}, v_{q2}, v_{q3}, v_{q4}]$. The goal of the EA is to find the best candidate that maximizes the WDO fitness function in expression (5.6), computed with the corresponding magnitudes quantized with B_{Ak} bits for the level difference and B_{Ak} bits for the phase difference. The steps of the EA, which is run for each proposed combination of B_{Ak} and B_{Pk}, are listed below.

1. For each frequency band and each combination of B_{Ak} and B_{Pk}, it generates an initial population with candidate solutions. The elements of the candidate solution are the weights $\mathbf{v}_k$ of the quadratic discriminant. After a number of experiments, the initial population size has been chosen as a tradeoff between computational complexity and performance. This size of the population has been found to be 200, and their individuals are generated at random.

2. After evaluating the fitness function for each candidate in the population, it selects a subpopulation of 20 candidate individuals (the 10% of the population) that best fit the WDO fitness function. These elite candidates are those that will survive to the next generation.

3. To recombine the parents by using the uniform crossover operator. Thus, a number of 180 novel candidates for the next generation are generated by random crossover with the previously selected 20 best candidates.

4. To mutate or randomly change the offspring, excluding the previous best candidate. Empirically, it has been observed that a mutation probability value of $P_M = 0.01$ gives good results. The mutation operator consists in adding a Gaussian value with variance 0.01 to the five weights of the mutated candidates.

5. To iterate steps 2-4 until a maximum number of generations (200) is reached or if the highest WDO value remains unchanged for a given number of iterations (20). These values have been found to be large enough to allow the algorithm to properly converge in the experiments.

Note that the best solution that maximizes the WDO fitness function over the design set is ulteriorly selected as the final solution for such frequency band and combination of B_{Ak} and B_{Pk}. The complete process, from step 1 to step 5, is iterated until all the frequency bands and combinations of proposed values of B_{Ak} and B_{Pk} are investigated.

Once the EA has been independently executed for all the combinations of B_{Ak} and B_{Pk}, the best combination of both values for each frequency band is obtained. Since the number of possible combinations for each frequency band is not excessive, and the computation time is low (the cost function is the WDO and it has already been computed), an exhaustive search has been carried out. The best combination is searched for different values of the total number of bits per frequency ($B_k = B_{Ak} + B_{Pk}$). This second step is performed due to the fact that the maximization of the WDO can be done independently for each frequency band.

This two-steps algorithm is labeled as $Evolutionary^1$.

5.3.2.2 Optimization of the bit distribution

In the algorithm proposed in the previous section, the problem of designing a speech separation system has been addressed separately for each frequency band and for each combination of number of bits. Now the approach is refined a little more in the sense that a more advanced version that aims at optimizing the bit distribution B_k among frequency bands is implemented. Obviously, the only design restriction remains still the total number of quantization bits, or equivalently, the required transmission rate that the algorithm tries to minimize to make the separation system more energy efficient.

To put into practice this global design, two evolutionary algorithms working in cascade are designed. For the sake of clarity when discussing the results achieved with the different approaches explored in this work, the complete double-EA-based method, has been named $Evolutionary^2$. It basically operates in this way:

- The first EA ($Evolutionary^1$) aims at designing the shape of the separation boundaries (or equivalently, the separation system itself): it computes those values for the coefficients $\mathbf{v}_k$ that optimize the WDO. It operates as the one described in the previous section.
- The second EA in the cascade is other EA that, by constraining the design to a small global number of bits per sample (for energy efficiency purposes), provides how many of these total number of bits per sample must be used for each frequency band (B_k), without degrading (and even enhancing) the potential of the system to properly separate speech.

This latter EA, not yet described, operates as listed:

1. It generates an initial population with a size of 2000 random candidate solutions. Each candidate solution contains K random integer values, which possible values are the values of B_k used in $Evolutionary^1$.
2. After evaluating the performance of each candidate in the population (i.e. evaluate the WDO factor), it selects the best subpopulation of 200 (10%) solutions. These elite candidates are those that will survive to the next generation, the remaining being thus removed.
3. To recombine the parents by using the uniform crossover operator. A number of 1800 novel candidates for the next generation are generated by random crossover with the previously selected 200 best candidates.
4. To mutate or randomly change the offspring, excluding the previous best candidate, with a mutation probability of $P_M = 0.01$. The mutation operator consists in adding a Gaussian value with unit variance, rounded towards nearest integer, to all the chromosomes of the mutated candidates.
5. The process iterates steps 2-4 until a maximum number of generations (500) is reached or if the highest WDO value remains unchanged for a given number of iterations (50). These values have been found to be large enough to allow the algorithm to properly converge in our experiments.

In this case the population is greater than in the first EA, since the number of parameters to be optimized is considerably greater. Furthermore, in those cases in which a large number of frequency bands is used (therefore more parameters must be optimized), a greater population should be more suitable.

5.3.3 Experimental work and results

5.3.3.1 Description of the experiments

The suitable database design plays a vital role in any kind of problem based on supervised machine learning. In order to validate the algorithms proposed in this chapter, a database of binaural speech and noise mixtures has been generated to design and test the algorithms. The speech sources have been selected from the TIMIT database, which contains a total of 626 speech male/female recordings with a duration of 4 seconds. Another 626 noise sources have been selected from an extensive database which contains both stationary and non-stationary noise. Stationary noise refers to homogeneous noisy environments, for instance, the aircraft cabin noise. Non-stationary noise to other non-homogeneous noises, for example, children shouting in a kindergarten. The noise database contains a variety of noise sources, including those from the following diverse environments: aircraft, bus, cafe, car, kindergarten, living room, nature, school, shops, sports, traffic, train, train station, etc. All the speech and noise signals have been initially normalized with power level of 0 dB.

A number of 1000 binaural mixtures are generated using the HRIR functions included in the CIPIC database. The mixtures are generated with the following setup: a speech source is placed in the front position (i.e. 0° in azimuth, 0° in elevation), and a different noise source is placed at each side of the head (there are two noise sources). The speech and noise sources are randomly selected from the TIMIT and noise databases respectively, and the positions of the noise sources are randomly selected among the positions defined in the CIPIC database, avoiding the front direction. The HRIR used in each mixture is also randomly selected among the HRIR functions of the different subjects contained in the CIPIC database. The database is generated with SNRs of 0, 3, and 5 dB, considering that the noise power is the addition of the power of both noise sources. Considering that $s(t)$ is the target speech signal and $n_1(t)$ and $n_2(t)$ are the two noise sources respectively, the binaural mixture is given by

$$\begin{aligned} x_L(t) &= s_L(t) + n_{1L}(t) + n_{2L}(t) \\ x_R(t) &= s_R(t) + n_{1R}(t) + n_{2R}(t), \end{aligned} \tag{5.8}$$

where the signals at the left and right ear (i.e. $s_{L/R}(t)$, $n_{1L/R}(t)$ and $n_{2L/R}(t)$) are obtained by filtering the original sources with the corresponding HRIR function. For properly designing and testing the speech separation system, the database is split into two different subsets, one for design and another for test. The design set contains the 70% of the 626 signals, and the test set the remaining 30%. It is very important to emphasize that the test sounds are not used in the design process. The sampling rate is 16 kHz and the signals are transformed into in the time-frequency domain with a STFT that uses a 128-points Hamming window with 50% of overlap, which corresponds with $K = 64$ frequency bands (DC component is not processed). The target IBM is calculated according to

$$T(k,l) := \begin{cases} 1, & |S_L(k,l)|^2 + |S_R(k,l)|^2 > |N_{1L}(k,l) + N_{2L}(k,l)|^2 + |N_{1R}(k,l) + N_{2R}(k,l)|^2 \\ 0, & \text{otherwise} \end{cases}. \tag{5.9}$$

Once the set up of the experimental work has been summarized, the sequence of experiments is the next:

1. As stated in expression (5.6), the separation problem can be analyzed independently for each frequency. So, in this group of experiments, the separation solutions for each fre-

quency band are implemented and evaluated. With this aim, the algorithm $Evolutionary^1$ is executed varying the values of the quantization bits B_{Ak} and B_{Pk} from 1 to 8.

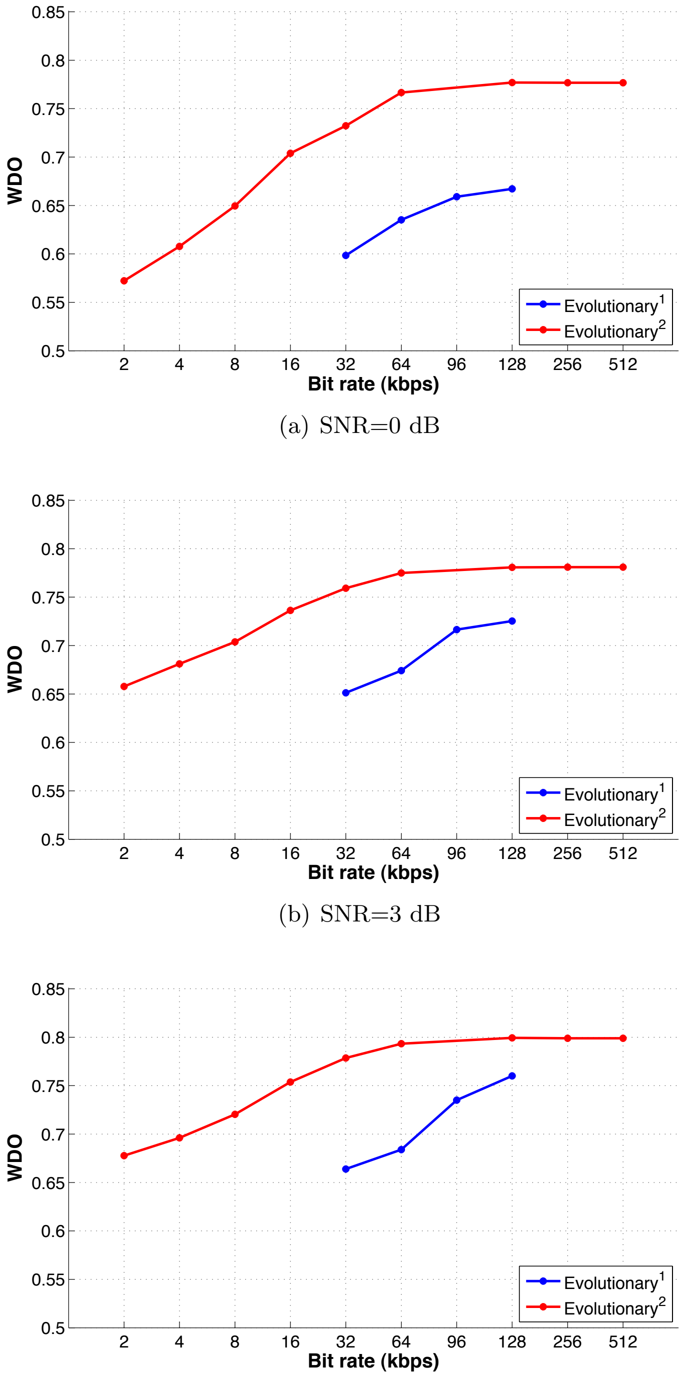
2. The second batch of experiments focuses on combining those separation systems designed in the first set of experiments, and aims at implementing the best complete separation system with a constrained maximum bit rate (or, equivalently, bits/sample), in the effort of elucidating how many of the global number of bits per sample should be used for quantizing the information of each frequency band. With this purpose, the $Evolutionary^2$ algorithm is executed, varying the transmission bit rate from 2 to 512 kbps. The transmission of 512 kbps is equivalent to quantify the amplitude and phase values with 8 bits (i.e. a total of 16 bits per frequency band).

5.3.3.2 Discussion of the results

Figure 5.2 represents the WDO values averaged over the test set in the database as a function of the number of bits transmitted from the left to the right device. The blue line corresponds with the WDO values obtained by the $Evolutionary^1$ algorithm and the red line the WDO values obtained by the $Evolutionary^2$ algorithm. The plot in (a) corresponds with a SNR of 0 dB, the plot in (b) with a SNR of 3 dB, and the plot in (c) with a SNR of 5 dB. The results obtained by the separation algorithm ($Evolutionary^1$) are good in terms of WDO, and, what is more interesting, its performance does not degrade excessively when the number of quantization bits decreases (i.e. the bit rate). For instance, in the case of SNR=0 dB (worst case), the system achieves an average WDO value of 0.67 transmitting at 128 kbps, and the value is only reduced to 0.60 (a 10 %) when the number of transmitted bits decreases to 32 kbps. Nevertheless, the improvement introduced by the $Evolutionary^2$ algorithm is amply noticeable for any SNR, but is even more noticeable in the case of SNR=0 dB, which is the most difficult case. In this case, for instance, the WDO value of 0.67 obtained by the $Evolutionary^1$ algorithm transmitting at 128 kbps is increased to 0.80 by the $Evolutionary^2$ algorithm transmitting at the same bit rate. And, what is even more important, the second algorithm allows for the design of a separation system that works at low bit rates, what makes it more energy efficient. The WDO values achieved with the maximum transmission bit rate (512 kbps) are still practically matched when the bit rate decreases down to 64 kbps. For lower bit rates, the WDO values start to decrease, but good separation results are obtained even only transmitting 2 kbps. The WDO values achieved when transmitting 512 kbps are decreased by 7% in the case of transmitting 16 kbps, and 18% in the case of transmitting 2 kbps, on average for the different SNRs.

5.4 Case 2: Designing a binaural speech separation system using the information of neighbor time-frequency points

The second approach proposed in this chapter uses a generalized version of the least squares linear discriminant analysis (LS-LDA)[Ye, 2007] which is trained to estimate the IBM defined in (2.18), using a novel set of features extracted from the current and neighbor time-frequency points of the signals received at both devices. The IBM is not necessarily the same for the left and the right devices. However, in order to preserve the binaural cues, it is assumed that the same mask is applied in the right and the left devices, and the IBM is calculated using the energy of the signals of both devices. In a first step, the classifier is designed assuming that all information has been exchanged between both devices (i.e. it uses the non-quantized version



(a) SNR=0 dB

(b) SNR=3 dB

(c) SNR=5 dB

Figure 5.2: WDO values averaged over the test set as a function of the number of bits transmitted from the right to the left device. The blue line corresponds with the WDO values obtained by the *Evolutionary*^{1} algorithm and the red line the WDO values obtained by the *Evolutionary*^{2} algorithm. The plot in (a) corresponds with a SNR of 0 dB, the plot in (b) with a SNR of 3 dB, and the plot in (c) with a SNR of 5 dB.

of $A_R(k,l)$ and $\phi_R(k,l)$), comparing different sets of proposed features. Unlike the previous approach, not only the ILD and the ITD are considered as input features, but also another different combinations of features are evaluated. The main novelty resides in the fact that the classifier considers not only the information of STFT of the current time-frequency point, but also the information related to the neighbor time-frequency points. This fact does allow to design the system independently for each frequency band, and the optimization is performed for all frequencies at once. Once the best set of features has been selected, a tailored evolutionary algorithm is designed to optimize the amount of information exchanged between both devices maximizing the performance of the speech enhancement algorithm at the same time.

5.4.1 Estimation of the IBM with a least squares generalized discriminant analysis (LS-GDA)

The computational cost associated to the estimation of the IBM must be relatively low, according to the low computational power available in hearing aids. In this approach, a low-cost classifier is proposed to decide whether a time-frequency point belongs to speech or noise, thus generating the time-frequency binary mask. The classifier uses a set of features extracted from the STFT of the left and right mixtures (each feature is a combination of the values $A_L(k,l)$, $\phi_L(k,l)$, $A_R^{B_{Ak}}(k,l)$ and $\phi_R^{B_{Pk}}(k,l)$).

The linear discriminant analysis (LDA) [Fisher, 1936] is a supervised pattern recognition method that uses a linear combination of a set of input features in order to tackle a classification problem, establishing linear decision boundaries between two or more classes. Let us consider the pattern vector $\mathbf{x_i}$ (i.e. the observations) containing P input features, $\mathbf{x_i} = [x_1, x_2, \ldots, x_P]^T$, which are extracted from the mixture signal in the problem at hand. Each pattern $\mathbf{x_i}$ can be assigned to one of the two possible classes defined in this work, speech or noise. The pattern matrix $\mathbf{P}$ of size PxL is defined as a matrix that contains the patterns $\mathbf{x_i}$ of a set of L data samples, $\mathbf{P} = [\mathbf{x_1}, ..., \mathbf{x_L}]$, and the matrix $\mathbf{Q}$ is defined as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{1} \\ \mathbf{P} \end{bmatrix}, \tag{5.10}$$

where $\mathbf{1}$ is a row vector of length L. The output of the LDA is obtained as a linear combination of the input features, according to

$$\mathbf{y} = \mathbf{v}^T \mathbf{Q}, \tag{5.11}$$

where the vector $\mathbf{v} = [v_0, v_1, v_2, \ldots, v_P]^T$ contains the bias v_0 and the weights applied to each of the P input features. In our specific case, there are a different vector $\mathbf{v}$ and matrix $\mathbf{Q}$ for each frequency band, which are denominated $\mathbf{v_k}$ and $\mathbf{Q_k}$ respectively, and the L data samples correspond with each of the time frames. Hence, the vector $\mathbf{y_k}$ is a vector of size 1xL containing the output of the LDA for the L input patterns in the k-th frequency band, $\mathbf{y_k} = [y(k,1), \ldots, y(k,l), \ldots, y(k,L)]^T$. For each of the patterns, the binary mask is generated according to

$$M(k,l) := \begin{cases} 1, & y(k,l) > y_0 \\ 0, & \text{otherwise} \end{cases}, \tag{5.12}$$

where y_0 is a threshold value. The output values of the classifier range from 0 to 1, so the threshold value is set to $y_0 = 0.5$.

The design of the classifier consists in finding the values $\mathbf{v_k}$ that minimize the estimation error. In supervised learning, the true values associated to each data sample are accessible, and they are used to train the classifier. These values are contained in the target vector defined as $\mathbf{t_k} = [t(k,1), t(k,2), \cdots, t(k,L)]^T$, with values of 1 in the case of speech and 0 in the case of noise. In this work, the target values are those corresponding to the IBM defined in (2.18). The estimation error is defined as the difference between the output values of the LDA (5.11) and the true values

$$\mathbf{e_k} = \mathbf{y_k} - \mathbf{t_k} = \mathbf{v_k}^T \cdot \mathbf{Q_k} - \mathbf{t_k}, \tag{5.13}$$

and the MSE for the k-th frequency band is computed according to

$$MSE_k = \frac{1}{L} \left\| \mathbf{y}_k - \mathbf{t}_k \right\|^2 = \frac{1}{L} \left\| \mathbf{v_k}^T \cdot \mathbf{Q_k} - \mathbf{t}_k \right\|^2. \tag{5.14}$$

In the least squares approach (LS-LDA) [Ye, 2007], the weights are adjusted in order to minimize the MSE. The minimization of the MSE is obtained by deriving the expression (5.14) with respect to every weight of the linear combination, giving raise to the following expression:

$$\mathbf{v_k} = \mathbf{t_k} \mathbf{Q_k}^T \left(\mathbf{Q_k} \mathbf{Q_k}^T \right)^{-1}. \tag{5.15}$$

The LDA is limited to separate both classes linearly. However, it is possible to discriminate classes with more complex decision boundaries by introducing non-linear transformations of the original input features. In the general case, the matrix $\mathbf{Q}$ can be defined as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{1} \\ f_1(\mathbf{P}) \\ \cdots \\ f_{N_T}(\mathbf{P}) \end{bmatrix}, \tag{5.16}$$

where $f_1, \ldots, f_{N_T}$ are N_T transformations performed over the original input features contained in $\mathbf{P}$. The weight vector is then defined as $\mathbf{v} = [v_0, v_1, \ldots, v_{N_T \cdot P}]^T$, and it can also be obtained using expression (5.15). Henceforth, this is denominated generalized discriminant analysis (GDA), and its least squares solution is labeled as LS-GDA.

The implementation of the proposed classifier is relatively simple, its computational cost being directly related to the number of features included in $\mathbf{Q}$. Considering that the selected data is consecutively stored in memory, and the processor performs the MAC operation in a single instruction, the number of instructions necessary to process each frequency band by the LS-GDA is approximately $P+1$, where P is the number of input features (the constant number of instructions necessary to generate the mask, which is a simple comparison, are not considered here). Hence, limiting the computational cost of the classifier is equivalent to limiting the number of features used for classification.

5.4.2 Weighted LS-GDA (W-LS-GDA)

The LS-GDA classifier proposed in the previous section estimates the IBM, but estimation errors are unavoidably introduced and they decrease the obtained WDO values, which implies a decrement in the output intelligibility. In order to increase the performance of the LS-GDA classifier, a weighted MSE metric that aims at maximizing the WDO factor is defined. The weighted LS-GDA, labeled as W-LS-GDA, finds the coefficients that minimize the weighted MSE, which is described below, instead of minimize the MSE of the estimation error. This allows estimating the IBM and maximizing the WDO factor at the same time.

