UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas

Técnicas de aprendizaje maquinal para separación ciega de fuentes sonoras con aplicación al reconocimiento automático del habla

Leandro Ezequiel Di Persia

Tesis remitida al Comité Académico del Doctorado
como parte de los requisitos para la obtención
del grado de
DOCTOR EN INGENIERIA
Mención Inteligencia Computacional, Señales y Sistemas
de la
UNIVERSIDAD NACIONAL DEL LITORAL

2009

Comisión de Posgrado, Facultad de Ingeniería y Ciencias Hídricas, Ciudad Universitaria, Paraje "El Pozo", S3000, Santa Fe, Argentina.

Doctorado en Ingeniería
Mención Inteligencia Computacional, Señales y Sistemas

Título de la obra:

Técnicas de aprendizaje maquinal para separación ciega de fuentes sonoras con aplicación al reconocimiento automático del habla

Autor: Leandro Ezequiel Di Persia
Director: Dr. Diego Milone
Codirector: Dr. Masuzo Yanagida

Lugar: Santa Fe, Argentina

Palabras Claves:

Separación ciega de fuentes sonoras
Análisis de componentes independientes
Reverberación, ruido del ambiente
Reconocimiento robusto del habla
Evaluación objetiva de calidad

To Maria Jose, for being my other half.
To my mother Silvia, my brother Gerardo and my father Danilo.
To my grandparents, for being a light in the darkness.

Acknowledgements

I want to express my deepest thanks to my thesis director, Dr. Diego Milone, for his complete support during my research. This thesis was enriched by his points of view, and the many hours of discussion with him have led to important improvements of the techniques presented here. I am very glad to have had the chance of working with him. I feel also deeply grateful to my thesis co-director, Dr. Masuzo Yanagida for his major contribution to the definition of the study subject and for his technical and material support for the experiments. His friendship and care were fundamental during my two stays in Japan.

I would also like to express my gratitude to all my lab collegues, in particular Leonardo, Leandro, Cesar, Marcelo and Maximiliano. Their friendship and support during my work were very important for my personal and professional development. My thanks also to Dr. Hugo Leonardo Rufiner and Dr. María Eugenia Torres, for giving me the opportunity to getting into the research world in an important project, and for being always willing to help with their enlightening advices and thoughts.

Finally, I thank my family for their support during all my work, in good and also bad moments, when I was far away from home. In particular, I thank Maria José, who was always by my side, helping me in all times, encouraging me and just being there for me, even when I sacrificed a lot of my time due to her in order to accomplish my work.

Special thanks to:

- **sinc**(*i*): Laboratory for Signals and Computational Intelligence, Faculty of Engineering and Water Sciences, National University of Litoral.
- Cybernetics Laboratory, Faculty of Engineering, National University of Entre Ríos.
- Intelligent Mechanisms Laboratory, Faculty of Engineering, Doshisha University.
- Informatics Department, Faculty of Engineering and Water Sciences, National University of Litoral.
- CONICET: National Scientific and Technical Research Council.

Leandro Ezequiel Di Persia

Department of Informatics
Santa Fe, March 2009.

Machine learning techniques for audio blind source separation applied to automatic speech recognition

Leandro Ezequiel Di Persia

Thesis Director: Diego H. Milone
Thesis Co-Director: Masuzo Yanagida
Department of Informatics, 2009

Abstract

In the last decades a new problem related to machine learning and signal processing has emerged in many disciplines: the blind source separation problem. One of the more succesfull examples of solutions to this problem is the Independent Component Analysis method. The blind source separation technique aims to segregate the sources that contribute to some variation of a physical quantity like sound, vibration or electric field, given a set of measurements of the global variation produced by all sources at a time. In particular, when applied to the separation of audio sources, this is known as the "cocktail party" problem, due to the ability of human beings to "concentrate in" or "separate" the conversation of interest from the rest of conversations and background noise.

One particular application of the blind source separation methods is the Automatic Speech Recognition, which can be defined as the task of determining the text that corresponds to a given spoken utterance. This kind of systems have being developed for many years, and although they have reached a maturity point in which some of them were implemented as stand alone applications they still suffer from a strong drawback: they cannot adequatelly manage the existence of noise or competing sources in the input. This aspect is particularly relevant in many applications where the microphones are located far from the speaker, like in teleconference systems.

This doctoral dissertation presents several advances in the technique of audio blind source separation in reverberat condition, from the approach of independent component analysis in the time-frequency domain. After identification of the main drawback of the standard approaches, three methods were developed in order to produce a better quality of separation and, at the same time, to reduce the processing times. The first method is based in the standard approach with improvements in the initialization and the permutation solution, the second is a multirresolution approach, and the third uses a simplified mixing model to produce an indeterminacy-free and very fast separation. The proposed algorithms were evaluated under realistic conditions such as different environments and different kind and power of competing sources. For this purpose we used two evaluation alternatives, objective quality measures of the resulting signal and the performance in the application of interest, that is, automatic speech recognition. The results for the different approaches show the possibility of getting through the dilemma between resulting quality and requiered processing time, converging to a very fast and high quality separation method that can separate sources in a variety of environments and yields a better quality and recognition rate than standard methods.

Técnicas de aprendizaje maquinal para separación ciega de fuentes sonoras con aplicación al reconocimiento automático del habla

Leandro Ezequiel Di Persia

Director de Tesis: Diego H. Milone
Co-Director de Tesis: Masuzo Yanagida
Departamento de Informática, 2009

Resumen

En las últimas décadas el problema de separación ciega de fuentes ha emergido en varias disciplinas relacionadas con el procesamiento digital de señales y el aprendizaje maquinal. En la resolución de este problema el objetivo es obtener por separado las fuentes que generaron en conjunto determinada variación de alguna cantidad física (tal como vibración, sonido o campo eléctrico), dado un conjunto de mediciones del efecto conjunto de todas las fuentes. En particular, en el contexto de fuentes sonoras, este problema se denomina "cocktail party", haciendo referencia a la situación de un ambiente con muchas personas hablando a la vez, y en el cual el oído humano tiene la capacidad de aislar la conversación de interés, del resto del ruido.

Entre las aplicaciones de interés para tal técnica está el reconocimiento automático del habla, en el cual se busca obtener una transcripción escrita a partir del habla emitida por una persona. Estos sistemas han estado en desarrollo por décadas, y aunque han alcanzado un grado de madurez tal que algunos han sido lanzados como aplicaciones finales, todavía sufren de una gran desventaja: no pueden manejar adecuadamente la existencia de ruido u otras fuentes sonoras que compiten en la entrada. Este problema es particularmente importante en aplicaciones como sistemas de teleconferencia, en los cuales los micrófonos están localizados a distancia del hablante.

En esta tesis se proponen tres técnicas basadas en el análisis de componentes independientes en el dominio frecuencial, para producir una efectiva separación de las fuentes sonoras presentes en un cuarto con reverberación. La primera de ellas es una variante del método estandar con mejoras en la inicialización y la resolución de permutaciones, la segunda es un enfoque multirresolución, y la tercera utiliza un modelo de mezcla simplificado para producir un algoritmo libre de indeterminaciones y muy rápido. Para la evaluación del desempeño de las mismas se realizó un estudio exhaustivo de medidas objetivas de calidad, y se desarrolló un protocolo experimental que permite una adecuada evaluación comparativa del desempeño de los algoritmos. Además, dada la aplicación de interés se realizó la evaluación de los mismos mediante la tasa de reconocimiento de un sistema de reconocimiento automático del habla, y como es de interés lograr aplicaciones en tiempo real, se evaluó también el tiempo de cálculo utilizado por los diversos métodos. Los resultados se contrastaron con los de métodos del estado del arte para esta tarea. Se verificó que todos los métodos propuestos produjeron importantes mejoras tanto en la calidad objetiva como en la tasa de reconocimiento. Por otro lado, dos de los métodos desarrollados logran significativas mejoras de desempeño con tiempos de cálculo un orden de magnitud menor a los métodos del estado del arte.

Técnicas de aprendizaje maquinal para separación ciega de fuentes sonoras con aplicación al reconocimiento automático del habla

Leandro Ezequiel Di Persia

Director de Tesis: Diego H. Milone
Co-Director de Tesis: Masuzo Yanagida
Departamento de Informática, 2009

Resumen extendido

En la presente tesis se estudian diferentes métodos para obtener la separación ciega de fuentes sonoras. En este resumen extendido se describirán en forma compacta el problema en cuestión, los objetivos de la investigación y las diferentes alternativas propuestas para su resolución, así como el protocolo experimental desarrollado para verificar las características de los métodos desarrollados. Se presentarán también los resultados obtenidos y las conclusiones generales del trabajo de tesis.

Separación Ciega de Fuentes mediante Análisis de Componentes Independientes

La separación ciega de fuentes (BSS, Blind Source Separation) es una técnica que apunta a obtener, a partir de múltiples registros de las señales mezcladas, las señales fuente por separado. Esta técnica explota características estadísticas y espaciales de las señales transmitidas para intentar aislar las fuentes sonoras originales, minimizando la información requerida sobre el tipo de fuentes y los canales de transmisión utilizados. Esto tiene aplicación directa en reconocimiento automático del habla (ASR, Automatic Speech Recognition) en situaciones realistas, en las que puede haber, por ejemplo, una fuente (hablante) de interés y simultáneamente diversas fuentes de ruido (aire acondicionado, televisor, otros hablantes) cercanas que interfieren con la señal deseada, y que deben ser eliminadas antes de utilizar la señal como entrada a un sistema de ASR.

En la transmisión de sonido en ambientes cerrados, las señales no sólo siguen el camino directo de la fuente hasta el sensor sino que también rebotan en todas las

superficies sólidas, llegando por lo tanto al sensor múltiples copias de la misma señal, con diferentes retardos y factores de atenuación. Este proceso se conoce como reverberación, y se puede modelar como la salida de un sistema lineal e invariante en el tiempo (LTI, Linear and Time Invariant). En este caso se aplica un modelo de mezcla convolutiva que se puede expresar en forma matricial como:

$$\mathbf{x}(t) = H(t) * \mathbf{s}(t) \tag{1}$$

donde la componente $h_{ij}(t)$ es la respuesta al impulso del ambiente, medida desde la posición de la fuente i hasta la posición del sensor j; $x_j(t)$ y $s_i(t)$ son la mezcla j-ésima y la fuente i-ésima respectivamente, y $*$ representa la operación de convolución.

La resolución de este problema es compleja ya que no se conocen ni las fuentes $s_i(t)$ ni las características de transmisión definidas por $h_{ij}(t)$. El problema se torna más complicado aún debido a las convoluciones involucradas. Una alternativa para la solución es aplicar una trasformada de Fourier de tiempo corto (STFT, Short Time Fourier Transform), con lo cual se obtiene:

$$\mathbf{x}(\omega, \tau) = H(\omega)\,\mathbf{s}(\omega, \tau) \tag{2}$$

Este modelo resulta ser una mezcla instantánea para cada frecuencia ω, es decir que las componentes de cada una de las fuentes sólo están escaladas por una constante antes de ser sumadas (a diferencia del caso convolutivo en el que estaban filtradas). Simplificando la notación, para un ω dado esto se puede expresar como $\mathbf{x}(\tau) = H\mathbf{s}(\tau)$. Este caso de mezclas sencillas se conoce en la bibliografía como modelo de mezcla instantánea y está prácticamente resuelto por diversos algoritmos de análisis de componentes independientes (ICA, Independent Component Analysis). Estos algoritmos utilizan como hipótesis la independencia estadística entre las fuentes, y por lo tanto, optimizan una función de costo para encontrar la matriz de separación que invierta el sistema anterior, maximizando la independencia estadística del resultado. Entonces, el método de separación denominado ICA en el dominio frecuencial (fd-ICA, frequency domain ICA) consiste en aplicar una STFT a las señales mezcladas, y luego en cada frecuencia resolver la separación mediante un algoritmo de ICA.

Si bien mediante la técnica de ICA en el dominio frecuencial (fd-ICA) el problema pareciera estar resuelto, existe una serie de problemas que impiden su aplicación práctica. Algunos de ellos son:

- Inicialización de los algoritmos: en general las funciones no lineales usadas en la resolución de los problemas de ICA presentan muchos extremos locales. La obtención de una "buena" solución depende en gran medida de una inicialización apropiada que lleve al extremo global.

- Indeterminaciones: todos los algoritmos de ICA son capaces de obtener una aproximación de las fuentes con un factor de escalado y una permutación arbitraria. Esto hace que en el método fd-ICA, cada frecuencia presente una permutación y escalado diferente de las otras frecuencias, lo que conduce a representaciones tiempo-frecuencia no consistentes. De esta forma la representación que se supone pertenece a una sola señal, en realidad tiene en algunas frecuencias informacion de una señal y en otras de otra señal diferente.

- Cantidad de datos necesarios: se requiere una gran cantidad de datos para estimar adecuadamente los estadísticos involucrados en la resolución del problema de ICA en cada frecuencia.

- Reverberación: al realizar la STFT se utilizan ventanas de una longitud que esta limitada por el tiempo en el cual la señal de voz puede considerarse estacionaria. Pero al aumentar la reverberación de un cuarto, la ventana a utilizar debería estar en relación a la longitud de los filtros implicados, y en general la señal de voz no puede considerarse estacionaria para períodos tan grandes. Esto tambien afecta a la cantidad de datos disponibles para estimar los métodos de ICA.

Para algunos de estos problemas existen soluciones parciales. Por ejemplo para el problema de las indeterminaciones se puede utilizar la correlación entre bandas de frecuencia sucesivas, reordenar las fuentes y así resolver las permutaciones con cierto grado de éxito, aunque persisten algunas residuales que degradan significativamente la calidad de la señal obtenida. Por otro lado, al tener que resolver tantos problemas de ICA como frecuencias utilizadas en la transformación, los tiempos de cómputo son considerables, lo que impide la aplicación práctica de los métodos. En base a un estudio extensivo del estado del arte en separación ciega de mezclas sonoras convolutivas, se plantearon para esta investigación doctoral los siguientes objetivos:

1. Diseñar un protocolo experimental unificado y consistente para la evaluación de algoritmos de BSS.

2. Estudiar medidas objetivas para evaluar la calidad de la separación y determinar las mejores para ser usadas dentro del protocolo de experimentación ya mencionado.

3. Desarrollar métodos de BSS basados en fd-ICA que permitan resolver las limitaciones de los métodos del estado del arte, permitiendo aumentar el rango de aplicabilidad.

4. Desarrollar algoritmos mas rápidos que los del estado del arte, lo que permitirá acercarse a la necesidad práctica de operación en tiempo real.

5. Producir una mejora significativa de la calidad de las señales separadas, evaluada mediante las medidas objetivas de calidad anteriormente mencionadas.

6. Obtener una mejora significativa en la tasa de reconocimiento de sistemas de reconocimiento automático del habla mediante la aplicación de los métodos de separación desarrollados.

A continuación se introducirán brevemente los tres métodos desarrollados en esta tesis, buscando hacer énfasis en la solución de los mencionados problemas.

Separación ciega de fuentes con mejora en la solución de permutaciones, inicialización y postfiltro

Durante experimentos previos con la metodología estándar de fd-ICA utilizando dos métodos de ICA instantáneo (JADE y FastICA), se comprobó que el método JADE era muy veloz, pero brindaba resultados de inferior calidad, mientras que el método FastICA podía generar resultados de mejor calidad, pero resultaba demasiado sensible a una apropiada inicialización. Se propuso entonces el uso del método JADE para obtener una matriz de separación inicial, que luego siguiera siendo refinada por FastICA para generar una salida de mejor calidad.

Por otro lado, se propuso una mejora en el método de solución de las permutaciones de cada frecuencia basado en correlaciones. En el método propuesto las envolventes acumuladas que se usan para determinar la permutación adecuada de cada frecuencia en base a las anteriormente clasificadas, se obtienen mediante una actualización autoregresiva, en lugar de un simple promediado. Esto permite resolver el problema de permutación con una mucho mayor exactitud y casi sin dejar permutaciones residuales.

Finalmente, debido al efecto de la reverberación se comprobó que la separación no era completa, quedando información residual de la fuente no deseada, así como de los ecos de la propia fuente. Para mejorar este aspecto, se introdujo el uso de filtros de Wiener tiempo-frecuencia, que permiten mejorar ambos aspectos, mejorando la separación al reducir los ecos de la fuente no deseada, y también los auto ecos de la fuente deseada. Este método se denominó JFW (Jade-FastICA-Wiener postfilter).

Separación ciega multiresolución

El segundo método propuesto en esta tesis utiliza un enfoque iterativo, basado en la selección óptima del tamaño de ventanas en la STFT. Para diferentes señales a separar, el tamaño óptimo de ventana a utilizar modifica la calidad de separación. Tamaños de ventana pequeños no capturan adecuadamente el efecto de reverberación, mientras que tamaños muy grandes generan muy pocos datos en cada frecuencia para permitir una adecuada separación mediante ICA.

En el enfoque propuesto se utiliza un conjunto de tamaños de ventana y se prueba la separación en ellos. Se mide el grado de separación obtenido mediante correlación y se selecciona el tamaño de ventana óptimo. A continuación, las señales separadas se usan para construir postfiltros de Wiener que mejoran la separación. La ventana utilizada es eliminada y se repite el proceso sobre las restantes longitudes de ventana, en una serie de iteraciones. En la siguiente iteración, al usar una ventana diferente con la salida anterior como entrada, se refina la separación sobre datos ligeramente distintos, lo que provee una separación adicional.

Para poder implementar este método se requiere que el algoritmo de separación sea rápido, de manera de poder evaluar la misma sobre todos los largos de ventana definidos, mitigando el incremento del costo computacional. Esto se consigue mediante una simplificación en el modelo de mezcla: se supone que la matriz de mezcla (y por lo tanto la de separación) es la misma para todas las frecuencias. De esta manera, la estimación de la misma se realiza mediante un método de ICA que debe ejecutarse una sola vez, a diferencia del método estándar que usa una estimación por ICA para cada frecuencia. Esto genera un método rápido para la estimación y la separación. Este método se denominó MRBSS (Multi-Resolution BSS).

Separación ciega basada en el modelo pseudoanecóico

El tercer método de separación desarrollado utiliza una nueva formulación para el modelo de mezclas. Si los micrófonos están situados suficientemente cercanos entre sí, la respuesta al impulso desde una fuente a todos los micrófonos puede escribirse como versiones escaladas y retardadas de la respuesta de uno de ellos. Es decir, bajo la condición de micrófonos cercanos, los canales de transmisión de una fuente a todos los micrófonos quedan caracterizados por una sola respuesta al impulso con escalados y retardos adecuados para representar las diferentes longitudes de los caminos de transmisión.

Este modelo, que llamamos pseudoanecóico (dado que resulta similar en estructura a un modelo anecóico pero teniendo en cuenta todos los ecos) simplifica el proceso de separación, ya que la matriz de mezcla queda expresada en función de constantes y su único parámetro variable esta en ángulos de fase, que varían linealmente con la frecuencia. De esta forma, estimando una sola matriz de mezcla para una frecuencia dada, se pueden determinar los parámetros de la mezcla y sintetizar las matrices de separación para todas las demás frecuencias. Al sintetizarse las matrices de separación con los mismos parámetros, se elimina la existencia de indeterminaciones de escalado y permutación.

Para obtener una estimación robusta de la matriz de mezcla mediante ICA se selecciona una frecuencia central y se usan los datos de un rango de frecuencias simétrico respecto a esa frecuencia central. Esto aumenta la cantidad de datos disponibles para

la convergencia del algoritmo de ICA, y a la vez brinda robustez frente a la posibilidad de que los datos en la frecuencia central no sean facilmente separables. También se utiliza un postfiltro de Wiener para mejorar aún más la separación. Este método se denominó PMBSS (Pseudoanechoic Model BSS).

Evaluación objetiva de calidad

Durante el desarrollo de los algoritmos de separación existe la necesidad de contar con medidas objetivas de calidad que permitan determinar el desempeño de los diferentes algoritmos (o de un mismo algoritmo con diferentes parámetros). Como en esta investigación la aplicación de interés es el reconocimiento del habla, resultó necesario contar con alguna medida objetiva que estuviera bien correlacionada con la tasa de reconocimiento de tales sistemas. Dado que no se encontraron en la bibliografía resultados de tal correlación, se decidió realizar un estudio de diversas medidas de calidad de la separación y su correlación con el desempeño de un sistema de reconocimiento.

Este estudio abarcó 13 medidas típicamente utilizadas en otras áreas del procesamiento de la voz, además de variantes y medidas compuestas derivadas de ellas. Del estudio producido se concluyó que la medida con mayor correlación es la PESQ (Perceptual Evaluation of Speech Quality) desarrollada en el estándar ITU P.862 para evaluación de voz sobre canales de transmisión, por lo que todos los métodos desarrollados se evaluaron con esta medida. La PESQ además ha demostrado estar altamente correlacionada con tests subjetivos en diversos estudios de métodos de codificación y realce de la señal de voz, con lo cual un valor alto de la misma además de ser indicador de una buena tasa de reconocimiento, indicaría también una buena calidad perceptual al escuchar la señal.

Como resultado de este estudio se elaboró un protocolo experimental adecuado para la evaluación de los algoritmos desarrollados. Este protocolo incluye tanto la forma adecuada de recolección de datos (determinación de condiciones de grabación, mezclas reales y artificiales, disposición de las fuentes originales), como la evaluación de cada algoritmo en forma tanto individual (para seleccionar el mejor conjunto de parámetros), como comparativa entre distintos algoritmos, incluyendo tanto las medidas objetivas de calidad, como la evaluación definitiva mediante la aplicación de interés.

Resultados

Como se mencionó anteriormente, los algoritmos desarrollados se compararon mediante la medida de calidad objetiva PESQ. Además, dado que la aplicación de interés es el ASR, se evaluó la tasa de reconocimiento utilizando la salida de los diferentes métodos. Como parte de este análisis también se evaluaron los tiempos de cómputo necesarios para cada método. Estos son relevantes porque la mayoría de las aplicacio-

Tabla 1. Calidad (medida como PESQ) para todos los método explorados y para las mezclas.

SNR	Ruido	Mezclas	Murata	Parra	JFW	MRBSS	PMBSS
0 dB	Habla	2.11	1.97	2.22	2.89	2.43	2.83
	Blanco	1.98	1.86	2.37	2.84	2.56	2.83
6 dB	Habla	1.73	1.71	2.19	2.54	2.37	2.59
	Blanco	1.64	1.67	2.16	2.47	2.28	2.54
Promedio		1.86	1.80	2.23	2.69	2.41	2.70

Tabla 2. Tasa de reconocimiento de palabras (%) para todos los métodos explorados y para las mezclas.

SNR	Ruido	Mezclas	Murata	Parra	JFW	MRBSS	PMBSS
0 dB	Habla	44.50	25.00	49.50	84.00	70.50	85.50
	Blanco	19.54	15.00	27.50	84.50	73.00	85.50
6 dB	Habla	30.00	27.00	49.00	79.50	70.50	83.00
	Blanco	7.20	11.00	20.00	69.00	36.00	67.50
Promedio		25.31	19.50	36.50	79.25	62.50	80.38

nes de interés requieren funcionamiento en tiempo real o casi real. Para el estudio se seleccionaron previamente, de acuerdo al protocolo experimental, los parámetros óptimos para cada algoritmo. Además de esto, se evaluó el desempeño con los datos sin procesar (es decir, las mezclas tal cual recibidas en los micrófonos) para determinar la mejora producida por los distintos métodos. También se utilizaron, sobre los mismos datos, dos métodos del estado del arte (llamados Parra y Murata por sus respectivos autores), para poder contrastar qué tanto se está mejorando respecto al estado del arte en separación ciega de fuentes. Las Tablas 1, 2 y 3 presentan respectivamente los resultados de PESQ, la tasa de reconocimiento porcentual, y el tiempo promedio de procesamiento. Se usaron dos tipos de ruido (habla y ruido blanco) y con dos niveles de relación señal-ruido (SNR, Signal to Noise Ratio). Las mezclas se realizaron en un cuarto con tiempo de reverberación de 195 ms.

Como puede observarse, los tres métodos propuestos proveen una mejora sustancial tanto en calidad medida por PESQ, como en la tasa de reconocimiento, respecto de las señales sin procesar. Más aún, todos ellos también superan ampliamente en ambos índices a los métodos del estado del arte. Respecto al tiempo de procesamiento, el algoritmo JFW produce su aumento de calidad a costa de un tiempo de procesamiento mayor. El método MRBSS produce una mejora de calidad un poco menor a JFW (aunque bastante mayor que la de los métodos del estado del arte), pero logra reducir

Tabla 3. Tiempo de procesamiento promedio en segundos para los métodos explorados y tiempo promedio de reconocimiento para las mezclas.

SNR	Ruido	Mezclas	Murata	Parra	JFW	MRBSS	PMBSS
0 dB	Habla	0.34	9.49	6.48	11.41	0.70	0.43
	Blanco	0.33	8.79	7.01	10.15	0.74	0.27
6 dB	Habla	0.37	9.56	6.60	11.63	0.73	0.42
	Blanco	0.66	8.98	6.48	17.55	0.69	0.28
Promedio		0.42	9.21	6.64	12.68	0.72	0.35

mucho el tiempo de cómputo, siendo sólo del 5.7 % del usado por JFW. Finalmente el método PMBSS presenta una calidad mayor aún a la de JFW, reduciendo aún más el tiempo de cómputo (solo 2.8 % del usado por JFW). Debe notarse que el tiempo de proceso del método PMBSS es comparable al del usado por el reconocedor para realizar el reconocimiento, para señales con una duración promedio de 3.55 segundos.