Let us recall the expression of the WDO factor in (5.6)

$$WDO = \sum_{(k,l)} M(k,l)K(k,l), \tag{5.17}$$

where $M(k,l)$ has values of 0 or 1. Since $K(k,l)$ was defined as $K(k,l) = (p_s(k,l) - p_n(k,l))/P_S$, it can be decomposed in its modulus and sign, according to $K(k,l) = t(k,l)|K(k,l)|$, where $t(k,l)$ are the target values, which are defined now as 1 or -1 (i.e. speech dominated or noise dominated time-frequency point, respectively). Thus, the WDO value is given by

$$WDO = \sum_{(k,l)} t(k,l)M(k,l)|K(k,l)|. \tag{5.18}$$

In order to have values of +1 and -1, the output of the classifier $y(k,l)$ is defined as

$$y(k,l) = 2M(k,l) - 1. \tag{5.19}$$

The next expressions, used to calculate the MSE, are equivalent

$$\sum_{(k,l)} (y(k,l) - t(k,l))^2 = \sum_{(k,l)} y(k,l)^2 + t(k,l)^2 - 2y(k,l)t(k,l) = \sum_{(k,l)} 2 - 2y(k,l)t(k,l), \tag{5.20}$$

and replacing $y(k,l)$ by its expression in (5.19) yields

$$\sum_{(k,l)} (y(k,l) - t(k,l))^2 = \sum_{(k,l)} 2 - 4M(k,l)t(k,l) - 2t(k,l). \tag{5.21}$$

Let us focus on the maximization of the WDO. Considering expression (5.18), the maximization is equivalent to

$$\begin{aligned} &\max_M\{WDO\} = \max_M\{\sum_{(k,l)} t(k,l)M(k,l)|K(k,l)|\} = \min_M\{\sum_{(k,l)} -4t(k,l)M(k,l)|K(k,l)|\} = \\ &\min_M\{\sum_{(k,l)} (2 - 4M(k,l)t(k,l) - 2t(k,l))|K(k,l)|\} = \min_y\{\sum_{(k,l)} (y(k,l) - t(k,l))^2|K(k,l)|\}. \end{aligned} \tag{5.22}$$

The previous mathematical manipulations show that the maximization of the WDO factor is equivalent to the minimization of a weighted MSE. Thus, the WDO maximization can be indirectly performed with a LS-GDA which minimizes a weighted version of the MSE, where the weighting function is $|K(k,l)|$:

$$\min_y\{\sum_{(k,l)} (y(k,l) - t(k,l))^2|K(k,l)|\}. \tag{5.23}$$

Introducing $\mathbf{y}_k = \mathbf{v_k}^T\mathbf{Q_k}$ and defining $\mathbf{k}_k = [K(k,1), \ldots, K(k,l), \ldots, K(k,L)]^T$, the weighted MSE, labeled as WMSE, is defined for the k-th frequency band as

$$WMSE_k = \frac{1}{L}||(\mathbf{v_k}^T\mathbf{Q_k} - \mathbf{t}_k)^2|\mathbf{k}_k||| = \frac{1}{L}||\mathbf{v_k}^T(\mathbf{Q}_k\sqrt{|\mathbf{k}_k|}) - \mathbf{t}_k\sqrt{|\mathbf{k}_k|})\}||^2, \tag{5.24}$$

where the root square operator is applied to each element of the matrix. According to this, the implementation of the proposed W-LS-GDA classifier is equivalent to the implementation of the LS-GDA proposed in the previous section, using a scaled version of the matrix $\mathbf{Q}_k$ and the targets $\mathbf{t}_k$, where the scaling factor is $\sqrt{|\mathbf{k}_k|}$.

5.4.3 Evolutionary algorithm to optimize the transmission bit rate

The low-cost classifiers (LS-GDA and W-LS-GDA) proposed in the previous sections provide an estimation of the IBM minimizing a MSE function. Both classifiers use a set of features calculated from the signals received at both ears, which implies that all the information is transmitted from the right device to the left one. Unfortunately, this is not an energy efficient system. The second step in the design of the binaural speech enhancement system proposed in this section is the reduction of the transmission bit rate, which implies a reduction in the power consumption, while minimizing the effect that quantization has in the output speech quality obtained by the speech enhancement system. Similarly than in the previous approach, the transmission rate is optimized by assigning a different number of bits B_{Ak} and B_{Pk} to quantize the values $A_R(k,l)$ and $\phi_R(k,l)$ of each frequency band. The number of bits B_{Ak} and B_{Pk} may also differ in the same frequency band. This transmission schema allows assigning more bits to the frequencies and values providing more information to the classifier.

In order to optimize the bit distribution, a tailored evolutionary algorithm is proposed, considering that the number of bits associated to the transmission of the data of each time frame (i.e. the bit rate) is constrained. The algorithm searches the best assignation of bits among frequency bands in order to minimize the fitness function, which is the MSE in the case of the LS-GDA classifier and the WMSE in the case of the W-LS-GDA classifier. The matrix $\mathbf{Q}$ is created including the selected set of features (that will be selected in the next section) calculated with the values $A_R^{B_{Ak}}(k,l)$ and $\phi_R^{B_{Pk}}(k,l)$ quantized with different number of bits B_{Ak} and B_{Pk}, considering all integer values from 0 to 8. The value $B_{Ak} = 0$ (or $B_{Pk} = 0$) means that no information from this value in the k-th frequency band is transmitted. Hence, the rows of $\mathbf{Q}$ contain the features quantized with different number of bits. The values $A_R^{B_{Ak}}(k,l)$ and $\phi_R^{B_{Pk}}(k,l)$ received by the left device are simulated by quantizing uniformly the values using $2^{B_{Ak}}$ and $2^{B_{Pk}}$ quantization steps, respectively. The dynamic range has been limited to 90 dB for the amplitude values (A_L and A_R are logarithmic values) and 2π for the phase values.

Each candidate solution is defined by a vector containing the number of bits (between 0 and 8) assigned to the level and phase values ($A_R(k,l)$ and $\phi_R(k,l)$) of each frequency band, a total of $2K$ values (K is the number of frequency bands). The search algorithm selects the quantized features among the rows of the matrix $\mathbf{Q}$ according to the bits of each candidate solution, and then evaluates the classifier using the quantized features. The complete steps of the search algorithm are the next:

1. The matrix $\mathbf{Q}$ is created containing the selected set of features calculated with the values $A_R^{B_{Ak}}(k,l)$ and $\phi_R^{B_{Pk}}(k,l)$ quantized with different number of bits, $B_k = 0, 1, \cdots, 8$.

2. An initial population of 100 candidate solutions is generated. Each solution contains $2 \cdot K$ random values between 0 and 8, which corresponds with a different number of bits for $A_R(k,l)$ and $\phi_R(k,l)$ for each frequency band.

3. The candidates of the population are validated to fulfill the constraint of total number of bits. If a candidate solution exceeds by N_D the maximum number of bits allowed, the number of bits of a number of N_D random positions of the candidate solution are decreased by one. In case that the number of bits of an element falls below 0, it is set to 0. The procedure iterates until the candidate solution fulfills the requirement.

4. The fitness function of the classifier is then evaluated for each candidate solution and frequency band, following the next steps:

(a) To extract the quantized version of the features from $\mathbf{Q}$, according to the current candidate solution.

(b) To weight the matrix $\mathbf{Q}_k$ and the targets $\mathbf{t}_k$, in case of W-LS-GDA.

(c) The weight values $\mathbf{v}_k$ are calculated for each frequency band, using expression (5.15).

(d) The MSE (or WMSE) of each solution and frequency band is calculated according to expressions (5.14) (or (5.24)).

(e) The MSE (or WMSE) associated to a candidate solution is the average of the MSE (or WMSE) obtained in all frequency bands.

5. A selection process is applied, using the MSE (or WMSE) of each solution as ranking. It consists in selecting the best 10% of the solutions of the population, removing the remaining solutions.

6. The remaining 90% solutions of the new generation are then generated by uniform crossover of the best candidates.

7. Mutations are applied to the 1% of the new population ($P_M = 0.01$), excluding the best obtained solution which is preserved. Mutations consist of increasing or decreasing by one the number of bits of random positions of the mutated candidate solution.

8. The process is repeated from step 3 to 7 until 100 generations are evaluated. Since the best solution of each iteration is not modified, the best solution obtained in the last iteration is considered the best solution.

The values of the parameters of the EA (population size, crossover rate, mutation scheme and number of generations) have been found to obtain a quite good tradeoff between design time and performance for the experiments carried out in this section.

5.4.4 Experimental work and results

The experiments and results described in this section are divided into two groups, according to the two stages of the algorithm described in this section. The first group of experiments corresponds with the experiments carried out to select the best input feature space, whereas the second group of experiments involves the optimization of the transmission rate, using the set of features selected in the previous step. All the experiments have been carried out with the database described in section 5.3.3.1.

5.4.4.1 Selection of the input feature space

In this section, different combinations of input features extracted from the STFT of the right and left signals are used, and its performance is evaluated using the proposed LS-GDA classifier. According to the limited computational resources of hearing aids, the most suitable feature set will be selected considering a tradeoff between the speech quality obtained by the enhancement system and the computational burden associated to the use of the selected feature set. The study in this section is carried out considering that the values $A_L(k,m)$, $\phi_L(k,m)$, $A_R(k,m)$ and $\phi_R(k,m)$ are available in the left device, that is, without taking into account quantization. Six different set of features are proposed, which are included in table 5.1. Each feature of the set is obtained by applying different transformations to the original amplitude and phase values, including squared amplitude, amplitude and phase differences, and amplitude product. The

Table 5.1: Proposed combination of features: A_L and ϕ_L are the amplitude and phase of the left ear respectively; A_R and ϕ_R are the amplitude and phase of the right ear respectively.

SET	$NFtSet$	Features
SET1	3	A_L, $(A_L - A_R)^2$, $(\phi_R - \phi_L)^2$
SET2	3	A_L, $abs(A_L - A_R)$, $abs(\phi_R - \phi_L)$
SET3	4	A_L, A_L^2, $(A_L - A_R)^2$, $(\phi_R - \phi_L)^2$
SET4	2	$(A_L - A_R)^2$, $(\phi_R - \phi_L)^2$
SET5	7	A_L, A_R, A_L^2, A_R^2, $A_L \cdot A_R$, $abs(A_L - A_R)$, $abs(\phi_R - \phi_L)$
SET6	6	A_L, A_L^2, $abs(A_L - A_R)$, $(A_L - A_R)^2$, $abs(\phi_R - \phi_L)$, $(\phi_R - \phi_L)^2$

different sets contain a variety of number of features, which is labeled as $NFtSet$, and it ranges from 2 to 7. Although quantization is not considered here, more importance is given to the features extracted from the left signal, which will be not quantized in the final system.

The classification of a time-frequency point into speech or noise can be performed using one of the proposed set of features, where the values are calculated from the STFT of the current time-frequency point. Additionally, it is proposed to include further information related to the neighbor time-frequency points, also calculating the proposed set of features from the STFT of these points. In this work, a constant pattern of features is considered, using a T-shaped time-frequency footprint, as it is shown in figure 5.3. The value $Nfreqs$ represents the number of neighbor frequencies taken in each direction (upper frequencies and lower frequencies), then $2Nfreqs$ is the total number of neighbor frequencies included. The number of previous time frames considered is $Nframes$. Hence, the total number of features P used by the classifier, which depends on the selected set, is given by

$$P = NFtSet(2Nfreqs + Nframes + 1). \tag{5.25}$$

The experiments carried out in this section have two objectives: first, the selection of the best set of features among the 6 proposed in table 5.1 and second, the selection of the optimum time-frequency footprint, finding the best values for $Nfreqs$ and $Nframes$. The two problems are solved separately in two different experiments described below.

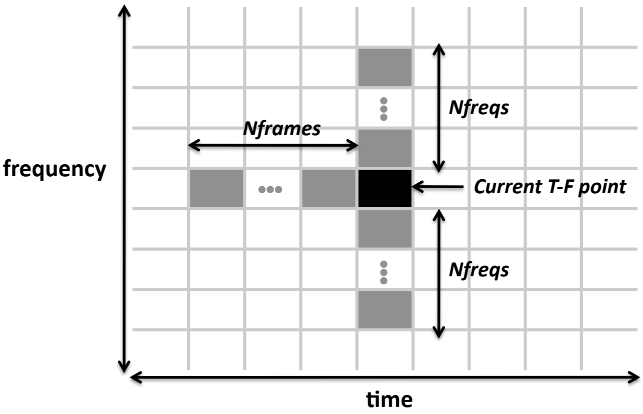


Figure 5.3: T-shaped time-frequency footprint.

5.4.4.1.1 Selection of the best set of features

The 6 different sets of features are evaluated using a time-frequency footprint with the same number of neighbor frequencies and time frames, $Nfreqs = Nframes$. The values of $Nfreqs$ (and $Nframes$) are varied from 0 to 10, which allows evaluating also the case of using only the information of the current time-frequency point (i.e. $Nfreqs = Nframes = 0$). The comparison is performed in terms of the WDO value, averaged over the test set, obtained by the separation algorithm when the classifier uses each set of features, using the database of speech and noise mixtures previously described, with SNR of 0, 3, and 5 dB . The steps carried out in this experiment are the next:

1. Create the matrix $\mathbf{Q}$ calculating the features corresponding to the evaluated set and time-frequency footprint, using the data from the design set.

2. Calculate the weights $\mathbf{v}$ of the LS-GDA classifier using equation (5.15).

3. Create the matrix $\mathbf{Q}$ calculating the features corresponding to the evaluated set and time-frequency footprint, using now the data from the test set.

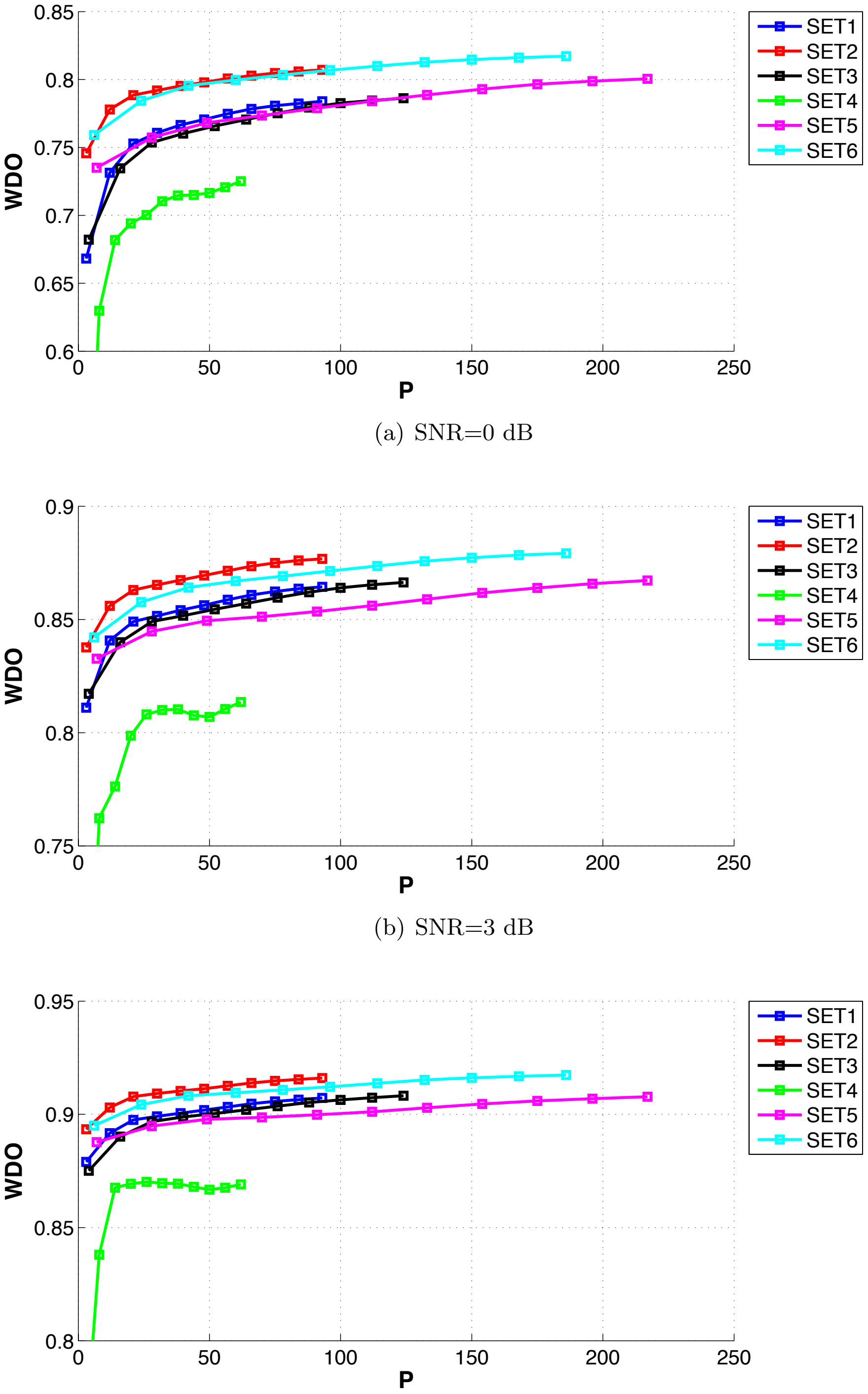
4. Generate the binary mask for each mixture of the test database, using the weights calculated in step 2, according to (5.12).

5. Compute the WDO value for all the mixtures of the test database using the obtained binary mask and the power of the original signals.

6. Repeat steps 1 to 5 for each set of features, time-frequency footprint and SNR.

The results of this experiment are shown in figure 5.4. The represented WDO values have been averaged over all the mixtures in the test database, and they are represented against the total number of features (P), which depends on the feature set and the time-frequency footprint. The different set of features are represented with lines of different colors, and the different values of $Nfreqs$ (and $Nframes$) are represented with squares over the lines. Analyzing the 3 subfigures, which corresponds with different levels of SNR, it is deduced that the relative behavior of the different set of features is the same for different SNRs, obtaining higher WDO values with higher SNRs, as it is expected. Additionally, the WDO obtained increases asymptotically with the number of features. It can be easily deduced that the SET2 (red line) represents the best tradeoff in terms of WDO and number of features, for any SNR. In the case of SNR=0dB (a), SET2 achieves WDO values around 0.8 with only 50 features. The feature set SET6 achieves the same levels of WDO but using a higher number of features. In the case of the set SET4, which only uses 2 features, the results are notably worse comparing to the rest of combinations. Adding more features to SET2, as in the cases of SET3, SET5 and SET6, does not bring any improvement. Another important result obtained from this experiment is the noticeable improvement achieved by the introduction of the information of neighbor time-frequency points.

The conclusion of this analysis is that the combination of features labeled as SET2 is the best solution among the evaluated. From here onwards, all the experiments will be carried out with this set of features.



(a) SNR=0 dB

(b) SNR=3 dB

(c) SNR=5 dB

Figure 5.4: WDO values averaged over the test set, as a function of the total number of features P, obtained by the non-quantified classifier using different combination of features and different sizes of the time-frequency footprint, with $Nfreqs$=$Nframes$. The different set of features are represented with lines of different colors, and the different values of $Nfreqs$ (and $Nframes$) are represented with squares over the lines.

5.4.4.1.2 Selection of the best time-frequency footprint

Unlike the previous experiment, which used a time-frequency footprint with the same number of neighbor frequencies and time frames, it is considered now that these two values may differ. The objective of the next experiment is to find the optimum values of $Nfreqs$ and $Nframes$, using the set of features selected in the previous experiment, SET2. For this purpose, the WDO value obtained by the separation algorithm is evaluated when the classifier uses different sizes of the time-frequency footprint, varying $Nfreqs$ and $Nframes$ from 0 to 6 independently, using the features defined in SET2. The steps of the experiment are:

1. Create the matrix $\mathbf{Q}$ with the features of SET2 and the time-frequency footprint evaluated (each pair of $Nfreqs$ and $Nframes$ values), using the data from the design set.

2. Calculate the weights $\mathbf{v}$ of the LS-GDA classifier using equation (5.15).

3. Create the matrix $\mathbf{Q}$ with the features of SET2 and the time-frequency footprint evaluated, using now the data from the test set.

4. Generate the binary mask for each mixture of the test database, using the weights calculated in step 2, according to (5.12).