Conclusiones

De los resultados presentados se desprende que se han logrado cumplir todos los objetivos propuestos, a la vez que se han desarrollado métodos muy novedosos que abren nuevas perspectivas para el estudio futuro en el área. A partir de esta investigación, se produjeron cuatro publicaciones en revistas internacionales con referato, dos solicitudes de patente en Japón, dos presentaciones en congresos internacionales y nueve presentaciones en congresos de alcance nacional y regional. Los resultados de reconocimiento automático son muy superiores a los de métodos del estado del arte, y en particular dicho aumento de calidad se ha logrado incluso reduciendo significativamente tiempo de procesamiento.

Contents

List of Tables

List of Figures

Preface

The Blind Source Separation (BSS) problem aims to obtaining an estimate of several sources that generate some changes in a given physical magnitude, using only information of several measurements of that physical magnitude. Since (almost) no information is used regarding the kind and characteristics of the sources and the mixture conditions, this process is termed "blind" [Hyvärinen et al., 2001]. In the particular case of audio sources, the magnitude of interest is the pressure variation, measured by means of microphones from a sound field generated by several sources arbitrarily distributed in a specific environment. The sources can be several speakers, some loudspeaker playing music, some machine producing noise, etc. In this way, the objective of the BSS methods is to be able to obtain each source that produces the whole sound field, by just some set of measurements of their joint effect in the room.

One of the most successfull approaches used to achieve this goal is Independent Component Analysis (ICA). In this approach, together with a model of the kind of mixture, an assumption of statistical independence is imposed on the sources. Using this hypothesis, the separation can be achieved by optimization of a properly designed cost function that measures the independence of the estimated sources. The technique of ICA is a machine learning technique rooted in neural networks research and the pioneering works of Jutten and Herault [Jutten, 1987; Jutten and Herault, 1991].

Human-machine interaction is a field of special interest for BSS. As speech is one of the most important means of communication among human beings, it is interesting to consider the possibility of developing methods for man-machine interfaces based on speech. This involves having a system capable of "translating" the sound pressure variation generated by speech in an equivalent transcription of the spoken words. This is the Automatic Speech Recognition (ASR) task.

In the last years, significant advances have been done in the development of such ASR systems, and they have reached a stage in which they can be reliably used in laboratory conditions, with close to the mouth microphones [Juang and Rabiner, 2005; O'Shaughnessy, 2008]. However, such systems present a series of problems when used in practical situations in real environments. It has been verified that the recognition rates of these systems fall very fast in the presence of noise [Lippmann, 1997]. This is particularly significant when the speaker and the microphone are located far from each other, in such a way that the speech of interest arrives after being modified by the room itself. Moreover, it arrives together with other competing signals whose powers are in the same order of magnitude. This is what motivates the use of robust ASR systems, in which the "robust" term makes reference to the capability of such a system

to work without considerable variation of its performance when used in situations that vary with respect to those for which the system was designed.

One method to provide robustness in the presence of noise is the speech enhancement approach. This method consists in a preprocessing step focused in producing a speech signal as similar to the clean ones used during training as possible. In other words, the idea is to clean up the signal before feeding it to the ASR system. This approach is exploited in the present thesis, where BSS techniques based on ICA are used to obtain the desired source separated from the competing noises. In particular the approach of frequency-domain ICA will be exploited, as detailed in the following.

The frequency-domain ICA approach has several drawbacks that prevent it from being of practical application. This thesis will be devoted to explore alternatives to overcome these limitations. The main limitation of standard methods is that they can handle only a very limited amount of reverberation, which is the main effect of a room for far located microphones in real rooms. This degrades the result of the separation, and as a consequence, the recognition rate of ASR systems. Another limitation for practical use of these algorithms is their computational cost, measured as a long processing time. Therefore this aspect will be important for this study since the lower computational cost solution will always be favored. Also, an important problem for the research was the lack of a consistent experimental framework that allows a valid comparisson of different algorithms, with reproducibility and in conditions as realistic as possible. We will also focus on the development of such framework.
Thus, the main goals of this dissertation are the following:

- To provide a standard and consistent framework for the experimentation and evaluation of BSS algorithms for audio source separation.
- To study and determine the best objective quality measures to be used in that framework for the evaluation purposes.
- To develop algorithms for the audio BSS problem in the frequency domain that overcome the limitations of the state-of-the-art methods, allowing a wider range of applications. In particular the algorithms should provide robustness to reverberation.
- To develop algorithms that are faster than the state-of-the-art methods, in order to make the applications more feasible.
- To produce a significant improvement in the quality on the separated speech signals, as evaluated by objective quality measures.
- To obtain a significant improvement in recognition rates of automatic speech recognizers under competing noise and reverberation.

Concerning to the structure of the present dissertation, an introduction to Acoustics is presented in Chapter 1. Since this work deals with speech signals modified by the effect of rooms, this introduction will be fundamental for understanding the involved physical phenomena. Then, a brief introduction to the problem of blind source separation of audio sources will be made, followed by a brief review of the automatic speech recognition methods used.

In Chapter 2, the different standard approaches used for audio BSS with special emphasis in the ICA-based ones will be presented. The standard technique of frequency-domain ICA (fd-ICA) will be described, including a detailed analysis of its weak points and drawbacks.

Chapter 3 deals with the important subject of performance evaluation for the BSS algorithms. When this research was started there was not a widely accepted method to evaluate the performance of algorithms of BSS for ASR tasks. After experimentation with a variety of objective quality measures, a framework for the evaluation was developed. This methodology is presented in detail in the chapter.

Chapter 4 presents the first separation algorithm developed in this thesis. It is a variant of the standard fd-ICA approach, in which some modifications were introduced to overcome the initialization problem and the permutation ambiguity of the standard methodology, and to enhance the rejection of echoes. The resulting algorithm shows an improvement in separation quality, at the cost of increased complexity.

In Chapter 5 the second algorithm developed in this thesis is presented. This algorithm introduces a simplification in the mixture model that allows to avoid the main drawback of standard fd-ICA approaches, the permutation ambiguity. Using a single separation matrix estimated over all the data, it is able to avoid the permutation problem, and yet produce good results in a very short time. That is, the quality of the algorithm is better than the state-of-the-art methods, and the computation times are significantly shorter.

Chapter 6 introduces the third algorithm developed in the thesis. In this case, a new mixture model based on closely spaced microphones leads to a simpler but very robust algorithm. This method also avoids the permutation problem, and is capable of producing a high quality separation, and at the same time, performing separation at a very high speed.

Finally, Chapter 7 summarizes the advances introduced in the dissertation, describes future research lines to improve even more the results, and present the general conclusions.

Chapter 1

Introduction

This doctoral dissertation explores solutions to the problem of blind sound sources separation. This problem takes place when there are several sound sources, scattered in a room, that produce a composite sound field to which each of them contribute. Usually, the microphones employed to capture this sound field are located far from the sources. This turns out the problem more complicated because now the sound travels not only from each source to the microphones, but also through its reflections of multiple orders that arrive at delayed times. It must be clear that the acoustic properties of the room will modify the sound field, and thus, it will be fundamental for the study of the separation problem to have some knowledge of these acoustic properties and how they are generated.

This chapter will start with a brief review of room acoustics, including the acoustical properties of interest and how can they be determined or measured. Then a brief introduction and general description of the blind source separation problem will be presented. The independent component analysis method for instantaneous mixtures will be briefly introduced in Section 1.3. This method will be extensively used in all this work as part of the algorithms for convolutive source separation. Then, since the application of interest is the automatic speech recognition task, the main concepts and algorithms used for it will be briefly revised, to end with a short introduction to speech quality measurement.

1.1. Principles of acoustics

In this section a brief summary of definitions, equations, properties and measurement techniques in acoustics will be presented. For more details in the subject, please refer to [Beranek, 2004, 1996; Everest, 2001; Kuttruff, 2000].

1.1.1. Physical characteristics of sound

Sound can be defined as a variation of pressure in some media. From a particle point of view, the sound is generated by oscillations of the particles of the media with respect to their stable position. In this way, sound cannot exist in vacuum, because by definition it is necessary to have some particle that can vibrate to produce sound. Therefore, given some medium like air or water or even a solid body, when some specific stimulus is applied, these particles vibrate. This oscillation is transmitted from one particle to the next one, due to the atomic forces involved, and this produces the travel of the vibration through the media [Beranek, 1996].

Normally, in absence of stimulus, all particles vibrate at their own rates with random directions, which translates macroscopically as no vibration at all, because the random movements cancellate each other. Nevertheless, given some specific stimulus, the particles can vibrate in a coordinate way. Although this produces no absolute movement (because after some time each particle will return to its equilibrium position), it will produce areas where there is in some moment more density of particles, and other areas where there is less density of them. This, translated to a global macromolecular view, is observed as areas of variations of pressure. In some moments and places, there is compression and augmented pressure, while in other areas and moments there is rarefaction and reduced pressure [Everest, 2001].

This pressure distribution in time and space is what our ears transduce as sound. The outer ear conduces the pressure to the middle ear, which in turn translates it from air pressure variation to mechanical vibration. Finally the inner ear is a very specialized organ that transduces these mechanical variations into electrical variations and neural impulses.

In this context, acoustics is the study of these pressure variations, how they are produced, transmitted and modified by the media, and how they can be measured and interpreted. This knowledge allows to characterize the sound sources, the transmitting media and the environment in which they were produced.

1.1.2. Sound wave generation

Even as sound is generated by individual particle movement, since the distances of the particles themselves are very small with respect to the macroscopic sizes, the media can be regarded as a continuum, and thus one may apply the continuum mechanics principles to analyze the sound propagation. Applying the fluid mechanics theory, the variation of pressure $p(\mathbf{v}, t)$ with respect to its steady value p_0, which is a function of the space coordinates $\mathbf{v}$ and time t, can be found to follow the wave equation [Kuttruff, 2000]:

$$\Delta p = \frac{1}{c^2} \frac{\partial^2 p}{\partial t} \,, \tag{1.1}$$

where Δ is the Laplacian operator[1] and c is a constant. An equivalent equation can be found for the velocity of the particles. It must be noted that this equation was obtained under the assumption of free propagation in an homogeneous, isotropic and non viscous gas at rest [Beranek, 1996].

Solutions of the wave equation: Planar waves

Let us assume that the transmission of sound is possible only in one direction. This is not necessary the same as saying that we are working in a one dimensional media, but that the sound cannot propagate in directions perpendicular to the one of interest. In this way, the wave equation can be written as $\frac{\partial^2 p}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}$. The general solution of this equation has the form [Kuttruff, 2000]:

$$p(x,t) = F(ct - x) + G(ct + x)\,, \tag{1.2}$$

where F and G are general functions only limited to have first and second derivatives. No matter what the shape of the functions F and G is, there are two aspects that must be noted. First, the quantity c has units of speed, and in fact it represents the propagation speed of the sound. Second, the two terms represent travelling waves, one propagating in the positive x direction, and the other in the negative one. To see this, suppose that at $x = 0$ and time t_1 the first term is $F(ct_1)$. After some time, $t = t_1 + t_2$, we have in general $F(c(t_1 + t_2) - x)$. Thus, if $x = ct_2$ the value of the function will be $F(ct_1)$, which means that the wave has travelled in some time t_2 a distance equal to ct_2 [Beranek, 1996]. The same reasoning can be presented for function G. This emphasize the interpretation of c as the speed of the travelling wave.

As known from Fourier analysis, any wave can be represented as a superposition of sinusoidal waves. Thus it is interesting to analyze the behaviour of these waves for each sinusoidal component. Let $p(x,t) = p(\omega)\cos(k(ct - x)) = p(\omega)\cos(\omega t - kx)$, which is a harmonic, progressive wave, for constants ω and k. For a fixed point in space, this is a sinusoidal with angular frequency given by $\omega = kc$ measured in rad/s, and frequency $f = \frac{\omega}{2\pi}$ measured in Hz. This wave have a period of $T = \frac{1}{f} = \frac{2\pi}{\omega}$. In a similar way, for a fixed time instant, it represents a sinusoidal function of space, with a spatial period $\lambda = \frac{2\pi}{k}$, called wavelength, measured in m. Combining these quantities one obtains a very usefull relation, $f\lambda = \frac{kc}{2\pi}\frac{2\pi}{k} = c$ which relates the frequency and the wavelength, with the speed of propagation [Kuttruff, 2000]. For propagation in air, the speed of sound varies mainly with the temperature, but it can be assumed to be around 343 m/s.

It must be noted that this planar wave propagates in x direction, with surfaces of constant pressures (called wavefronts) that are parallel to the yz plane, as shown in Figure 1.1.

[1] This operator is also denoted by ∇^2.

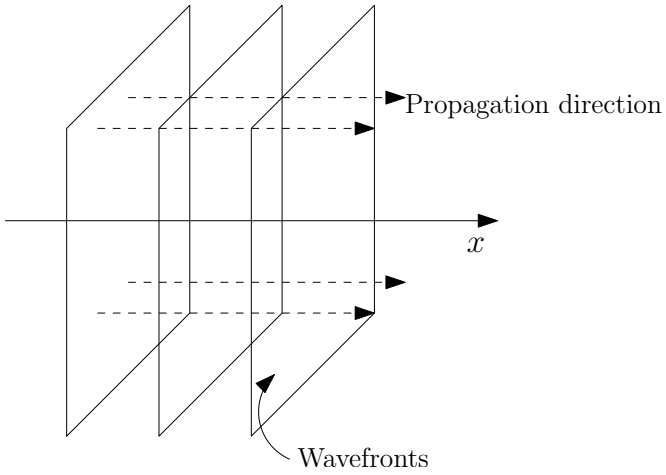


Figure 1.1. Planar waves.

Solutions of the wave equation: Spherical waves

Let us now assume that one has a source of excitation that is infinitesimal in size, and that irradiates sound in all directions. This is called a "point source" in the literature [Kuttruff, 2000]. Moreover, assume that this source can irradiate sound in all directions with the same power (this is called an omnidirectional source). In this case the solution of the wave equation gives rise to spherical waves, that is, sound waves that propagates radially from the point source with equal power in all directions. Thus, the wave fronts are spheres of increasing size, that propagates radially with center in the source. This kind of wavefront is presented in Figure 1.2.

It must be noted that in both types of solutions free propagation was assumed, meaning that the traveling wave found no obstacles during its propagation. However, in most real cases (particularly in room acoustics) the sound will experiment reflections, transmissions and diffractions when it interacts with solid surfaces like furniture, walls, etc. Nevertheless, at least in the period before its interaction with solid surfaces, many sound sources can be approximated as point sources with spherical wave propagation.

Propagation of waves in closed environments

As already mentioned, as the sound propagates in a room it will reach obstacles with different densities than that of the transmission media (in the case of our interest, the transmission media is air). The sound speed is given by $c^2 = \frac{\gamma P_0}{\rho_0}$, where γ is the adiabatic exponent (computed as the ratio of specific heats defined as $\gamma = C_p/C_v$,

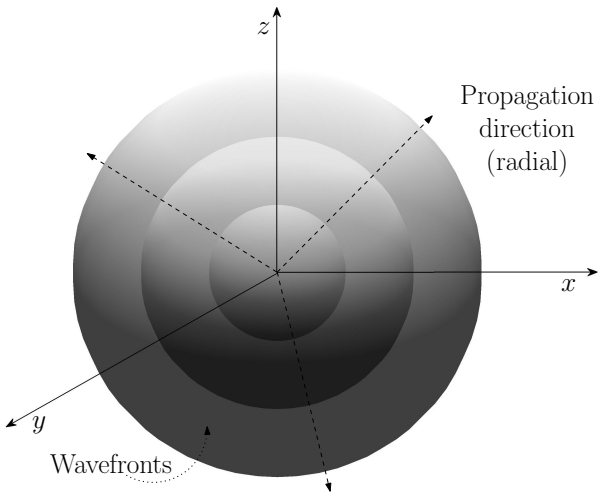


Figure 1.2. Spherical waves.

where C_p is the specific heat at constant pressure, and C_v the specific heat at constant volume), P_0 is the static pressure in the media (that is, the pressure of the media when no sound energy is present), and ρ_0 is the static density of the media. It can be seen that, for different materials, the speed of propagation will be different. At an interface between different materials, the acoustic impedance changes abruptly. This produces some phenomena that are similar to these of light transmission: there is some proportion of the energy that is transmitted through the new media, and there is a fraction of it that is reflected back to the original media [Kuttruff, 2000].

In Figure 1.3 a simulation of 2D propagation of sound in a closed environment is presented. The simulated environment has rectangular shape, with dimensions in m. At some point an impulsive source is simulated. The wave equation was solved using a finite difference time domain method to discretize the pressure and velocities involved [Sakamoto et al., 2002]. In the figure, six different instants from the simulation are shown, and it can be clearly seen how the travelling waves are reflected in the boundaries. As time advances, the number of reflections is increased and the sound field becomes more complex, although its energy decreases due to energy dissipation at the media and in the walls.

Suppose that one tries to meassure this dynamic sound field generated by a point source, in a fixed location in the room, as in the case shown in Figure 1.4. The microphone located there will receive first the direct wave from the source. But after

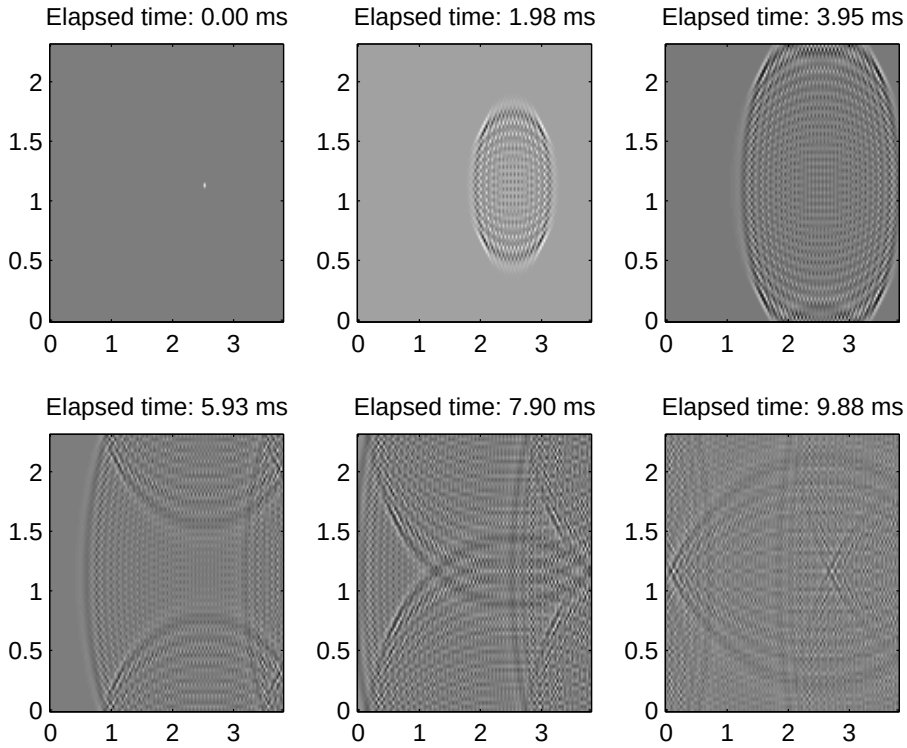


Figure 1.3. Simulation of sound propagation in a closed 2D environment.

some time, some echoes originated in reflections in the walls will start to arrive to the microphone. Thus, in response to an isolated impulse, the measured signal will not be just a delayed impulse, but a series of waveforms generated by each reflection.

This effect is shown in schematic way in Figure 1.5. This figure was generated by the image source method of room simulations [Allen and Berkley, 1979]. The first peak represents the propagation of the direct wave to the microphone. The time between zero and its arrival is the delay related with the distance travelled by the wave. The subsequent impulses are the effect of echoes in the walls.

As the walls and air attenuate differently each frequency component, the arriving components will not be impulses, but rather some complex shapes. Assuming that one can throw away all nonlinear effects (as was done in the presentation of the wave equation), the room can be characterized as a linear system. Moreover, if the source

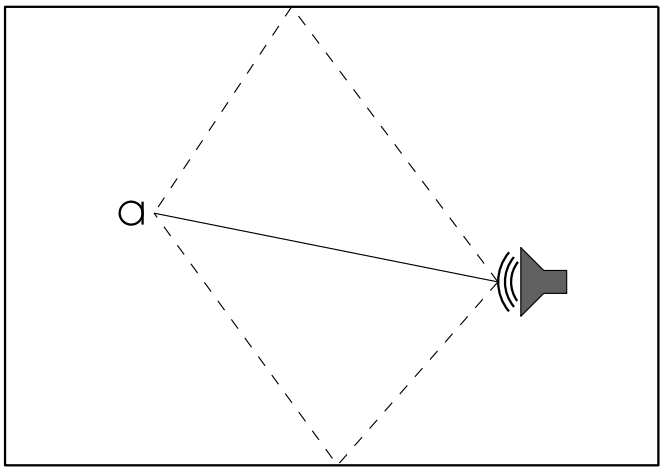


Figure 1.4. Propagation of sound in closed environment: reflections in walls.

and receiver locations are fixed, and the temperature and all conditions are assumed to be static, this system can be also considered time invariant. Thus, the system is linear and time invariant (LTI) and it will be completely characterized by its impulse response (IR) [Kahrs and Brandenburg, 2002]. The knowledge of the impulse response allows the synthesis of the room effect for any input. In equation form, given the impulse response $h(t)$ measured form the source location to the sensor location, the received signal at the sensor, $x(t)$, produced by a sound source $s(t)$ is given by

$$x(t) = s(t) * h(t) \,, \tag{1.3}$$

where $*$ stands for convolution. This method is used in several techniques of artificial reverberation and spacialization of sound.

Of course that, from each possible source location to each possible sensor location, a different impulse response will be produced, because the patterns of echoes will differ from point to point. Thus, a complete characterization of a room would require the knowledge of the impulse responses in all possible combinations of source and receiver positions. This seems discouraging, as it would be impossible to fully characterize a room in this way. There are, however, some properties of the rooms, like the reverberation time, that are quite independent of the location, and can be used to study their global acoustic properties.

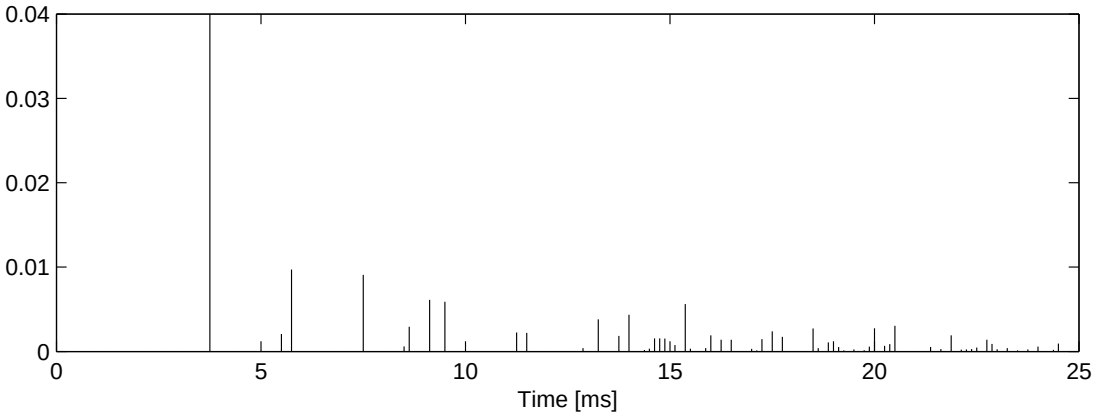


Figure 1.5. Schematic impulse response.

Near and far field

As explained before, two important types of wave propagation are planar and spherical waves. Ideally, a point source will generate spherical waves. When these waves travel far away from the source (thus the radius of the sphere is larger), its curvature will be decreased. At points far enough from the source, this curvature will be very small and can be disregarded. This means that, for sensors located far from the source, the wavefronts are seen as planar. This has important implications, because planar waves are easier to analyze. There is a critical distance that marks the limit of validity of the planar approximation. When the receiver is located farther than this critical distance, it is said to be in the "far field", and for shorter distances, it is in the "near field" [Everest, 2001].

This concept is important for microphone array processing. A linear microphone array is a set of microphones arranged in a straigth line, with equal spacing among them. If this array is located in the "far field" of a sound source, it can be assumed that the wavefront is planar, and the delay in the arrival of the wavefront to two adjacent microphones can be calculated using simple trigonometry. For linear microphone arrays, the critical distance can be calculated with the following formula [Doclo and Moonen, 2003]:

$$r = \frac{d_t^2 f_s}{c}\,, \tag{1.4}$$

where d_t is the total length of the array, f_s is the sampling frequency, and c the sound speed.

1.1.3. Techniques for impulse response measurement

The more direct way of performing this measuremet is by using an impulse-like source and recording the resulting response in some location. Among the used sources, one can cite clapping, exploding balloons, starting pistols firing blanks, etc. Although this seems easy, it is difficult to control both, the characteristics of the source and its location. This lack of control affects the reproducibility of the experiment, as each time the source characteristics will be different. It is clear that some ways to give more control on the source production are needed.