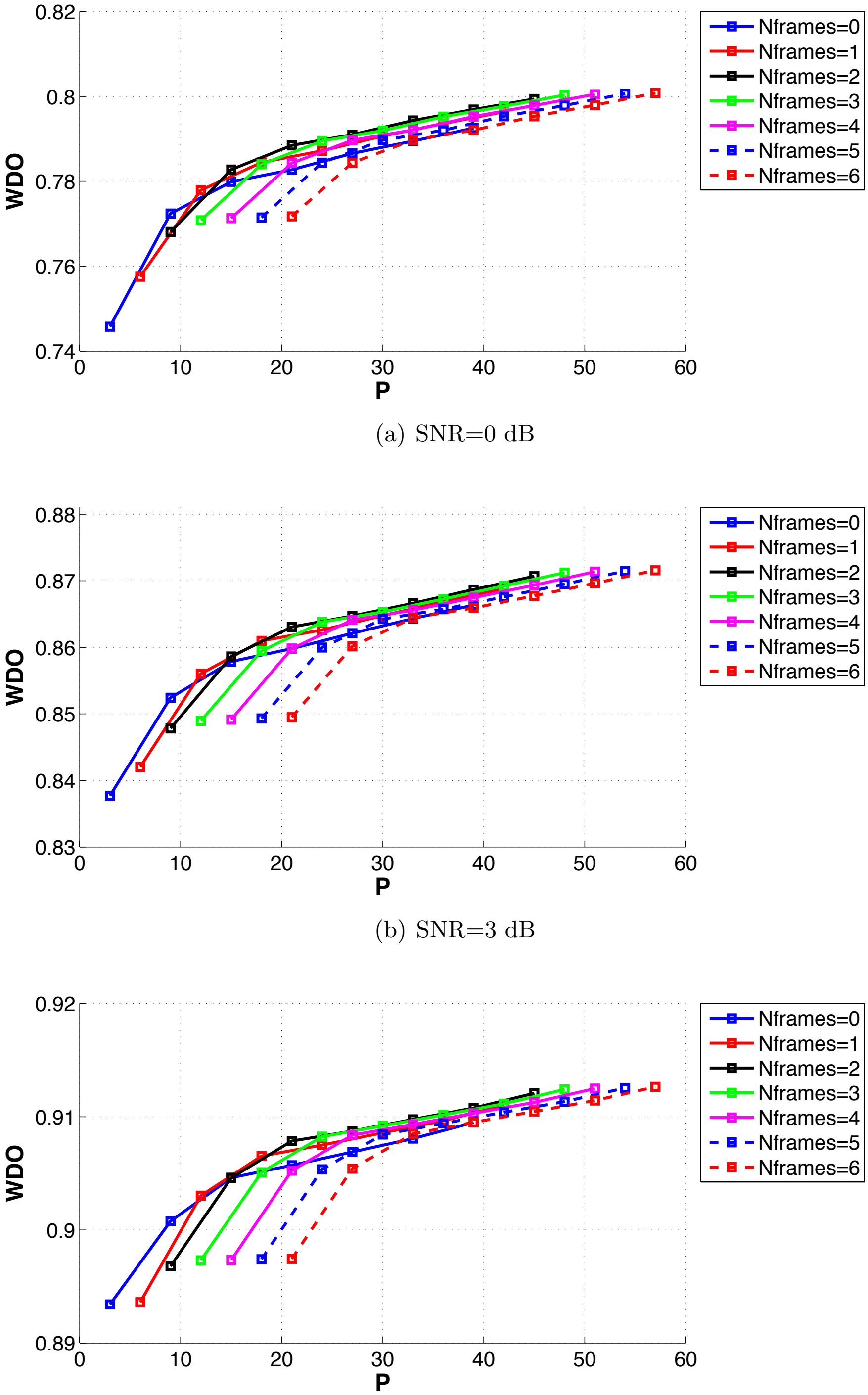
5. Compute the WDO value for all the mixtures of the test database using the obtained binary mask and the power of the original signals.

6. Repeat steps 1 to 5 for each value of $Nfreqs$ and $Nframes$ and each SNR.

Figure 5.5 shows the results of this experiment. The WDO values have been averaged over all the mixtures in the test database, and they are represented against the total number of features P, which depends on the values given to $Nfreqs$ and $Nframes$ (see expression (5.25)). The different values of $Nframes$ are represented with lines of different colors, and the different values of $Nfrecs$ with squares over the lines. The relative behavior of the different sizes of the time-frequency footprint is the same for the different SNRs. Concerning the number of previous time frames ($Nframes$), the higher WDO values are generally obtained when using only 2 time frames. Regarding the number of neighbor frequencies, the increment of the WDO values is more noticeable for values up to $Nfreqs = 3$, whereas the amount of increment decreases with higher number of frequencies. Finally, the improvement obtained by the introduction of the information of neighbor time-frequency points is clearly demonstrated.

From the analysis of the results obtained with this experiment it is deduced that a footprint with $Nfreqs = 3$ and $Nframes = 2$ represents a good tradeoff between speech quality and computational cost. The proposed solution obtains an WDO value of 0.79 for mixtures at 0 dB, using only 27 features to classify each time-frequency point. According to the study carried out in section 4.3, and considering a sampling rate of 16 kHz, analysis window of 128 samples and 65 frequency bands, the number of instructions available to process each frequency band of a frame is IPF= 308. The number of instructions necessary to process each frequency band by the LS-GDA is approximately $P + 1$, excluding the computation of the features. Therefore, the number of instructions associated to the proposed solution represents a 9 % of the available number of instructions. This result supports the feasibility of implementing the proposed speech enhancement algorithm in a commercial hearing aid.

Finally, it is worth mentioning that a square-shaped time-frequency footprint has been also considered. However, it does not outperform the results of the T-shaped footprint due to the notably higher number of required features.



(a) SNR=0 dB

(b) SNR=3 dB

(c) SNR=5 dB

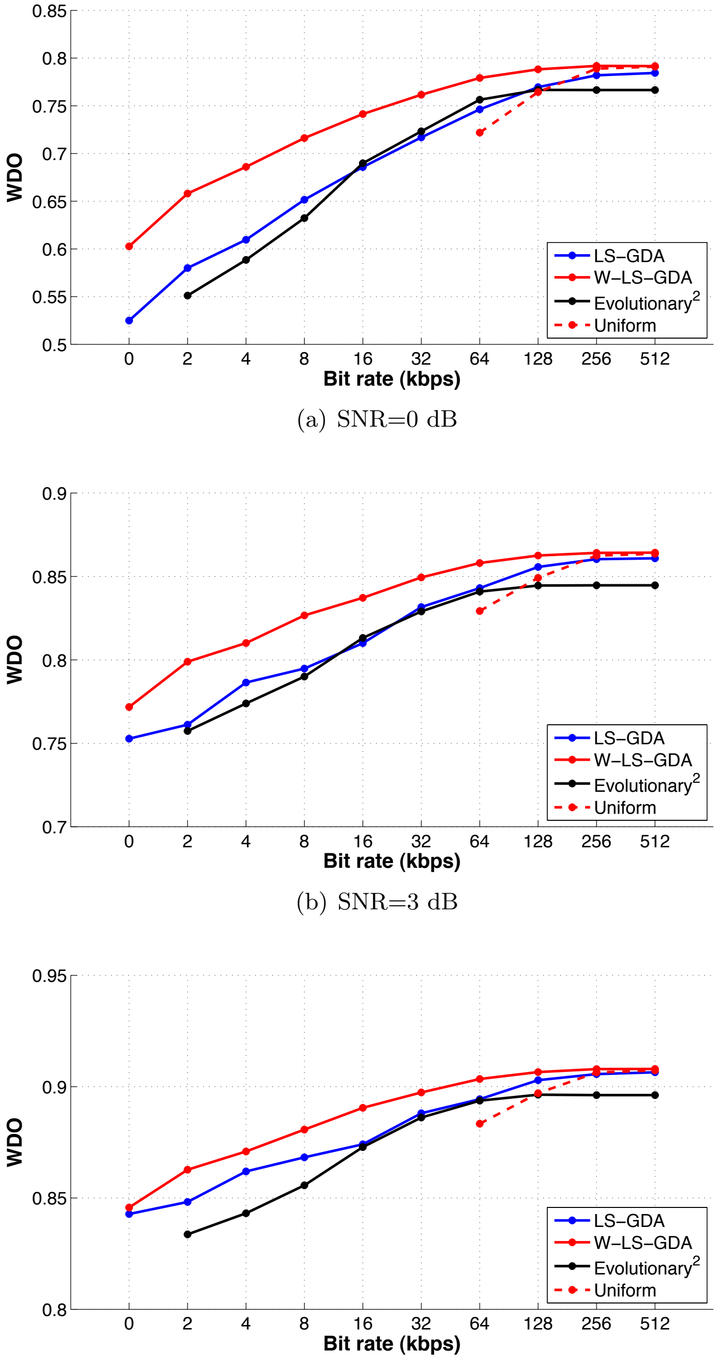
Figure 5.5: WDO values averaged over the test set, as a function of the total number of features P, obtained by the non-quantified classifier with the selected combination of features varying the number of neighbor frequencies ($Nfrecs$) and previous time frames ($Nframes$) of the time-frequency footprint. The different values of $Nframes$ are represented with lines of different colors, and the different values of $Nfrecs$ with squares over the lines.

5.4.4.2 Optimizing the transmission rate

The features for the classifier without considering quantization have already been selected. The second step in the design is to optimize the transmission schema between both devices introducing quantization in the values transmitted from the right to the left device, considering the features selected in the previous step. The proposed EA to optimize the bit distribution described in section 5.4.3 has been executed different times varying the transmitted bit rate from 0 to 512 kbps and using the two different classifiers, LS-GDA and W-LS-GDA. In the case that the bit rate is 512 kbps, all the quantized data is transmitted with the maximum number of bits, $B_{Ak} = B_{Bk} = 8$ (i.e. 16 bits per frequency band), hence the optimization is not required. In order to compare the effectiveness of the proposed optimization algorithm, the performance obtained by an uniform distribution of bits is also evaluated, assigning a constant number of bits to the amplitude and phase values of each frequency band. The values assigned in this case are 1, 2, 4 and 8, which correspond with transmission rates of 64, 128, 256 and 512 kbps respectively. Additionally, the results are compared with the ones obtained by the *Evolutionary*^{2} algorithm previously described in this chapter.

Figure 5.6 represents the WDO values averaged over the test set as a function of the number of bits transmitted from the right to the left device. The blue line corresponds with the LS-GDA solution, the red line with the W-LS-GDA solution, the black line with the *Evolutionary*^{2} algorithm, and the dashed red line is obtained with an uniform distribution of bits using the W-LS-GDA solution. The plot in (a) corresponds with a SNR of 0 dB, the plot in (b) with a SNR of 3 dB, and the plot in (c) with a SNR of 5 dB. Comparing the results obtained by the two different classifiers, the WDO values obtained by the W-LS-GDA clearly outperforms the ones obtained by the LS-GDA in any case, as it was expected. In the case of transmitting the quantized values with the maximum number of bits (512 kbps), the WDO values obtained practically match the WDO values in case of non-quantization (i.e. using $A_R(k, m)$ and $\phi_R(k, m)$). The performance is nearly unaffected when the transmission rate is decreased down to 128 kbps, but the decrease begins to be noticeable for lower bit rates. Nevertheless, in the case of SNR= 0 dB (worst case), the performance of the W-LS-GDA is only reduced by 1.5% in the case of transmitting 64 kbps, 3.8% in the case of transmitting 32 kbps, and 6.3% in the case of transmitting 16 kbps, which are acceptable transmission rates for hearing aids. Additionally, the figure also shows the case in which no information is transmitted from the right to the left device (0 kbps). In such a case, the features are calculated only using the information available in the left ear, and the performance clearly drops to WDO values around 0.6 for SNR=0 dB, which supports the use of binaural separation. The WDO values obtained by the previous *Evolutionary*^{2} algorithm are clearly improved by the optimization scheme proposed in this section, in the case of using the W-LS-GDA classifier. However, the LS-GDA classifier only obtains higher WDO values than the *Evolutionary*^{2} algorithm for the lower and highest bit rates, but the results are quite similar for bit rates between 16 and 64 kbps. Finally, it is noticeable that the results obtained by the optimized distribution outperforms the results obtained by the uniform distribution, using the two classifiers, and the difference between them increases when the number of transmitted bits decreases. Nevertheless, the use of an uniform distribution does not allow reducing the transmission rate below 64 kbps.

Finally, figure 5.7 illustrates the bit distribution obtained by the W-LS-GDA method in the case of a transmission rate of 64 kbps and SNR of 0 dB. The blue bars represent the number of bits assigned to the amplitude values (B_{A_k}), the red bars represent the number of bits assigned to the phase values (B_{P_k}), and the dashed black line the total number of bits assigned to each frequency band ($B_k = B_{A_k} + B_{P_k}$). In the lower frequency bands, the bits are mainly assigned to



(a) SNR=0 dB

(b) SNR=3 dB

(c) SNR=5 dB

Figure 5.6: WDO values averaged over the test set as a function of the number of bits transmitted from the right to the left device. The blue line corresponds with the LS-GDA solution, the red line with the W-LS-GDA solution, the black line with the *Evolutionary*^{2} algorithm, and the dashed red line is obtained with an uniform distribution of bits using the W-LS-GDA solution. The plot in (a) corresponds with a SNR of 0 dB, the plot in (b) with a SNR of 3 dB, and the plot in (c) with a SNR of 5 dB.

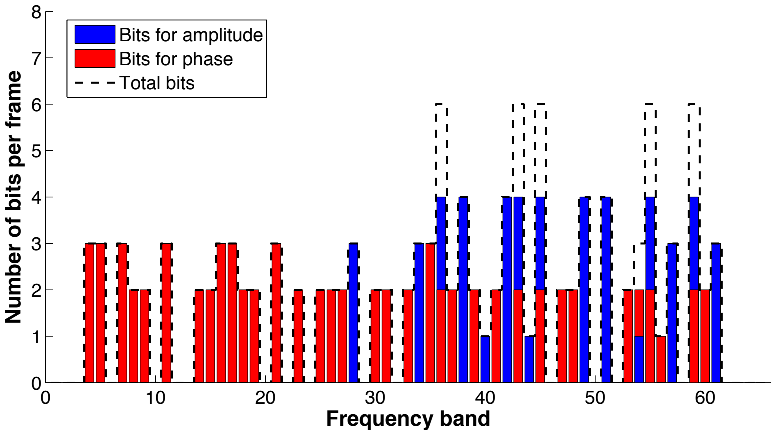


Figure 5.7: Distribution of the number of bits per frame among the frequency bands, obtained by the W-LS-GDA in the case of a transmission rate of 64 kbps and SNR of 0 dB. The blue bars represent the number of bits assigned to the amplitude values (B_{A_k}), the red bars represent the number of bits assigned to the phase values (B_{P_k}), and the dashed black line the total number of bits assigned to each frequency band ($B_k = B_{A_k} + B_{P_k}$).

the phase values, whereas in the higher frequency bands, more bits are assigned to the amplitude values. This behavior is expected due to the fact that interaural time differences predominates in the lower frequencies and interaural level differences predominates in the higher frequencies. The optimization algorithm clearly allows an efficient bit distribution.

5.5 Discussion

This chapter has presented two different approaches to design binaural source separation systems that increase the energy efficiency of the wireless-communicated binaural hearing aids. The first approach, which aims at maximizing the WDO, obtains good separation results using extremely few computational resources, even for low SNRs. The system is able to work at low bit rates, what makes it more energy efficient. The second approach, which estimates the IBM, notably improves the separation performance of the first approach. Although it uses more computational resources than the first approach, they are less than the 10% of the available computational resources for signal processing in hearing aids, which is reasonably affordable. The improvement associated to the introduction of the information of neighbor time-frequency points in the decision whether a time-frequency point belongs to speech or noise has been proved. The performance of the algorithm using the proposed WMSE in terms of WDO is only reduced by 1.5% in the case of transmitting 64 kbps, 3.8% in the case of transmitting 32 kbps, and 6.3% in the case of transmitting 16 kbps, in the worst case (SNR=0 dB). These transmission rates are feasible for hearing aids. The optimization algorithm allows distributing the bits efficiently.

In summary, the two approaches proposed in this chapter obtain good separation results, using few computational resources and working at low bit rates. The first approach uses less computational resources and it is easier to implement that the second approach, due to the fact that it only uses the information of the current time-frequency point. The second approach notably improves the performance of the previous approach and, although requires more computational resources, its implementation is completely feasible in commercial hearing aids.

5.6 Summary of contributions

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

1. A extremely low-cost binaural speech separation system that maximizes the WDO has been proposed. It is based on a quadratic discriminant that uses the ILD and ITD to classify each time-frequency point between speech or noise. The weights are calculated with a tailored evolutionary algorithm that aims at maximizing the WDO factor.

2. A generalized version of the LS-LDA has been proposed. The LS-GDA allows applying any type of transformation to the input features, in order to obtain non-linear separation boundaries.

3. A low-cost binaural speech separation system that estimates the IBM has been proposed. It is based on the LS-GDA and its computational cost directly depends on the number of input features considered for classification.

4. A weighted MSE (WMSE) metric has been introduced into the LS-GDA. The metric allows estimating the IBM and maximizing the WDO factor at the same time.

5. A novel set of features to classify a time-frequency point between speech or noise is presented. The main novelty resides in that the information related to the neighbor time-frequency points is also considered by the classifier.

6. A transmission schema to enhance the energy efficiency of the wireless-communicated binaural hearing aids has been introduced. The schema quantizes the amplitude and phase values of each frequency band with a different number of bits. The bit distribution among frequency bands is optimized by means of evolutionary computation.

The contributions obtained in this chapter have originated the publications [Gil-Pita et al., 2012] and [Ayllón et al., 2013c].

Chapter 6

Design of microphone arrays for hearing aids optimized to unknown subjects

6.1 Introduction

The improvement of speech intelligibility in hearing aids is still an unsolved problem. Modern devices, either monaural or binaural, may include microphone arrays to provide directivity by means of spatial filtering. The design of a microphone array system for hearing aids is limited by two engineering constraints. First, the reduced power of the processor limits the computational cost of the algorithms used for speech enhancement, which must be very low. Unlike adaptive beamforming, fixed beamforming techniques only need to compute the coefficients once. Hence, fixed beamforming is a real-time computable solution for speech enhancement in hearing devices, considering that the filter coefficients that satisfy the design constraints can be previously computed and easily included as constant values in the embedded algorithm. The second limitation is related to the reduced dimensions of such devices, which limits the number of microphones of the array. This chapter considers in-the-canal (ITC) devices, which have an ample role in the market. The shape of this kind of device can be approximated by a cylinder 1.5 cm in length and 1 cm in diameter, and common omnidirectional microphones placed in hearing devices have a diameter of 0.25 cm, which is relatively a large portion of the overall available area. To be realistic, this work considers that there can not be more than 4 microphones assembled in each device.

Most commercial hearing aids are still monaural systems, in which case the microphone array is composed of microphones placed in the same device. However, many people have bilateral hearing loss and they are forced to wear a hearing device in both ears. Commonly, both devices work independently, but there is a new trend of binaural hearing aids that allows them to be connected in order to exchange data and thus preserve spatial information. In such a case, the microphones of the array are placed in both sides of the human head. In any case, the signals that arrive at each element of the array do not differ only in time differences, which depend on the relative position between the source and the microphone, but it also undergo amplitude distortions due to the well-known head shadow effect, which must be considered in the design. The fact that this effect is highly dependent on a person causes the design of an array customized for a subject to need a correct measurement of such effect, which is not practical in real scenarios. The lack of information about the head of the hearing aid user causes directivity reduction and distortions. Many attempts of improving the intelligibility in both monaural and binaural hearing aids by means of superdirective beamforming can be found in the literature. In some cases, the filter coefficients are calculated without considering the head shadow effect, thus obtaining non-realistic results. Some examples are [Gordy et al., 2008; Peterson and Zurek, 1987]. In other cases, the head shadow effect is considered in the design, then assuming that it has been measured or modeled somehow, for instance in [Lotter and Vary, 2006; Rohdenburg et al., 2007; Welker et al., 1997].

The main objective of this chapter is the design of superdirective beamformers for monaural and binaural hearing aids considering the head shadow effect but assuming unavailable head measurements of the subjects. When a correct characterization of the head of the subject is available, the spatial filter can be optimized to that person, causing gain reduction and distortions when the system is worn by a different one. The characterization of the head effects of one subject can be obtained either from theoretical models that use anthropometric measurements of the body, which are not very accurate since they are based on physical approximations, or from experimental measurements carried out with microphones in the eardrum of the subject. None of these methods seems appropriate to be performed whenever a hearing-impaired person needs a hearing aid. Consequently, this work suggests finding optimized filter coefficients that minimize the effects of using non-customized arrays, these coefficients being valid for any person. Three different methods to compute optimized filter coefficients with the purpose of maximizing the array gain and minimizing distortions, and with the objective that these filter coefficients are valid for all users, are proposed. Furthermore, measurements of the amplitude and phase distortions caused by a lack of head information in the design are introduced. In addition, the methods proposed in this chapter are evaluated in 13 different array configurations in an ITC hearing aid.

6.2 Microphone array signal processing

Spatial filtering, commonly known as beamforming, is a common approach applied in digital signal processing to reduce noise and thus it is useful to improve the quality and intelligibility of the desired speech. Using a microphone array, it is possible to spatially filter the signal coming from the desired direction by applying different complex weights to each input channel in order to coherently combine the signals coming from the steered direction and incoherently combine the signals coming from different directions. There are different criteria for the computation of the filter coefficients: frequency invariant response, superdirectivity, robustness against steering vector errors, etc. The beamforming techniques explored in this chapter are based on the MVDR filter [Capon, 1969], which is a superdirectivity beamformer that attenuates noise coming from other directions than the desired, whilst the target speech signal is not modified, obtaining a distortionless response. The basis of beamforming techniques are described in this section.

6.2.1 Directivity pattern

Let us consider the general case of a tridimensional microphone array composed of M sensors arranged with any geometry in the coordinate system represented in figure 6.1. The angle θ varies between 0 and 2π and represents azimuth, and the angle ϕ varies between 0 and π and it is related to elevation through $\pi/2 - \phi$. The directivity pattern or frequency response $D(k, \theta, \phi)$ of the array for the k-th frequency bin ($k = 0, \cdots, K-1$) depends on the direction of arrival DOA of the incident wave, and, in the ideal case where the array is composed of similar

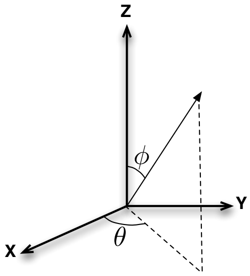


Figure 6.1: Coordinate system.

omnidirectional microphones placed in the open-air, and assuming far-field sources (i.e. planar incident waves), it is given by

$$D(k,\theta,\phi) = \sum_{m=1}^{M} W_m(k) \exp\left(i \frac{\pi k f_s}{c(K-1)} \mathbf{r}_m^T \cdot \mathbf{u}_{\theta\phi} \right), \tag{6.1}$$

where θ and ϕ are the angles defining the DOA, $W_m(k)$ are the complex weights applied to the m-th element of the array, $\mathbf{r}_m$ is a column vector that represents the position of the m-th sensor with respect to the origin of coordinates, $\mathbf{r}_m = [r_{m_x}, r_{m_y}, r_{m_z}]^T$, $\mathbf{u}_{\theta\phi}$ is a column vector pointing at the DOA determined by the azimuth θ and the elevation ϕ, $\mathbf{u}_{\theta\phi} = [\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\phi]$, K is the total number of frequency bands, f_s is the sampling frequency and c the speed of sound. For the sake of simplicity, the directivity pattern is usually formulated using matrix notation by

$$D^*(k,\theta,\phi) = \mathbf{w}_k^H \cdot \mathbf{d}(k,\theta,\phi), \tag{6.2}$$

where $(.)^*$ and $(.)^H$ represent the matrix conjugate and Hermitian transpose operators respectively, $\mathbf{w}_k = [W_1(k), \cdots, W_M(k)]^T$ is the array weight vector, and $\mathbf{d}(k,\theta,\phi)$ is the steering vector, which for a given DOA is defined as [Krim and Viberg, 1996]

$$\mathbf{d}(k,\theta,\phi) = \left[1, \cdots, \exp\left(-i \frac{\pi k f_s}{c(K-1)} \mathbf{r}_m^T \cdot \mathbf{u}_{\theta\phi} \right), \cdots, \exp\left(-i \frac{\pi k f_s}{c(K-1)} \mathbf{r}_M^T \cdot \mathbf{u}_{\theta\phi} \right) \right]. \tag{6.3}$$

The sources are considered to be in the near-field when

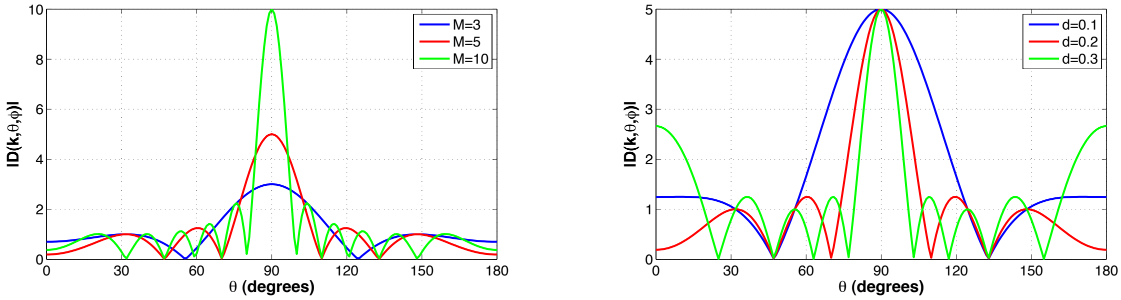
$$|r| < \frac{2L_a^2}{\lambda}, \tag{6.4}$$

where r is the distance of the source to the array, L_a is the effective length of the array, and λ the wavelength. In such a case, the source position can be defined by its cartesian coordinates instead of its DOA, and the steering vector defined in expression (6.3) becomes into

$$\mathbf{d}(k,\theta,\phi) = \left[1, \cdots, \exp\left(-i \frac{\pi k f_s}{c(K-1)} ||\mathbf{r}_m - \mathbf{r}_n|| \right), \cdots, \exp\left(-i \frac{\pi k f_s}{c(K-1)} ||\mathbf{r}_M - \mathbf{r}_n|| \right) \right], \tag{6.5}$$

where $\mathbf{r}_n$ is a column vector that represents the position of the n-th source with respect to the origin of coordinates.

In the case of acoustic signals, the steering vector is normally referred as the acoustic transfer function (ATF), which represents the frequency response between a microphone and a determined source or the frequency response between a microphone and a determined DOA. Henceforth, the frequency response of the m-th microphone for a given direction (θ, ϕ) is represented



(a) Number of microphones

(b) Distance between microphones

Figure 6.2: Directivity pattern of an unsteered uniform linear array composed of omnidirectional microphones placed along the x-axis, in the case of varying the number of elements with a fixed distance of 0.2 m (a) and in the case of varying the distance between the sensors with a fixed number of elements of $M = 5$ (b), for a frequency of 1 kHz.

by $A_m(k, \theta, \phi)$. Defining $\mathbf{a}_{k\theta\phi}$ as a column vector containing the M microphone responses, $\mathbf{a}_{k\theta\phi} = [A_1(k, \theta, \phi), \cdots, A_M(k, \theta, \phi)]^T$, the directivity pattern, considering the ATFs, is expressed as

$$D^*(k, \theta, \phi) = \mathbf{w}_k^H \cdot \mathbf{a}_{k\theta\phi}. \tag{6.6}$$

Figure 6.2 represents the directivity pattern for a constant elevation (i.e. ϕ does not vary) of an unsteered uniform linear array (i.e. $\mathbf{w}_k = \mathbf{1}, \ \forall k$) composed of omnidirectional microphones placed along the x-axis, in the case of varying the number of elements with a fixed distance of $d = 0.2m$ (a) and in the case of varying the distance between sensors with a fixed number of elements of $M = 5$ (b), for a frequency of 1 kHz. When the linear array is not steered to any special direction it shows a maximum directivity for a DOA of 90°, which corresponds to perpendicular incident waves. An increment in the number of elements of the array clearly results in a narrower main beam and lower sidelobe level (see figure 6.2 (a)), which means that the array is more selective and has a higher ability rejecting noise and interferences. The increment of the distance between sensors (see figure 6.2 (b)) also leads to narrower main beams but the difference between the levels of main beam and the sidelobes does not vary in this case. The two directivity plots in figure 6.2 are represented for a frequency of 1 kHz, but the directivity pattern of the array varies with frequency. Figure 6.3 shows the directivity pattern of an unsteered uniform linear array of 5 elements separated a distance of 0.1 m, where the frequency has been varied from 400 Hz to 4 kHz. The result is that the main beam width decreases as the frequency increases, and the sidelobe level and null positions also vary with frequency. This fact may cause distortions and unwanted signal cancelation for wideband signals such as speech. In such cases, it is desirable to design arrays with invariant frequency response.

6.2.1.1 Array gain and directivity factor

The array gain is defined as the improvement in SNR between a reference sensor and the array output, and it can be expressed as $G = G_d/G_n$, where G_d is the gain to the desired signal (i.e. the power of the directivity pattern in the steering direction) and G_n is the average gain to all noise sources, which depends on the nature of the noise field. In the case of a diffuse noise field, in which the energy radiated by the noise is the same over all directions and all times, the denominator term G_n is calculated by averaging the received power over the whole sphere. In

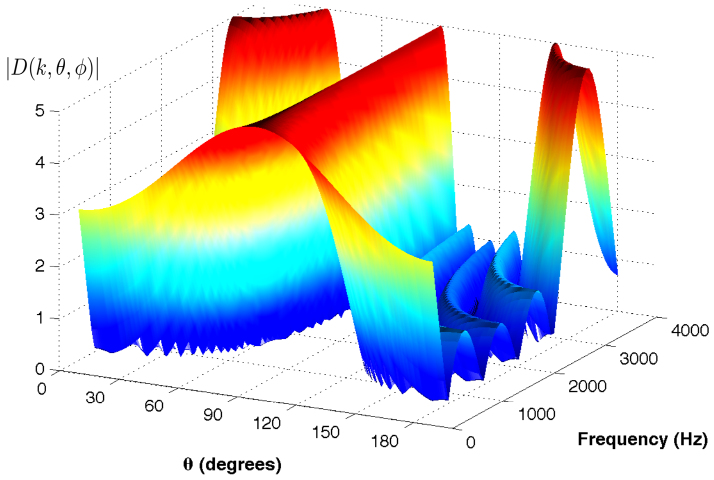


Figure 6.3: Variations with frequency of the directivity pattern of an unsteered uniform linear array of 5 elements separated a distance of 0.1 m.

this case, the array gain is also known as directivity factor and it is calculated according to

$$G(k,\theta_0,\phi_0) = \frac{|D(k,\theta_0,\phi_0)|^2}{\frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}|D(k,\theta,\phi)|^2 sin\phi\, d\phi\, d\theta} = \frac{|\mathbf{w}_k^H \cdot \mathbf{b}_k|^2}{\mathbf{w}_k^H \cdot \mathbf{H}_k \cdot \mathbf{w}_k}, \tag{6.7}$$

where $\mathbf{b}_k = \mathbf{a}_{k\theta_0\phi_0}$ represents the k-th frequency responses of the M microphones in the steering direction, and $\mathbf{H}_k$ represents the cross-spectral density of the diffuse noise between sensors for the k-th frequency band, which is calculated as

$$\mathbf{H}_k = \frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}\mathbf{a}_{k\theta\phi}\cdot\mathbf{a}_{k\theta\phi}^H sin\phi\, d\phi\, d\theta, \tag{6.8}$$

where the integrals have been applied to each element of the matrix.

6.2.2 Spatial Aliasing

Spatial aliasing is a problem related to the distance between the sensors of the array, and it is equivalent to the frequency aliasing problem which establishes the minimum sampling frequency (Nyquist frequency) to avoid aliasing in the sampled signal (i.e. the appearance of grating lobes). Sensor arrays perform spatial sampling, and an analogous requirement exists to avoid grating lobes in the directivity pattern, which are sidelobes substantially larger in amplitude and approaching the level of the main lobe. This problem can be seen in figure 6.3, where grating lobes appear around 0° and 180° at high frequencies. Spatial aliasing causes phase uncertainty when calculating the phase difference between the signals received by two adjacent sensors. Considering a pure sinusoidal signal of wavelength λ, phase uncertainty is avoided for a distance between sensors of $d < \lambda/2$. Hence, for a wideband signal, which can be decomposed into the addition of complex sine waves of different frequencies, the requirement to avoid spatial aliasing is given by

$$d < \frac{\lambda_{min}}{2} = \frac{c}{2f_{max}}. \tag{6.9}$$

6.2.3 Beamforming

The complex weights $W_m(k)$ in expression (6.1) can be used to modify the directivity pattern of the array with the aim to enhance the desired signal. In general, the complex weights can be expressed as

$$W_m(k) = a_m(k)e^{i\varphi_m(k)}, \tag{6.10}$$

where a_m and φ_m are the frequency dependent amplitude and phase weights respectively. The shape of the directivity pattern is changed by modifying the amplitude weights and the angular position of the main lobe is controlled with the phase weights. Beamforming techniques aim to calculate the value of the complex weights in order to obtain the desired shaping and steering of the directivity pattern.

This concept is illustrated here using the simplest of all beamforming techniques, the delay and sum (DS) beamformer, and assuming an array composed of omnidirectional microphones in the open air (i.e. the ATF corresponds with the steering vector defined in (6.3)). The amplitude of the weights is set to one, $a_m(k) = 1$, and the phase is changed to steer the main beam to the desired DOA given by (θ_0, ϕ_0), which leads to

$$W_m(k) = \exp\left(-i\frac{\pi k f_s}{c(K-1)}\mathbf{r}_m^T \cdot \mathbf{u}_{\theta_0\phi_0}\right). \tag{6.11}$$

According to expression (6.1), the steered directivity pattern is given by

$$D(k, \theta_0, \phi_0) = \sum_{m=1}^{M} \exp\left(i\frac{\pi k f_s}{c(K-1)}\mathbf{r}_m^T \cdot (\mathbf{u}_{\theta\phi} - \mathbf{u}_{\theta_0\phi_0})\right), \tag{6.12}$$

and the array output $Y(k)$ is expressed as the sum of the weighted channels as

$$Y(k) = \frac{1}{M}\sum_{m=1}^{M} W_m(k)X_m(k), \tag{6.13}$$

where $X_m(k)$ is the signal received by the m-th microphone. The DS beamformer is equivalent to combine an aligned version of the input signals in the time domain, where the delay applied to the m-th channel (τ_m) corresponds with the time-difference of arrival (TDOA) between the m-th sensor and the reference channel for the desired DOA, that is

$$y(t) = \frac{1}{M}\sum_{m=1}^{M} x_m(t - \tau_m). \tag{6.14}$$

An example of the DS beamformer is shown in figure 6.4 which plots the directivity pattern of an unsteered uniform linear array composed of 5 omnidirectional microphones separated a distance of 0.2 m (blue line) and the same array steered to 60° (red line) with the DS beamformer technique.

6.2.4 MVDR beamformer

The MVDR beamformer, also known as Capon filter [Capon, 1969], is perhaps the most widely used superdirective beamformer. The basic idea is to maximize the array gain (i.e. the output SNR) by finding the filter coefficients that minimize the output power with the constraint that the desired signal is not affected. This is equivalent to minimize the denominator in expression (6.7) with the constraint that the numerator has a constant value. In the case of wideband

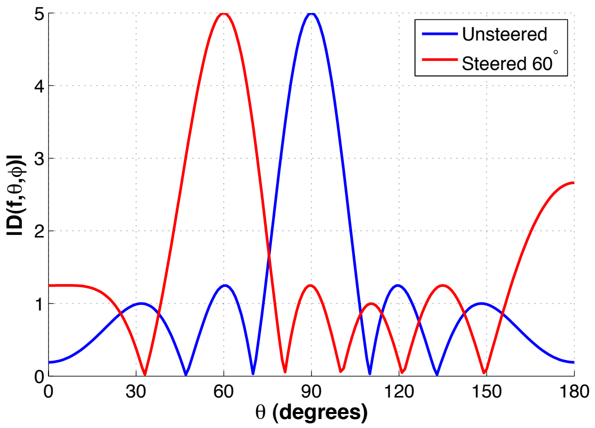


Figure 6.4: Directivity pattern of an uniform linear array composed of 5 omnidirectional microphones separated a distance of 0.2 m when it is unsteered (blue line) and when it is steered to 60° (red line) with a DS beamformer.

signals, the optimization problem can be solved independently for each frequency band, and it is written as

$$\begin{aligned} &\underset{\mathbf{w}_k}{min} \left\{ \mathbf{w}_k^H \cdot \mathbf{H}_k \cdot \mathbf{w}_k \right\} \\ &subject\ to\ \mathbf{w}_k^H \cdot \mathbf{b}_k = 1. \end{aligned} \tag{6.15}$$

The optimization problem is solved by applying the method of Lagrange multipliers, resulting in

$$\mathbf{w}_k = \frac{\mathbf{H}_k^{-1}\mathbf{b}_k}{\mathbf{b}_k^H \mathbf{H}_k^{-1} \mathbf{b}_k}. \tag{6.16}$$

This optimization provides an invariant frequency response for the steering direction, ensuring a distortionless filtered target signal. The MVDR filter is an adaptive beamformer that can adapt itself to the noise environment for maximum noise reduction. This is achieved by updating the cross-spectral density matrix of the noise $\mathbf{H}_k$ to recalculate the filter coefficients. Nevertheless, in the case of stationary diffuse noise field the matrix $\mathbf{H}_k$ does not change with time, and the MVDR filter can be considered to be a fixed beamformer.

6.3 Problem formulation

6.3.1 Signal model considering the head distortions

Let us consider a hearing aid with the microphone array composed of M sensors in the coordinate system represented in figure 6.5. The directivity pattern is given by expression (6.6) and the array gain, assuming diffuse noise field, is given by expression (6.7). The matrix $\mathbf{H}_k$ represents the cross-spectral density of the noise between sensors for the k-th frequency band, and it is calculated by expression (6.8). In those cases where the microphone responses $\mathbf{a}_{k\theta\phi}$ are spatially sampled in azimuth with $2N_\theta + 1$ angles $(\theta_{-N_\theta}, \cdots, \theta_0, \cdots, \theta_{N_\theta})$, and in elevation with $2N_\phi + 1$ angles $(\phi_{-N_\phi}, \cdots, \phi_0, \cdots, \phi_{N_\phi})$, the result is that the whole sphere is sampled by $(2N_\theta + 1)(2N_\phi + 1)$ directions, and the integrals of expression (6.8) are approximated by the

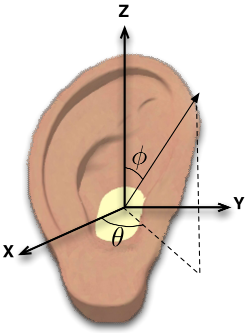


Figure 6.5: Hearing aid position within the coordinate system.

weighted summations of expression (6.17), where the weighting term Δ_{mn} is calculated by using equation (6.18).

$$\mathbf{H}_k = \sum_{m=-N_\theta}^{N_\theta} \sum_{n=-N\phi}^{N_\phi} \mathbf{a}_{k\theta_m\phi_n} \cdot \mathbf{a}_{k\theta_m\phi_n}^H \Delta_{mn}. \tag{6.17}$$

$$\Delta_{mn} = \frac{1}{4\pi} \int_{\frac{\phi_{n-1}+\phi_n}{2}}^{\frac{\phi_n+\phi_{n+1}}{2}} \int_{\frac{\theta_{m-1}+\theta_m}{2}}^{\frac{\theta_m+\theta_{m+1}}{2}} \sin\phi \, d\phi \, d\theta. \tag{6.18}$$

An important conclusion that arises from the expression of the array gain in (6.7) is that the array gain for the k-th frequency band is not modified when a scale factor is applied to the filter coefficients. If the array weighting vector $\mathbf{w}_k$ is scaled by a complex constant value d_k, the array gain does not change, since the variations in the numerator term are compensated by the equivalent variations in the denominator term. This fact is important, and as it will be shown in further sections, it can help in the optimization of the array response.

The MVDR described in section 6.2.4 is the beamforming technique selected for speech enhancement in this chapter. Assuming that the hearing aid user usually faces the desired speaker, the beamformer is steered to the straight-ahead direction, which corresponds with 0° in azimuth and 0° in elevation.

In case that the microphones of the array are placed within the hearing aid, the received signals are distorted by the head shadow effect, which is described by the HRTF. The HRTFs are measured with a microphone placed in the canal of the ear. When the ITC hearing device is centered in the origin of coordinates (see figure 6.5), the HRTF provides the microphone response of an element placed just in the center of the device. In order to obtain the frequency response of a microphone placed in a different position inside the device, it is necessary to add some additional attenuations to the HRTF, as well as time delays according to its position. For the sake of simplicity, the additional attenuation is neglected in our model, introducing only an extra delay to the HRTF function according to each microphone position. This is not very far from reality since the distance between microphones in the same device is very small, the level differences between them being negligible. Since the response of a microphone placed in the origin of coordinates, which corresponds with the center of the ITC device, is given by $H_{L/Rs}(k, \theta, \phi)$, and assuming that the microphones are close enough to disregard amplitude differences between them, the frequency response $A_{ms}(k, \theta, \phi)$ of a microphone with position

$\mathbf{r}_m$, for the s-th subject, can be obtained using the next expression:

$$A_{ms}(k,\theta,\phi) = H_{L/Rs}(k,\theta,\phi)\exp\left(-i\frac{\pi k f_s}{c(K-1)}\mathbf{r}_m^T \cdot \mathbf{u}_{\theta\phi}\right). \tag{6.19}$$

This model allows obtaining the directivity pattern of an array of microphones placed in the left ear, right ear, or in both (binaural system), by only selecting the appropriate HRTF of each microphone.

6.3.2 Real scenario: unavailable HRTFs

With the MVDR beamformer the filter coefficients can be easily optimized to reduce noise whilst obtaining a distortionless response for the steering direction, as long as the HRTF of the subject is known, calculating the microphone response according to expression (6.19). In this case, the array is customized to that person. Unfortunately, this scenario is seldom found in real situations. When a hearing-impaired person goes to the audiologist to acquire a new hearing aid, the audiologist does not have any knowledge of his HRTF. If that person wears a device with coefficients that are not fitted to his head, it will cause gain reduction and distortions in the filtered target signal. The lack of an accurate mathematical model along with the tedious and expensive measurements required to obtain the HRTF for every person who needs a hearing aid, encourage the idea of designing an optimized array that minimizes the gain reduction and distortions regardless of the subject. This design can be based on the optimization of the filter coefficients for the subjects of a HRTF database.

For a set of S subjects, the average array gain $\overline{G}$ in dB, for the speech frequency range, can be expressed by:

$$\overline{G} = \frac{1}{LS}\sum_{s=1}^{S}\sum_{k=k_{min}}^{k_{max}} 10\log_{10} G_{sk}(\theta_0,\phi_0), \tag{6.20}$$

where $G_{sk}(\theta_0,\phi_0)$ is the array gain for the s-th subject of the database and for the k-th frequency, $k = k_{min}, ..., k_{max}$ are the frequency bands belonging to the speech frequency range, and $L = k_{max} - k_{min} + 1$ is the number of frequency bands contained in the speech frequency range. In this work, the speech frequency range considered goes from 20 Hz to 3.5 kHz.