To overcame this problem, some wideband noise sources have been proposed. In 1979, Schroeder proposed a method for IR measurement using maximum length sequences (MLS). These are pseudo-random signals with long periods generated from a shift register using feedback loops in the output of specific registers. For a shift register with n stages, the period of the pseudo-random noise is $L = 2^n - 1$, which can be made longer enough for the measurement needed. These MLS sequences have the property that its one period autocorrelation equals 1 for lag 0 and $1/L$ in other case. If L is large enough, this $1/L$ part will be small and the correlation approaches an impulse. Then, the measurement method consists of exciting the room with the pseudo-random sequence using a loudspeaker, capturing the response with a microphone, and then calculating the autocorrelation with the original sequence. This autocorrelation will produce approximately the IR of the room [Schroeder, 1979].

Another measurement method is based in time stretched pulses (TSP) [Aoshima, 1981]. In this method the excitation sequence is a sinusoidal whose frequency is increased in time to cover the whole frequency range [Berkhout et al., 1980]. This "sine sweep" has a simple inverse, which is the same signal but reversed, so that its frequency is diminished in time. The convolution of both signals is an impulse. Hence, the method consists in exciting the room with the sweep using a loudspeaker, recording the response, and then deconvolving the original sequence with its inverse, producing the desired IR [Suzuki et al., 1995].

A technique used to improve the signal to noise ratio (SNR), consists of reproducing the excitation signal not only one, but several times, and to coherently average them before the deconvolution. As the excitation and the room IR are the same, the only difference among the repetitions should be due to random noise in the measurement, and so averaging them will reduce the noise level. Figure 1.6 presents an example IR measured with the technique of TSP, using three repetitions of the excitation source to improve the SNR.

With respect to the effects of the reverberation produced by the IR, from a point of view of the signal characteristics, it produces a "propagation" (smearing) of the information along the time. It changes the energy of the temporal signal, and produces a change in the frequency magnitudes (phenomena often called coloration).

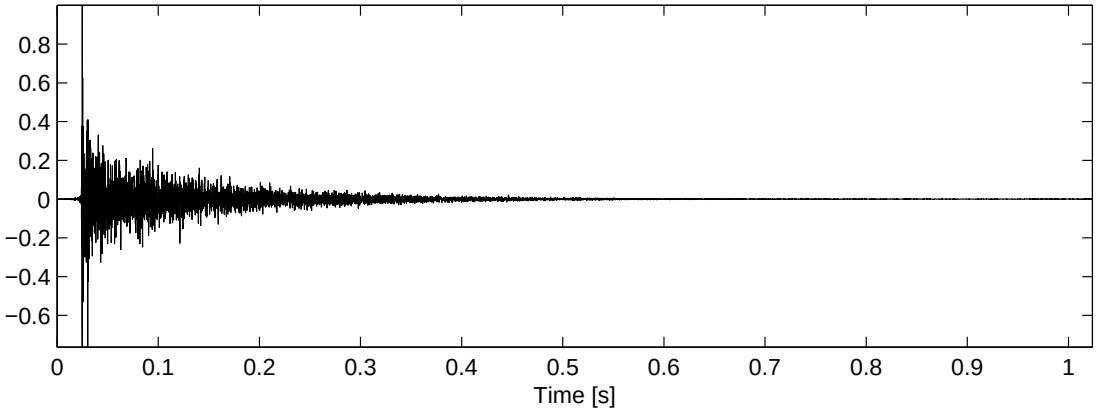


Figure 1.6. Impulse response of a room measured by the TSP method.

In Figure 1.7, the IR of Figure 1.6 was applied by filtering as in (1.3) to a clean speech signal. In the left side, the original signal and its spectrogram are presented. In the right, the reverberant signal and its spectrogram show the main effects, both in time and time-frequency domains. They clearly show how some of the properties of the signal, like pauses and gaps, have dissapeared due to the smearing effect, and how the magnitude has been changed, with some bands having less power than before. It should be clear that some simple processings (like segmentation) would be more difficult in this case, because the short time, transient properties of certain speech sounds have been masked by the smearing process.

From the point of view of psychoacoustics, this reverberation phenomenon produces echoes and spectral distortion affecting the spatial perception of sound [Blauert, 1997; Gilkey and Anderson, 1997; Kahrs and Brandenburg, 2002] and the intelligibility of speech [Crandell and Smaldino, 2000; Finitzo-Hieber and Tillmann, 1978]. These phenomena are worsened when the amount and duration of the reverberation is increased.

1.1.4. Reverberation time

One of the most important measurements to characterize a room is the "reverberation time" τ_{60} which is the time needed by the sound field in a room to decay to one millionth of its initial power ($\Delta P = 60$ dB) [Everest, 2001; Kuttruff, 2000]. Roughly speaking, is the time needed for the sound to become inaudible. Usually, it is assumed that the sound decay follows an exponential law (assuming that there is only one dominant decay mode), then the power in logarithmic scale would represent a straight line.

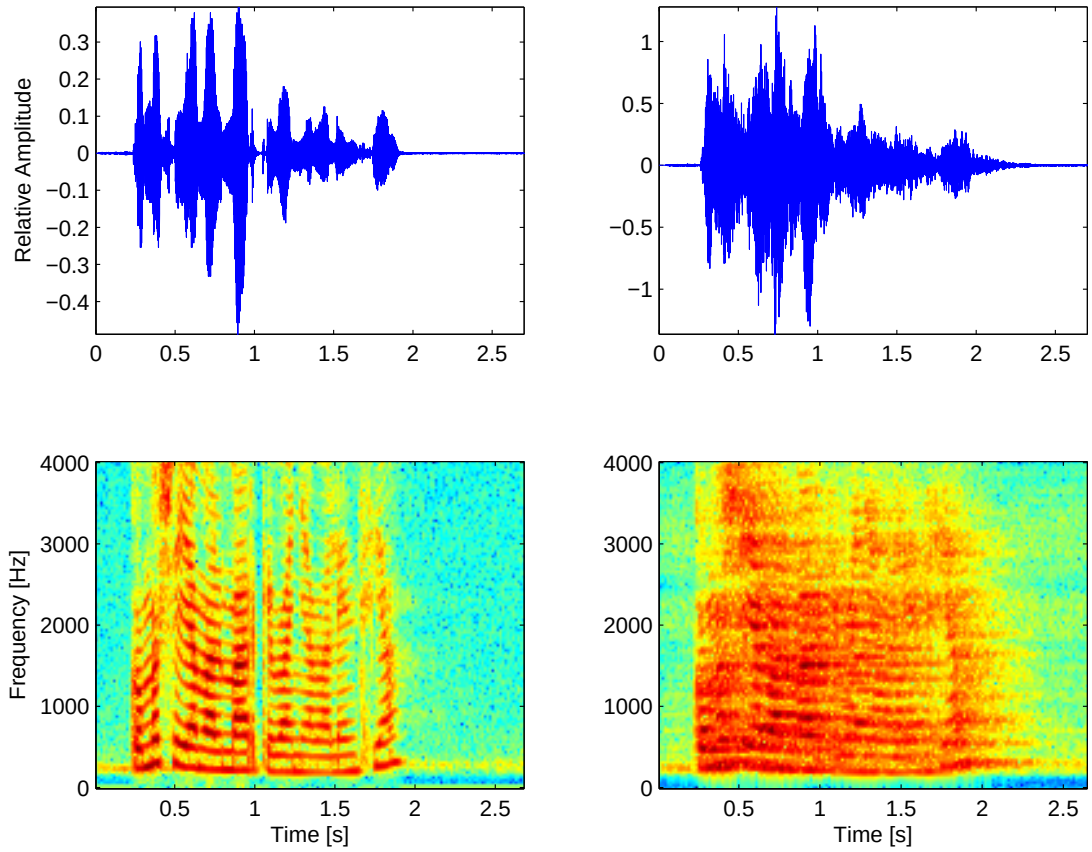


Figure 1.7. Effects of reverberation on speech. Left top, original signal, and bottom, its spectrogram. Right top, reverberated signal, and bottom its spectrogram.

The slope of this line will vary with the decay constant.

The power of the IR has many abrupt variations due to the peaks corresponding to strong echoes, therefore is very difficult to determine the straight line of decay. The most used method for this is the Schroeder backintegration method [Schroeder, 1965]. It evaluates the expected value of the power over an ensemble of measures, using a unique impulse response. If the impulse response is $h(t)$, then Schroeder proved that the expected value of its power can be evaluated as:

$$\mathcal{E}\{h^2(t)\} = \int_t^\infty h^2(t)dt\,, \tag{1.5}$$

where $\mathcal{E}\{\cdot\}$ stands for the expectation operator.

The result of this equation is a smoothed curve, more suitable for the evaluation of

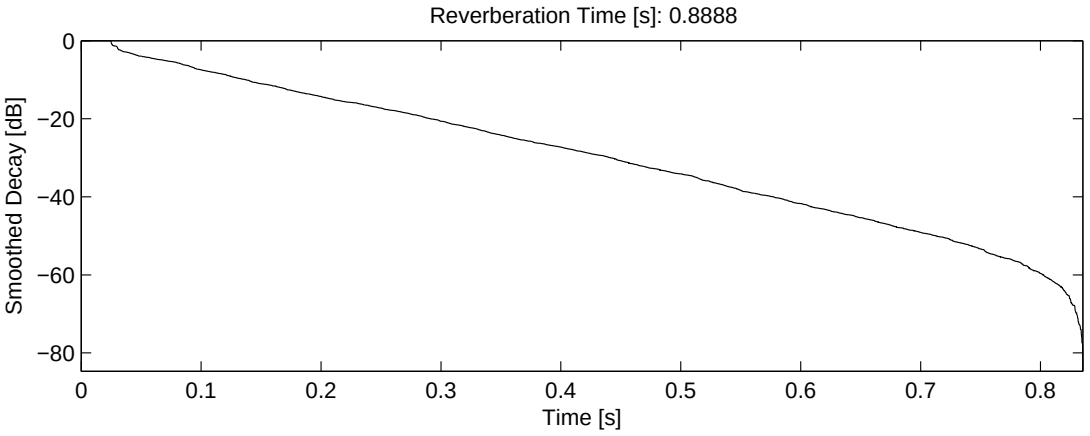


Figure 1.8. Decay curve of an impulse response obtained by Schroeder backintegration method. The estimated τ_{30} value is also shown.

decay and reverberation time. As the measures are subject to measurement noise, it is difficult for the curve to reach the -60 dB level. For this reason, it is usual to extrapolate the value, assuming a linear decay. The first 5 dB drop is disregarded because usually is a jump due to the energy of the direct sound. Then the time is measured for a drop from -5 dB to -25 dB. The reverberation time is found by multiplication of this value by a factor of 3. This produces an estimation called τ_{20} to emphasize that it was obtained for a 20 dB drop and extrapolated to 60 dB. If a value of 30 dB is used for the decay instead, the estimation is obtained after multiplication by 2, and the estimation is called τ_{30}

In Figure 1.8, the Schroeder backintegration method was applied to the IR of Figure 1.6, and the reverberation time was estimated for a 30 dB drop. In this case the reverberation time is quite long, and it can be seen that the decay in logarithmic scale is quite linear.

1.2. The problem of blind source separation of audio sources

Up to now, in all the presentation on acoustics it was assumed that there was only one active sound source in the environment. Although this is useful when theoretical analysis are required, it does not corresponds to reality. In general, there are several sources of sound acting in any environment, including voices of human speakers, noises coming from the street, noises due to machines working nearby, sound from musical instruments, radios, audio equipment, etc. All these sound sources interact in any

environment, producing a sound field which is influenced by all of them.

1.2.1. Convolutive model

Under the assumption of absence of nonlinear effects in the transmission (as was done in the previous development of the wave equation), the room can be regarded as an LTI system, as already said. This means that the effect of each source will contribute additively point-to-point to the general sound field.

The problem, then, is related to the possibility of extraction of the contributions from each source, starting from knowledge of the general sound field only. This scenario is known as the *cocktail party* problem in the literature, due to its similarity to the real situation, in which several people are in a room, and all of them are chatting in groups [Divenyi, 2004]. In that environment, there are many sound sources, and even so, any particular person has the ability of concentrating in a particular conversation, somehow rejecting the other sources and focusing in the one of its interest. This ability has been linked (among other phenomena) to the binaural hearing (that is, to the fact that we have two sensors). There are studies that show that binaural hearing provides for better source localization [Hawley et al., 1999] and lowers the audibility threshold [Beranek, 1996]. Also, the intellegibility and separation of sources was shown to be improved when comparing binaural to monoaural hearing [Hawley et al., 1999; Lavandier and Culling, 2008]. This knowledge suggests the use of multiple microphones to capture the sound field and somehow combine them to improve the signal. Thus, the problem is approached from the perspective of microphone array processing.

Consider the case in which there are M active sound sources, and the sound field generated by them is captured by N microphones, as shown in Figure 1.9. From source j to microphone i, an impulse response h_{ij} characterizes the room. Using the notation s_j for the sources and x_i for the microphone signals, with $i = 1, \ldots, N$ and $j = 1, \ldots, M$, the mixture can be represented at each instant t as [Cichocki and Amari, 2002]:

$$x_i(t) = \sum_j h_{ij}(t) * s_j(t) \, . \tag{1.6}$$

Let us form a vector of sources, $\mathbf{s}(t) = [s_1(t), s_2(t), \cdots, s_M(t)]^T$, and the same for the vector of mixtures $\mathbf{x}(t) = [x_1(t), x_2(t), \cdots, x_N(t)]^T$ measured by the microphones, where $[\cdot]^T$ stands for transposition. Then the previous equation can be written (with a little abuse of notation) as:

$$\mathbf{x}(t) = H * \mathbf{s}(t) \tag{1.7}$$

where the "matrix" H has as each element a filter given by the impulse response from one source location to one microphone location. The equation must be understood as a simple matrix-vector product, but replacing the multiplications by a filtering operation via convolution.

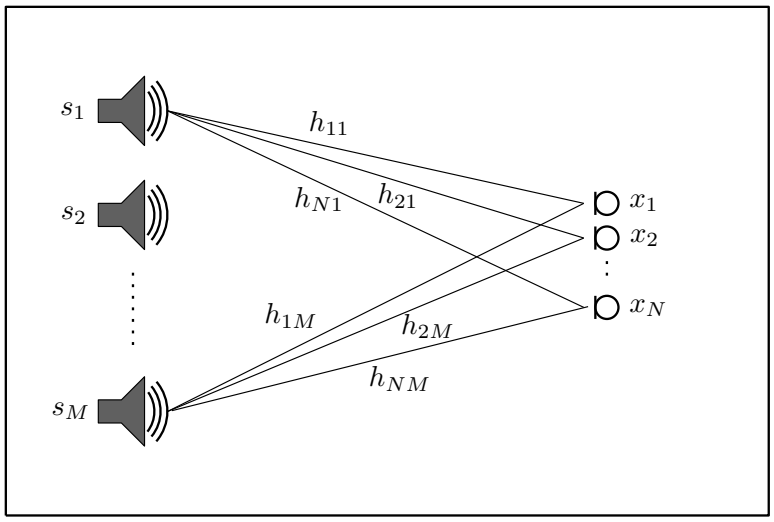


Figure 1.9. A case of cocktail party with M sources and N microphones.

1.2.2. The convolutive separation problem

Equation (1.7) represents the ideal mixture process and shows the problem of blind source separation: to find the vector of sources $\mathbf{s}(t)$, having only knowledge of the vector of mixtures $\mathbf{x}(t)$.

It must be noted that the main difficulty related to this is that, in general, one has no knowledge regarding what the sources are, nor the characteristics of the filters involved. In other words, both H and $\mathbf{s}(t)$ are unknown, and the objective of the algorithm is to obtain them blindly, that is, using the less amount possible of information regarding the mixing process. The length of the filters involved adds complexity to the problem. If the room has a reverberation time of say, 0.5 s, and a sampling frequency $f_s = 8000$ Hz is used, the filters would have 4000 taps each one. That is, for 4 microphones and 4 sources the unknown matrix H is of size $4 \times 4 \times 4000$.

The mixing model proposed assumes the absence of noise. Usually there is some extra noise, for example the microphones include some amount of noise in their measurements, which is related to the thermal movement of the molecules of the materials that compose them. This noise can be modeled as an additive random process in the previous equation. In all this document it will be assumed that, either there is no noise, or that it is so small that its effect is negligible, and thus the noiseless case will be analyzed.

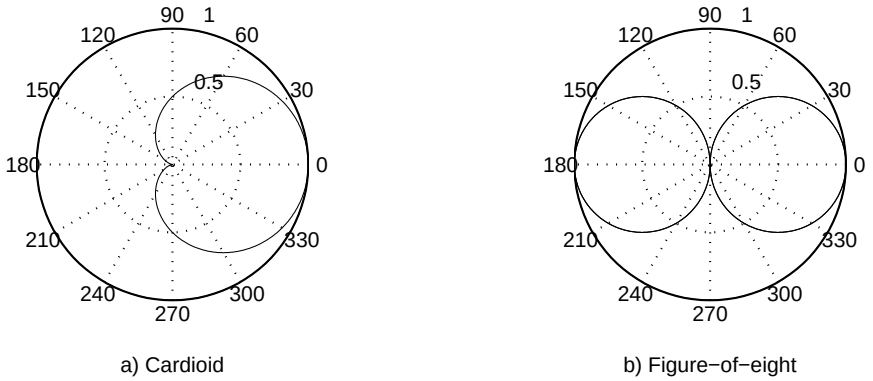


Figure 1.10. Directivity patterns for microphones pointing towards zero degrees.

1.2.3. Characteristics of the sensors

There are many kinds of michophones available for capturing the sound field. In general they can be divided into two classes according to their working principle: dynamic and capacitive[2]. In dynamic microphones, the pressure produces displacement of a diaphragm that is connected to a moving coil, which moves inside a magnetic field. This generates a potential difference which is related to the amount of displacement, which in turn is related to the force acting on the diaphragm, and this force is proportional to the pressure. In a capacitive microphone, the diaphragm also works as one of the plates of a capacitor, and thus when the pressure produces a displacement of it, the space among the plates is varied, which changes the capacity and can be measured as a change in electrical charge or voltage. In general the capacitive microphones are more sensitive than the dynamic ones, and can be built with smaller sizes (by example, 1/2" diameter) [Beranek, 1996]. Given these advantages, capacitive microphones will be used in this work.

The microphones also can be classified according to their directivity pattern. Omnidirectional microphones are equally sensitive to sound coming from all directions. Directional microphones are more sensitive in specified directions (angles of incidence). There are many kind of directional microphones, for example cardioid and figure-of-eight, names that make reference to the shape of the directivity pattern plotted in polar coordinates. Figure 1.10 shows the directivity patterns for cardioid and figure-of-eight microphones pointing in the zero degree direction. For the subject of this research all

[2]There are also piezoelectric microphones but they are not widely available, and thus we will not discuss this type

sources must be captured no matter their direction relative to the microphones. This has a practical reason: the applications of this kind of methods should work no matter where in the room the desired source is located. For this reason omnidirectional microphones were used in this research.

1.3. Independent component analysis

One of the first approaches to the general problem of blind source separation was independent component analysis. This technique has its origin in PCA techniques and neural networks. The first works in the area were produced by french researchers in the 1980s and were related to nonlinear decorrelation [Jutten, 1987; Jutten and Herault, 1991]. Following the precursor works of Oja with neural networks that can perform decorrelation (PCA) [Oja, 1982, 1989], they proposed a new method in which a nonlinear function was applied to the outputs before training. This produced nonlinear decorrelation, and was shown to yield source separation by statistical independence. A lot of new techniques based on different ways of measuring the statistical independence have been proposed since then.

1.3.1. Mixing model

In ICA, a central aspect is the mixture model assumed, which is linear and instantaneous. The linear part means that the sources are additively combined after some arbitrary scaling, to form the mixtures. The instantaneous part means that the sources are mixed as they are, affected only by some scaling, but not by delays, convolutions or other effects. Using the same notation as for the convolutive case, with $\mathbf{s}(t)$ for the vector of sources and $\mathbf{x}(t)$ for the mixtures, the mixing process can be written as

$$\mathbf{x}(t) = A\mathbf{s}(t)\,, \tag{1.8}$$

where A is a scalar matrix, and the operation is a standard matrix-vector multiplication [Hyvärinen et al., 2001].

Comparing this model with the one in (1.7), it is clear that this is a special case of the convolutive model, that occurs if the filters are just scaled impulses at time zero. This is clearly the simplest case to solve.

The problem of ICA is the following: given a number N of signals (measures or mixtures), and assuming that they were generated by the linear instantaneous mixture model, obtain a number M of sources as statistically independent as possible.

Note that the statistical independence is measured among different sources. Another level of independence is usually assumed among the succesive samples of the process. It is assumed that each source is generated by an independent and identically

distributed (iid) random process. This means that there is no temporal correlation among samples. There are some ICA methods that assume iid sources and use this to produce independent sources, and some methods that explote the temporal correlation to produce the desired independent sources.

1.3.2. Solution alternatives

The basic idea underlying the solution alternatives is the incorporation of the important assumption of statistical independence of the sources. This assumption can be very plausible in many situations in which the sources are not linked and each one acts independently of the others. In this case, the solution to the problem consist of finding a matrix W such that the separated sources

$$\tilde{\mathbf{s}}(t) = \mathbf{y}(t) = W\mathbf{x}(t) \tag{1.9}$$

result as statistically independent as possible. In the ideal case, it would result that $W = A^{-1}$ and thus the sources will be recovered perfectly (of course that it was assumed that the mixing matrix was invertible).

Thus the key part of the process is to produce some cost function that measures, in a direct or indirect way, how much statistical independence one gets. Once produced such measure, the process of finding W involves the solution of an optimization problem, maximizing the independence via the cost function. This optimization usually involves an iterative process of refinement of the solution to provide a better W [Hyvärinen et al., 2001].

There are many measures that can quantify the amount of independence. Some of them are mutual information, Kullback-Leibler divergence, log likelihood, negentropy and kurtosis [Lee, 1998]. There are in general two kinds of measures: those that need knowledge of the probability distribution function (pdf) of the sources (like mutual information or divergence), and those that do not use this information (like kurtosis and negentropy approximations based on nonlinear functions). The last ones introduce a measure of independence by elimination of higher-order correlations, which appears in their series expansion, and thus is an indirect method. Nevertheless they are robust and often quite used in the practice [Hyvärinen and Oja, 2000]. Other approaches measure the independence by means of high order statistical measures, like cumulants. This approach is exploited in the joint approximated diagonalization of eigenmatrices (JADE) approach [Cardoso and Souloumiac, 1993], in which the independence is obtained by diagonalizing fourth-order cumulant tensors.

1.3.3. Ambiguities of the method

If one examines carefully the structure of the proposed solution, it is easy to find out that there are some ambiguities in the solution. It is clear that if two of the sources are interchanged, it will result in an interchange of the two corresponding columns of the mixing matrix, and it will produce the same mixtures. This implies that the sources can be recovered up to an arbitrary permutation of them, which is called the permutation ambiguity in the literature.

In second place, it should be clear that if the sources are scaled with a diagonal matrix, then there is a mixing matrix which is scaled by the inverse of the diagonal matrix, which will produce exactly the same mixtures. This means that the sources can only be recovered up to an arbitrary scaling. This is called the scaling ambiguity.

These ambiguities can be expressed in equation form. If the mixing matrix is A and the separation matrix found is W, in the ideal case of perfect separation the application in cascade of both would give

$$WA = PD\,, \tag{1.10}$$

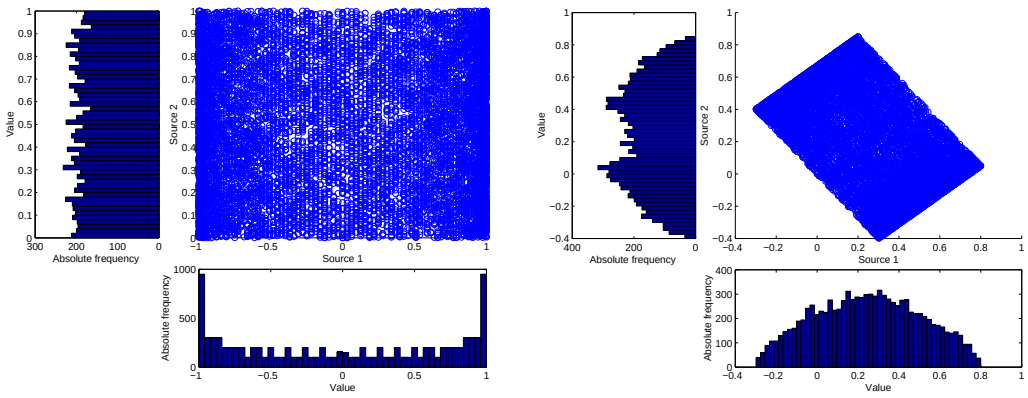
where P is a permutation matrix and D is a diagonal scaling matrix with arbitrary values in the diagonal [Cichocki and Amari, 2002].