In case that the filter coefficients are optimized to maximize $\overline{G}$, they will not be completely fitted to any of the S subjects, then introducing gain reduction and distortions in the filtered speech. In order to measure the amplitude and phase distortions, the next two parameters, $\overline{AD}$ and $\overline{PD}$ are proposed. $\overline{AD}$ measures the average amplitude distortion for the S subjects computing the root mean square (RMS) value of the power of the frequency response for the target direction, calculated in dB and averaged in the speech frequency range and for all subjects:

$$\overline{AD} = \sqrt{\frac{1}{LS}\sum_{s=1}^{S}\sum_{k=k_{min}}^{k_{max}} \left(10\log_{10}|D_s(k,\theta_0,\phi_0)|^2\right)^2}, \tag{6.21}$$

where $D_s(k,\theta_0,\phi_0) = \mathbf{w}_k^H \cdot \mathbf{a}_{sk\theta_0\phi_0} = \mathbf{w}_k^H \cdot \mathbf{b}_{sk}$ is the directivity pattern for the s-th subject and for the k-th frequency band, in the steering direction given by θ_0 and ϕ_0. Note that the power of the frequency response for the steering direction has a value of 0 dB for distortionless responses, so, in fact, expression (6.21) is actually computing the root mean square error (RMSE).

In the case of phase distortion, it is considered that a filter with linear phase frequency response does not distort the signal. $\overline{PD}$ measures this distortion with the RMS value of the second derivative numerical estimation of the phase of the frequency response, which is 0 for linear phases. This estimation is given by

$$\gamma_k = \frac{\Phi_{k+1} + \Phi_{k-1} - 2\Phi_k}{h^2}, \tag{6.22}$$

where Φ_k is the phase of the k-th frequency of the frequency response, and h is the differentiation step, which is considered to be one sample. $\overline{PD}$ is obtained by averaging along the frequencies of the speech frequency range and the S subjects:

$$\overline{PD} = \sqrt{\frac{1}{LS}\sum_{s=1}^{S}\sum_{k=k_{min}}^{k_{max}} \gamma_{sk}^2}. \tag{6.23}$$

The beamformers designed in this work are optimized to the straight-ahead direction. In case of existing errors in the steering direction, for instance due to a slight head movement, they also cause gain reduction and distortions in the response. This effect is measured by calculating the gain reduction between the true steering direction and the assumed steering direction (0°, 0°), and averaging these errors over a range of deviations of $\pm D$ angles in azimuth and $\pm D$ angles in elevation, according to the next expression:

$$\overline{GL}^{ev} = \frac{1}{LS(2D+1)^2}\sum_{s=1}^{S}\sum_{k=k_{min}}^{k_{max}}\sum_{m=-D}^{D}\sum_{n=-D}^{D}(10\log_{10} G_{sk}(\theta_0,\phi_0) - 10\log_{10} G_{sk}(\theta_m,\phi_n)). \tag{6.24}$$

The following section describes a novel approach to solve the optimization problem composed of maximizing $\overline{G}$ whereas minimizing $\overline{AD}$ and $\overline{PD}$.

6.4 Optimization of the filter coefficients using a database of microphone responses

This section describes the procedure followed to find the filter coefficients that maximize the average gain while minimizing speech distortions for the subjects of a database. The design is based on a HRTF database containing measurements for different spatial directions and subjects. The database is split into two sets, one for the design and another for the test, in order to overcome a generalization loss of the results. The HRTFs that belong to the design set are used to compute the optimized filter coefficients, which are tested with the subjects of the test set. For convenience, the HRTFs of the database are normalized to the HRTF corresponding to the straight-ahead direction (0°, 0°), which is the steering direction.

Considering that the design set is composed of S subjects, the average gain in dB for the k-th frequency band $\overline{G_k}$, is given by

$$\overline{G_k} = \frac{1}{S}\sum_{s=1}^{S} 10\log_{10} G_{sk}(\theta_0,\phi_0). \tag{6.25}$$

Introducing expression (6.7) and applying logarithmic identities yield the next expression

$$\overline{G_k} = 10\log_{10}\frac{\sqrt[S]{\prod_{s=1}^{S}|\mathbf{w}_k^H\cdot\mathbf{b}_{sk}|}^{2}}{\sqrt[S]{\prod_{s=1}^{S}\mathbf{w}_k^H\cdot\mathbf{H}_{sk}\cdot\mathbf{w}_k}}, \tag{6.26}$$

where $\mathbf{b}_{sk}$ and $\mathbf{H}_{sk}$ are the frequency responses of the M microphones in the steering direction and the cross-spectral density of the noise between microphones, for the s-th subject and for the k-th frequency band, respectively. Let us assume that the weighting vector $\mathbf{w}_k$ can be expressed as a complex constant d_k multiplied by a normalized weighting vector $\mathbf{v}_k$, $\mathbf{w}_k = d_k \cdot \mathbf{v}_k$. The fact that scaling the filter coefficients by d_k does not modify the array gain causes that the maximization of (6.26) does not have a unique solution. Consequently, an additional constraint must be introduced to solve the optimization problem, finding a determined value of d_k. In the case of the MVDR filter (6.15), the array gain is maximized by minimizing the output noise power, which corresponds to the denominator, with the constraint of keeping the numerator to a constant value, ensuring a distortionless response for the steering direction. In case of optimizing the averaged array gain in (6.26), it is not possible to find a set of filter coefficients that ensures a distortionless response for all the subjects at the same time. However, it is proposed here to find the value of d_k that introduces the minimum average amplitude distortion $\overline{AD}$. Replacing $D_s(k, \theta_0, \phi_0)$ by $\mathbf{w}_k^H \cdot \mathbf{b}_{sk} = d_k \cdot \mathbf{v}_k^H \cdot \mathbf{b}_{sk}$ in the expression of $\overline{AD}$ in (6.21) yields the next expression:

$$\overline{AD} = \sqrt{\frac{1}{LS}\sum_{s=1}^{S}\sum_{i=k_{min}}^{k_{max}} \left(10\log_{10}|d_i|^2 + 10\log_{10}|\mathbf{v}_i^H \cdot \mathbf{b}_{si}|^2\right)^2}. \tag{6.27}$$

In order to determine the values d_k, $k \in \{k_{min}, ..., k_{max}\}$, that minimize $\overline{AD}$, the following system of equations must be solved

$$\frac{\partial \overline{AD}}{\partial |d_k|} = 0,\ k \in \{k_{min}, ..., k_{max}\}, \tag{6.28}$$

each equation being obtained as

$$\frac{\partial \overline{AD}}{\partial |d_k|} = \frac{20}{\overline{AD} LS |d_k| \log 10}\left(20S\log_{10}|d_k| + \sum_{s=1}^{S} 20\log_{10}|\mathbf{v}_k^H \cdot \mathbf{b}_{sk}|\right) = 0. \tag{6.29}$$

The solution to this trivial system of equations leads to

$$|d_k| = \left(\prod_{s=1}^{S} |\mathbf{v}_k^H \cdot \mathbf{b}_{sk}|\right)^{\frac{-1}{S}}. \tag{6.30}$$

Regarding the phase of d_k, it is proposed here to use a null average phase response for the steering direction, though it does not imply minimum $\overline{PD}$. The proposed value of d_k is

$$d_k = \left(\prod_{s=1}^{S} |\mathbf{v}_k^H \cdot \mathbf{b}_{sk}|\right)^{\frac{-1}{S}} e^{-i\frac{1}{S}\sum_{s=1}^{S}\Phi\{\mathbf{v}_k^H \cdot \mathbf{b}_{sk}\}}. \tag{6.31}$$

Finally, the optimization problem of maximizing expression (6.26) can be solved by minimizing the denominator of the expression and imposing a constraint according to

$$\begin{aligned} &\underset{\mathbf{w}_k}{min} \left\{ \sqrt[S]{\prod_{s=1}^{S} \mathbf{w}_k^H \cdot \mathbf{H}_{sk} \cdot \mathbf{w}_k} \right\} \\ &subject\ to \left(\prod_{s=1}^{S} |\mathbf{w}_k^H \cdot \mathbf{b}_{sk}| \right)^{\frac{1}{S}} e^{i \frac{1}{S} \sum_{s=1}^{S} \Phi\left\{\mathbf{w}_k^H \cdot \mathbf{b}_{sk}\right\}} = 1, \end{aligned} \tag{6.32}$$

where the filter coefficients $\mathbf{w}_k$ are obtained by $\mathbf{w}_k = d_k \cdot \mathbf{v}_k$, with d_k calculated according to (6.31). It is worth clarifying that the filter coefficients $\mathbf{w}_k$ obtained by solving the problem in (6.32) are not optimized to minimize distortion, but the selected scaling factor d_k is the one that introduces the lowest average amplitude distortion.

The remainder of this section describes three different methods proposed to solve the optimization problem formulated in (6.32).

6.4.1 Standard optimization using an evolutionary algorithm

In this first approach, the coefficients of each frequency band are optimized independently using a tailored evolutionary algorithm designed to solve the optimization problem formulated in (6.32), which maximizes the average array gain for a set of subjects, by scaling the coefficients to reduce amplitude distortion. Each candidate solution is an array weight vector $\mathbf{w}_k$, and the average array gain $\overline{G_k}$ is the cost function. The steps of the EA are described as follows:

1. An initial population of 50 candidate solutions is generated. Each candidate solution is defined by an array weight vector $\mathbf{w}_k$. In a first approach, the values of the coefficients obtained by applying the standard MVDR fitted to each subject in the database are used to initialize the candidates of the population.

2. The coefficients of each candidate solution are scaled using the factor defined by equation (6.31), in order to satisfy the constraint included in equation (6.32).

3. The average array gain $\overline{G_k}$ is then evaluated for each solution, and it is used as a ranking in order to determine the best solution of the population.

4. After evaluating the performance of each candidate solution in the population, a selection process is applied. It consists of selecting the best 10% solutions of the population by removing the remaining solutions.

5. The 90% remaining solutions of the new population are then generated by uniform crossover of the best candidates.

6. Mutations are applied to the whole new population, excluding the best solution, adding a random Gaussian complex value to each element of the candidate solution ($P_M = 1$). The standard deviation of this random mutation factor is updated in every iteration. When the best gain achieved in the current iteration is higher than the one obtained in the previous iteration, the standard deviation is increased by a 20%, otherwise it is reduced by 50%.

7. The process is reiterated 100 times from step 2 to 6.

The candidate solution that achieves the highest average array gain over the design set is selected as the final solution. The values of the parameters of the evolutionary algorithm (population size, crossover rate, mutation scheme and number of generations) have been found to obtain a quite good tradeoff between design time and array gain for the experiments carried out in this paper. This first approach has been labeled in this work as maximum gain with minimum amplitude distortion response (MGMADR).

6.4.2 Solution approximated by averaging the filter coefficients

The second proposed approach approximates the solution of (6.32) by a two-step procedure. Considering that the anthropometric differences between the heads of different subjects are relatively small, it can be assumed that the individual filter coefficients obtained by solving (6.15) do not differ much from subject to subject. Furthermore, taking into account that the average array gain in (6.25) is composed of a summation of individual terms, an estimation of the filter coefficients can be obtained by averaging the individual coefficients of each subject. The two-step procedures is:

1. Estimation of the coefficients according to $\mathbf{v}_k^{1)} = \frac{1}{S}\sum_{i=1}^{S} \mathbf{w}_k|_{S=s_i}$ where $\mathbf{w}_k|_{S=s_i}$ are the filter coefficients fitted to the i-th subject.

2. Scale the estimated coefficients $\mathbf{v}_k^{1)}$ by d_k according to (6.31), $\mathbf{v}_k^{2)} = d_k \cdot \mathbf{v}_k^{1)}$, to ensure that the approximated coefficients satisfy the constraint in (6.32). This step is very important in order to minimize speech distortion, due to the fact that a simple average of the filter coefficients leads to a high distortion solution.

This second optimization method does not assure a maximum average array gain, but it is much faster than the solution proposed using other methods and, as it will be shown in the results of the next section, it practically matches the results obtained from the first approach based on evolutionary computation. This second approach has been labeled as approximated maximum gain minimum amplitude distortion response (AMGMADR).

6.4.3 Multifrequency optimization with minimum phase distortion

The optimization problem proposed in (6.32) enables to maximize the average array gain, finding the scaling factor d_k that provides the minimum average amplitude distortion $\overline{AD}$. However, the filter coefficients are not directly optimized to minimize amplitude and phase distortions. In this third approach, the optimization problem is reformulated in order to maximize the average array gain and, at the same time, minimize both the average amplitude and phase distortions $\overline{AD}$ and $\overline{PD}$. For this purpose, the next multi-objective minimization problem is formulated

$$\begin{aligned} &\underset{\mathbf{w}_{K_{min}},\ldots,\mathbf{w}_{K_{max}}}{max} \left\{\overline{G} - \alpha\overline{AD} - \beta\overline{PD}\right\} \\ &subject\ to \left(\prod_{s=1}^{S} \left|\mathbf{w}_k^H \cdot \mathbf{b}_{sk}\right|\right)^{\frac{1}{S}} = 1 \ \forall k \in k_{min},\ldots,k_{max}, \end{aligned} \tag{6.33}$$

where α and β are the weighting values that control the influence of the different terms in the optimization. Different values of α and β have been considered, finding $\alpha = 3$ and $\beta = 10$ as a correct tradeoff between gain maximization and distortion reduction. Note that the constraint in (6.33) is only applied to the modulus in this case, but according to (6.30), the value of d_k with minimum $\overline{AD}$ is still ensured. In addition, measuring the linearity of the phase implies all frequency bands, so it is necessary to perform a multifrequency optimization for all the frequencies at once, making the solution to this problem more difficult. A tailored evolutionary algorithm is proposed to solve this particular problem. The novelty in this case is that the best candidates are selected not only considering those that minimize the output noise power, but also those that obtain lower amplitude and phase distortions. The differences with the previous EA described in section 6.4.1 are:

- In step 1, the solution to the second approach, AMGMADR, is also introduced as an initial solution to the population.
- In step 2, only the modulus of the coefficients is scaled.
- In step 3, the average array gain, amplitude distortion $\overline{AD}$ and phase distortion $\overline{PD}$ are computed to evaluate the multi-objective cost function.
- In step 7, the process is reiterated now 1000 times (from step 2).

The candidate solution that achieves the highest value of $\overline{G} - \alpha\overline{AD} - \beta\overline{PD}$ over the design set is selected as the final solution. Again, the values of the parameters of the evolutionary algorithm (population size, crossover rate, mutation scheme and number of generations) have been found to obtain a fairly good tradeoff between the design time and the array gain achieved from the experiments carried out in this paper. This third approach has been labeled as maximum gain with minimum distortion response (MGMDR).

6.4.4 Binaural output beamformer

The output of the MVDR beamformer is a monaural enhanced signal. However, in the case of binaural hearing aids, the enhanced monaural signal must be further converted into a binaural signal, in order to preserve the spatial information. The method proposed in this work to obtain the binaural signal is based on the method described in [Lotter and Vary, 2006], which proposes to use the beamformer monaural output for the calculation of spectral weights that are applied to the array input channels. Unlike the devices considered in that work, which only contain a single microphone, the devices in the current work contain multiple microphones. The original method is extended to this case.

Figure 6.6 shows a block diagram of the proposed method. The output of the MVDR beamformer, $Y(k)$, is obtained by combining the input signals of both the left and right devices, $X_{L1}(k)...X_{LM}(k)$ and $X_{R1}(k)...X_{RM}(k)$ respectively. The enhanced monaural signal $Y(k)$ is used to compute the weights $G_b(k)$, which are the same for both ears in order to preserve the spatial cues. In the original method proposed in [Lotter and Vary, 2006], the weights $G_b(k)$ are computed using the input signals received by the microphones placed in each device, and the output is obtained by weighting these signals with the weights calculated. In order to adapt this schema to the multichannel input case, it is proposed that the signals that are used to calculate the weights and to produce the binaural output are an aligned combination of the input signals received by the microphones of the same device. The aligned combination is obtained by computing a simple DS beamformer with the input channels of each device. The outputs of the left and right DS beamformers, $X_L^{DS}(k)$ and $X_R^{DS}(k)$ respectively, are multiplied

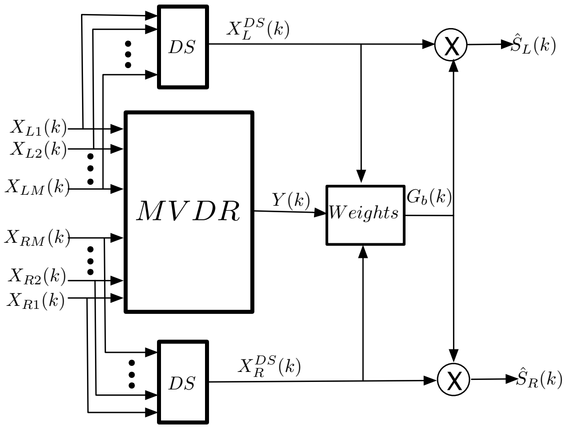


Figure 6.6: Binaural output beamformer.

by the weighting coefficients $G_b(k)$, that are calculated according to expression (6.34). The output is the enhanced binaural signals, $\hat{S}_L(k)$ and $\hat{S}_R(k)$.

$$G_b(k) = \frac{|Y(k)|}{|X_L^{DS}| + |X_R^{DS}|}. \tag{6.34}$$

6.5 Experimental work and results

This section describes the experiments conducted to valuate the performance of the different design methods proposed in this chapter, as well as the results obtained. All the experiments have been performed with the HRTFs of the CIPIC database described in section 2.5.3.3, using a sampling rate of $f_s = 16$ kHz and $K = 129$ frequency bands, which corresponds to the use of a 256-point DFT decomposition. The idea of splitting the database into two different sets of subjects, one for design and another for test, is worthwhile when there is an extensive database available. However, the CIPIC database only contains 45 subjects, which is a reduced number to halve the database without running the risk of over-fitting. Nevertheless, the database can be extended with bootstrapping techniques, using the 'leave-one-out cross-validation' technique (LOOCV) [Efron, 1979]. According to this paradigm, the test set is reduced to only one subject, designing the coefficients with the remaining 44 subjects, repeating the design and evaluation process 45 times, changing the test subject in every iteration. Using this technique, the database limitation is overcome.

In order to compare different array configurations, 13 different microphone arrangements, 6 monaural and 7 binaural, have been assessed. The placement of the microphones within the hearing device is shown in figure 6.7, which has scaled dimensions. The yellow rectangle represents the ITC device contour of 1.5 x 1 cm and the black circles represent the microphones of 0.25 cm of diameter. Binaural arrays are symmetric, containing the same microphone arrangement in both ears, while monaural configurations only consider the left ear, using $H_{Ls}(k, \theta, \phi)$ in that case. Each configuration is labeled according to N-CX-M or N-CX-B, N being the total number of microphones of the array, X a number that indicates the configuration of the microphones, which are represented in figure 6.7, and M and B indicate whether the configuration is monaural or binaural, respectively. Note that the configuration 'C0' is only binaural, the configuration 2-C0-B being an array composed of a microphone placed in each eardrum. The evaluation of these 13 arrays will not provide exactly the best solution, but it is a rough approximation,

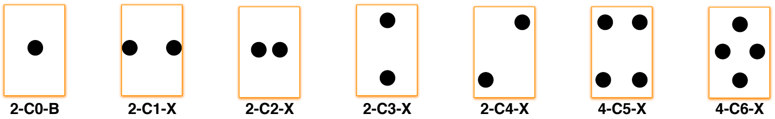


Figure 6.7: Microphone array configurations. Binaural arrays contain the same microphone in both devices. The first number of the label represents the total number of microphones, and the last letter whether it is a monaural or binaural array. Configuration 2-C0-B is only binaural.

due to the fact that the proposed arrays have different orientations and number of elements and they are inside a very small device, where the alternatives for microphone placement are limited. Moreover, the advantages and disadvantages of using binaural microphone arrays and increasing the number of microphones can be evaluated.

To evaluate the different design approaches and array configurations, a set of experiments has been carried out in two different design scenarios: known test HRTF and unknown test HRTF. In the first case, the average array gain is computed according to expression (6.20). In the second case, in addition to the average array gain, the amplitude and phase distortions according to expressions (6.21) and (6.23) and the gain reduction caused by errors in the steering vector according to expression (6.24) are calculated, for each of the microphone configurations shown in figure 6.7. Finally, the speech intelligibility obtained by the different methods and array configurations has been evaluated using the PESQ parameter in two different scenarios, white noise and babble noise.

6.5.1 Known test HRTFs

When the HRTF of the intended user of the array is known, the coefficients can be easily fitted to each subject using the standard MVDR given by equation (6.16). This case is labeled as *Known HRTF-MVDR*. In this case, the gain shown is an average of the gain obtained for all the subjects of the database, calculating the frequency response with their fitted filter coefficients. This is the maximum gain achieved for the arrays with the MVDR method, obtaining a distortionless response. These values will be the reference values to measure gain reduction and distortions when the HRTF is unknown.

Table 6.1 contains the mean value, maximum value, minimum value and standard deviation of the array gain in the speech frequency range for all of the subjects with fitted filter coefficients. Looking at the mean values it can be generalized that the 4-C6-M and 8-C6-B configurations obtain the best array gain for monaural and binaural configurations respectively, regardless of the subject. These configurations have also the lowest deviation from the mean value. Regarding the gain variability from subject to subject, binaural configurations obtain lower deviation than the monaural ones. The improvement introduced by the binaural configurations in comparison to the same monaural configurations is 2.83 dB on average. Furthermore, comparing the different binaural configurations with configuration 2-C0-B, it is noticeable that the increment of the number of microphones leads to meaningful higher gains.

6.5.2 Unknown test HRTFs

In the most common case in which the HRTF of the final user of the hearing aid is unknown, the use of a non-fitted array design causes gain reduction, and amplitude and phase distortions. In a first approach, this performance reduction can be deduced by evaluating the average gain reduction and distortions caused in case a person wears a device containing the filter coefficients

Table 6.1: Array gain (dB) in the speech frequency band averaged over S subjects, in the case of *Known HRTF-MVDR*.

ARRAY	Mean	Max	Min	Std
2-C1-M	5.36	7.30	4.41	0.61
2-C2-M	5.34	7.29	4.40	0.61
2-C3-M	0.51	2.94	-0.53	0.76
2-C4-M	1.72	3.95	0.78	0.67
4-C5-M	5.59	7.48	4.65	0.59
4-C6-M	7.87	9.78	7.02	0.58
2-C0-B	3.35	4.98	2.57	0.51
4-C1-B	8.06	9.10	7.35	0.40
4-C2-B	8.05	9.10	7.33	0.40
4-C3-B	3.52	4.98	2.71	0.52
4-C4-B	4.74	6.27	4.00	0.45
8-C5-B	8.35	9.45	7.70	0.38
8-C6-B	10.62	11.62	9.94	0.36

fitted to another person. This experiment is carried out by evaluating equations (6.20), (6.21) and (6.23) for the test subject using the coefficients fitted to each subject of the design dataset. The gain reduction and distortions of one subject of the test set are calculated by taking the average of the gain reduction and distortions obtained when this subject uses the fitted coefficients to each subject of the design set. This solution is labeled as *Unknown HRTF-MVDR*. In order to minimize the gain reduction and distortions, the three approaches described in section 6.4 have been implemented. They are labeled as *Unknown HRTF-MGMADR, -AMGMADR* and *-MGMDR*, respectively. Furthermore, the results obtained by the proposed optimization methods are compared to the results obtained in the case of computing the microphone responses with the HRTF model proposed in [Brown and Duda, 1997], which is a simple model that approximates the head by a sphere and it only needs a body measurement of the subject. In this work, the radius of the sphere has been obtained by averaging the anthropometric measurements of the distance between ears and the distance between the forehead and the nape, which are also provided in the CIPIC database. This method is labeled as *Unknown HRTF-MODEL*.

Table 6.2 shows the average array gain reduction (%) in the speech frequency range for the five different methods with unknown HRTF, in relation to the array gain obtained with known HRTF (table 6.1). The values have been averaged for all of the subjects in the 5 different methods. The standard MVDR introduces up to 18% of gain reduction, due to the ignorance in the values of the HRTF of the test subject. However, the gain reduction obtained by the HRTF model is notably higher, reaching values up to 42%. Furthermore, the gain reduction is higher in the case of binaural configurations. With the proposed methods, the relative average gain reduction is drastically reduced, being approximately divided by a factor of 2 compared to the standard MVDR with unknown HRTF. There are only slight differences among the gain reduction of the three proposed approaches, MGMADR being the best method in terms of gain reduction, and MGMDR method the worst of these three.

Concerning the amplitude distortion, table 6.3 shows the corresponding averaged results for each method with unknown HRTF. Analyzing the results it is deduced that the uncertainty in the HRTF of the final subject causes an average amplitude distortion of 3.4 dB for all the array configurations (column labeled as MVDR). However, the optimization methods proposed achieve

Table 6.2: Average relative array gain reduction (%) in the speech frequency band, averaged over S subjects, in the case of *Unknown HRTF*.

ARRAY	MVDR	MODEL	MGMADR	AMGMADR	MGMDR
2-C1-M	0.28%	6.53%	0.14%	0.14%	0.18%
2-C2-M	0.28%	6.58%	0.14%	0.15%	0.17%
2-C3-M	8.44%	57.5%	4.37%	4.55%	4.56%
2-C4-M	1.01%	3.56%	0.51%	0.53%	0.56%
4-C5-M	1.32%	10.55%	0.68%	0.71%	0.75%
4-C6-M	0.45%	1.62%	0.23%	0.23%	0.28%
2-C0-B	18.32%	36.03%	9.11%	9.23%	9.40%
4-C1-B	7.93%	21.12%	3.94%	4.00%	4.10%
4-C2-B	7.95%	21.17%	3.94%	4.01%	4.11%
4-C3-B	18.84%	42.82%	9.39%	9.52%	9.67%
4-C4-B	13.40%	27.37%	6.65%	6.76%	6.88%
8-C5-B	8.56%	23.79%	4.30%	4.33%	4.43%
8-C6-B	6.10%	13.03%	3.01%	3.07%	3.15%

Table 6.3: Amplitude distortion $\overline{AD}$ (dB) in the speech frequency band, averaged over S subjects, in the case of *Unknown HRTF*.

ARRAY	MVDR	MODEL	MGMADR	AMGMADR	MGMDR
2-C1-M	3.3802	3.2574	2.4172	2.4172	2.4172
2-C2-M	3.3802	3.2574	2.4172	2.4172	2.4172
2-C3-M	3.3802	3.2574	2.4172	2.4172	2.4172
2-C4-M	3.3802	3.2574	2.4172	2.4172	2.4172
4-C5-M	3.3802	3.2574	2.4172	2.4172	2.4172
4-C6-M	3.3802	3.2574	2.4172	2.4172	2.4172
2-C0-B	3.3659	3.1995	2.3590	2.3767	2.3683
4-C1-B	3.3660	3.2289	2.3589	2.3767	2.3671
4-C2-B	3.3667	3.2288	2.3597	2.3743	2.3644
4-C3-B	3.3695	3.1993	2.3606	2.3778	2.3665
4-C4-B	3.3676	3.2009	2.3593	2.3767	2.3711
8-C5-B	3.3634	3.2290	2.3608	2.3770	2.3715
8-C6-B	3.3669	3.2353	2.3600	2.3751	2.3614

a lower amplitude distortion around 2.4 dB for all the array configurations included in this work. The amplitude distortion obtained by the HRTF model is slightly lower than the one obtained by the standard MVDR, but in any case around 1 dB higher than the proposed methods. Note that in the case of monaural configurations, the amplitude distortion is constant regardless of the array configuration. In the monaural case, the term of the HRTF introduced in the approximation carried out in equation (6.19) is the same for all the microphones, and therefore the constraints imposed in equations (6.32) and (6.33) make the array response independent of the array configuration. This fact causes that the amplitude distortion is the same for all monaural configurations.

The results corresponding to phase distortion are included in table 6.4. It is appreciated that the uncertainty in the HRTF of the final subject causes phase distortions between 0.141 and 0.145 radians for all of the array configurations, the phase distortion obtained by the model being slightly lower. In the case of the methods proposed, MGMADR and AMGMADR obtain phase distortions between 0.106 and 0.108 radians, and MGMDR, which includes phase distortion reduction in the optimization, achieves phase distortions between 0.099 and 0.101 radians.

Table 6.4: Phase distortion $\overline{PD}$ (radians) in the speech frequency band, averaged over S subjects, in the case of *Unknown HRTF*.

ARRAY	MVDR	MODEL	MGMADR	AMGMADR	MGMDR
2-C1-M	0.1410	0.1049	0.1077	0.1077	0.1008
2-C2-M	0.1410	0.1049	0.1076	0.1077	0.1009
2-C3-M	0.1410	0.1049	0.1077	0.1077	0.1009
2-C4-M	0.1410	0.1049	0.1075	0.1077	0.1009
4-C5-M	0.1410	0.1049	0.1078	0.1077	0.1008
4-C6-M	0.1410	0.1049	0.1076	0.1077	0.1009
2-C0-B	0.1441	0.1363	0.1061	0.1067	0.0996
4-C1-B	0.1444	0.1496	0.1061	0.1069	0.0999
4-C2-B	0.1443	0.1496	0.1063	0.1069	0.0997
4-C3-B	0.1440	0.1363	0.1060	0.1067	0.0998
4-C4-B	0.1436	0.1368	0.1061	0.1067	0.0999
8-C5-B	0.1442	0.1489	0.1063	0.1069	0.0998
8-C6-B	0.1450	0.1428	0.1063	0.1071	0.0999

Table 6.5: Gain reduction (dB) due to errors in the steering vector in the speech frequency band, averaged over S subjects, in the case of *Known* and *Unknown HRTF*.

	Known HRTF	*Unknown HRTF*				
ARRAY	MVDR	MVDR	MODEL	MGMADR	AMGMADR	MGMDR
2-C1-M	0.22	0.22	0.21	0.22	0.22	0.22
2-C2-M	0.22	0.22	0.21	0.22	0.22	0.22
2-C3-M	0.11	0.11	0.21	0.11	0.11	0.11
2-C4-M	0.19	0.19	0.19	0.19	0.19	0.18
4-C5-M	0.24	0.24	0.26	0.23	0.23	0.23
4-C6-M	0.38	0.38	0.38	0.38	0.38	0.38
2-C0-B	1.08	0.79	0.87	0.81	0.81	0.81
4-C1-B	1.23	0.91	1.04	0.93	0.93	0.93
4-C2-B	1.23	0.91	1.04	0.93	0.93	0.93
4-C3-B	1.09	0.80	0.88	0.82	0.82	0.82
4-C4-B	1.17	0.87	0.97	0.89	0.89	0.89
8-C5-B	1.25	0.93	1.09	0.94	0.94	0.94
8-C6-B	1.40	1.08	1.20	1.09	1.09	1.09

The effects that the errors in the steering direction have on the array gain are included in table 6.5, in both known HRTF and unknown HRTF cases. This table shows the average gain reduction related to the expected steering direction, computed according to equation (6.24). The value of D used in this work is 2, corresponding to variations of ±10 in azimuth and ±11.25 in elevation, according to the angles provided in the CIPIC database. The first column contains the gain reduction in the case of using the standard MVDR with known HRTF. It is easily deduced that for all the methods binaural configurations are clearly more affected than monaural configurations. The gain reduction introduced by the HRTF model is higher, as it was expected, but the gain reduction introduced by the proposed methods is even lower than the one obtained by the fitted filter coefficients with known HRTF, this reduction being higher in binaural arrays. This unexpected behavior makes the proposed optimization methods more robust against errors in the steering vector. Finally, it is worth clarifying that the gain reduction values shown in this table are relative to the gain obtained for the steering direction of each method.

Table 6.6: Average and standard deviation of the PESQ value calculated in 100 speech and white noise mixtures with a SNR of 10 dB, in the case of *Known* and *Unknown HRTF*.

	Known HRTF		*Unknown HRTF*									
ARRAY	*MVDR*		*MVDR*		*MODEL*		*MGMADR*		*AMGMADR*		*MGMDR*	
	Mean	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*
2-C1-M	3.0	0.19	2.4	0.18	2.9	0.19	3.0	0.19	3.0	0.19	3.0	0.19
2-C2-M	3.0	0.19	2.4	0.18	2.9	0.19	3.0	0.19	3.0	0.19	3.0	0.19
2-C3-M	2.7	0.20	2.2	0.19	2.6	0.19	2.7	0.19	2.7	0.19	2.7	0.19
2-C4-M	2.7	0.20	2.2	0.19	2.7	0.19	2.7	0.19	2.7	0.19	2.7	0.19
4-C5-M	3.0	0.19	2.4	0.18	2.9	0.19	3.0	0.19	3.0	0.19	3.0	0.19
4-C6-M	3.1	0.19	2.6	0.18	3.1	0.19	3.1	0.19	3.1	0.19	3.1	0.19
2-C0-B	2.9	0.19	2.3	0.18	2.9	0.17	2.8	0.19	2.8	0.19	2.8	0.19
4-C1-B	3.2	0.18	2.5	0.18	3.1	0.18	3.2	0.18	3.2	0.18	3.2	0.18
4-C2-B	3.2	0.19	2.5	0.18	3.1	0.18	3.2	0.18	3.2	0.19	3.2	0.19
4-C3-B	2.9	0.19	2.3	0.18	2.8	0.17	2.9	0.19	2.7	0.19	2.9	0.19
4-C4-B	2.9	0.19	2.3	0.18	2.8	0.17	2.9	0.19	2.7	0.19	2.9	0.19
8-C5-B	3.2	0.19	2.5	0.18	3.1	0.18	3.2	0.19	3.2	0.19	3.2	0.19
8-C6-B	3.3	0.19	2.7	0.18	3.2	0.18	3.3	0.19	3.3	0.19	3.3	0.19

Table 6.7: Average and standard deviation of the PESQ value calculated in 100 speech and babble noise mixtures with a SNR of 10 dB, in the case of *Known* and *Unknown HRTF*.

	Known HRTF		*Unknown HRTF*									
ARRAY	*MVDR*		*MVDR*		*MODEL*		*MGMADR*		*AMGMADR*		*MGMDR*	
	Mean	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*
2-C1-M	2.9	0.14	1.4	0.16	2.8	0.13	2.9	0.14	2.9	0.14	2.9	0.14
2-C2-M	2.9	0.14	1.4	0.16	2.8	0.14	2.9	0.14	2.9	0.14	2.9	0.14
2-C3-M	2.3	0.16	1.2	0.18	2.3	0.15	2.3	0.16	2.3	0.16	2.3	0.16
2-C4-M	2.4	0.15	1.2	0.17	2.4	0.14	2.4	0.15	2.4	0.15	2.4	0.15
4-C5-M	2.8	0.14	1.4	0.16	2.8	0.14	2.9	0.14	2.9	0.14	2.9	0.14
4-C6-M	3.0	0.14	1.5	0.16	2.9	0.14	3.0	0.14	3.0	0.14	3.0	0.14
2-C0-M	2.4	0.15	1.2	0.17	2.3	0.15	2.4	0.15	2.4	0.15	2.4	0.15
4-C1-M	3.0	0.14	1.5	0.16	2.9	0.14	3.0	0.14	3.0	0.14	3.0	0.14
4-C2-M	3.0	0.14	1.5	0.16	2.9	0.13	3.0	0.14	3.0	0.14	3.0	0.14
4-C3-M	2.4	0.16	1.2	0.17	2.4	0.15	2.4	0.15	2.4	0.15	2.4	0.15
4-C4-M	2.5	0.15	1.3	0.17	2.5	0.15	2.5	0.14	2.5	0.14	2.5	0.14
8-C5-M	2.9	0.14	1.4	0.16	2.8	0.13	2.9	0.14	2.9	0.14	2.9	0.14
8-C6-M	3.1	0.14	1.5	0.16	3.0	0.14	3.1	0.14	3.1	0.14	3.1	0.14

In addition to compute the amplitude and phase distortions to measure the speech quality, the PESQ parameter is calculated in two different scenarios: white noise and babble noise. The PESQ evaluation is performed by mixing a speech source located in the straight ahead direction with the two types of noise, with a SNR of 10 dB. The first type of noise is spatially uncorrelated white noise coming from all the other directions defined in the CIPIC database, and the second type of noise is babble noise generated by random speech sources coming also from the directions defined in the CIPIC database. A total of 100 simulations changing the target speech source have been executed for each method and array. The speech sources have been selected randomly from the TIMIT database. The averaged PESQ values and standard deviations are included in table 6.6 for white noise and table 6.7 for babble noise. Comparing the results obtained with the standard MVDR with known HRTF with the ones obtained with the standard MVDR in the case of unknown HRTF, an average PESQ reduction of 20% for white noise and 50% for babble noise is deduced. The values obtained by the HRTF model are slightly lower (around 2.5% in both noises) than the ones obtained with the fitted coefficients, but they are acceptable. However, the PESQ values obtained by the proposed optimization methods are similar than the ones obtained in the case of known HRTF, with an unappreciable decrease around 0.26% in the case of MGMADR and MGMDR, and 1.33% in the case of AMGMADR. The relative behavior of the different methods in terms of PESQ is the same for both types of noise, but in the case of babble noise all PESQ values are around 9% lower in average. The low standard deviations obtained in both cases guarantee that there is not large variations of speech quality between the different subjects.

6.6 Discussion

This chapter tackles the problem of designing microphone arrays for speech enhancement in monaural and binaural ITC hearing aids, considering the head shadow effects, which are modeled with the so-called HRTFs. When the HRTF of the subject is known, the filter coefficients are easily optimized using a standard MVDR beamformer. However, the availability of these HRTFs in real scenarios is limited, and the filter coefficients cannot be fitted to the subject, thus causing gain reduction and introducing both amplitude and phase distortions in the speech signal. The parameters $\overline{AD}$ and $\overline{PD}$ that measure these two distortions over a set of subjects have been introduced. Furthermore, three different approaches to optimize the filter coefficients by maximizing the average array gain while minimizing the average distortions have been proposed, using a design dataset. In addition, the proposed methods have been evaluated with 13 different array configurations, 6 monaural and 7 binaural.

The results of the experiments carried out in this work demonstrate that the proposed optimization methods reduce significantly the gain reduction and distortions caused by computing the filter coefficients with unknown HRTF of the subject. The methods are also more robust to errors in the steering vector. Furthermore, it has been shown that the use of a simple HRTF model that does not entail tedious anthropometric measurements notably decreases the array gain, although the amplitude and phase distortions do not worsen significantly. Comparing the three optimization methods proposed, it is deduced a tradeoff between distortion reduction and computational cost of the design stage. When computation is not a constraint, the MGMDR solution is the best one, even though it obtains slightly higher gain reduction than the other two options, but with the smallest distortions. On the other hand, if the computation time is a restriction, the AMGMADR solution is the most suitable. Finally, regarding the different array configurations, it has been found that binaural arrays improve 2.83 dB on average the gain obtained by the monaural arrays, as well as an increase in the number of microphones in binaural arrays from 2 to 8 raises drastically the gain in 7.3 dB.

6.7 Summary of contributions

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

- A simple expression to calculate the directivity pattern introducing the HRTF of the subject is proposed. The frequency response of any microphone placed inside a hearing aid centered in the origin of coordinates is approximated by multiplying the free-field microphone response by the HRTF.

- Measurements for gain reduction and amplitude and phase distortions caused by the use of filter coefficients that are not fitted to a specific subject of a database have been proposed. Additionally, a measure of the gain reduction due to errors in the steering direction has also been proposed.

- Three different approaches to optimize the filter coefficients in case of unknown HRTF have been proposed. The three methods aim at maximizing the average array gain while minimizing the average distortions, using a design dataset. The results of the experiments carried out in this work have demonstrated that the proposed optimization methods reduce significantly the gain reduction and distortions caused by computing the filter coefficients with unknown HRTF of the subject. The methods are also more robust to errors in the steering vector. Additionally, the proposed methods have been compared to a simple HRTF model that does not entail tedious anthropometric measurements. The HRTF model notably decreases the array gain, although the amplitude and phase distortions do not worsen significantly.

- A schema to generate a binaural output from two multichannel arrays, one placed in each ear, has been proposed. The schema combines the input channels of each device with a DS beamformer and applies to their output a spectral gain that is calculated from the monaural output of the MVDR beamformer.

- Different microphone array configurations have been compared, including monaural and binaural arrays. The improvement obtained by binaural arrays has been amply demonstrated. The increment in the number of microphones that compose the array has also important benefits in the array gain.

These contributions have been published in [Ayllón et al., 2013a] and [Ayllón et al., 2011b].

Chapter 7

Conclusions

This chapter summarizes the main contributions of the thesis and analyses the results obtained along the thesis. It also includes a description of the future research lines. The chapter concludes with a list of the publications obtained from the main contributions of this thesis.

7.1 Summary of conclusions

In this thesis, sound source separation methods and microphone array signal processing techniques have been combined with machine learning and evolutionary computation, with the main objective of enhancing corrupted speech in audiological applications. First, a comprehensive review of the state of the art in this field has been carried out, as well as the the basis for speech enhancement in the time-frequency domain have been studied.

In this section, the contributions and conclusions derived from the research work carried out to fulfill each of the goals of this thesis are described. Each of the successive subsections is related to one of the goals of the thesis.

7.1.1 Improvement of time-frequency SSS methods based on clustering

The DUET algorithm is one of the most important algorithms for SSS in the time-frequency domain, performing separation using only two mixtures. One of the steps of the algorithm is the clustering of estimates of the level and time differences between two microphones, which is performed by a two-dimensional weighted histogram. The clustering step allows estimating the mixing parameters by identifying the peaks of the clusters. Unfortunately, in some cases, the clusters are not well defined and the performance of the algorithm is drastically affected. In order to improve the performance of the DUET algorithm, the next contributions have been made.

- The first step of the research has been the evaluation of the DUET algorithm for the separation of different types of mixtures and sources, including linear and binaural anechoic mixtures, echoic mixtures, and mixtures of speech with other types of sources such as noise and music. These experiments demonstrate the need for more advanced clustering techniques.