Example: FastICA

One of the most widely accepted algorithms for ICA is the FastICA method [Hyvärinen, 1999], that uses the nongaussianity as a measure of independence. Suppose that each source was produced by a random process with some specific pdf. As all mixture of random variables has a pdf that is more similar to a Gaussian than the pdf of any of them (by application of the central limit theorem), maximization of nongaussianity of the obtained signals will provide separation.

To show this principle, Figure 1.11 presents in the left a scatter plot of $\mathbf{s}(t)$, for two sources. Source 1 is a sinusoidal function of time, and source 2 is a random signal with uniform distribution in $[0, 1]$. The figure also shows the marginal histograms for these sources. In the right, the same is presented for $\mathbf{x}(t) = A\mathbf{s}(t)$ for a given mixing matrix. It can be seen that the marginal histograms are more similar to gaussian densities, than the original histograms.

In the following, we will drop the sample index t in the equations for simplicity, but all signals must be interpreted as samples in time. An assumption that simplifies the processing is that the data is centered (that is, has zero mean), and its covariance is the identity. This assumption can be satisfied by centering the data (substracting the mean), and using a sphering or whitening transformation as follows. The data is



a) b)

Figure 1.11. Example of the nongaussianity principle for ICA. Scatter plots and marginal histograms of: a), independent sources, and b) mixtures produced using the mixing model $\mathbf{x}(t) = A\mathbf{s}(t)$.

transformed as $\mathbf{z} = Q\mathbf{x}$, with $Q = D^{-1/2}E^T$, where $C_\mathbf{x} = EDE^T$ is an eigendecomposition of the covariance matrix of the data, $C_\mathbf{x} = \mathcal{E}\left\{\mathbf{x}\mathbf{x}^T\right\}$. It is easy to show that under this transformation, the covariance matrix of $\mathbf{z}$ will be the identity $C_\mathbf{z} = \mathcal{E}\left\{\mathbf{z}\mathbf{z}^T\right\} = I$.

Now the problem can be restated as finding a matrix W such that $\tilde{\mathbf{s}} = \mathbf{y} = W\mathbf{z}$ result as statistical independent as possible. The advantage of using $\mathbf{z}$ instead of $\mathbf{x}$ is that now the matrix W will be orthogonal. Of course that working with W on $\mathbf{z}$ is the same as working with WQ on the original (centered) data $\mathbf{x}$. In this sense, instead of finding the matrix W as a whole, the problem can be solved by finding each row of the matrix, $\mathbf{w}_i^T$, subject to orthogonality restriction among the succesive estimated rows. This is equivalent to extract the sources, one by one, in a deflationary approach. Then, for each source we search for $s_i = \mathbf{w}_i^T\mathbf{z}$, where $\mathbf{w}_i$ is obtained maximizing the nongaussianity of the extracted source s_i. During this process, a unitary norm condition is imposed (as necessary to produce an orthogonal matrix W), and $\mathbf{w}_i$ is forced to be orthogonal to all previously estimated $\mathbf{w}_j$. This orthogonality is imposed by a Gram-Schmidt like method.

The solution then consists of maximizing the absolute value of negentropy, which will ensure nongaussianity of the signals. Instead of using negentropy directly, which would requiere knowledge of the pdf of the sources, some approximations to it based on nonlinear functions are used. In equations, after the prewhitening step, each row of matrix W is found by the following estimation problem: optimize the contrast $J(y_i) = \mathcal{E}\left\{G(\mathbf{w}_i^T\mathbf{z})\right\}$ subject to $\|\mathbf{w}_i\|^2 = 1$. In this equations, $G(\cdot)$ is a properly chosen nonlinear function, for example, $G(y) = \frac{1}{a}\log\cosh(ay)$, that can work as a robust approximation of the negentropy, as shown in [Hyvärinen, 1999].

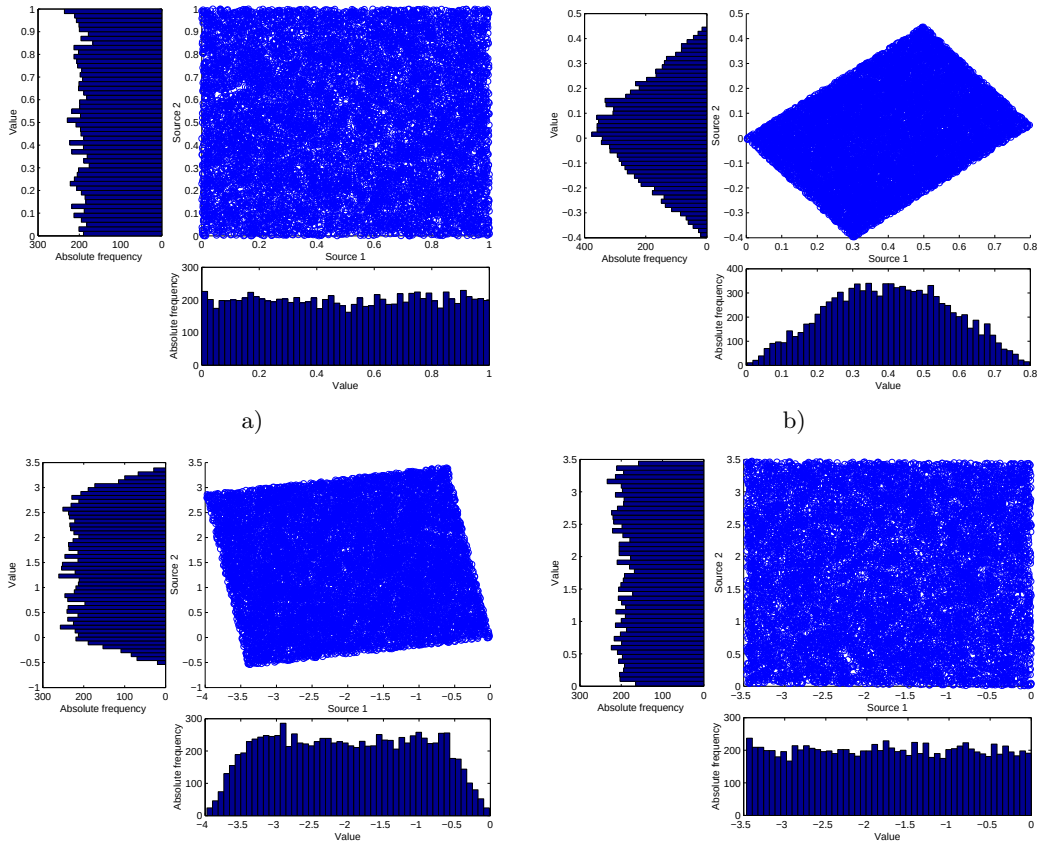


Figure 1.12. Example of the application of FastICA algorithm. The image shows scatter plots and marginal histograms for: a), the sources $\mathbf{s}$, b) the mixed signals $\mathbf{x}$, c) the whitened signals $\mathbf{z}$, and d) the obtained estimated sources $\mathbf{y}$.

Each independent component is extracted one by one, and a restriction of orthogonality is applied among them to prevent the extraction of previous sources. The optimization is made by means of a quasi-newton approach which yields very fast convergence, this is the reason for the name of the algorithm.

As an example to illustrate the method, Figure 1.12 shows scatter plots and marginal histograms at the different stages of the separation process, for two random sources that are independent and have uniform pdf. In part a), the original sources are shown. Part b) presents the mixed signals $\mathbf{x}$, mixed with an arbitrary mixing matrix. It can be seen that the effect of mixing is a deformation and rotation of the joint distribution. In part c), the whithening transformation was applied to produce $\mathbf{z}$. The

deformation was corrected by this process but the variables are not independent. Finally, in part d), the optimization using nongaussianity corrects the rotation producing independent signals.

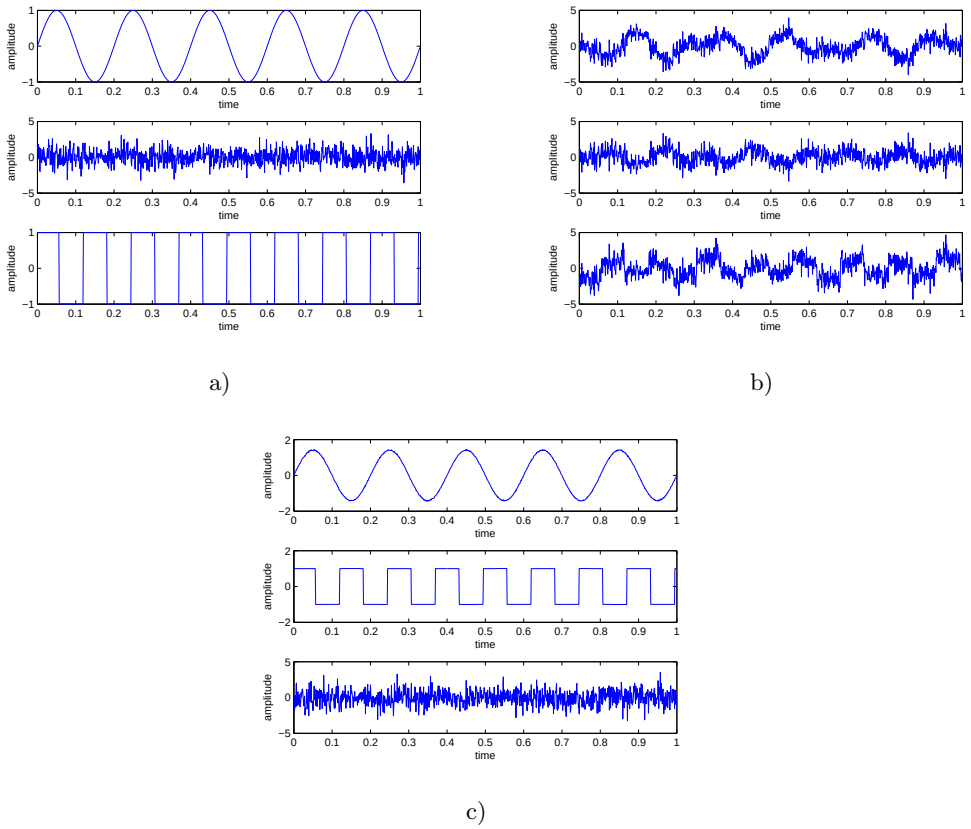


Figure 1.13. Example of separation of three sources. In part a) the original sources are drawn. Part b) shows the mixed signals. In c) the separated sources are presented.

As another example, Figure 1.13 presents the case of three sources. Part a) shows the original sources, part b) presents the mixtures produced by a given mixing matrix, and part c) the separated sources obtained using FastICA. This example also illustrates on the ambiguities of the method: it can be seen that the extracted sources are in a diferent order than the original ones and with some arbitrary scaling.

Up to now, it was assumed that the data were real valued, but all the previously stated will be still valid for complex data, provided that some specific adaptations to cope with the nature of the complex signals are made. In this way, there are complex versions of many of the real ICA algorithms, as for example a complex FastICA method

[Bingham and Hyvärinen, 2000]. In the case of complex signals the ambiguities are the same as for the real ones, with the exception that the scaling can now be complex, so it includes a magnitude and an argument.

1.4. Automatic speech recognition

Automatic speech recognition (ASR) is the general name of a series of tasks related to the translation of a speech signal into the transcription of its text content. Given the kind of spoken content, ASR can be classified into isolated word recognition, continuous speech recognition and spontaneous speech recognition. The first involves words separated by well defined pauses among them. The second involves a continuous stream of voice, in a similar way than real speech is produced. The spontaneous ASR takes into account the case of a real conversation, with partial phrases, hesitation, repetitions, misspronunciations, etc. Of course they involve increasing complexity [Furui, 1989].

Also ASR systems can be distinguished by the size of the vocabulary they can recognize. There are small vocabulary recognizers (up to about 100 words), developed to be used in very specific task (like recognition of numbers, or commands for a machine). Large recognition systems can deal with more complex tasks like simple dialog systems (up to 1000 words). Very large systems can deal with everyday conversations (tens of thousands words).

With respect to the number of the speakers allowed, ASR systems can be built for a single speaker, for a group of speakers, or independent of the speaker. In the first two cases, the system is guaranteed to work only for the person or group of persons for which it was trained. The last option, on the other hand, means that the recognizer must maintain its recognition rate no matter who is using it [Deller et al., 1993], which is clearly a more difficult task.

In general, the more complex the ASR system is, the more sensitive will it be to variations on the working characteristics respect to the training stage. For the application of interest, with an algorithm of BSS as a preprocessing stage, using a simpler recognizer could mask the capabilities of the separation methods. That is, if a too simple ASR task is selected, the ASR system will be able to solve the problem even if the method used for audio source separation was not too good. For this reason, a large vocabulary, continuous speech, speaker independent recognizer will be used to test the capabilities of our algorithm.

1.4.1. Statistical ASR systems

Among the paradigms for speech recognition, the pattern recognition method is the more studied and succesfull [Rabiner and Juang, 1993]. In this paradigm, there are

many technologies used for speech recognition. In the early stages of ASR technology, template based recognizers with dynamic time warping alignment was the standard for isolated word recognition [Juang and Rabiner, 2005]. Then, hidden Markov models (HMM) caused a revolution in the '80s.

The use of HMMs is based on statistical pattern recognition methods. Here a very brief descrition will be made. Suppose that one has a sequence of acoustic evidences (information obtained from the speech source) in $A = \mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ and a sequence of words $W = w_1, w_2, \ldots, w_q$ where each word belongs to a finite vocabulary $\mathcal{W}$. Using the notation $P(W|A)$ to represent the probability of a word sequence given the acoustic evidence, the recognition system should choose the best candidate word sequence as [Huang et al., 1991]:

$$\tilde{W} = \arg \max_{W} P(W|A) \, . \tag{1.11}$$

Now, applying the Bayes theorem to the probability of the words given the acoustic evidence

$$P(W|A) = \frac{P(W)P(A|W)}{P(A)} \, . \tag{1.12}$$

Since the denominator does not depend on W, for the purpose of speech recognition it is not important (will be a constant scaling for all the candidate word sequences), and can be disregarded. Then, the recognizer would select the word sequence according to

$$\tilde{W} = \arg \max_{W} P(W)P(A|W) \, . \tag{1.13}$$

The two terms in this equation have a direct interpretation. The *a priori* probability $P(W)$ is the probability of each possible sequence W in the language. Evaluating this requieres a precise *language model*. On the other side, the conditional probability of the acoustic evidence for a given word sequence needs the existence of appropiate models for the acoustic evidence. This is provided by an *acoustic model* [Jelinek, 1998].

In this context the HMM provides the acoustic models. A Markov model is a finite state automaton that follows the markovian hypotesis that the state in a time instant depends only on the previous state. At each state the model emits or observes a specific symbol or pattern, and thus given a sequence of these patterns there is a direct correspondence among the sequences and the states of the model. A hidden Markov model is a markov model which in each state emits a symbol or pattern with certain probability. In this way, given a sequence of observations it is impossible to know exactly which sequence of states generate these observations. The observed or emited pattern can be a discrete object from a finite set, or a sample from a continuous distribution. This separates the discrete and continuous hidden Markov models.

In the context of continuous speech recognition, the HMM are used to model individual low level speech units like phonemes or triphones. Then, these small models

are linked together to form higher level structures like syllables and words. The low level markovian models are called the acoustics models, and they are trained with the help of large databases of speech. One particular property of Markov models is that they can manage the length variability of speech. For example, the duration of an "a" is not uniform in all utterances, and so the additional observations produced by the length variation are simply absorbed in transitions to the same state [Huang et al., 1991].

In addition to the acoustic models, there is a language model, that has the information of how the low-level models must be linked to form significant words, and how these words can be linked into significant sentences. This information is added as transition probabilities among the structures, and so a very large model is built with all possible combinations of small speech units, words and sentences [Jelinek, 1998]. There are efficient algorithms to find the best (more probable) sequence in this huge model, and thus the most probable sentence can be obtained for a given utterance.

This is the state of the art in speech recognition. Although there are other approaches like neural networks, HMM technology is still the standard method used, and so this is the kind of recognizer that will be used in this work. These kind of systems have achieved a very high rate of success [Juang and Rabiner, 2005; O'Shaughnessy, 2008]. Nevertheless, it is also known that the performance of these systems is degraded in the presence of noise [Rufiner et al., 2004]. Moreover, when compared to human listeners, the ASR systems are found to be very sensitive to noise, with a large degradation for cases were humans can still perform almost perfect recognition [Lippmann, 1997].

To illustrate this point, Figure 1.14 presents a comparisson of the performance of humans with normal hearing, hearing impaired humans, and an ASR system, when the speech is contaminated with white noise at different signals-to-noise ratios (SNR) [3]. It can be seen that normal hearing humans can mantain a high recognition rate even when the noise power equals the desired source power (0dB), while both hearing impaired persons and ASR systems loose a lot of recognition capabilities with increasing noise power.

This clearly shows the need for robust speech recognition systems, where the word "robust" makes reference to the ability of the system to maintain its performance when the usage conditions differ from those of the training.

There are basically three approaches for producing robust ASR systems [Grimm and Kroschel, 2007], taking into account the level in the system in which the method is implemented. The first method is based in noise compensation, either at the level of

[3]The normal hearing humans data was adapted from [Lippmann, 1997], the hearing impaired data was adapted from [Finitzo-Hieber and Tillmann, 1978], and the ASR system performance from [Rufiner et al., 2004]

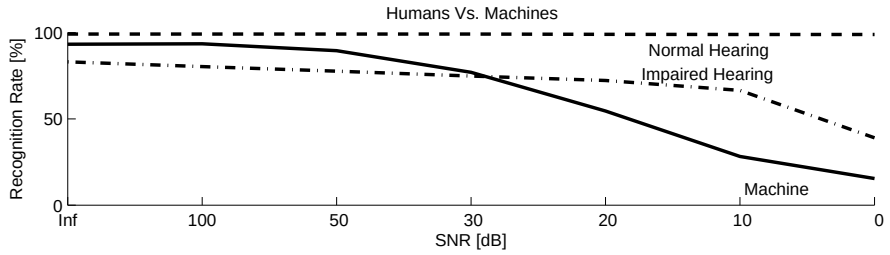


Figure 1.14. Recognition rates under different white noise powers for normal hearing humans, hearing impaired humans and ASR systems.

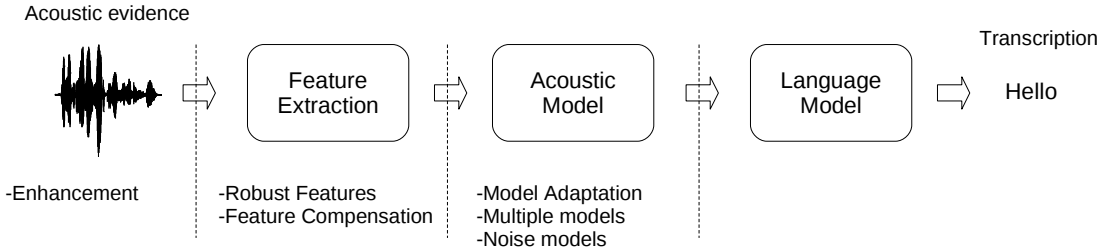


Figure 1.15. Block diagram of an ASR system, including the levels at which robustness can be added.

the speech itself where it is called enhancement [Benesty et al., 2005], or at the level of the characteristics, where it is called feature compensation [Gong, 1995]. The second technique is based on the use of robust features, that are intrinsically less sensitive to noise [Hermansky and Morgan, 1994; Kinsbury and Morgan, 1997]. The third is the adaptation of the clean models to the acoustic environment. This can be done for example, by taking into account the noise characteristics as a paralell noise model, or having several models trained in different kind of environments and selecting which models fits better the data [Gales, 1998]. Figure 1.15 presents a scheme of an ASR system, from the acoustic evidence (speech) to the resulting transcription, including the levels at which the methods to produce robustness are introduced.

This investigation deals with the case of speech captured remotely with respect to the source. In this case, there are two basic sources of noise to consider: uncorrelated noise, due to other sound sources, and correlated noise, due to reverberation. In this work we will use an approach of noise compensation at the level of the speech signal itself. This means that our objective will be to use the noisy speech signal captured by several microphones, to produce a version of it as clean as possible, before introducing it to the ASR system.

1.4.2. Performance evaluation for speech recognition

Being ASR a very specific task related to transcription of speech to text, there are objective ways to evaluate the performance. These are based on measures of how many mistakes and of what type are made by the recognizer. One kind of measures acts on the whole sentence, counting the rate of errors (or the successful recognition) of the sentences as a whole. Although this will be the final evaluation, it is not so informative, as sentences have many words, and an error on one of them will make the whole sentence incorrect. It is not the same if the recognizer fails in every word, or just in one of them.

To overcome this drawback, usually the recognition (or error) rate are evaluated at the level of words. When counting word errors, there are several kinds of mistakes. Suppose that for a given utterance, there is a reference transcription (made by an expert) of the spoken words. The ASR system under study produces an output sentence. This output sentence is aligned with the real transcription by means of a dynamic programming method. In the ASR aligned transcription, some words match those of the real transcription while other words are deleted, inserted or substituted by others. One of the most used measures of performace is the word recognition rate (WRR), defined as [Young et al., 2005]:

$$WRR\% = \frac{T - D - S}{T} 100 \tag{1.14}$$

where T is the total number of words in the reference transcription, D is the number of deletion errors and S is the number of substitution errors. Of course, if there is a database with many utterances and their corresponding reference transcriptions, these quantities can be measured in a cummulative way for all the database, generating an overall WRR value. This measure will be used in all the work.

When these evaluations are done, usually a huge database of speech is necessary. This database is partitioned in a training set, from which the method is trained, and a test set in which the performance of the trained recognizer is evaluated. To produce statistically significant results, with good sensitivity, it would be necessary a huge database. A method to produce statistical significant results when the database size is limited, is to create partitions of the data, in a standard technique called k-folding. In this method, with k partitions, each partition is produced by division of the database randomly into training and test sets. Then k systems are trained with the training sets, and tested in the test set. The average of the k partitions gives a result that is equivalent to that of a database of k-times the size of the one used. This technique can be applied to the test of acoustic models, language models, or both.

1.5. Quality evaluation for speech signals

Quality evaluation of speech after application of an enhancement method is a complex problem that depends on the application field. In some cases, the main interest is not recovering the original signal but preserving some characteristics that are required for the task concerned. For example, when retrieval of a voice to be used in a hearing aid device is desired, a perfect reconstruction of the original waveform is not as important as a good perceptual quality. In the same way, for ASR systems, auditory perception is not as important as preserving some acoustic cues that are used by the system to perform the recognition. On the contrary, in other situations the aim is to recover the original signal as exactly as possible, such as a waveform coder. So far, few works have been presented with specific proposals for quality evaluation in the field of BSS. Particularly, in the context of automatic speech recognition, the only available way to evaluate the performance of some blind source separation algorithm is through a speech recognition test.

There are basically three approaches for this evaluation: objective quality measures, subjective quality evaluation and task-specific evaluation. Objective quality measures are designed to produce a numerical value that correlates well with the quality of the signal. The importance of this kind of measures is that they allow for a direct and objective comparison of different processing alternatives. In general they are presented as a measure of discrepancies between the clean original source, and the processed one.

On the other side, subjective quality evaluation uses a group of people that present, in some numerical scale, their opinion with respect to some aspects of the quality, like overall quality, background residual noise, etc. A large number of these opinions are averaged to produce a measure of the aspect under evaluation. Finally, in task-specific evaluations, there is a system which will use the processed speech, and thus one can measure the quality of the speech by means of some measure of the behaviour of this system. This is the case of using the output of an ASR system to evaluate how good was the enhancement method employed as pre-processing.

Among the objective quality measures, there is a group that has gained popularity in the last years. They are the perceptually derived quality measures. The main idea behind this kind of measures is to mimic the processing done by the human hearing system, and then evaluate the distortion of the processed speech with respect to the original clean speech in this perceptual domain. These measures have shown to be highly correlated to subjective quality tests and are therefore very attractive. Some examples of this kind of measures are the meausuring normalizing blocks measure (MNB) [Voran, 1999a] and the perceptual evaluation of speech quality measure (PESQ) [ITU, 2001]. In Chapter 3, a deep study of objective quality measures an their application to ASR will be presented.

1.6. Concluding remarks

In this first chapter, a brief introduction to the theoretical bases over which all the dissertation will be developed was presented. The concepts of sound transmission in enclosed environments, convolutive mixtures and reverberation time from acoustics will be extensively used. The basic ideas of ICA processing will be applied as principle of separation in the developed algorithms, although it will be necesary to introduce some changes to adapt it to the convolutive case. The evaluation of the improvements obtained by the developed techniques will be measured, among other methods, with specific tasks of ASR.

Chapter 2

Standard approaches for BSS of audio sources

Several possible solutions to the problem of audio sources separation have been presented in the past. They include a variety of methods, ranging from microphone array processing techniques related to those applied in radar and medical ultrasound imaging devices, to very recent ones like the ones based on sparsity and disjoint orthogonality. Some of them will be discussed in the following. This analysis will start with the classical approach of spatial filtering and array processing using beamforming. It will provide some background and usefull equations for dealing with microphone arrays, which are applicable for all kinds of array processing. Next, the method of separation based in sparsity will be presented, with an analysis of its limitations, followed by a comment on the time-domain ICA approach. Finally the frequency-domain ICA approach will be presented, with a detailed analysis of its working principles, advantages and limitations. This approach will be used in this work, and thus a good comprehension of its working principles will be very important to understand the methods proposed in the next chapters.

2.1. Beamforming

The basic method of spatial filtering is that of array processing and beamforming. The terms “spatial filtering” make reference to the capability of enhancing or rejecting the signals arriving from specific directions, expressed as angles respect to the array axis. This technique is based on special combinations of the signals of several microphones to produce a result in which the signals arriving from some direction are enhanced. Thus, these techniques transform an array of omnidirectional microphones into a directional sensor with a specific directivity pattern (that is, with sensitivity dependent in the angle of arrival of the sound source) [Brandstein and Ward, 2001].

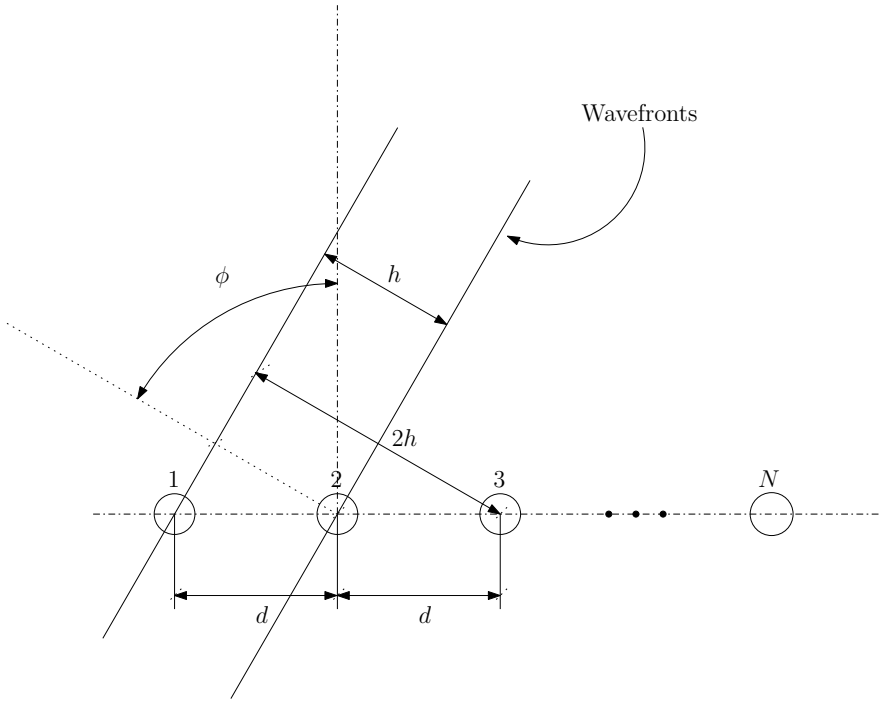


Figure 2.1. Transmission model for the uniform linear array in the far field.

Figure 2.1 shows a schema of an uniform linear array, in which the spacing among microphones is always d. There is a wavefront produced by a source, assumed to be in the far field (and thus, assumed planar propagation), arriving with a direction of propagation that forms an angle ϕ with respect to the axis perpendicular to the array. In this situation, the wavefront arrives at microphone 2 after travelling a distance h longer than the one for arriving at microphone 1. This difference in distance can be written as $h = d\sin(\phi)$, and given the travelling speed c of the wavefront, the wave arrives at microphone 2 after a delay of $\tau = \frac{d\sin(\phi)}{c}$ with respect to microphone 1. Also, the wavefront looses energy with the inverse of the distance to the source. Let be r the distance from the source to microphone 1 and p_0 the pressure level at the source. Then, the pressure level at microphone 1 will be $p_1 = \frac{p_0}{r}$, the pressure level at microphone 2 will be $p_2 = \frac{p_0}{r+h}$, and so on. If $r \gg h$, that is, in the far field the microphone spacing is very small compared to the distance to source, one can eliminate the h terms in the denominator and $p_1 \approx p_2 \approx \cdots \approx p_N$. Under this approach then the signal measured

by the array $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_N(t)]^T$ can be written as:

$$\mathbf{x}(t) = [x_0(t), x_0(t-\tau), x_0(t-2\tau), \cdots, x_0(t-(N-1)\tau)]^T \,, \tag{2.1}$$

where all the signals are referred to that measured at microphone 1. After a Fourier transform, this can be written as [Van Veen and Buckley, 1988]:

$$\begin{aligned} \mathbf{X}(f) &= X_0(f) \left[1, e^{\frac{-j2\pi f d \sin(\phi)}{c}}, e^{\frac{-j2\pi f 2d \sin(\phi)}{c}}, \cdots, e^{\frac{-j2\pi f (N-1) d \sin(\phi)}{c}}\right]^T \\ &= X_0(f)\mathbf{v}(f,\phi) \,, \end{aligned} \tag{2.2}$$

where $\mathbf{v}(f,\phi)$ is the array response vector as a function of frequency f and arrival angle ϕ. It must be noted that this array response characterizes how the microphones capture the sound coming from sources located at different angles, and thus it is the frequency response of a single-input multiple-output (SIMO) system. It is a property of the array that depends on its geometry and the source location.

The simplest of the beamforming approaches is the delay-and-sum beamformer, in which combining specific delays with scalings the array can be steered in one direction. The delay-and-scale operation corresponds to a complex scaling in the frequency domain. Figure 2.2 presents a scheme of the operation of a delay-and-sum beamformer in the frequency domain. Let us use the notation $\mathbf{a} = [\alpha_1 e^{j2\pi f \tau_1}, \alpha_2 e^{j2\pi f \tau_2}, \ldots, \alpha_n e^{j2\pi f \tau_n}]^T$ to represent the delay and scale operations. This vector of coefficients characterize the behaviour of the beamformer. The parameters α_i and τ_i must be adjusted to produce the desired directional response, and so they are expressed as functions of the desired steering direction θ.

For a given set of beamformer coefficients $\mathbf{a}$, the output of the beamformer $Y(f)$ can be written as

$$Y(f) = \mathbf{a}^H \mathbf{X}(f) = X_0(f)\mathbf{a}^H \mathbf{v}(f,\phi) \tag{2.3}$$

where $[\cdot]^H$ represents the conjugate transpose operation. The quantity $\mathbf{r}(f,\phi) = \mathbf{a}^H \mathbf{v}(f,\phi)$ is called the beamformer response, and for fixed coefficients is a function of the frequency and angle of arrival. Another useful characterization of a beamformer is its beampattern, defined as $|\mathbf{r}(f,\phi)|^2$ also called the directivity pattern of the beamformer. This directivity pattern is plotted, usually in polar forms and in decibel scale, and represents the power gain of the beamformer for each frequency and for all the possible arrival angles.

For example, to design a beamformer to enhance the source in direction θ, the delays τ_i in $\mathbf{a}$ should be adjusted as $\tau_i = (i-1)\frac{d \sin(\theta)}{c}$. For an angle of arrival $\theta = 30$, with a microphone spacing of $d = 0.04$ m and using the speed $c = 343$ m/s, the directivity pattern for a frequency $f = 2000$ Hz is shown in Figure 2.3. In this directivity pattern the scaling factors were $\alpha_i = \frac{1}{N} \, \forall i$. In the left, the array has $N = 5$ elements,

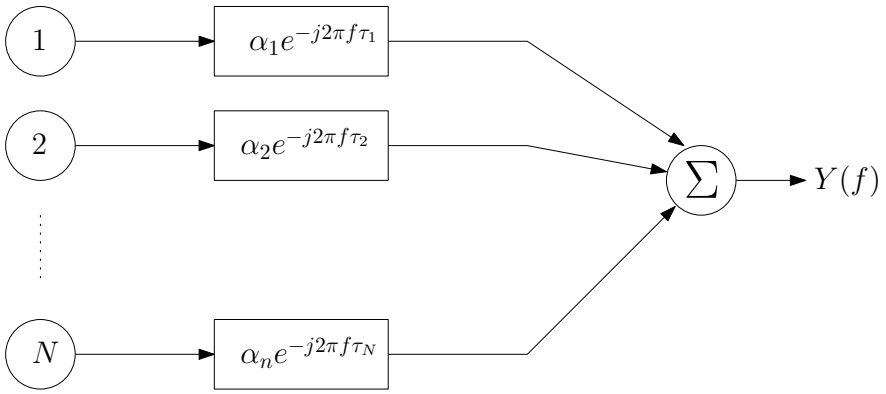


Figure 2.2. Delay and sum beamformer in the frequency domain.

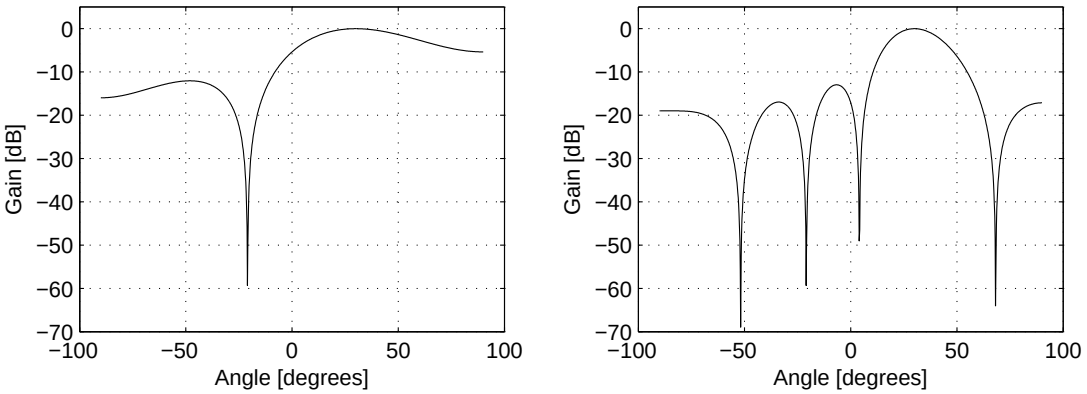


Figure 2.3. Directivity patterns for a delay and sum beamformer. Left, 5 element array. Right, 10 element array.

and in the right $N = 10$ elements. It is clear that increasing the number of elements in the array reduces the width of the main lobe and improves the rejection in other directions.

This example was designed using only the information of the direction of arrival of the desired source. If there are other sources, the beamformer coefficients can be optimized to not only enhance the source from the desired direction(s), but also to reject those coming from other specific directions. Of course, the problem of finding the proper beamformer coefficients is more difficult to solve, and is usually stated as the solution of an overcomplete set of equations.

Other variants exist, like the more advanced filter-and-sum, in which each mi-

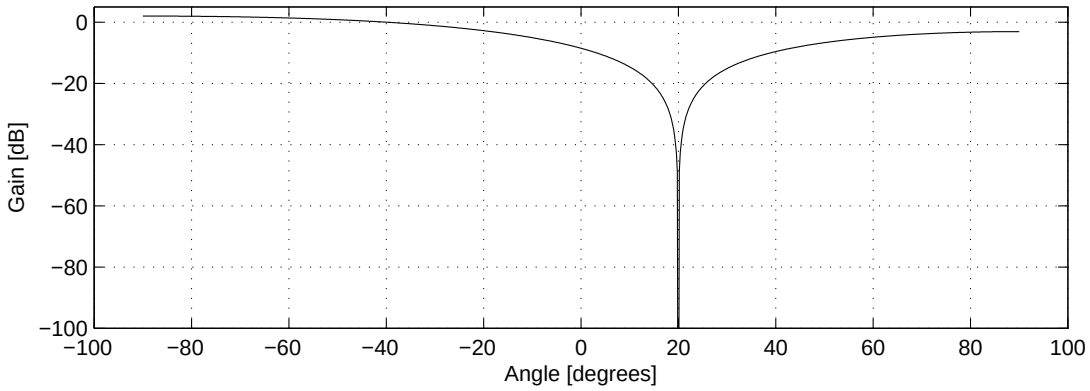


Figure 2.4. Example beampattern of a null beamformer.

crophone input, before being added, is filtered with some specifically designed filter. This allows for additional enhancement of the desired source and/or specific rejection of sources from undesired directions.

A different kind of beamformers, called null beamformers, instead of enhance the signal from some direction, are designed to reject the signals in some specified angle. This is particularly interesting for our work because, as will be seen soon, some frequency domain methods for convolutive BSS can be interpreted as sets of adaptive null beamformers. Suppose that one needs to design a beamformer to work at a frequency f_1, that enhances the signals from direction θ_1 but rejects those coming from direction θ_2. Moreover, assume that $N = 2$ microphones are used. The design problem can be stated as a set of two equations and two unknowns, as following:

$$\begin{cases} \mathbf{a}^H \mathbf{v}(f_1, \theta_1) = & 1 \\ \mathbf{a}^H \mathbf{v}(f_1, \theta_2) = & 0 \end{cases} \tag{2.4}$$

Note that the rigth-hand side was set to provide a magnitude of 1 in the desired direction and a 0 on the rejection direction, but nothing was imposed for the in-between directions, and thus they will have other magnitudes. Figure 2.4 shows an example of null beamformer calculated using this technique for a $N = 2$ element array, with spacing $d = 0.04$ m, for frequency $f = 2000$ Hz, designed to pass the signal coming from $\theta_1 = -40$ degrees and with a null at $\theta_2 = 20$ degrees.

Combining a beamformer for the desired direction with several beamformers for the undesired directions, one can produce a device called *generalized sidelobe canceller* (GSC). The bemformers for the undesired directions are used as an estimation of the undesired residual noise sources, and thus the output of the beamformer is improved by

eliminating the remaining noises using, for example, a Wiener filter [Manolakis et al., 2005].

There are some drawbacks for the beamformers. In first place, they require knowledge of the desired direction of the source (the angle θ). Also, the pattern produced is not uniform with frequency, and this is key in signals that have wideband content, like speech. These two aspects have been subject of important research, aiming to obtain constant directivity and adaptive beamformers, that can track variations of source locations, with some degree of success.

Another subject that needs to be addressed is the spatial aliasing problem. As the microphones are sampling the space, care must be taken to avoid the occurrence of spatial aliasing, which is a phenomenon in which additional sidelobes appear in places they were not supposed to be, degrading the performance of the beamformer [Brandstein and Ward, 2001]. This subject of spatial aliasing is specially important, as it is common to many array processing techniques. If a signal of frequency f and wavelength λ is sampled, to avoid the spatial aliasing the maximum allowed space d should be lower than half the wavelength. Thus, using the relation $f\lambda = c$, we find the relation between the microphone spacing and the frequency [Manolakis et al., 2005]

$$d_{max} = \frac{\lambda}{2} = \frac{c}{2f_{max}} \,. \tag{2.5}$$

This equation depends on the frequency, so if the array processing is supposed to work properly for all the frequency range, then the maximum frequency present on the signals must be used to produce the maximum allowable spacing. For example, for speech sampled at 8000 Hz, the maximum frequency is 4000 Hz, and using $c = 343$ m/s, one gets $d_{max} = 0.042875$ m, about 4.3 cm.

To show an example of spatial aliasing, let us repeat the same example used before for the delay and sum beamformer (Figure 2.3), but for a spacing that produces aliasing. Using $f = 2000$ Hz in (2.5), the maximum spacing to avoid space aliasing in this frequency is $d = 0.085750$ m. Using a spacing of 12 cm, larger than the critical one, the beampatterns are as shown in Figure 2.5. The spatial aliasing phenomenon is clearly seen, as now there is a second lobe enhancing the signal with a different direction than that for which the beamformer was designed. This means that if a competing source arrives with that specific direction the array will be useless for rejecting it.

2.2. Sparsity-based BSS

This is a recent method and a very successful one, as long as the mixing process behaves according to its main hypothesis of sparsity. This hypothesis assumes that at each time and frequency, only one source is active. In this way if one can classify all the time-frequency slots where a source is active, then that source can be reconstructed.

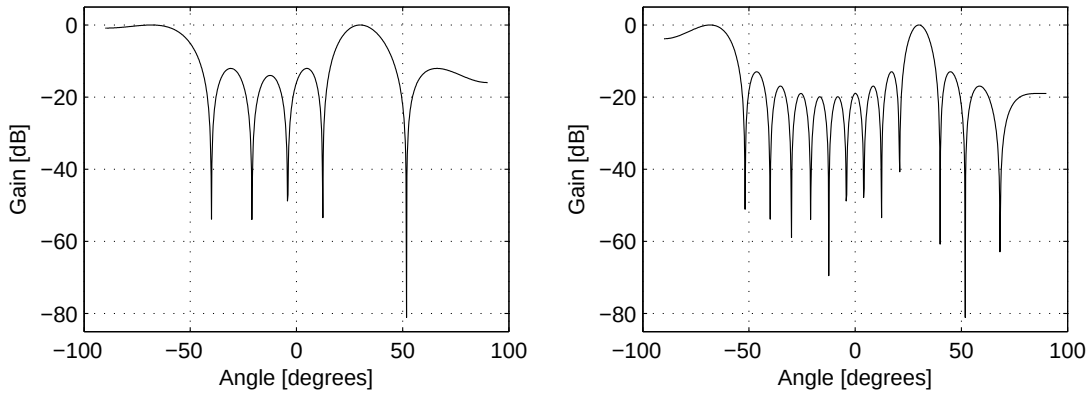


Figure 2.5. Beampattern for a delay and sum beamformer with spatial aliasing. Left, for a 5 element array. Right, for a 10 element array.

The method is strongly related to the computer auditory scene analysis (CASA) alternatives of masking, that are based in the ideas of the masking thresholds in human audition [Wang and Brown, 2006]. In that context, the segregation is achieved by application of a binary mask that contains ones only in the positions of the time-frequency domain corresponding to the main source. The key aspect is how these masks are generated. If some auditory cues and auditory models are used to generate the masks, the approach is more related to CASA, whereas if the generation is more "ad-hoc" or based just in signal processing, more blind in nature, it is classified as a BSS method.

In this context, the idea of BSS exploiting sparsity is to make use of the property of sound signals of being sparse in the time-frequency plane. This means that there are a lot of areas of the plane, in which the signal has no energy. When there are two or more signals, there is some probability that the time-frequency slots in which one of them has significant energy will not coincide with those for the other signals. This is the principle called "disjoint orthogonality". Using the notation $s(\omega, \tau)$ to refer to the short-time Fourier transformed signal, disjoint orthogonality can be stated as [Yilmaz and Rickard, 2004]

$$s_i(\omega, \tau)s_j(\omega, \tau) = 0 \; \forall i, j, \omega, \tau, \tag{2.6}$$

where s_i and s_j are any pair of signals that compose a mixture.

In other terms, the main hypothesis is that at each time-frequency slot, only one of the sources is active. So, if one can classify all the time-frequency slots and cluster them according to the source they belong to, putting all other slots to zero, this will allow the reconstruction of each source. This is based in a binary masking process: some masks are constructed with a value of one for the slots that belong to a given source and zero for the others. By constructing these binary masks, all sources can be

separated, with some success. Moreover, this method is capable of working in the case of less microphones than sources, and even for only one micophone, as long as some method to produce the proper binary masks can be developed.

There are different methods according to the strategy used for clustering the bins into different sources. The first step is to produce a vector of characteristics in which the sources could be clearly clustered. For this task, usually the difference in angle and the ratio of magnitudes between a given microphone and another microphone selected as reference are used. Then once these vectors of characteristics are formed, they should be clustered to decide to which source each bin belongs. Using all bins belonging to a cluster, the masks are generated.

This method can work as long as the hypothesis of disjoint orthogonality remains valid. But it is known that it is only partially true [Yilmaz and Rickard, 2004], and moreover, that it collapses when the reverberation time of the room is increased (due to the smearing effect of reverberation that increases the time duration of all phenomena), and when the number of sources is increased [Araki et al., 2007]. To illustrate these aspects, Figure 2.6 shows the definition of disjoint orthogonality applied to a mixture of two signals. In the left, the mixture is instantaneous, and in the right convolutive. In each column, the top two images show the spectrogram of the signals prior to the mixture, and the last image shows the values where the product of spectrograms has significant values. This was calculated as the points where the product $s_1(\omega,\tau)s_2(\omega,\tau)$ normalized by its maximum, is larger than 0.001, that is, the points that are up to 30 dB attenuated with respect to the maximum. This measures the points for which the main assumption is not satisfied. As can be seen in the left part, the hypothesis is not completely true, although the number of points that are common to the two sources is only 3.42% of the total time-frequency points. In the right it is clearly seen how the smearing effect introduced by reverberation increases this value to 12.46% of the total points. This means that the number of points for which the disjoint orthogonality principle cannot be applied is 3.64 times larger under this reverberation condition.

When the main assumption collapses, each slot belongs to several sources, and thus what is needed is a quantification of belongings of each slot to each source. Some approaches have started to appear which use continuous masks instead of binary ones, with some degree of success [Park and Stern, 2006; Srinivasan et al., 2006]. Another problem of the binary mask method is the occurrence of the so-called “musical noise” due to the discontinuities introduced by the binary mask in the time-frequency plane. The use of continuous mask can also mitigate to some degree this effect [Araki et al., 2004].

The main advantage of this sparsity approach is that it can work when there are less microphones than sources, a case called “undercomplete” or “underdetermined”, that cannot be solved by other methods. The main disadvantage is the negative effect

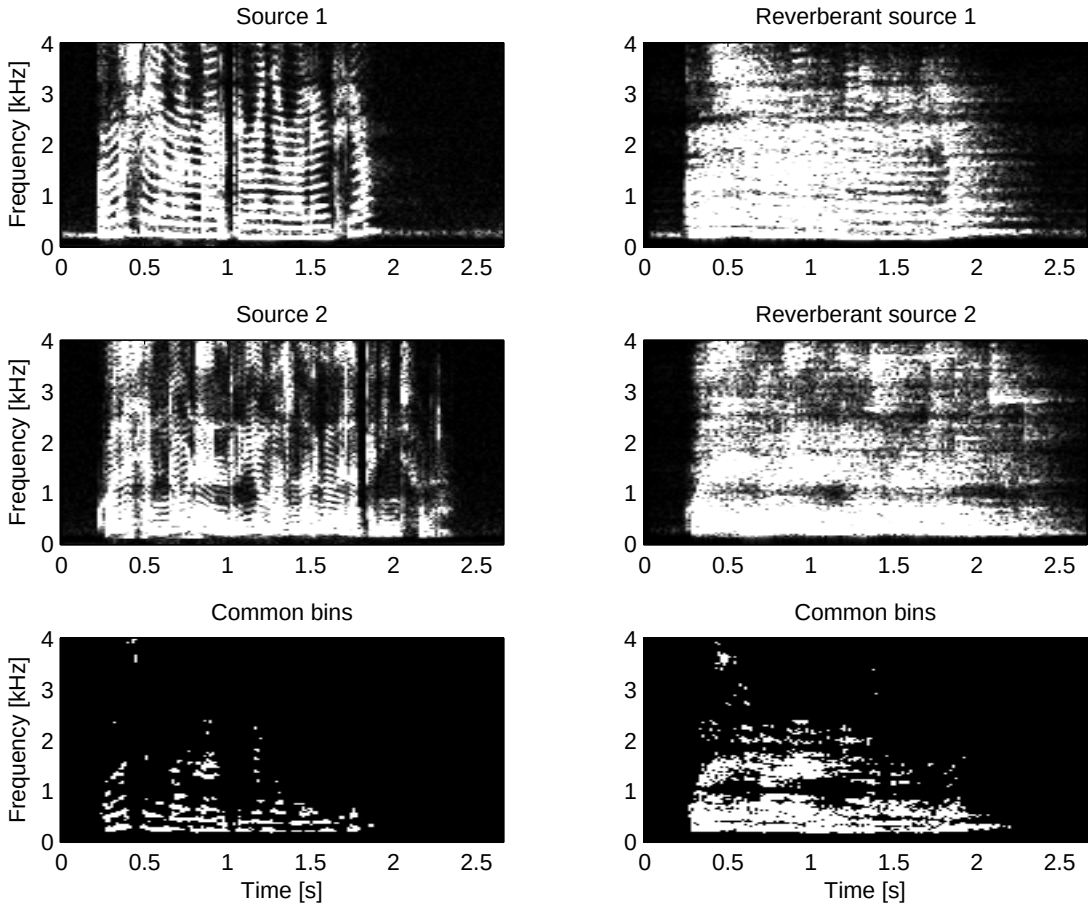


Figure 2.6. Disjoint orthogonality for instantaneous and convolutive mixtures. Left column, instantaneous mixture. Right column, convolutive mixture.

of reverberation on its performance, which implies a strong limitation for practical applications where it is very likely that reverberation effects will happen.

2.3. Time-Domain ICA approach

In the same way that with the ICA case for instantaneous mixtures, a formulation in term of independent components analysis can be proposed for the convolutive case. The procedure is similar to the one described in Chapter 1, but adapted to the convolutive case [Cichocki and Amari, 2002]. This adaptation means that in all places where a product with an element of the mixing matrix was done (the cost functions and the

separation equations), now it has to be replaced by a convolution. The optimization of cost functions with convolutive terms produces very difficult to evaluate and very complex equations for the updating of the separating filters.

In general, the time-domain approaches use the relative gradient or the natural gradient update to improve the convergence properties. In these approaches, the usual calculated gradient of the cost function is modified to adapt it to the local Riemannian structure of the parameter space [Amari et al., 1997]. Several algorithms have been proposed using cost functions based on Kullback-Leibler divergency [Amari et al., 1997], mutual information [Douglas and Sun, 2003], or other "ad-hoc" cost functions [Kawamoto et al., 1998].