- After finding the limitations of the DUET algorithm in some situations, a novel source separation algorithm that combines the mean shift clustering technique with the basis of DUET has been proposed. The clustering step in DUET, which is based on a weighted

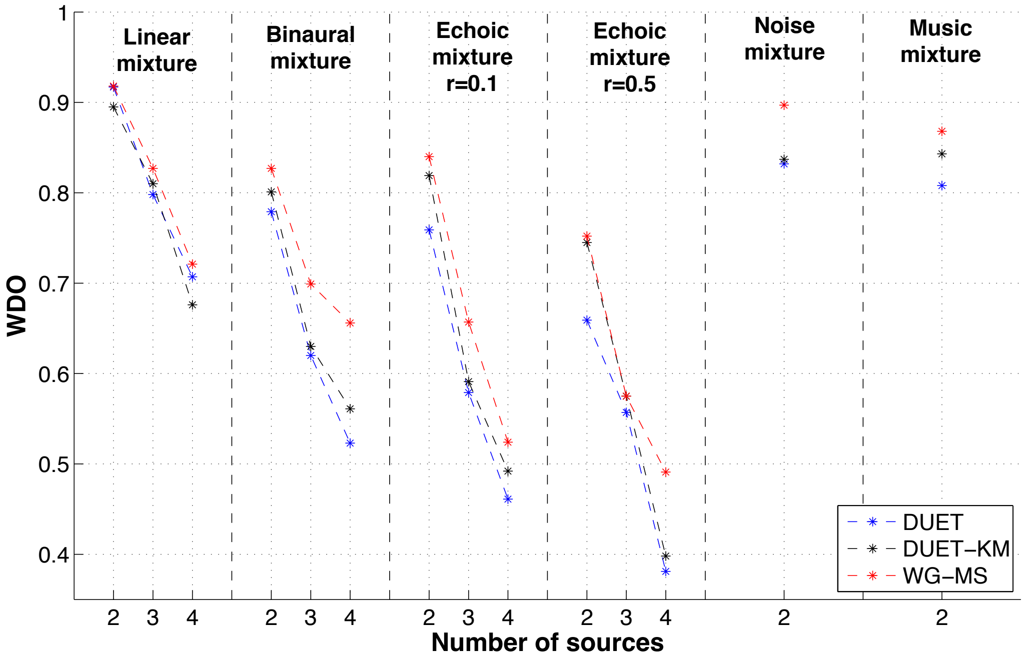


Figure 7.1: Average WDO values obtained by the DUET, DUET-KM and the proposed WG-MS algorithms for the separation of 2, 3 and 4 sources in different types of mixtures.

histogram, is replaced by a generalized version of the mean shift algorithm. A weighted-Gaussian kernel mean shift vector has been inferred and included in an iterative process to clusterize the bi-dimensional feature input space composed of the level and time differences between the two microphones. The proposed WG-MS algorithm has been tested in different scenarios: linear and binaural anechoic speech mixtures, echoic speech mixtures with different reverberation coefficients, and anechoic mixtures of speech with noise and speech with music. The WG-MS algorithm has been compared with the original DUET algorithm and with a modification thereof which introduces the k-means algorithm in the clustering step.

- The WG-MS algorithm, which is originally proposed for two microphones, has been extended to the case of any number of microphones and array geometry. The weighted Gaussian mean shift algorithm previously proposed allows clustering a feature space of any dimension. However, it is necessary to infer a new expression for the ML source estimator to consider any number of microphones. Several experiments varying the number of microphones support the suitability of the method, which shows a special robustness in the case of echoic mixtures.

The main results obtained by the WG-MS are summarized in figure 7.1, which represents the average WDO values obtained by the DUET, DUET-KM and the proposed WG-MS algorithms for the separation of 2, 3 and 4 sources in different types of mixtures. The results demonstrate that the WG-MS algorithm clearly outperforms the original DUET and its modification using k-means in the clustering step (DUET-KM). The improvement is amply noticeable in case of binaural and echoic mixtures and when speech is mixed with non-speech sources, cases in which the clusters are not so well defined in comparison to the linear anechoic case. The replacement of the DUET histogram by a simple clustering technique such as k-means clearly improves the results, which are largely outperformed by the proposed WG-MS algorithm in any case.

In summary, the proposed WG-MS algorithm has obtained excellent results, notably improving the ones obtained by the original DUET algorithm. The generalization of the proposed algorithm for an array of any number of microphones and geometry increases the robustness of the method in reverberant mixtures.

7.1.2 Solutions for the speech source enumeration problem

Determining the number of speech sources is a critical first step for some SSS algorithms, which sometimes assume to know the number of sources in advance. A novel speech source enumeration algorithm has been proposed in this thesis. The algorithm uses an AR model to estimate the PDF of the time differences between two microphones. An information theoretic criteria is used to estimate the order of the AR model, which determines the number of sources in the mixture.

The results obtained by the algorithm for the enumeration of 2, 3, 4 and 5 speech sources in anechoic mixtures are shown in figure 3.10. The number of sources is estimated with high accuracy for 2 and 3 sources in the mixture. When the number of sources increases, the estimation error also increases, but it is as low as 20% for 5 sources in the mixture, which is a noticeable good value for speech enumeration.

In short, the algorithm has obtained very good results and it has shown good robustness in the enumeration of anechoic mixtures up to 5 speech sources. Additionally, the potential of the algorithm to enumerate sources in echoic mixtures has been demonstrated.

7.1.3 Speech enhancement to improve intelligibility in monaural hearing aids

The main goal of this thesis is the design of speech enhancement algorithms for audiological applications, specifically for hearing aids. Traditional single-channel speech enhancement algorithms may improve the SNR, but they can not yet prove to enhance the speech intelligibility. However, the requirements in hearing aids are higher that the ones found in other type of applications: the improvement of intelligibility is crucial and the computational cost of the algorithms should be relatively low.

Considering these two requirements, the first goal of the thesis related to hearing aids is the development of an algorithm that increases the intelligibility of corrupted speech in monaural hearing aids. The proposed algorithm combines supervised machine learning and evolutionary computation to estimate a time-frequency mask that aims at maximizing an objective measurement correlated with intelligibility. The main contributions related to this objective are listed below.

- Three different time-frequency masks have been proposed and compared to the CASA IBM for single-channel noise reduction. One of them is also a binary mask and the other two are soft masks. The experiments carried out have demonstrated that the use of soft masks instead of binary masks is beneficial for single-channel speech enhancement.

- It has been proved that the performance of the time-frequency masks depends on the frequency resolution of the STFT used to compute the time-frequency representation of the signals.

- A study of the computational resources available for signal processing in state-of-the-art commercial hearing aids has been carried out. The result of this study has been used to limit the computational cost of the speech enhancement algorithm for hearing aids proposed in this thesis.

- A generalization of the least squares estimator (GLSE) is proposed. The estimator is adapted to use any transformation of the input features.
- A novel set of features to estimate the time-frequency mask has been proposed. The main novelty resides in the fact that the information of neighbor time-frequency points is included as input features. Additionally, two different alternatives to introduce the information regarding to the previous time frames have been proposed: the use of instantaneous values and the use of different EWMA.
- A low-cost algorithm for single-channel speech enhancement in monaural hearing aids has been proposed. The algorithm aims at maximizing the output PESQ score with a tailored optimization algorithm that uses a previous estimation of the proposed Wiener soft mask, which is estimated with the GLSE. In order to reduce the computational cost of the proposed speech enhancement algorithm, a feature selection algorithm based on evolutionary computation has also been proposed. The results obtained with the proposed algorithm have shown good intelligibility increase using a small part of the available computational resources.
- A relationship between the PESQ score and the SNR has been obtained, using the NOIZEUS database.

The results of the proposed algorithm for monaural hearing aids are summarized in table 4.1. The table contains the PESQ values corresponding to the unprocessed mixture and the ones obtained by the proposed algorithm in the test set using different number of features. The proposed algorithm clearly improves the output speech quality, even for a low SNR of -5 dB. An average improvement of 0.31 in the output PESQ score is obtained using only 7.4 % of the computational resources available for signal processing, which is equivalent to an increment of 6 dB in the SNR. Additionally, the computational cost can be further reduced to only a 5.1% obtaining an average improvement of 0.19 in the output PESQ score, which is still a good value and equivalent to 4.5 dB.

In conclusion, the proposed algorithm has obtained good speech intelligibility improvement using a very small part of the available computational resources in hearing aids. The use of supervised machine learning has been found to be a computationally efficient alternative to the use of traditional CASA techniques in the estimation of time-frequency masks.

7.1.4 Energy efficient speech enhancement in binaural hearing aids

Binaural hearing aids provide important benefits associated to binaural hearing, but they require the exchange of information between the left and the right devices. Due to aesthetic reasons, the best solution is the use of a wireless link for data transmission, which increments the power consumption.

One of the objectives in this thesis is the design of speech enhancement systems that increase the energy efficiency of wireless-communicated binaural hearing aids. In addition to the requirements of monaural devices, binaural systems should optimize the data transmission in order to reduce the power consumption associated to the wireless link. The proposed algorithms are based on supervised machine learning and they are optimized by means of evolutionary computation.

Two different algorithms for speech enhancement in binaural hearing aids have been proposed, originating the next contributions.

- An extremely low-cost binaural speech separation system that maximizes the WDO has been proposed. It is based on a quadratic discriminant that uses the ILD and ITD to classify each time-frequency point between speech or noise. The weights are calculated with a tailored evolutionary algorithm that aims at maximizing the WDO factor.

- A generalized version of the LS-LDA has been proposed. The LS-GDA allows applying any type of transformation to the input features, in order to obtain non-linear decision boundaries.

- A low-cost binaural speech separation system that estimates the IBM using the information of the current and neighbor time-frequency points has been proposed. It is based on the LS-GDA and its computational cost directly depends on the number of input features considered for classification.

- A weighted MSE metric has been introduced into the LS-GDA. The metric allows estimating the IBM and maximizing the WDO factor at the same time.

- A novel set of features to classify a time-frequency point between speech or noise is presented. The main novelty resides in considering the information related to the neighbor time-frequency points for classification.

- A transmission schema to enhance the energy efficiency of the wireless-communicated binaural hearing aids has been introduced. The schema quantizes the amplitude and phase values of each frequency band with a different number of bits. The bit distribution among frequency bands is optimized by means of evolutionary computation.

The first proposed algorithm needs lower computational resources than the second one, but the latter obtains better results and it is clearly feasible to be implemented in a hearing aid. The results of both algorithms are compared in figure 5.6, which represents the WDO values averaged over the test set as a function of the number of bits transmitted from the right to the left device. The results show excellent separation performance even for low SNR when transmitting at 512 kbps, but the performance is still very good when the bit rate is reduced down to 16 kbps, and acceptable even only transmitting 2 kbps.

To summarize, the two proposed algorithms have obtained good speech intelligibility improvement using low computational resources and low transmission bit rate, which supports the feasibility of the algorithms to be implemented in commercial hearing aids.

7.1.5 Design of optimized microphone arrays for unknown subjects

Concerning the design of microphone arrays for hearing aids, an interesting problem has been detected: how to customize a spatial filter to a determined person when the information related to his head (i.e. HRTF) is not available. The objective in this thesis is to generalize the design of customized microphone arrays for speech enhancement in monaural and binaural hearing aids. With this respect, three different optimization methods have been proposed, aiming at minimize the gain reduction and distortions introduced by the lack of such information.

The next contributions related to this problem have been made in this thesis.

Table 7.1: Array gain obtained in the case of known HRTF, and averaged gain reduction, averaged amplitude distortion and averaged phase distortion in the case of unknown HRTF using a standard MVDR, and the proposed MGMDR method, for the monaural array 4-C6-M and the binaural arrays 2-C0-B and 8-C6-B.

ARRAY	$\overline{G}$	$\overline{G}$ loss		$\overline{AD}$		$\overline{PD}$	
	Known HRTF	MVDR	MGMDR	MVDR	MGMDR	MVDR	MGMDR
4-C6-M	7.87 dB	0.45 %	0.28 %	3.38 dB	2.42 dB	0.141 rad	0.100 rad
2-C0-B	3.35 dB	18.32 %	9.40 %	3.37 dB	2.37 dB	0.144 rad	0.099 rad
8-C6-B	10.62 dB	6.10 %	3.15 %	3.37 dB	2.36 dB	0.145 rad	0.099 rad

- A simple expression to calculate the directivity pattern of an array considering the HRTF of the subject is proposed. The frequency response of any microphone placed inside a hearing aid centered in the origin of coordinates is approximated by multiplying the free-field microphone response by the HRTF function.

- Measurements for gain reduction and amplitude and phase distortions caused by the use of beamforming filter coefficients that are not fitted to a specific subject of a database have been proposed. Additionally, a measure of the gain reduction due to errors in the steering direction has also been proposed.

- Three different approaches to optimize the beamforming filter coefficients in case of unknown HRTF have been proposed. The three methods aim at maximizing the average array gain while minimizing the average distortions, using a design dataset. The results of the experiments carried out in this work have demonstrated that the proposed optimization methods reduce significantly the gain reduction and distortions caused by computing the filter coefficients with unknown HRTF of the subject. The methods are also more robust to errors in the steering vector. Additionally, the proposed methods have been compared to a simple HRTF model that does not entail tedious anthropometric measurements. The HRTF model notably decreases the array gain, although the amplitude and phase distortions do not worsen significantly.

- A schema to generate a binaural output from two multichannel arrays, one placed in each hearing aid, has been proposed. The schema combines the input channels of each device with a DS beamformer and applies to their outputs a spectral gain that is calculated from the monaural output of the MVDR beamformer.

- Different microphone array configurations have been compared, including monaural and binaural arrays, in terms of the array gain obtained by a MVDR beamformer. The improvement obtained by binaural arrays has been amply demonstrated. The increment in the number of microphones that compose the array has also important benefits in the array gain.

Table 7.1 summarizes the main results obtained by the proposed optimization method MGMDR, which represents the best tradeoff between gain reduction and distortions among the three proposed methods. The gain reduction and the amplitude and phase distortions caused by the lack of knowledge of the HRTF are clearly reduced by the proposed optimization algorithm. The increment in the array gain obtained by the use of binaural arrays is also demonstrated, as well as the increment obtained by the use of arrays with higher number of microphones.

In brief, the results obtained show an important improvement in terms of array gain and speech distortions, leading to a noticeable improvement in the speech intelligibility of the output signal. Additionally, the use of binaural arrays has been justified.

7.2 Future research lines

The research carried out in this thesis is not intended to close any research line, but tries to expand the knowledge in the research field of speech enhancement and its application to engineering constrained problems such as hearing aids. Indeed, some of the algorithms proposed in this thesis open new and promising research lines. Considering the theoretical studies and the results derived in this thesis, the next future research lines are proposed:

- The WG-MS algorithm for SSS proposed in chapter 3 was generalized to perform separation with microphone arrays of any number of elements and geometry. Although the results of the experiments carried out with echoic mixtures are promising, the study should be extended with different array configurations and with real measurements.

- The speech source enumeration method proposed in chapter 3 has shown excellent performance for anechoic mixtures and demonstrated its potential to perform source enumeration in echoic mixtures. The presented results are promising, but further research is necessary in order to determine the robustness in echoic and noisy environments, the dependence on the relative positions of speech sources as well as their energy.

- The performance of different time-frequency masks in terms of PESQ score has been evaluated for single-channel speech enhancement. The time-frequency decomposition has been based on the STFT, but the effects of using other transformations such as a cochleagram or a discrete wavelet transform should be further studied.

- The proposed speech enhancement algorithm for monaural hearing aids has been designed to maximize the PESQ score, which is a metric correlated with intelligibility. However, some other metrics higher correlated with the intelligibility exist, and the same design approach can be applied to maximize any other metric.

- The metrics that correlate with intelligibility in the case of normal hearing people do not necessarily correlate in the same way for hearing-impaired people. In order to ensure that the designed algorithms also improve the speech intelligibility in people with hearing loss, new experiments should be carried out with hearing-impaired people.

- The binaural speech enhancement systems proposed in chapter 5 use time-frequency binary masks. The effect of using soft masks can be further studied, although it would imply higher data transmission.

- The algorithms designed in chapter 6 are useful to optimize the beamformer filter coefficients when the HRTF of the subject is not available. Although they are designed to minimize the gain reduction and distortions, they showed an increase in the robustness against errors in the steering vector. However, this robustness should be increased using robust beamforming techniques, and the proposed optimization techniques can be expanded to design this kind of beamformers.

7.3 List of publications

Due to the contributions done in this thesis, several international journal and conference proceeding papers have already been published and submitted. The works accepted and published in journals and conference proceedings, as well as those under review, that support the main contributions of this thesis, are listed below.

- **International journals**

  - D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera. (2013). An algorithm that improves the intelligibility of speech in noise for monaural hearing aids. *Speech Communication.* (Submitted).
  - D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera. (2013). Rate-constrained source separation for speech enhancement in wireless-communicated binaural hearing aids. *Eurasip Journal on Advances in Signal Processing.* (Submitted).
  - D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera. (2013). Design of microphone arrays for hearing aids optimized to unknown subjects. *Signal Processing*, 93(11), 3239-3259.
  - D. Ayllón, R. Gil-Pita, P. Jarabo-Amores, and M. Rosa-Zurera. (2012). Speech source separation using a generalized mean shift algorithm. *Signal Processing*, 92(9), 2248-2252.
  - R. Gil-Pita, L. Cuadra, E. Alexandre, D. Ayllón, L. Álvarez, and M. Rosa-Zurera. (2012). Enhancing the energy efficiency of wireless-communicated binaural hearing aids for speech separation driven by soft-computing algorithms. *Applied Soft Computing*, 12(7), 1939-1949.

- **International conferences**

  - D. Ayllón, R. Gil-Pita, M. Rosa-Zurera, and H. Krim. (2013, May). An information theoretic approach for speech source enumeration. In *IEEE International Conference in Acoustics, Speech and Signal Processing 2013 (ICASSP'13).* IEEE.
  - D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera. (2013, October). Speech separation with microphone arrays using the mean-shift algorithm. In *Audio Engineering Society Convention 133.* Audio Engineering Society.
  - D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera. (2011, December). Optimum microphone array for monaural and binaural in-the-canal hearing aids. In *4-th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP 2011).* IEEE.
  - D. Ayllón, R. Gil-Pita, P. Jarabo-Amores, M. Rosa-Zurera, and C. Llerena-Aguilar. (2011, June). Energy-weighted mean shift algorithm for speech source separation. In *IEEE Statistical Signal Processing Workshop (SSP 2011).* IEEE.
  - D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera. (2010, November). Estimation of the Probability Density Function of the Interaural Level Differences for Binaural Speech Separation. In *Audio Engineering Society Convention 129.* Audio Engineering Society.

Some other journal and conference papers also related to the content of this thesis are the next:

- **International journals**

  - D. Ayllón, R. Gil-Pita, M. Utrilla-Manso and M. Rosa-Zurera. (2013). An evolutionary algorithm to optimize the microphone array configuration for speech acquisition in vehicles. *Engineering Applications of Artificial Intelligence.* (Submitted).

- **International conferences**

  - D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera. (2013, May). Monaural Speech Source Separation by Estimating the Power Spectrum Using Multi-Frequency Harmonic Product Spectrum. In *Audio Engineering Society Convention 134.* Audio Engineering Society.
  - D. Ayllón, R. Gil-Pita, M. Utrilla-Manso and M. Rosa-Zurera. (2012, June). Noise model for car-embedded speech acquisition system design. In *IEEE Intelligent Vehicles Symposium (IV).* IEEE.
  - D. Ayllón, V. Benito-Olivares, C. LLerena-Aguilar, R. Gil Pita, and M. Rosa Zurera. (2012, March). 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.
  - C. Llerena, R. Gil-Pita, L. Álvarez, L. Cuadra, and D. Ayllón. (2012, March). 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.
  - D. Ayllón, A. Levi, and H. Silverman. (2011). Real-time phase-isolation algorithm for speech separation. In *19th European signal processing conference (EUSIPCO).* Eurasip.
  - R. Ahmed, R. Gil-Pita, D. Ayllón, and L. Alvarez-Perez. (2011, May). Speech Source Separation Using a Multi-Pitch Harmonic Product Spectrum Based Algorithm. In *Audio Engineering Society Convention 130.* Audio Engineering Society.

Bibliography

F. Abrard and Y. Deville. A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. *Signal Processing*, 85(7):1389–1403, 2005.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

A. N. Akansu and P. R. Haddad. *Multiresolution signal decomposition: transforms, subbands, and wavelets.* Academic Press, 2000.

E. Alexandre, E. Cuadra, M. Rosa, and F. López-Ferreras. Feature selection for sound classification in hearing aids through restricted search driven by genetic algorithms. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2249–2256, 2007.

V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 99–102, 2001.

J. Allen. Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(3):235–238, jun 1977.

J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65:943, 1979.

J. B. Allen, D. A. Berkley, and J. Blauert. Multimicrophone signal-processing technique to remove room reverberation from speech signals. *The Journal of the Acoustical Society of America*, 62:912, 1977.

ANSI. *American National Standard: Methods for Calculation of the Speech Intelligibility Index.* Acoustical Society of America, 1997.

S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada. Underdetermined blind separation for speech in real environments with sparseness and ICA. In *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, volume 3, pages 881–884, 2004.

S. Araki, H. Sawada, R. Mukai, and S. Makino. Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. *Signal Processing*, 87(8):1833–1847, 2007.