In all these methods, the update equations include convolutions that must be calculated for each tap of the filters and for each iteration of the algorithm. Thus, the main drawback of these approach is their computational cost. For audio sources separation, the filters need to capture the real room characteristics, and thus very long filter (thousands of taps) are needed.

One variant of this approach was proposed in [Lee et al., 1997]. In this method, the nonlinear functions involved in the updates are calculated in the time domain, but all the convolutions are calculated in the frequency domain using finite impulse response (FIR) matrix algebra [Lambert, 1996]. This reduces the complexity due to the fact that convolutions are converted into multiplications in the frequency domain. Nevertheless, the computational requirements are still high, as there is a need to perform direct and inverse transforms in each step of the iteration, which makes the approach only useful for short filters.

Recently, a convolutive time-domain version of FastICA algorithm was presented and shown a performance equivalent to that of frequency domain approaches. Nevertheless the complexity of the solution is high (for example, the authors report a processing time of 2.75 seconds for iteration for a two sources / two mixtures problem) [Douglas et al., 2007].

2.4. Frequency-Domain ICA approach

This approach is one of the most successful ones, and is the kind of algorithms that will be explored in this dissertation. For this reason, it will be presented in more detail, with an analysis of its weaknesses and advantages.

2.4.1. Statement of the problem

To start the development, let us refresh the convolutive mixture (1.7) for the case of M sources and N microphones, and restrict to the case of $M = N$, rewritten here

for clarity:

$$\mathbf{x}(t) = H * \mathbf{s}(t) = \sum_{p=0}^{L} H_p \mathbf{s}(t-p), \tag{2.7}$$

where now we have expanded the convolution involved, using a FIR matrix H_p of $N \times N$ which contains the filter coefficients for lag p. In this equation, L represents the length of the mixture filters. This is a LTI FIR system with multiple inputs and multiple outputs (MIMO).

At this point it is important to remember the hypothesis employed in this formulation:

- The mixture process is linear.
- The room response is time-invariant, or equivalently, the sources and microphones are fixed.
- There is no noise, or its effect is so small that it can be disregarded.
- The sources that produce the mixture are independent.
- Each source is generated by an iid random process.
- There are as many microphones as sources, thus the mixing matrices are square.
- The mixing filters are invertible.

The hypothesis about the number of sources and microphones is very important. If there are more microphones than sources, the method can be applied after some dimension reduction to make the problem square. But if there are more sources than microphones the mixture is called underdetermined and cannot be solved with the methods that will be presented here (instead, methods based on sparsity as previously seen can be used).

Also it was assumed that the mixing system was invertible, in other words, that there exists a FIR MIMO system that can reverse the process. It has been shown that, under certain assumptions, this inverse system exists. This was enunciated in the multichannel inverse theorem (MINT) in [Miyoshi and Kaneda, 1988]. So the problem can be stated as the search for a FIR MIMO system W such that

$$\mathbf{y}(t) = W * \mathbf{x}(t) \approx \mathbf{s}(t - \mathbf{d}) \,, \tag{2.8}$$

where $\mathbf{d} = [t_1, \ldots, t_N]$ is a delay vector, with different values of delay in each component. In the ideal case of perfect inversion and no indeterminacies, this inverse system would exactly reproduce the sources, delayed as a consequence of the cascaded filters.

Nevertheless, in this model there are also ambiguities that are produced in the same way than those for the linear instantaneous case. The global effect of the mixing and separation system expressed in the z-domain would be:

$$G(z) = W(z)H(z) = D(z)\,, \tag{2.9}$$

where $D(z)$ is a diagonal matrix with values $d_i = z^{-t_i}$, where t_i represents the delay produced in the i-th source by the mixing-separating system. But given the ambiguities already discussed for the instantaneous case, the best one can expect to obtain is

$$G(z) = P(z)\Lambda(z)D(z), \tag{2.10}$$

where $P(z)$ is a permutation matrix, $\Lambda(z)$ is a diagonal scaling matrix, and $D(z)$ is a diagonal arbitrary filtering matrix [Cichocki and Amari, 2002]. In the case of perfect invertibility of the system, $D(z)$ would contain only delay elements, but for a more realistic case it will have arbitrary filters. The meaning of this is that we can expect to find unsorted, arbitrarily scaled, and filtered versions of the sources, and in the best case, delayed versions of them.

2.4.2. Solution in the time-frequency domain

The idea of using a frequency domain approach is to exploit the convolution property of Fourier transform, that a convolution in time domain will become a product in the frequency domain. Given the time-varying characteristics of the speech and general sound sources, the proper tool to apply a frequency analysis is the short-time Fourier transform (STFT). Let $x(n)$ be a digital signal and $x(m,\tau) = \phi(m)x(m+\tau R)$ the windowed and time-shifted version of $x(n)$. The STFT $\mathcal{X}(\omega_k,\tau)$ is given by

$$\mathcal{X}(\omega_k,\tau) = \sum_{m=0}^{N-1} x(m,\tau)e^{-j\omega_k m}\,, \tag{2.11}$$

where $\omega_k = \frac{2\pi k}{N}$ is the discrete normalized frequency, with bin index $k = 0,\ldots,N-1$, frame index $\tau = 0,\ldots,L-1$, $\phi(n)$ is a Hanning window of length N and R is the frame shifting interval for the analysis. If a STFT is applied to (2.7), the mixture can be written as [Benesty et al., 2005, chapter 13]

$$\mathbf{x}\,(\omega,\tau) = H\,(\omega)\,\mathbf{s}\,(\omega,\tau)\ , \tag{2.12}$$

where the variable τ represents the time localization given by the sliding window in the STFT, and ω is the frequency. It should be noted that, as the mixing system was assumed to be LTI, the matrix $H(\omega)$ is not a function of the time. What is important

about this equation is that, for a fixed frequency ω_0, the equation can be written as an instantaneous linear mixture problem:

$$\mathbf{x}_{\omega_0}(\tau) = H_{\omega_0}\mathbf{s}_{\omega_0}(\tau) \ . \tag{2.13}$$

In this way, the problem is simplified from the solution of a very complicated convolutive ICA problem in the time domain, to a set of simpler instantaneous ICA subproblems, one for each frequency bin of the STFT. Then, any ICA algorithm capable of handling complex data can be used to solve each subproblem. For example, there exist the complex FastICA already discussed [Bingham and Hyvärinen, 2000], the complex version of JADE [Cardoso and Souloumiac, 1993], and some versions of complex Infomax algorithm using natural gradient [Sawada et al., 2004]. As this approach uses ICA but is formulated in the frequency domain, it is called frequency domain ICA (fd-ICA) or frequency domain BSS (fd-BSS) method in the literature. However, the term fd-BSS can lead to confusions, as the masking methods previously presented can also be termed fd-BSS. To avoid confusions, we will refer to this approach as fd-ICA from now on.

After separation the time-domain signals can be recovered by an inverse short-time Fourier transform (ISTFT). One way of performing this inverse transform is by the overlap-and-add method. For a given $\mathcal{Z}(\omega_k, \tau)$, the windowed and time-shifted inverse for each frame is obtained as

$$z(m,\tau) = \frac{1}{N}\sum_{k=0}^{N-1} \mathcal{Z}(\omega_k,\tau)e^{j\omega_k m} \ . \tag{2.14}$$

Using each reconstructed frame we form the sum of all them, correcting the time-shift

$$\begin{aligned} z^*(n) &= \sum_{\tau=0}^{L-1} z(n-\tau R,\tau) \\ &= \sum_{\tau=0}^{L-1} \phi(n-\tau R) z(n-\tau R+\tau R) \\ &= z(n)\sum_{\tau=0}^{L-1}\phi(n-\tau R) \\ &= z(n)\Phi(n) \end{aligned} \tag{2.15}$$

where $\Phi(n)$ denotes the shifted sum of the windows. From the first equation to the second, the windowed and delayed signal $z(n-\tau R,\tau)$ was replaced by the product of the delayed window and signal, in a similar way that it was used in equation (2.11). Using this, the signal can be reconstructed as $z(n) = z^*(n)/\Phi(n)$. For some windows,

with certain frame shifting intervals, $\Phi(n) = C$ where C is a constant (except at the start and the end of the signal), but for general windows and shifting intervals, $\Phi(n)$ will have important oscillations. As can be seen, these formulas are valid only for frame shifting intervals smaller than the window length (otherwise $\Phi(n) = 0$ for some ranges of n and the division will produce an indeterminacy). Also, if the window used tapers to zero, the beginning and ending part of $\Phi(n)$ will produce numerical problems. This is solved in practice by setting the first and last R values of Φ to the value of $\Phi(R+1)$. Although this produces attenuation of the signal at the boundaries, for speech signals this is not a problem because they usually start and end with silence.

Although the problem seems to be solved now, there are some negative aspects in this formulation. First of all, the ambiguities of the ICA solution affect negatively the approach. The problem is that, for each subproblem solved in each frequency bin, there will be an arbitrary permutation and scaling. The permutation in one bin will be completely independent of the permutation of another one, and thus the order of the extracted sources will differ for each frequency bin. But in order to be able to reconstruct the sources, all the frequency bins must belong to the same source, and so the permutations should be solved before reconstruction.

Also, the arbitrary scaling would introduce a filtering effect by changing the frequency gain in each bin. In this way, the permutation and scaling ambiguities are one of the main drawbacks of this separation method. Figure 2.7 illustrates these problems for a 2-by-2 mixture. As it can be seen, for the bin corresponding to frequency ω_1, the information of the two input STFT is used to solve a problem of ICA. The estimated sources s_1 and s_2 are affected by arbitrary scalings α_1 and β_1, and they are assigned to each of the separated sources. At the bin corresponding to ω_2, the process is repeated, but now the output signals are in reverse order, s_2 and s_1, and also affected by arbitrary scalings (different with respect to the previous ones). As there is no way to known whether the permutation occurred or not it is not possible to correct it, and the bins are wrongly assigned to the output STFT.

2.4.3. Solutions to the permutation problem

The permutations of the fd-ICA method constitute a quite difficult to solve problem. Basically there have been four approaches for its solution (also some variations or combinations of them). The first proposed method uses the temporal structure of the speech signals. As utterances occur, there are modifications in their energy that are quite common for all frequencies. That is, when there is a pause, the energy falls in all frequencies, and when there is an emphasized sound, the energy rises all along the frequency contents. In this way, there is some kind of amplitude modulation or envelope that should follow a similar pattern for all the frequency contents. On the other side, for two different signals, the amplitude modulation patterns should be different.

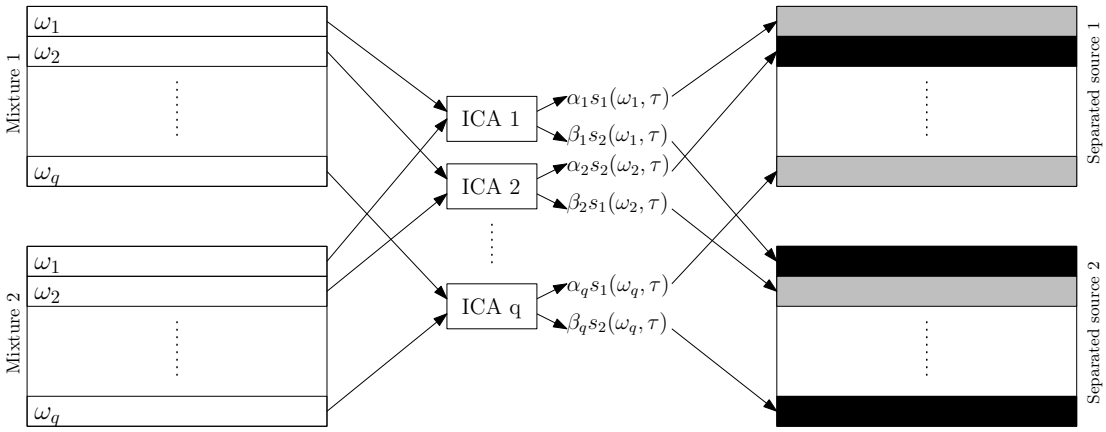


Figure 2.7. Illustration of the permutation and scaling ambiguity in fd-ICA. The extracted sources are obtained with arbitrary permutation respect to the other frequency bins, and with arbitrary scalings.

Therefore, a method for classification of frequency bins can be based on measuring the similitude of the envelope amplitude of a new bin to that of the previously classified bins. This method was first proposed in [Ikeda and Murata, 1998; Murata et al., 2001]. The algorithm is initialized with the first bin, and an envelope in each bin is estimated with this arbitrary permutation. The next bin is classified according to the higher correlation among its envelope and the previous ones, then its envelope is combined with the previous to form a kind of global average of the envelope for that class. The algorithm procedes in that way for all bins.

The second method is based on directions of arrival, assuming an anechoic model of transmission. The estimated mixing matrix for each bin can be interpreted as a set of null beamformers, each one rejecting one of the sources. Then, each column contains information regarding the direction of arrival (DOA) of one of the sources (coded in the null location for that source). The algorithm proposed in [Ikram and Morgan, 2002] clusters that information and classifies the bins according to the distance of each DOA to the cluster centers. Sawada et al. [2004] proposed a method that combines the correlation approach with the DOA approach to improve the results.

The third method includes imposing constrains in the time domain. This method was introduced in [Parra and Spence, 2000]. They observed that the existence of permutations produces longer filters in the separation system (that is, the filters have more nonzero elements in the time domain). If a constrain in the length of the filter is imposed in the time domain during the learning, this forces a coupling of the separation matrices at each frequency bins, and eliminates the permutations. The filters generated by the separation matrix are, in each iteration, transformed back to time domain, and

the last k coefficients are made zero, for a properly chosen value of k. Then these filters are transformed back to the frequency domain and the iterations proceed until convergence. The main drawback of this method is that the time domain filters need to be long enough to capture the reverberation characteristics. Also, the filters need more length to impose the time-domain constrains, and therefore a larger window is needed in the STFT processing. This implies a larger computational cost. Furthermore, in each iteration there is the need to transform the filters to the time domain and back to impose the length constrains, which adds to the computational cost.

Finally, recently some methods have been proposed that produce an optimization of all the separation matrices in a coupled way, using a multivariate pdf defined on all the time-frequency plane. This effectively couples all the separation and no permutations are produced, although the complexity of the approach is quite high [Lee et al., 2007].

2.4.4. Solutions to the scaling problem

Regarding the scaling problem, there are some proposed approaches. One is based in the minimum distortion principle [Matsuoka, 2002], and is basically the used in [Murata et al., 2001]. In this case, after separation, one of the signals is set at zero, and then the mixture system is applied. Suppose that for a 2-by-2 case, in the bin corresponding to frequency ω_r the separation matrix $W(\omega_r)$ was found, and its inverse is the estimated mixing matrix $H(\omega_r)$. For this bin, the separated estimation of sources are $\mathbf{Y}(\omega_r, \tau) = [Y_1(\omega_r, \tau), Y_2(\omega_r, \tau)]^T$. Let us form an auxiliary vector $\mathbf{Y}^1 = [Y_1(\omega_r, \tau), 0]^T$. Applying the estimated mixing matrix to this vector we obtain

$$\mathbf{Z}(\omega_r, \tau) = H(\omega_r)\mathbf{Y}^1(\omega_r, \tau) = [h_{11}(\omega_r)Y_1(\omega_r, \tau), h_{21}(\omega_r)Y_1(\omega_r, \tau)]^T . \tag{2.16}$$

In this way, the obtained signals can be interpreted as the effect of the separated source in each of the microphones, and the scalings h_{ij} compensate the arbitrary scalings introduced in the estimated separation matrix, thus eliminating the scaling ambiguity. It should be noted that the method will not reduce the reverberation effect, because it will produce an estimation of the sources as arrived at the microphones, after the coloration effect of reverberation.

For the case of the permutation correction with the method using DOA, the principle used is based on the estimated beampatterns. As the separation matrix consists of a set of null beamformers, and after the permutation correction the DOAs are known, the method consists in estimating the beamformer gain at each DOA for the beampattern, and scaling the source to have a unitary gain in the beampattern in its direction [Sawada et al., 2004]. For the other two methods of permutation correction, there is no scaling ambiguity, and so it is not necessary to correct it.

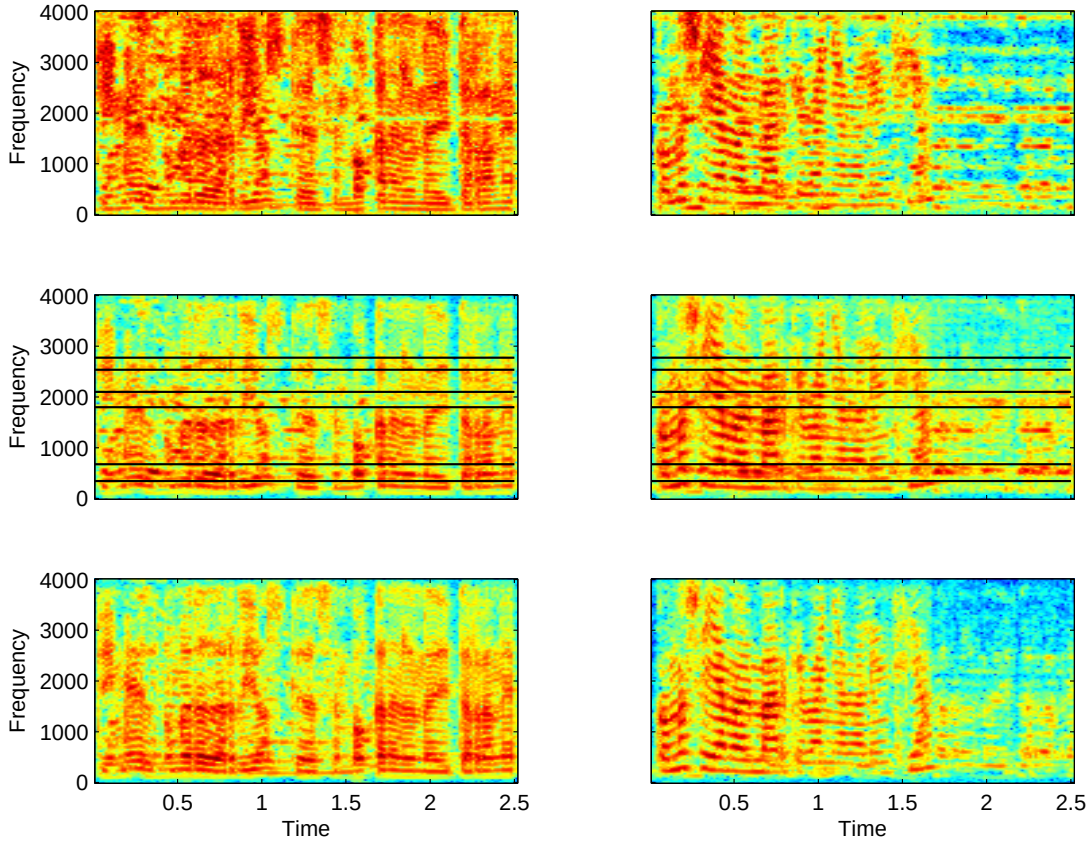


Figure 2.8. Example of the permutation and scaling ambiguities and their correction.

To illustrate on the importance of these problems, Figure 2.8 shows an example of the effect of the scaling and permutation ambiguities for a 2-by-2 case. In the top row, the two separated signals before correction of the scaling and permutation problems are presented. It is clear that some areas have unusual amplitudes compared to their neighbors. In the middle row, the scaling problem has been corrected (but not the permutation problem) by using the minimal distortion principle. The marked areas show clear cases of permutation missalignment. The last row is the case where permutations have been solved by the correlation method. It seems that the problems were solved, and the STFT representation presents consistent permutations and scalings in all bins, although for the low energy bins it is difficult to say whether there are some residual permutations or not.

2.4.5. Main advantages and drawbacks of the fd-ICA method

The main advantage of the fd-ICA method is its simplicity. Instead of a very complex problem, it reduces the solution to several simpler subproblems. This has one important advantage. When a problem is too complex, the solutions are usually harder to find (one example of this is the curse of dimensionality usually seen in pattern recognition techniques). In this way, it can be expected that the solutions of the simpler problems will provide for a higher quality solution of the whole problem. Also, it produces solutions that are practical from the computational complexity point of view.

There are, however, some important drawbacks that were clearly explained in [Araki et al., 2003]. We have implemented a standard method for fd-ICA, using FastICA as the main separation principle and the correlation-based approach to solve the permutation problem, with the minimal distortion principle to solve the amplitude ambiguity [Handa et al., 2004; Tsutsumi et al., 2004]. We evaluated the performance of this method for different mixture conditions and under reverberation. This study allowed us to identify the main drawbacks of the method, which are summarized as follows.

In first place, there exist ambiguities produced by ICA, that cause suboptimal solutions when cannot be successfully corrected, as already explained. In second place, the methods of ICA used in each bin need a large amount of data to be correctly estimated. In the STFT analysis, the most important parameters for the transformation are the window size and the step size for the window. Most of the ICA methods use an assumption of random iid process, but if the step size is too small, the bins in the STFT will not fulfill this hypothesis. Then a larger step is needed. However, for a given signal length, a larger step implies less points in the time dimension of the STFT, which means less data to produce a reliable estimation of the separation parameters.

A third aspect that degrades the separation quality is the initialization of the algorithms. Different initial conditions lead to different solutions, and some algorithms are very sensitive to initial conditions. In this way, good methods for initialization are required. This aspect is related also to the convergence quality of the separation in all bins. There is nothing that guarantees that the convergence in all bins will produce an equivalent quality of separation for all of them. In fact, in the above cited methods, there are usually several bins where the ICA algorithms fail to converge. It would be desirable to have an equivalent separation quality for all frequencies. But this is difficult to produce, particularly because the signals are not homogeneous in their statistical properties along different frequencies. For example, it is possible that in a particular frequency bin, one of the signals had not enough energy, and thus for that frequency there is only one effective signal. Therefore, as the approach assumes that all signals are present for all bins, it will produce an erroneous estimation.

Finally, the reverberation is a degrading aspect, which affects the algorithm in several ways. First, the smearing effect reduces the applicability of the iid hypothesis. Second, as the reverberation is increased, larger windows are required to properly capture its effects. However, at the same time, larger step sizes are required to improve the validity of the iid hypothesis, and all this produces a reduction of the data available for each ICA subproblem (which affects the convergence quality). This degrades the quality of separation, although the fd-ICA approach has shown the capability to separate signals for a wider range of reverberant conditions than the other methods. And third, other aspect by which reverberation affects the separation quality is related to the directivity patterns generated by the separation matrices. As clearly shown in [Araki et al., 2003] for a 2-by-2 case, the method can be interpreted as a pair of null beamformers, where each beamformer rejects the sound coming from one specific direction. This means that the spatial filtering implemented by them can only reject the direct sound but not the reverberant echoes arriving from other directions. Thus, when the reverberation time is increased the performance will be degraded. The beampatterns obtained for reverberant cases show two degrading aspects. In the first place, the direction of the null is not exactly the same for all the frequencies, which shows that the convergence is not uniform. And in second place, the deep of the null also varies with the frequency, usually with a reduction of the attenuation at low frequencies. These two effects of reverberation reduce the separation quality.

To sum up, the following drawbacks are important limiting aspects of the fd-ICA approach for real applications:

- The permutation and scaling ambiguities.
- The large amount of data needed for proper convergence of the ICA algorithm in each bin.
- The need for proper initialization.
- The lack of uniform performance of the ICA algorithms for all bins.
- The degrading effects of reverberation, related to:
  - Reduction of the iid hypothesis validity.
  - The compromise of window length/amount of data needed in each frequency bin.
  - The capability of the method to reject only the direct sound but not the reflections.

2.5. Concluding remarks

In this chapter, a brief review of the main approaches for convolutive audio sources separation was presented. The beamformer methods are the simpler ones, but require a knowledge of the mixture conditions, which is not available in most cases. The sparsity and time domain ICA approaches are not appropriate as they cannot handle reliably even small reverberation times. The fd-ICA methods can be interpreted as sets of blind null beamformers, and thus are an improvement respect to static beamformers. They have shown to be able to work for a wider range of reverberation conditions, but there are several aspects that degrade their performance. Nevertheless, as fd-ICA is the most successfull approach up to now, this doctoral dissertation will be focused on producing methods and alternatives to overcame the limitations of the state-of-the-art methods presented here.

Chapter 3

Performance evaluation of BSS algorithms

Quality evaluation for speech processing is a very important step in the development of advanced algorithms. This is specially important for the field of blind source separation of speech, which has emerged in the last years. In particular, the main interest in this research is the quality evaluation of BSS algorithms for the specific task of automatic speech recognition. In this context, it is desirable to have some objective way of quantify the quality improvement produced by separation algorithms, that is, to have some objective quality measure.

The objective of this chapter is to find objective quality measures that correlate well with automatic speech recognition rate when using blind source separation as a mean to introduce robustness into the recognizer. To accomplish this objective, some set of potentially good measures need to be selected. We will first present a brief review of methods of quality evaluation that have been successfully used both for BSS and for other areas related to speech processing. Based on that review, a set of candidate measures will be selected. Then an experiment to evaluate the amount of correlation of each quality measurement with the recognition rate of an ASR system will be presented. The chapter will end with the description of the experimental protocol to be used in the evaluation of the developed separation methods. This protocol will be applied in the following chapters of this work.

3.1. Brief review of quality evaluation

The blind source separation task has gained importance in the last years, with a large presence in the most important conferences and journals in the signal processing, machine learning and computer sciences areas. The increase in the number of papers

published shows, however, a bias towards the presentation of new separation algorithms, but only a few of them discusse effective ways to evaluate the quality of the algorithms [Schobben et al., 1999]. As an example, the work by Vincent et al. [2006], where several measures are specifically designed to take into account the different factors that can affect the result of BSS algorithms can be cited. As clearly stated in [Divenyi, 2004, chapter 20], the evaluation is a very complex task, and it must be matched to the application for which the algorithm is produced.

As the application of our interest is ASR, in this case the ultimate measure of quality would be the output of a speech recognition system. Nevertheless, there are two factors that make the evaluation using a recognizer undesirable. First of all, the method will have very little sensitivity with respect to the parameters of the algorithm. This means that perhaps a considerable improvement in quality would be needed before obtaining a difference in the recognizer output, and so two BSS algorithms would get the same recognition rate even if there are differences in quality between them. The second aspect is that obtaining statistically significant results from the output of a speech recognizer would imply the use of a large number of signals, making this impractical at least in the first stages of the research. These problems are very similar to the ones that are present in the case of subjective evaluation tests.

In this way, it is desirable to have some objective quality measure, that can provide a better precision in the comparisson of separation algorithms without the mentioned problems. One of the possible practical applications of quality measures in the field of BSS for ASR is in algorithm selection/tuning. In early research stages, where a particular separation algorithm and its parameters should be selected, direct evaluation by means of recognition rate would be prohibitive as a result of the large amount of tests over complete databases. The alternative is to use one of the objective quality measures to select some candidate algorithms and their parameters, and then perform a fine tuning with the complete ASR system.

Despite the importance of the subject, as mentioned previously there are not many research works in the area of quality evaluation of BSS algorithms, and there is not a standard methodology to evaluate it in the general case, and this is worst in relation to ASR, in which the only option used is the evaluation with an ASR system. In the following sections a review of the main approaches used in other areas will be presented. This will allow for a selection of the quality measures candidates to study.

3.1.1. Quality evaluation for BSS

In the particular case of evaluating BSS algorithms, many different alternatives have been used, generally derived from other areas of signal processing. Those methods can be classified into two main areas: *subjective assessment*, where some appreciation is used regarding either subjective perceived quality of resulting sound [Ikeda and Murata,

1998; Mitianoudis and Davies, 2001], or visual differences between waveforms of separated signal and original ones [Gotanda et al., 2003; Ikeda and Murata, 1998; Lee et al., 1997], or visual differences of spectrograms of separated signals and original ones [Mitianoudis and Davies, 2001]; and *objective evaluation*, where some numerical quantity directly associated to separation quality is used, permitting an objective comparison among different algorithms. Subjective tests are complex and difficult to reproduce, and moreover, they must be performed very carefully if one wants to be able to compare the results of different algorithms. Thus we will focus on objective assessment using objective quality measures.

Regarding objective measures that have been applied to BSS problem, they can be divided into three kinds:

1. Measures that require knowledge about transmission channels: These measures use information about impulse responses between each sound source and each microphone, or require knowledge of individual signals arriving at each microphone. These kinds of measures are hard to apply to realistic environments as they depend on factors that may vary from one experiment to another. Among them, it can be mentioned: Multichannel inter symbol interference (MISI) [Cichocki and Amari, 2002]; Signal to interference ratio (SIR) [Douglas and Sun, 2003; Parra and Alvino, 2002; Parra and Spence, 2000] and Distortion - Separation [Schobben et al., 1999].

2. Measures that use information about the sound sources: In this case some measures of discrepancy between the separated signal and the original source signal are used. One drawback of these measures is that, by comparing with original sources, the algorithms that perform separation but not reverberation reduction will yield poorer results as the resulting signal will be always distorted, even for perfect separation. Some measures of this kind commonly used in BSS are: Total relative distortion (TRD) proposed in [Gribonval et al., 2003; Vincent et al., 2006], and segmental Signal to noise ratio (segSNR) [Douglas and Sun, 2003; Murata et al., 2001].

3. Indirect measures: In this case the processed signal is used as input to another system with which the result can be evaluated in an objective way. The most typical example of this is an ASR system with which evaluation is made on recognition rate obtained after separation [Asano et al., 2001, 2003].

All of these show an important lack of experimentation in the area of quality evaluation for algorithms of BSS in realistic environments. Problems for quality measure proposals came mainly from two aspects that must be taken into account for a correct evaluation of such algorithms: *reality level* required in the experiments, which is

necessary for the results to be directly extrapolated to practical situations, and *task complexity*, since for example, some BSS algorithms search only for separated signals while others try to eliminate reverberation effects too. These aspects also need to be considered carefully for choosing a suitable kind of evaluation.

3.1.2. Quality measures applied to other areas

In applications where the final result will be listened by humans, the ideal way for quality assessment is by means of subjective evaluation of perceptual quality [Deller et al., 1993]. Many standardized tests allow the evaluation with subjective measures. For example, the composite acceptability (CA) of diagnostic acceptability measure (DAM) [Voiers, 1977] consists of a parametric test where the listener has to evaluate acceptability of sound based on 16 categories of quality. Other widely used subjective measure is the mean opinion score (MOS), a measure where each subject has to evaluate the perceived quality in a scale of 1 to 5. This kind of tests has high cost both in time and resources.

Several objective quality measures have been proposed to overcome this drawback [Gray and Markel, 1976; Gray et al., 1980; Klatt, 1982]. In [Barnwell, 1980] the correlation between a large number of objective measures and the subjective measure CA-DAM was studied evaluating it under several speech alterations, contamination noises, filters and coding algorithms. In that work, weighted spectral slope measure (WSS) was the best predictor of subjective quality.

In the last years, some objective measures that use perceptual models have been introduced. The first widely adopted one was perceptual speech quality measure (PSQM) [Beerends and Stemerdink, 1994] and more recently, the audio distance (AD) based on measuring normalizing blocks (MNB) was proposed [Voran, 1999a]. This measure presents a correlation with MOS of more than 0.9 for a wide variety of coding algorithms and languages [Voran, 1999b].

Of particular interest is the perceptual evaluation of speech quality (PESQ) measure. This measure is defined in the standard ITU P.862 as the mean for evaluating the quality of speech transmitted over communication channels [ITU, 2001]. It has been widely studied and a very high correlation with subjective quality for a wide variety of transmission channels, distortions, codification algorithms, and languages has been shown. Also, a recent study [Hu and Loizou, 2008] has shown its good properties for evaluating speech enhancement algorithms in terms of subjective perceptual quality.

3.1.3. Quality measures in ASR

Regarding ASR, objective quality measures have been used at two different levels. In template-based isolated word recognizers, a measure of distance between the test

Table 3.1. Performance of quality measures for different tasks (see full explanation in text). Second column: $|\rho|$ DAM, correlation as predictor of CA-DAM. Third column: WER%, word error rate in isolated word recognition. Fourth column: r_{err}, prediction error as predictor of recognition rate in robust continuous ASR.

Measure	$\|\rho\|$ DAM	WER%	r_{err}
segSNR	0.77	–	4.71
IS	–	11.35	0.41
LAR	0.62	–	0.88
LLR	0.59	8.45	–
LR	–	8.63	–
WLR	–	9.15	–
WSS	0.74	8.45	1.72
CD	–	8.88	–
LSD	0.60	–	–

signal and the stored templates is needed [Rabiner and Juang, 1993]. Several objective measures originally proposed for speech enhancement or coding have been successfully used within this context [Itakura, 1975; Mansour and Juang, 1989; N. Nocerino et al., 1985]. On the other hand, the capability of those measures to predict recognition rate of a robust speech recognizer has been studied [Hansen and Arslan, 1995].

As a summary of the background available for our work, Table 3.1 shows a comparison of results obtained by various researchers in different tasks[1]. First column lists the objective quality measures used [Basseville, 1989; Gray and Markel, 1976; Gray et al., 1980; Itakura, 1975; Klatt, 1982]:

- Segmental signal to noise ratio (segSNR).
- Itakura-Saito distance (IS).
- Log-area ratio (LAR).
- Log-likelihood ratio or Itakura distance (LLR).
- Likelihood ratio (LR).
- Weighted likelihood ratio (WLR).

[1] As the application fields and contexts are different respect to this work, these results are not directly applicable to our research, but can give some cues on potentially interesting measures to evaluate.

- Weighted spectral slope (WSS).
- Cepstral distortion (CD).
- Log-spectral distortion (LSD).

The second column presents the absolute value of correlation coefficient between quality measures and subjective test CA-DAM [Deller et al., 1993], evaluated over a set of speech modified with 322 different distortions. It should be noted that correlation for segSNR was calculated using only a subset of 66 distortions produced by waveform coders, for whom a measure based in waveform similarity has sense. Excluding this case, a high correlation between subjective quality level and WSS can be noted.

The third column presents the percentage of word error rate (WER%) for an isolated word recognition system, in which the measures where applied as selection criteria for classification, for a set of 39 words of a telephone recording database [N. Nocerino et al., 1985]. A good performance for recognizers based on LLR and WSS measures can be noted.

Finally, the fourth column shows the performance of measures as predictors of recognition rate in a continuous speech recognition system using a robust set of features, on speech contaminated with additive noise [Hansen and Arslan, 1995]. The presented value (r_{err}) is the mean squared prediction error, averaged over all sentences in the database of processed speech. In this case the best performance is obtained by LAR and IS measures.

With respect to perceptually designed measures, in [Sun et al., 2004] the PESQ measure was evaluated to predict recognition rate of an ASR system, showing good results. In that work, however, the correlation is evaluated with respect to the noisy speech (with additive and convolutive noise), without using any separation algorithm. On the contrary, in [Yamada et al., 2006] the performance of three measures as predictors of recognition rate of an ASR system for speech with additive noise, after the application of single channel speech enhancement algorithms, is evaluated. PESQ is shown to have the higher correlation with recognition rate. Finally, in [Fox et al., 2007], the perceptual evaluation of audio quality (PEAQ) measure, which is a perceptual measure similar to PESQ but designed for general wideband audio sources and not only for speech, is evaluated as predictor of subjective quality. The authors used music sources contaminated by convolutive noise and enhanced by BSS algorithms. The evaluation was done with respect to human listeners and subjective scores, comparing PEAQ with several other objective quality measures.

3.2. Selected measures

We will focus our attention on objective measures that make use of sound source information. This kind of measures attempt to evaluate some "distance" or "distortion" of separated signal with respect to original signal and have been selected for three reasons. First, by using this approach experiments can be performed with mixtures recorded in real rooms (this gives the experiment a high level of realism) and there is no need to know any information about transmission channels between sources and sensors. Second, as the sources must be available, the experiments could be extended to other mixing conditions. Third, as in general the ASR systems are trained with clean speech, using a method that permits to compare algorithm output with the clean signals is very reasonable, particularly if the ASR system was not constructed to be noise-robust.

Based on the analysis presented in Section 3.1 of previous works, a set of 11 single objective quality measures, plus several composite measures, was selected for this study. They are explained in detail in the next subsections. The following notation will be used: let the original signal be $\mathbf{s}$ and separated signal $\widehat{\mathbf{s}}$, both of M samples. Frame m of length N of original signal is defined as $\mathbf{s}_m = [s[mQ], \ldots, s[mQ + N - 1]]$, where Q is the step size of the window in a short-time analysis, and with analogous definition for corresponding frame of the separated signal. In the case of measures derived from linear prediction, a system order P is assumed.

3.2.1. Segmental signal to noise ratio (segSNR)

Given a frame of original signal and corresponding frame of separated signal, segSNR is defined as [Deller et al., 1993]:

$$d_{SNR}(\mathbf{s}_m, \widehat{\mathbf{s}}_m) = 10 \log_{10} \frac{\|\mathbf{s}_m\|^2}{\|\widehat{\mathbf{s}}_m - \mathbf{s}_m\|^2} , \tag{3.1}$$

where $\|\cdot\|$ is the 2-norm defined as usual, $\|\mathbf{x}\| = \left(\sum_{n=1}^{N} x[n]^2\right)^{1/2}$.

This measure is included because it is widely used due to its simplicity. Besides this, it has been used in the context of BSS to evaluate separation algorithms, as mentioned in Section 3.1 [Deller et al., 1993].

3.2.2. Itakura-Saito distortion (IS)

Given LP coefficients vector of original (clean) signal, $\mathbf{a_m}$, and LP coefficient vector for the corresponding frame of separated signal, $\widehat{\mathbf{a}}_\mathbf{m}$, IS distortion is defined as [Hansen

and Pellom, 1998; Rabiner and Juang, 1993]:

$$d_{IS}(\mathbf{a}_m, \widehat{\mathbf{a}}_m) = \frac{\sigma_m^2}{\widehat{\sigma}_m^2} \frac{\widehat{\mathbf{a}}_m^T \mathbf{R} \widehat{\mathbf{a}}_m}{\mathbf{a}_m^T \mathbf{R} \mathbf{a}_m} + \log\left(\frac{\widehat{\sigma}_m^2}{\sigma_m^2}\right) - 1 \,, \tag{3.2}$$

where $\mathbf{R}$ is the autocorrelation matrix, and σ^2, $\widehat{\sigma}^2$ are the all-pole system gains.

This measure is derived from linear prediction (LP) analysis [Hansen and Pellom, 1998; Rabiner and Juang, 1993]. Its good performance as predictor of recognition rate for signals with additive noise in continuous speech recognition systems makes this measure a good candidate for the present research.

3.2.3. Log-area ratio distortion (LAR)

It is also derived from LP coefficients. Given reflection coefficient vector for an LP model of a signal, $\mathbf{k}_m = [\kappa(1;m), \ldots, \kappa(P;m)]^T$, the Area Ratio vector is defined as $\mathbf{g}_m = [g(1;m), \ldots, g(P;m)]^T$, where $g(l;m) = \frac{1+\kappa(l;m)}{1-\kappa(l;m)}$. These coefficients are related to the transversal areas of a variable section tubular model for the vocal tract. Using these coefficients, for a frame of original signal, and corresponding frame of separated signal, LAR distortion is defined as [Deller et al., 1993; Hansen and Pellom, 1998]:

$$d_{LAR}(\mathbf{g}_m, \widehat{\mathbf{g}}_m) = \left\{ \frac{1}{P} \left\| \log \mathbf{g}_m - \log \widehat{\mathbf{g}}_m \right\|^2 \right\}^{\frac{1}{2}} \,. \tag{3.3}$$

where the log is applied to each element of the vectors. This measure has been selected given its good performance as predictor of recognition rate in continuous speech recognition systems, as it can be seen in Table 3.1.

3.2.4. Log-likelihood ratio distortion (LLR)

This measure is calculated similarly to IS distortion. Given LP coefficient vector of a frame of original and separated signal, $\mathbf{a_m}$ and $\widehat{\mathbf{a}}_\mathbf{m}$ respectively, LLR distortion is given by [Hansen and Pellom, 1998; Itakura, 1975]:

$$d_{LLR}(\mathbf{a}_m, \widehat{\mathbf{a}}_m) = \log \frac{\widehat{\mathbf{a}}_m^T \mathbf{R} \widehat{\mathbf{a}}_m}{\mathbf{a}_m^T \mathbf{R} \mathbf{a}_m} \,, \tag{3.4}$$

where $\mathbf{R}$ is the autocorrelation matrix.

Its good performance as a dissimilarity measure in isolated word recognition systems makes interesting its application in the context of this research.

3.2.5. Weighted spectral slope distortion (WSS)

Given a frame of signal, the spectral slope is defined as $SL[l;m] = S[l+1;m] - S[l;m]$, where $S[l;m]$ is a spectral representation (in dB), obtained from a filter bank using B critical bands in Bark scale (with index l referring to position of filter in filter bank). Using this, WSS between original signal and separated one is defined as [Hansen and Pellom, 1998; Klatt, 1982]:

$$\begin{aligned} d_{WSS}(\mathbf{s}_m, \widehat{\mathbf{s}}_m) &= K_{spl}(K - \widehat{K}) + \\ &\sum_{l=1}^{B} \bar{w}[l] \left(SL[l;m] - \widehat{SL}[l;m] \right)^2 , \end{aligned} \tag{3.5}$$

where K_{spl} is a constant weighting global sound pressure level, K and $\widehat{K}$ are sound pressure level in dB, and weights $w[l]$ are related to the proximity of band l to a local maximum (formant) and global maximum of spectrum, as $\bar{w}[l] = (w[l] + \widehat{w}[l])/2$, with:

$$w[l] = \left(\frac{C_{loc}}{C_{loc} + \Delta_{loc}[l]} \right) \left(\frac{C_{glob}}{C_{glob} + \Delta_{glob}[l]} \right) , \tag{3.6}$$

with a similar definition for $\widehat{w}[l]$, where C_{glob} and C_{loc} are constants and Δ_{glob}, Δ_{loc} are the log spectral differences between the energy in band l and the global or nearest local maximum, respectively. This weighting will have larger value at spectral peaks, especially at the global maximum, and so it will give more importance to distances in spectral slopes near formant peaks (for more details, see [Klatt, 1982; N. Nocerino et al., 1985; Rabiner and Juang, 1993]).

This measure is mainly related to differences in formant locations [Klatt, 1982], and was selected because of its relative good performance in all cases presented in Table 3.1.

3.2.6. Total relative distortion (TRD)

It is based on an orthogonal projection of the separated signal on the original signal [Vincent et al., 2006]. The separated source can be decomposed as $\widehat{\mathbf{s}} = \mathbf{s}^D + \mathbf{e}^I + \mathbf{e}^N + \mathbf{e}^A$, where $\mathbf{s}^D = \langle \widehat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s} / \|\mathbf{s}\|^2$ is the part of $\widehat{\mathbf{s}}$ perceived as coming from the desired source, and $\mathbf{e}^I$, $\mathbf{e}^N$ and $\mathbf{e}^A$ the error parts coming from the other sources, sensors noises and artifacts of the algorithm. For each frame m of these components, TRD is defined as [Gribonval et al., 2003; Vincent et al., 2006]:

$$d_{TRD}(\mathbf{s}, \widehat{\mathbf{s}}; m) = \frac{\left\| \mathbf{e}^I_m + \mathbf{e}^N_m + \mathbf{e}^A_m \right\|^2}{\left\| \mathbf{s}^D_m \right\|^2} . \tag{3.7}$$

As this measure is specific for performance evaluation of BSS algorithms, it was considered appropriate to include it in this work.

3.2.7. Cepstral distortion (CD)

This measure is also known as truncated cepstral distance [Gray and Markel, 1976]. Given the vectors of cepstral coefficients $\mathbf{c}_m$ and $\widehat{\mathbf{c}}_m$, corresponding to a frame of original signal and corresponding separation result, CD for the first L coefficients is defined as [Rabiner and Juang, 1993]:

$$d_{CD}(\mathbf{s}_m, \widehat{\mathbf{s}}_m) = \sum_{l=1}^{L} (c_m[l] - \widehat{c}_m[l])^2 \ . \tag{3.8}$$

As ASR systems for continuous speech make use of cepstral-based feature vectors, it is reasonable to include some measures using distances calculated in the cepstral domain.

3.2.8. Mel cepstral distortion (MCD)

This measure is calculated in a similar way as CD, but the energy output of a filter bank in mel scale is used instead of spectrum of signals. Given mel cepstral coefficients $\mathbf{c}_m^{mel}$ and $\widehat{\mathbf{c}}_m^{mel}$ corresponding to original and resulting separated signal respectively, calculated using a filter bank of B filters in mel scale, MCD for the first L coefficients is defined as [Deller et al., 1993; Rabiner and Juang, 1993]:

$$d_{MCD}(\mathbf{s}_m, \widehat{\mathbf{s}}_m) = \sum_{l=1}^{L} \left(c_m^{mel}[l] - \widehat{c}_m^{mel}[l] \right)^2 \ . \tag{3.9}$$

Also, as many ASR systems use mel cepstral coefficients, it is reasonable to use a distance measure based on them as a predictor of recognition rate.

3.2.9. Measuring normalizing blocks (MNB)

This technique applies a simple model of auditory processing but then evaluates the distortion at multiple time and frequency scales with a more sophisticated judgment model [Voran, 1999a]. This measure is more complex than the previous ones, so it will be only outlined here. It includes first a time-frequency representation, that is transformed to Bark scale to obtain a representation more closed to the auditory mapping. After this transformation the auditory time-frequency representations of the reference $S(t, f)$ and test $\widehat{S}(t, f)$ are analyzed by a hierarchical decomposition of measuring normalizing blocks in time (tMNB) and frequency (fMNB). Each MNB produces a series of measures and a normalized output $\widehat{S}'(f, t)$. For a tMNB, the normalization

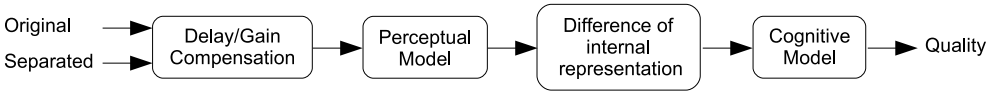


Figure 3.1. Scheme of the stages in PESQ calculation.

is done by:

$$\begin{aligned} e(t,f_0) &= \frac{1}{\Delta f}\int_{f_0}^{f_0+\Delta f}\widehat{S}(t,f)df - \frac{1}{\Delta f}\int_{f_0}^{f_0+\Delta f}S(t,f)df\ , \\ \widehat{S}'(f,t) &= \widehat{S}(t,f) - e(t,f_0)\ , \end{aligned} \tag{3.10}$$

where f_0 and Δf define a frequency band for the integration. By integration of $e(t,f_0)$ over time intervals, a group of measures for this tMNB is obtained. The same is used for each fMNB, with the roles of t and f interchanged. So, the hierarchical decomposition proceeds from larger to smaller scales, for frequency and time, calculating distances and removing the information of each scale. After this process, a vector of measures $\boldsymbol{\mu}$ is obtained. Then a global auditory distance (AD) is built by using appropriate weights $AD = \sum_{i=1}^{J} w_i\mu_i$. Finally, a logistic map is applied to compress the measure and adjust it to a finite interval, given by $L(AD) = \frac{1}{1+e^{aAD+b}}$. The authors have proposed two different hierarchical decompositions, called structure 1 and 2, that use different tMNB and fMNB decompositions. For more details, refer to [Voran, 1999a].

It is a modern approach including perceptual information and, due to his high correlation with MOS scores, was considered as a good candidate for this study.

3.2.10. Perceptual evaluation of speech quality (PESQ)

This measure uses several levels of analysis in an attempt to mimic the human perception. Figure 3.1 presents a block diagram of PESQ calculation. The first stage is a gain/delay compensation. The gain for both signals is adjusted to produce a constant prescribed power after bandpass filtering. The delay compensation follows at two levels. First at a general level, cross-correlation of envelopes is used to estimate a general delay between the original and the processed (separated) signal. Then a finer correlation/histogram-based algorithm is used to estimate delays for individual utterances. This produce a time-varying delay estimation for segments of the sentences.

The second stage is a transformation to a perceptual domain. This is made by a short-time Fourier transform, followed by a band integration using a Bark scale filterbank, to get a warped pitch-scaled representation. Then, a time-variant gain compensation is applied. The power densities of the original and the separated speech are transformed to the Sone loudness scale using the Zwicker's law.

In the third stage, the disturbance density is calculated by difference of the distorted and reference loudness density.

For the final stage, some cognitive models are applied. The disturbance density $D(f,t)$ is thresholded to account for masking thresholds of the auditory system. Also a second disturbance distribution is calculated, the asymmetrical disturbance density $DA(f,t)$, to take into account that some speech codec introduce time-frequency information in places were it was not present. To account for this, an asymmetry factor is calculated, that has a large value if the separated signal is markedly different than the original one. The asymmetrical disturbance density is calculated as the product of the disturbance density with the asymmetry factor.

Both densities are integrated over frequencies using two different L_p norms.They are aggregated in segments of 20 frames using an L_6 norm and after that, again aggregated for the whole signal using an L_2 norm, thus producing two values, one for the disturbance D and other for the asymmetrical disturbance DA. The final score is calculated as $PESQ = 4.5 - \alpha D - \beta DA$, with $\alpha = 0.1$ and $\beta = 0.0309$. This produces a value that is between -0.5 and 4.5 and is called raw PESQ score [ITU, 2001].

The correlation of this measure with the subjective perceived quality measured using a MOS scale was evaluated over a wide range of distortions and speech codecs and in different languages, yielding correlation values larger than 0.90 in most cases [Rix et al., 2001].

This measure uses a complex cognitive model to evaluate different types of distortions that have importance at the perceptual level. The model also includes a compensation for time-varying delays and time-varying scalings [ITU, 2001]. This is a very successful measure in several areas related to speech, and thus is a good candidate for the evaluation.