S. Arberet, R. Gribonval, and F. Bimbot. A robust method to count and locate audio sources in a multichannel underdetermined mixture. *IEEE Transactions on Signal Processing*, 58(1): 121–133, 2010.

D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera. Estimation of the probability density function of the interaural level differences for binaural speech separation. In *129th Audio Engineering Society Convention*. Audio Engineering Society, 2010.

D. Ayllón, R. Gil-Pita, P. Jarabo-Amores, M. Rosa-Zurera, and C. Llerena-Aguilar. Energy-weighted mean shift algorithm for speech source separation. In *IEEE Statistical Signal Processing Workshop (SSP)*, pages 785–788, 2011a.

D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera. Optimum microphone array for monaural and binaural in-the-canal hearing aids. In *4th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 177–180, 2011b.

D. Ayllón, A. Levi, and H. Silverman. Real-time phase-isolation algorithm for speech separation. 2011.

D. Ayllón, R. Gil-Pita, P. Jarabo-Amores, and M. Rosa-Zurera. Speech source separation using a generalized mean shift algorithm. *Signal Processing*, 92(9):2248–2252, 2012a.

D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera. Speech separation with microphone arrays using the mean shift algorithm. In *133th Audio Engineering Society Convention*. Audio Engineering Society, 2012b.

D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera. Design of microphone arrays for hearing aids optimized to unknown subjects. *Signal Processing*, 93(11):3239–3259, 2013a.

D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera. An algorithm that improves the intelligibility of speech in noise for monaural hearing aids. *Speech Communication*, (Submitted), 2013b.

D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera. Rate-constrained source separation for speech enhancement in wireless-communicated binaural hearing aids. *Advances in Signal Processing*, (Submitted), 2013c.

D. Ayllón, R. Gil-Pita, M. Rosa-Zurera, and H. Krim. An information theoretic approach for speech source enumeration. In *IEEE International Conference in Acoustics, Speech and Signal Processing (ICASSP)*, pages 4300–4304, 2013d.

J. Benesty, J. Chen, Y.A. Huang, and J. Dmochowski. On microphone-array beamforming from a MIMO acoustic signal processing perspective. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1053–1065, 2007.

J. V. Berghe and J. Wouters. An adaptive noise canceller for hearing aids using two nearby microphones. *The Journal of the Acoustical Society of America*, 103:3621, 1998.

M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 208–211, 1979.

P. Bofill and M. Zibulevsky. Underdetermined blind source separation using sparse representations. *Signal processing*, 81(11):2353–2362, 2001.

S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2):113–120, 1979.

B.R. Breed and J. Strauss. A short proof of the equivalence of LCMV and GSC beamforming. *IEEE Signal Processing Letters*, 9(6):168–169, 2002.

A.S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

C.P. Brown and R.O. Duda. An efficient HRTF model for 3-D sound. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 1997.

G.J. Brown and M. Cooke. Computational auditory scene analysis. *Computer speech and language*, 8(4):297–336, 1994.

D.S. Brungart, P.S. Chang, B.D. Simpson, and D.L. Wang. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America*, 120:4007, 2006.

K. Buckley. Broad-band beamforming and the generalized sidelobe canceller. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(5):1322–1323, 1986.

D.R. Campbell and P.W. Shields. Speech enhancement using sub-band adaptive Griffiths-Jim signal processing. *Speech communication*, 39(1):97–110, 2003.

J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408 – 1418, 1969.

O. Cappé. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, 2(2):345–349, 1994.

T. Chabuk, J. Reggia, J. Lohn, and D. Linden. Causally-guided evolutionary optimization and its application to antenna array design. *Integrated Computer-Aided Engineering*, 19(2): 111–124, 2012.

J. Chen, J. Benesty, Y. Huang, and S. Doclo. New insights into the noise reduction Wiener filter. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1218–1234, 2006.

W. Cheng, S. Lee, Z. Zhang, and Z. He. Independent component analysis based source number estimation and its comparison for mechanical systems. *Journal of Sound and Vibration*, 2012.

Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:790–799, 1995. ISSN 0162-8828.

E.C. Cherry. 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, 1953.

K. Chung. Challenges and recent developments in hearing aids part I. Speech understanding in noise, microphone technologies and noise reduction algorithms. *Trends in Amplification*, 8(3): 83–124, 2004.

D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:603–619, 2002.

D. Comaniciu, V. Ramesh, and P. Meer. The variable bandwidth mean shift and data-driven scale selection. In *Eighth IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 438–445, 2001.

P. Comon. Independent component analysis, A new concept. *Signal Processing*, 36(3):287–314, 1994.

H. Cox, R. Zeskind, and T. Kooij. Practical supergain. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(3):393–398, 1986.

A. De Cheveigné. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *The Journal of the Acoustical Society of America*, 93:3271, 1993.

A. De Cheveigné and H. Kawahara. Multiple period estimation and pitch perception model. *Speech Communication*, 27(3):175–185, 1999.

M. Dendrinos, S. Bakamidis, and G. Carayannis. Speech enhancement from noise: A regenerative approach. *Speech Communication*, 10(1):45 – 57, 1991.

H. Do and H. F. Silverman. A robust sound-source separation algorithm for an adverse environment that combines MVDR-PHAT with the CASA framework. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 273–276, 2011.

S. Doclo, T. Van den Bogaert, J. Wouters, and M. Moonen. Comparison of reduced-bandwidth MWF-based noise reduction algorithms for binaural hearing aids. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 223–226, 2007.

R.O. Duda, C. Avendano, and V.R. Algazi. An adaptable ellipsoidal head model for the interaural time difference. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, March 1999.

H. Duifhuis, L. F. Willems, and R.J. Sluyter. Measurement of pitch in speech: An implementation of goldstein's theory of pitch perception. *The Journal of the Acoustical Society of America*, 71:1568, 1982.

D. J. Edelblute, J. M. Fisk, and G. L. Kinnison. Criteria for optimum-signal-detection theory for arrays. *The Journal of the Acoustical Society of America*, 41(1):199–205, 1967.

B. Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26, 1979.

Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(6):1109–1121, 1984.

Y. Ephraim and H.L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 3(4):251–266, 1995.

M. Er and A. Cantoni. Derivative constraints for broad-band element space antenna array processors. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 31(6):1378–1393, 1983.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7 (2):179–188, 1936.

W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall. The DARPA speech recognition research database: specifications and status. In *Proceedings of the DARPA Workshop on speech recognition*, pages 93–99, 1986.

J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko. Computer-steered microphone arrays for sound transduction in large rooms. *The Journal of the Acoustical Society of America*, 78: 1508, 1985.

O.L. Frost. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8):926–935, 1972.

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

S. Gannot, D. Burshtein, and E. Weinstein. Signal enhancement using beamforming and non-stationarity with applications to speech. *IEEE Transactions on Signal Processing*, 49(8): 1614–1626, 2001.

R. Gil-Pita, L. Cuadra, E. Alexandre, D. Ayllón, L. Álvarez, and M. Rosa-Zurera. Enhancing the energy efficiency of wireless-communicated binaural hearing aids for speech separation driven by soft-computing algorithms. *Applied Soft Computing*, 12(7):1939–1949, 2012.

E.N. Gilbert and S.P. Morgan. Optimum design of directive antenna arrays subject to random variations. *Bell Systems and Technology Journal*, 34:637–663, 1955.

K. D. Gilbert and K. L. Payton. Source enumeration of speech mixtures using pitch harmonics. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 89–92, 2009.

B.R. Glasberg and B.C.J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1):103–138, 1990.

Z. Goh, K.C. Tan, and B. T. G. Tan. Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model. *IEEE Transactions on Speech and Audio Processing*, 7(5): 510–524, 1999.

J. D. Gordy, M. Bouchard, and T. Aboulnasr. Beamformer performance limits in monaural and binaural hearing aid applications. In *Canadian Conference on Electrical and Computer Engineering (CCECE).*, pages 381–386, 2008.

J. E. Greenberg. Modified LMS algorithms for speech processing with an adaptive noise canceller. *IEEE Transactions on Speech and Audio Processing*, 6(4):338–351, 1998.

J. E. Greenberg and P. M. Zurek. Preventing reverberation-induced target cancellation in adaptive-array hearing aids. *The Journal of the Acoustical Society of America*, 95:2990, 1994.

L. Griffiths and C.W. Jim. An alternative approach to linearly constrained adaptive beamforming. *EEE Transactions on Antennas and Propagation*, 30(1):27–34, 1982.

E. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski. New insights into the MVDR beamformer in room acoustics. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1):158–170, 2010.

E.A.P. Habets, J. Benesty, S. Gannot, P.A. Naylor, and I. Cohen. On the application of the LCMV beamformer to speech enhancement. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 141–144, 2009.

V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass. Signal processing in high-end hearing aids: state of the art, challenges, and future trends. *EURASIP Journal on Applied Signal Processing*, 2005:2915–2929, 2005.

W.W. Hansen and J.R. Woodyard. A new principle in directional antenna design. *Proceedings of the Institute of Radio Engineers*, 26(3):333–345, 1938.

F. J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

R. L. Haupt and S. E. Haupt. *Practical genetic algorithms*. Wiley-Interscience, 2004.

D. B Hawkins and W. S Yacullo. Signal-to-noise ratio advantage of binaural hearing aids and directional microphones under different levels of reverberation. *Journal of Speech and Hearing Disorders*, 49(3):278, 1984.

M. L. Hawley, R. Y. Litovsky, and J. F Culling. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, 115:833, 2004.

H. G. Hirsch and D. Pearce. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Automatic Speech Recognition: Challenges for the new Millenium. ISCA Tutorial and Research Workshop (ITRW)*, 2000.

M.W. Hoffman, T.D. Trine, K.M. Buckley, and D.J. Van Tasell. Robust adaptive microphone array processing for hearing aids: realistic speech enhancement. *The Journal of the Acoustical Society of America*, 96:759, 1994.

G. Hu and D. Wang. Auditory segmentation based on onset and offset analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):396–405, 2007.

G. Hu and D. Wang. Segregation of unvoiced speech from nonspeech interference. *The Journal of the Acoustical Society of America*, 124:1306, 2008.

G. Hu and D.L. Wang. Speech segregation based on pitch tracking and amplitude modulation. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics, 2001*, pages 79–82, 2001.

G. Hu and D.L. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks*, 15(5):1135–1150, 2004.

Y. Hu and P.C. Loizou. A generalized subspace approach for enhancing speech corrupted by colored noise. *IEEE Transactions on Speech and Audio Processing*, 11(4):334–341, 2003.

Y. Hu and P.C. Loizou. Subjective comparison and evaluation of speech enhancement algorithms. *Speech communication*, 49(7):588–601, 2007.

Y. Hu and P.C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:229–238, 2008.

A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492, 1997.

Recommendation P ITU-T. 862-perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)*, 2001.

G. J. Jang and T. W. Lee. A probabilistic approach to single channel blind signal separation. In *Advances in neural information processing systems*, pages 1173–1180, 2002.

R.E. Kalman and R. Emil. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

S. Kamath and P. Loizou. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *EEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 4, pages 4164–4164, 2002.

Y. Kaneda and J. Ohga. Adaptive microphone-array system for noise reduction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(6):1391–1400, 1986.

J. M. Kates and M. R. Weiss. A comparison of hearing-aid array-processing techniques. *The Journal of the Acoustical Society of America*, 99:3138, 1996.

J.M. Kates. *Digital Hearing Aids*. Plural Pub, 2008.

H. Katmeoka, T. Nishimoto, and S. Sagayama. Separation of harmonic structures based on tied gaussian mixture model and information criterion for concurrent sounds. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, volume 4, pages 297–300, 2004.

S. Kay. Model-based probability density function estimation. *IEEE Signal Processing Letters*, 5(12):318–320, 1998.

G. Kim and P.C. Loizou. Improving speech intelligibility in noise using a binary mask that is based on magnitude spectrum constraints. *IEEE Signal Processing Letters*, 17(12):1010–1013, 2010.

G. Kim, Y. Lu, Y. Hu, and P.C. Loizou. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 126(3): 1486, 2009.

A. P. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing*, 11(6):804–816, 2003.

A. P. Klapuri. A perceptually motivated multiple-f0 estimation method. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*, pages 291–294, 2005.

Z. Koldovsky, P. Tichavsky, and E. Oja. Efficient variant of algorithm fastica for independent component analysis attaining the cramér-rao lower bound. *IEEE Transactions on Neural Networks*, 17(5):1265–1277, 2006.

D. Kolossa and R. Orglmeister. Nonlinear postprocessing for blind speech separation. *Independent Component Analysis and Blind Signal Separation*, pages 832–839, 2004.

H. Krim and J. H. Cozzens. A data-based enumeration technique for fully correlated signals. *IEEE Transactions on Signal Processing*, 42(7):1662–1668, 1994.

H. Krim and M. Viberg. Two decades of array signal processing research: the parametric approach. *IEEE Signal Processing Magazine*, 13(4):67–94, 1996.

K. D. Kryter. Validation of the articulation index. *The Journal of the Acoustical Society of America*, 34:1698, 1962.

R. Kumaresan and A. Rao. Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications. *The Journal of the Acoustical Society of America*, 105:1912, 1999.

N.M. Kwok, D.K. Liu, and G. Dissanayake. Evolutionary computing based mobile robot localization. *Engineering Applications of Artificial Intelligence*, 19(8):857–868, 2006.

A. Levi and H. F. Silverman. An alternate approach to adaptive beamforming using SRP-PHAT. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 2726–2729, 2010.

N. Li and P.C. Loizou. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *The Journal of the Acoustical Society of America*, 123(3):1673, 2008.

Y. Li and D.L. Wang. On the optimality of ideal binary time-frequency masks. *Speech Communication*, 51(3):230–239, 2009.

J.S. Lim and A.V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, 1979.

C. Liu and S. Sideman. Simulation of fixed microphone arrays for directional hearing aids. *The Journal of the Acoustical Society of America*, 100:848, 1996.

P.C. Loizou. Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum. *IEEE Transactions on Speech and Audio Processing*, 13(5):857–869, 2005.

P.C. Loizou and G. Kim. Reasons why current speech enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):47–56, 2011.

T. Lotter and P. Vary. Dual-channel speech enhancement by superdirective beamforming. *EURASIP Journal on Applied Signal Processing*, 2006:175–175, 2006.

D. Luengo, I. Santamaria, L. Vielva, and C. Pantaleón. Underdetermined blind separation of sparse sources with instantaneous and convolutive mixtures. In *IEEE 13th Workshop on Neural Networks for Signal Processing (NNSP)*, pages 279–288, 2003.

J. Ma, Y. Hu, and P.C. Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America*, 125(5):3387, 2009.

N. Ma, M. Bouchard, and R.A. Goubran. Speech enhancement using a masking threshold constrained kalman filter and its heuristic implementations. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):19–32, 2006.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Transactions onSpeech and Audio Processing*, 13(5):845–856, 2005.

I. A. McCowan and H. Bourlard. Microphone array post-filter based on noise field coherence. *IEEE Transactions on Speech and Audio Processing*, 11(6):709–716, 2003.

T. Melia and S. Rickard. Underdetermined blind source separation in echoic environments using DESPRIT. *EURASIP Journal on advances in Signal Processing*, 2007, 2006.

NIA, 2005. URL http://www.nia.nih.gov. National Institute of Aging.

A. Nuttall. Some windows with very good sidelobe behavior. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(1):84–91, 1981.

P. D. O'Grady and B. A. Pearlmutter. The LOST algorithm: finding lines and separating speech mixtures. *EURASIP Journal on Advances in Signal Processing*, 2008(1):784296, 2008.

L. Parra and C. Spence. Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, 8(3):320–327, 2000.

T. W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *The Journal of the Acoustical Society of America*, 60:911, 1976.

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

P. M. Peterson and P. M. Zurek. Multimicrophone adaptive beamforming for interference reduction in hearing aids. *Journal of Rehabilitation Research and Development*, 24(4), 1987.

R. Plomp. A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *Journal of Speech, Language and Hearing Research*, 29(2):146, 1986.

L. Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(1):24–33, 1977.

J. Ramírez, P. Yélamos, J.M. Górriz, J.C. Segura, and L. Garcıa. Speech/non-speech discrimination combining advanced feature extraction and SVM learning. In *International Conference on Spoken Language Processing*, pages 1662–1665, 2006.

L. Rayleigh. XII. On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232, 1907.

I. Recommendation. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. *ITU-T Recommendation*, page 862, 2001.

S. Rickard. *"Blind Speech Separation"*, chapter 8: "The DUET Blind Source Separation Algorithm", pages 217–241. SpringerLink, 2007.

S. Rickard and O. Yilmaz. On the approximate W-disjoint orthogonality of speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 529–532, May 2002.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ) -a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 749–752, 2001.

RNID, 2008. URL http://www.rnid.org.uk. Royal National Institute for Deaf People.

T. Rohdenburg, V. Hohmann, and B. Kollmeier. Robustness analysis of binaural hearing aid beamformer algorithms by means of objective perceptual quality measures. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 315–318, 2007.

N. Roman, D. Wang, and G. J. Brown. Speech segregation based on sound localization. In *International Joint Conference on Neural Networks (IJCNN)*, volume 4, pages 2861–2866, 2001.

N. Roman, S. Srinivasan, and D. Wang. Binaural segregation in multisource reverberant environments. *The Journal of the Acoustical Society of America*, 120:4040, 2006.

M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837, 1956.

E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E Urbanek, and M. Weinstock. Ieee recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246, 1969.

S. T. Roweis. One microphone source separation. *Advances in neural information processing systems*, pages 793–799, 2001.

O. Roy and M. Vetterli. Rate-constrained beamforming for collaborating hearing aids. In *IEEE International Symposium on Information Theory*, pages 2809–2813, 2006.

N. Ruiz-Reyes, P. Vera-Candeas, S. García-Galán, and J. E. Muñoz. Two-stage cascaded classification approach based on genetic fuzzy learning for speech/music discrimination. *Engineering Applications of Artificial Intelligence*, 23(2):151–159, 2010.

J. C. Rutledge. A computational auditory scene analysis-enhanced beamforming approach for sound source separation. *EURASIP Journal on Advances in Signal Processing*, (2009), 2009.

M. Sambur. Adaptive noise canceling for speech signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(5):419–423, 1978.

G. H. Saunders and J. M. Kates. Speech intelligibility enhancement using hearing-aid array processing. *The Journal of the Acoustical Society of America*, 102:1827, 1997.

M. R. Schroeder. Period histogram and product spectrum: New methods for fundamental-frequency measurement. *The Journal of the Acoustical Society of America*, 43(4):829–834, 1968.

J. C. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. Wiley, 2005.

A. Spriet, M. Moonen, and J. Wouters. Robustness analysis of multichannel Wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications. *IEEE Transactions on Speech and Audio Processing*, 13(4):487–503, 2005.

S. Srinivasan and A. C. Den Brinker. Rate-constrained beamforming in binaural hearing aids. *EURASIP Journal on Advances in Signal Processing*, 2009:8, 2009.

R. W. Stadler and W. M. Rabinowitz. On the potential of fixed arrays for hearing aids. *The Journal of the Acoustical Society of America*, 94:1332, 1993.

H. L. Van Trees. *Detection, Estimation and Modulation Theory, PartI*, pages 198–206. Wiley, 1968.

S. Valaee and P. Kabal. An information theoretic approach to source enumeration in array signal processing. *IEEE Transactions on Signal Processing*, 52(5):1171–1178, 2004.

E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3): 528–537, 2010.

N. Virag. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on Speech and Audio Processing*, 7(2):126 –137, mar 1999.

T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007.

D. L. Wang. Time-frequency masking for speech separation and its potential for hearing aid design. *Trends in amplification*, 12(4):332–353, 2008.

D.L. Wang. On ideal binary mask as the computational goal of auditory scene analysis. *Speech separation by humans and machines*, pages 181–197, 2005.

D.L. Wang and G.J. Brown. *Computational auditory scene analysis: Principles, algorithms, and applications.* IEEE Press, 2006.

M. Wax and I. Ziskind. Detection of the number of coherent signals by the MDL principle. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(8):1190–1196, 1989.

D. P. Welker, J. E. Greenberg, J. G. Desloge, and P. M. Zurek. Microphone-array hearing aids with binaural output. II. A two-microphone adaptive system. *IEEE Transactions on Speech and Audio Processing*, 5(6):543–551, 1997.

N. Wiener. *The interpolation, extrapolation and smoothing of stationary time series.* Wiley, 1949.

T. Wittkop, S. Albani, V. Hohmann, J. Peissig, W. S. Woods, and B. Kollmeier. Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction. *Acta Acustica united with Acustica*, 83(4):684–699, 1997.

J. Woodruff and B. Pardo. Using pitch, amplitude modulation, and spatial cues for separation of harmonic instruments from stereo music recordings. *EURASIP Journal on Advances in Signal Processing*, 2007, 2006.

J. Ye. Least squares linear discriminant analysis. In *24th International Conference on Machine learning*, pages 1087–1093, 2007.

O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.

R. Zelinski. Noise reduction based on microphone array with LMS adaptive post-filtering. *Electronics letters*, 26(24):2036–2037, 1990.

E. Zwicker and T. Beckenbauer. Hearing aid with locating microphones, 1988. US Patent 4,773,095.