3.2.11. Nonlinear PESQ (NLP)

In several works, a nonlinear mapping is used to improve the correlation of the PESQ with the target measure. The standard ITU P.862 [ITU, 2001] recommends a logistic function that maps the raw PESQ scores to another measure that correlates very well with subjective quality (as measured by MOS tests). Motivated by this, we propose here to use a nonlinear mapping of the form

$$NLP = \alpha_1 + \frac{\alpha_2}{1 + \exp^{(\alpha_3 PESQ + \alpha_4)}} , \tag{3.11}$$

where the coefficient vector $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3, \alpha_4]$ is adjusted from the data to yield maximum correlation. This adjustment can be obtained by nonlinear regression methods.

3.2.12. Composite measures

To improve the correlation even more, a linear combination of several measures can be constructed. This composite measure is defined as [Hu and Loizou, 2008]:

$$CM_K = \omega_0 + \sum_{i=1}^{K} \omega_i M_i, \tag{3.12}$$

where $\mathcal{S}_{CM_K} = \{M_i\}$ for $1 \leq i \leq K$ denotes the set of measures that will be used for the linear combination, K denotes the number of measures combined, and ω_i are coefficients adjusted to maximize the correlation. The values of the coefficients can be obtained by multivariate linear regression.

3.2.13. Parameters for the measures

With the exceptions of MNB, PESQ and NLP, all the selected measures are frame based, therefore each of them yields a vector. However, for the evaluation a unique value for each sentence is needed. To achieve this, median value has been employed, as suggested in [Hansen and Pellom, 1998], because in general the measures are affected by outliers corresponding to silence segments at the beginning and the end of each original sentence (in which only noise is observed).

For the analysis, the following parameters were used:

- Frame length $N = 512$ samples (32 ms of signal).
- Step size for analysis window $Q = 128$ samples (8 ms of signal).
- Order for LP models $P = 10$.
- WSS: $B = 36$, $K_{spl} = 0$, $C_{loc} = 1$ and $C_{glob} = 20$ as recommended by author in [Klatt, 1982].
- CD: truncation at $L = 50$ coefficients.
- MCD: number of filters $B = 36$, number of coefficients $L = 18$.
- MNB (structure 1): $a = 1.0000$ and $b = -4.6877$ as suggested in [Voran, 1999a]. The signals were subsampled to 8000 Hz before applying this measure.

For PESQ evaluation, the standard implementation available at the ITU web site was used[2]

[2] http://www.itu.int/rec/T-REC-P.862/en

3.3. Experimental setup

In order to evaluate the performance of selected measures as predictors of recognition rate, an experimental setup was designed. This consists of the reproduction in a room of pre-recorded clean speech sentences and noise, to obtain the mixtures to be used in the evaluation. Reproduction was made through loudspeakers with flat frequency response in the range from 20 Hz to 20 kHz. In all experiments, two sources where used and the resulting sound field was picked-up at some selected points by two Ono Sokki MI 1233 omnidirectional measurement microphones, with flat frequency response from 20 Hz to 20 kHz and with preamplifiers Ono Sokki MI 3110. In the following sections, brief descriptions of the speech database, spatial location of sources and microphones, separation algorithm and speech recognizer employed in this work will be given.

3.3.1. Speech database

In this study a subset of a database generated by the authors is used. It consists of recordings of 20 subjects, 10 male and 10 female, each pronouncing 20 sentences selected for a specific task (remote controlling of a TV set using voice commands). These sentences, in Japanese language, were recorded in an acoustically isolated chamber using a close contact microphone with sampling frequency of 44 kHz, later downsampled to 16 kHz with 16 bit quantization. From this database, one male and one female speakers were selected for this study. In consequence, the original sources consist of 40 utterances, 20 from a male speaker and 20 from a female speaker. The corpus contains an average of 1.4 words/sentence, with average duration of 1.12 s.

Three kinds of interfering signals were selected. One is a signal obtained from recording the noise in a room with a large number of computers working. Spectral and statistical characteristics of this noise source can be seen in Figure 3.2. The second kind of noise is a speech signal, pronouncing a sentence different from those used as desired sources. In the case of sentences spoken by female speakers, utterances from male speakers were used as noise and vice versa. The third noise employed is a recording of sound emitted by a TV set. This noise includes speech simultaneously with music. The same TV sound was used to interfere with spoken sentences of both speakers. More data on this recordings can be found in Appendix A.1.

3.3.2. Spatial setup

All the mixtures were performed in an acoustically isolated chamber as shown in Figure 3.3. This setup includes two loudspeakers and two microphones with or without two reflection boards (used to modify reverberation time). As it can be seen in the figure, there are three locations for microphones, a, b and c. In addition, the speech source

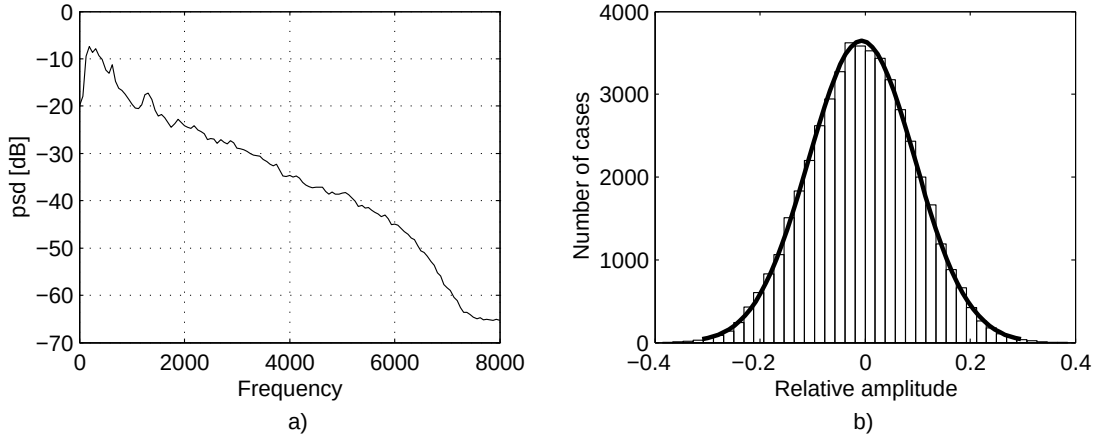


Figure 3.2. Computer noise characteristics. a) shows power spectral density (psd) estimated by Welch method, and b) shows an estimation of probability density function (pdf)(the line is a normal fit for histogram).

and noise can be reproduced by loudspeakers in the way shown in Figure 3.3, named "position 1", or they can be exchanged to "position 2" (that is, playing the speech in the speakerphone labeled "noise" and vice versa). Powers of reproduced signals were adjusted in such a way to get a power ratio of speech and noise at loudspeakers output of 0 dB or 6 dB.[3].

Each of these spatial-power (SP) combinations will be referred as an "SP-Case" in the following. Table 3.2 shows the codes assigned to each of the SP-Cases, with explanation of the parameters used in each case.

In order to analyze the changes in reverberation properties of the room, with the same positions explained before, there were one or two reflection boards added. Without reflection boards, the measured reverberation time was $\tau_{60} = 130$ ms, whereas with one reflection board this time increased to $\tau_{60} = 150$ ms, and with two reflection boards the time was $\tau_{60} = 330$ ms. The same 10 SP-Cases previously mentioned were repeated in each of the reverberation conditions, giving three sets of experiments which would be referred from now on as Low ($\tau_{60} = 130$ ms) , Medium ($\tau_{60} = 150$ ms) and High ($\tau_{60} = 330$ ms)[4].

Briefly, there are three reverberation conditions. For each one of them, 10 SP-Cases

[3]We use the terms "power ratio" instead of the standard SNR because we can adjust the power of the sources at the loudspeakers, but the room will modify it in the mixing and we cannot determine the power produced by each source in the microphones

[4]This naming convention is just to distinguish relative duration of reverberation times in this set of experiments, but this does not imply that the case named "High" actually corresponds to very long reverberation time.

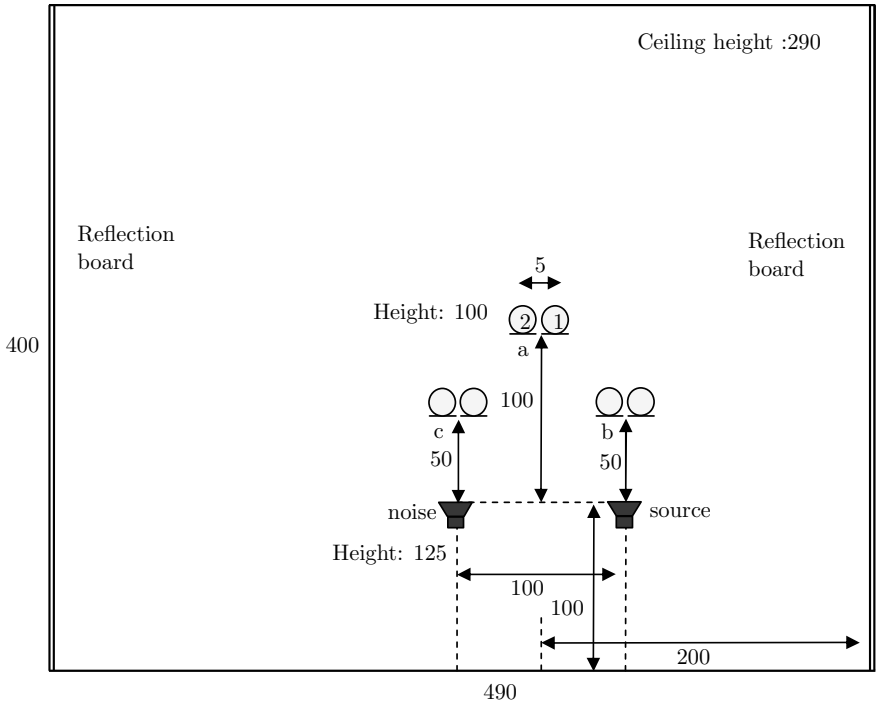


Figure 3.3. Room used for all recordings in the quality evaluation experiments. All dimensions are in cm.

were performed, with different combinations of microphone and source locations and different power ratios. Each of these cases consists of 20 utterances from a male and 20 from a female speaker, mixed with each one of the three kinds of noise employed, adding to a total of 120 utterances for each SP-Case. In total, separation and recognition over 3600 experimental conditions were evaluated.

3.3.3. Separation algorithm

The BSS algorithm is based on independent component analysis (ICA) in the frequency domain [Cichocki and Amari, 2002; Hyvärinen et al., 2001], as explained in Chapter 2. We have used a STFT with a Hamming window of 256 samples. In order to have enough training data to perform ICA on each frequency bin, a window step of 10 samples was used. For each frequency band, a combination of JADE [Cardoso and Souloumiac, 1993] and FastICA [Bingham and Hyvärinen, 2000] algorithms was used to achieve separation. FastICA is sensitive to initial conditions, because it is a Newton-like algorithm. For this reason, JADE was applied to find an initial approximation to separation matrix, and then FastICA was employed to improve the results (with the

Table 3.2. Signal-Power experimental case codes and their meanings.

SP-Case	Microphones	Source-Noise	Power ratio
a10	a	position 1	0 db
a16	a	position 1	6 db
a20	a	position 2	0 db
a26	a	position 2	6 db
b10	b	position 1	0 db
b16	b	position 1	6 db
b20	b	position 2	0 db
c10	c	position 1	0 db
c20	c	position 2	0 db
c26	c	position 2	6 db

JADE guess as initial condition). For FastICA we have used the nonlinear function $G(y) = log(a + y)$ with its derivative $g(y) = \frac{1}{a+y}$. Both complex versions of JADE and FastICA were obtained from the websites of their authors. Permutation and amplitude indeterminacies are solved by the algorithm proposed in [Murata et al., 2001]. After the separation and the solution of the indeterminacies, the overlap-and-add method of reconstruction was used to obtain the time-domain signals [Allen and Rabiner, 1977] [5].

For each one of the reverberation conditions, mixture signals captured by microphone in each combination of sentences and noises were processed with this algorithm. From each pair of separated signals, the signal more likely to represent the desired source was selected by means of a correlation. In this way, a database with separated signals corresponding to each one of the sentences in each experimental condition was generated. Before the application of quality measures, correlation was used to compensate any possible delay between separated and original signals, and to detect possible signal inversions (if maximum correlation is negative, the signal is multiplied by -1). Also all signals were normalized to minimize the effect of magnitude indeterminacies. This was done by dividing both separated and original signals by their respective energy.

3.3.4. Recognizer

The recognizer used was the large vocabulary continuous speech recognition system Julius [Lee et al., 2001], based on hidden Markov models (HMM). This is a standard recognition system widely used for Japanese language. The decoder performs a two-pass

[5] The separation algorithm used in this chapter is the same one that will be presented in Chapter 4 but without the Wiener postfilter. For more details on the algorithm, please refer to that chapter.

Table 3.3. Word recognition rates (WRR%) with only one source in the real room. In this case, there is only reverberation effect. This can be considered an upper limit of the obtainable recognition rate in each case.

Mic.	Low Rev.	Med. Rev.	High Rev.
1cm	83	80	79
a	80	75	53
b	75	66	66
c	56	49	44

search, the first with a bi-gram and the second with a tri-gram language model. This system was used with acoustic models for continuous density HMM in HTK [Young et al., 2005] format. The models were trained with two databases provided by the Acoustic Society of Japan (ASJ): a set of phonetically balanced sentences (ASJ-PB) and newspaper article texts (ASJ-JNAS). About 20000 sentences uttered by 132 speakers of each gender were used.

The recognizer used 12 mel frequency cepstral coefficients (MFCC) computed each 10 milliseconds, with temporal differences of coefficients ($\Delta MFCC$) and energy (ΔE) for a total of 25 feature coefficients. Also cepstral mean normalization was applied to each utterance. Phonetic tied-mixture triphones are used as acoustic models. The full acoustic model consists of 3000 states tying 64 Gaussian from a base of 129 phonemes with different weights depending of the context. For the language model, both bi-gram and tri-gram models were generated using 118 million words from 75 months newspaper articles, which were also used to generate the lexicon [Kawahara et al., 2000].

Furthermore, the WRR of this system was evaluated on the source sentences reproduced in the room but without any interfering noise, with microphones in location a, b and c, and also with microphones located at 1 cm from source. This allows to evaluate the degradation effect caused on the recognizer by reverberation, even without interfering noise. These results are shown in Table 3.3. As the algorithm generally does not reduce –in a great amount– the reverberation effect, these values can be taken as a baseline limit for obtainable recognition rate in each case[6].

Word recognition rates for both mixtures and BSS separated signals were also evaluated for reference, as shown in Figure 3.4. For this figure, the SP-Cases where grouped according to location of microphones relative to sources, as equal distance (a10+a20 and a16+a26), nearer to desired source (b10+c20 and b16+c26), and nearer

[6]It must be noted that feeding the ASR system with the original clean sentences yielded a word recognition rate of 100%.

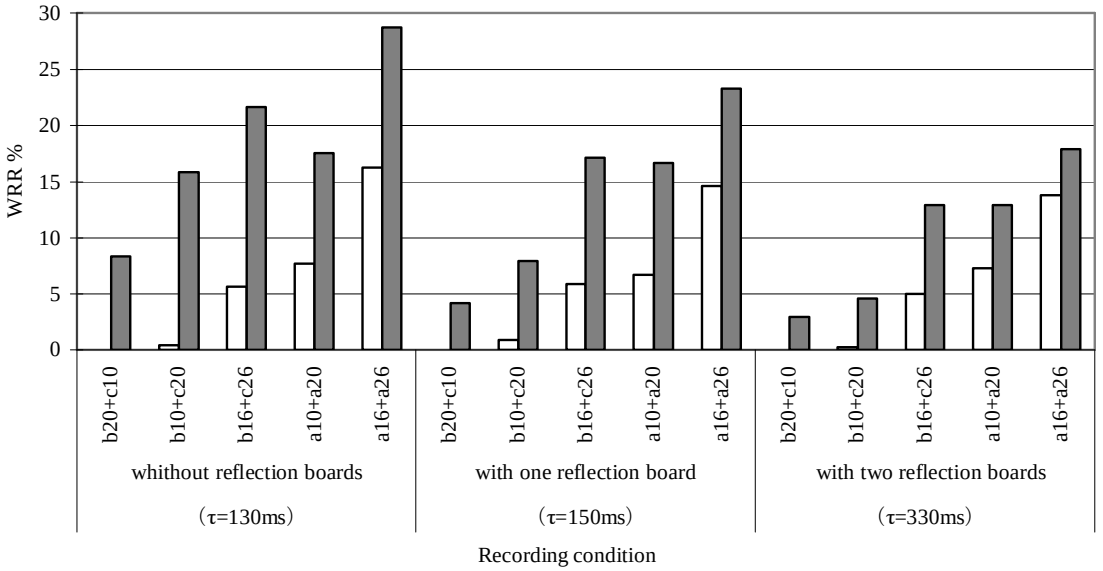


Figure 3.4. Word recognition rates using mixtures and separated sources for different reverberation conditions. White bars: mixed (recorded) sound; Gray bars: sound separated by BSS. Label b20+c10 mean average of results of these SP-Cases (with similar meaning for the other labels).

to noise source (b20+c10).

3.4. Results

For each one of the reverberation conditions and for each SP-Case, quality measures have been calculated for all sentences and all noise types. Then, an average of each measure for all utterances has been taken, grouping them for noise kind. In this way, for each combination of reverberation condition/SP-Case, three values of quality were generated corresponding to average quality in each noise kind. In the same way, for each combination of reverberation condition and SP-Case, recognition rate was also evaluated, separated by the kind of noise, obtaining three values of recognition rate for each case.

Table 3.4. Correlation coefficient $|\rho|$ for all experiments. Best value for each case has been marked in boldface. "All" includes in the sample all noise kinds for a given reverberation condition, and "ALL" includes all noise kinds and all reverberation conditions. Last row shows the standard deviation of the regression residual.

Reverb	Noise	segSNR	IS	LAR	LLR	WSS	TRD	CD	MCD	MNB	PESQ	NLP
Low	Comp.	0.80	0.44	0.70	0.70	0.92	0.88	0.81	0.87	0.90	0.95	**0.96**
	TV	0.68	0.13	0.76	0.72	0.84	0.77	0.78	0.80	0.78	0.89	**0.90**
	Speech	0.64	0.77	0.74	0.65	**0.84**	0.62	0.79	**0.84**	0.56	0.77	0.76
	All	0.61	0.39	0.73	0.71	0.86	0.75	0.77	0.80	0.62	**0.87**	**0.87**
Medium	Comp.	0.78	0.43	0.58	0.62	0.88	0.82	0.74	0.85	0.85	0.93	**0.94**
	TV	0.76	0.31	0.92	0.91	0.90	0.86	0.91	0.85	0.85	0.91	**0.96**
	Speech	0.78	0.76	0.82	**0.85**	0.66	**0.85**	0.76	0.62	0.64	0.71	0.77
	All	0.76	0.46	0.75	0.74	0.77	0.83	0.74	0.72	0.78	0.87	**0.91**
High	Comp.	0.77	0.53	0.74	0.75	0.83	0.85	0.81	0.80	0.83	0.85	**0.88**
	TV	0.81	0.71	0.92	0.92	0.93	0.90	0.93	0.90	0.87	0.94	**0.97**
	Speech	0.74	0.33	0.75	0.74	0.77	0.72	0.79	0.75	0.66	0.79	**0.82**
	All	0.75	0.50	0.78	0.79	0.81	0.84	0.84	0.79	0.75	0.84	**0.87**
$\lvert\rho\rvert$	ALL	0.74	0.43	0.73	0.71	0.83	0.84	0.76	0.77	0.75	0.88	**0.90**
σ_r	ALL	5.94	9.70	6.68	7.36	4.98	5.74	6.42	5.77	7.86	5.02	**3.59**

With these data three analyses were made. Table 3.4 presents Pearson correlation coefficient (absolute value) for the analyses, defined as [Montgomery and Runger, 2003]:

$$\rho_{xy} = \frac{\sum_i \left[(x_i - \bar{x}) (y_i - \bar{y}) \right]}{\left[\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2 \right]^{1/2}} \tag{3.13}$$

where x_i represents the quality measure to be used as predictor, y_i the WRR%, and $\bar{x}$, $\bar{y}$ the corresponding estimated mean values. First, for each reverberation condition, correlation of quality measures as predictors of recognition rate was evaluated, discriminated for each kind of noise, in such a way that the sample consists of 10 pairs of quality measures/recognition rates (each pair is a SP-Case). Second, the same analysis was performed considering all kind of noises for a reverberation condition, that is taking in the sample the 30 pairs of quality measures/recognition rates, giving as a result one value of general correlation for each reverberation condition. Third, the correlation was evaluated without data segregation, that is, including in the sample all kinds of noise and all the reverberation conditions.

The NLP measure and the composite measures depend on coefficients that must be adjusted from the data. For this task, we used the data corresponding to all combinations of noise and reverberation conditions. After obtaining the coefficients, the correlation of groups of data corresponding to each reverberation condition and each noise kind was evaluated. In the case of NLP, this method was used to solve a nonlinear regression problem using least squares. The best coefficient vector for the nonlinear equation obtained was $\boldsymbol{\alpha} = [-1.7558, 63.2960, -1.6962, 5.2361]$.

After this first experiment, the best five single measures were selected as candidates to evaluate the use of composite measures. The best measure is NLP, followed very closely by PESQ, which is not a surprise as NLP is just a fine tuned version of PESQ. Without taking into account these two measures, there is no clear winner among the rest. For the lowest reverberation time, WSS is clearly superior to the other measures. In the intermediate reverberation case, the best performance is for TRD but sharing the success for different noises with LAR, LLR and WSS. In the High reverberation case, CD measure seems to behave better, but closely followed by TRD. Thus, the five measures selected to evaluate the composite ones were WSS, TRD, CD, PESQ and NLP.

All possible combinations of the measures, taking them by pairs, by terns, by quartets, and all five (WSS, TRD, CD, PESQ and NLP), for a total of 26 composite measures were evaluated. The best set of measures for each case of $K = 2, 3, 4, 5$ was selected. The sets of measures used to construct the best composite measures were: $\mathcal{S}_{CM_2} = \{WSS, NLP\}$, $\mathcal{S}_{CM_3} = \{WSS, TRD, NLP\}$,$\mathcal{S}_{CM_4} = \{WSS, TRD, PESQ, NLP\}$ and all $\mathcal{S}_{CM_5} = \{WSS, TRD, CD, PESQ, NLP\}$. The results are summarized in Table 3.5.

Table 3.5. Correlation coefficient $|\rho|$ for the best five single measures and the composite measures. "All" includes in the sample all noise kinds for a given reverberation condition, and "ALL" includes all noise kinds and all reverberation conditions. Last row shows the standard deviation of the regression residual.

Reverb.	Noise	WSS	TRD	CD	PESQ	NLP	CM_2	CM_3	CM_4	CM_5
Low	Comp	0.92	0.88	0.81	0.95	**0.96**	0.96	0.97	0.97	0.97
	TV	0.84	0.77	0.78	0.89	**0.90**	0.89	0.91	0.91	0.91
	Speech	**0.84**	0.62	0.79	0.77	0.76	0.77	0.78	0.78	0.79
	All	0.86	0.75	0.77	**0.87**	**0.87**	0.87	0.89	0.89	0.89
Medium	Comp	0.88	0.82	0.74	0.93	**0.94**	0.93	0.95	0.95	0.95
	TV	0.90	0.86	0.91	0.91	**0.96**	0.96	0.97	0.97	0.97
	Speech	0.66	**0.85**	0.76	0.71	0.77	0.76	0.72	0.73	0.73
	All	0.77	0.83	0.74	0.87	**0.91**	0.91	0.91	0.91	0.91
High	Comp	0.83	0.85	0.81	0.85	**0.88**	0.88	0.87	0.88	0.88
	TV	0.93	0.90	0.93	0.94	**0.97**	0.97	0.97	0.97	0.97
	Speech	0.77	0.72	0.79	0.79	**0.82**	0.82	0.82	0.83	0.82
	All	0.81	0.84	0.84	0.84	**0.87**	0.87	0.87	0.87	0.87
$\|\rho\|$	ALL	0.83	0.84	0.76	0.88	**0.90**	0.90	0.91	0.91	0.91
σ_r	ALL	4.98	5.74	6.42	5.02	**3.59**	3.57	3.53	3.51	3.50

For clarity we have repeated the best five single measures, with the best single measure for each case marked in bold letters.

Figure 3.5 shows a scatter graph for the case of all variables together for the five single measures considered in this study, and the best composite measure CM_5, for the case of all noise kinds and all reverberation conditions (last rows in Table 3.4). A dispersion graphic was drawn for each measure including a total least squares regression line and two lines that mark regression value plus/minus two times the standard deviation of residual. This standard deviation was estimated according to $\sigma_r^2 = \sum_{i=1}^{N} (y_i - \widehat{y}_i)^2 / (N-2)$ where y_i is the true WRR% and $\widehat{y}_i$ the predicted value by regression [Montgomery and Runger, 2003]. Also, this figure shows the values of $|\rho|$ and σ_r.

3.5. Discussion

As it can be seen in the results, the NLP measure was the best single measure in almost all cases, having the lowest variance in the global evaluation. The global correlation is 0.90, which is a very good value for this measure. The next best measure is the raw PESQ score, which achieved a correlation of 0.88, although it has a larger variance than the nonlinear version.

Table 3.4 also shows that both PESQ and NLP measures have a very stable performance when varying the mixing conditions and with different noises, and so the

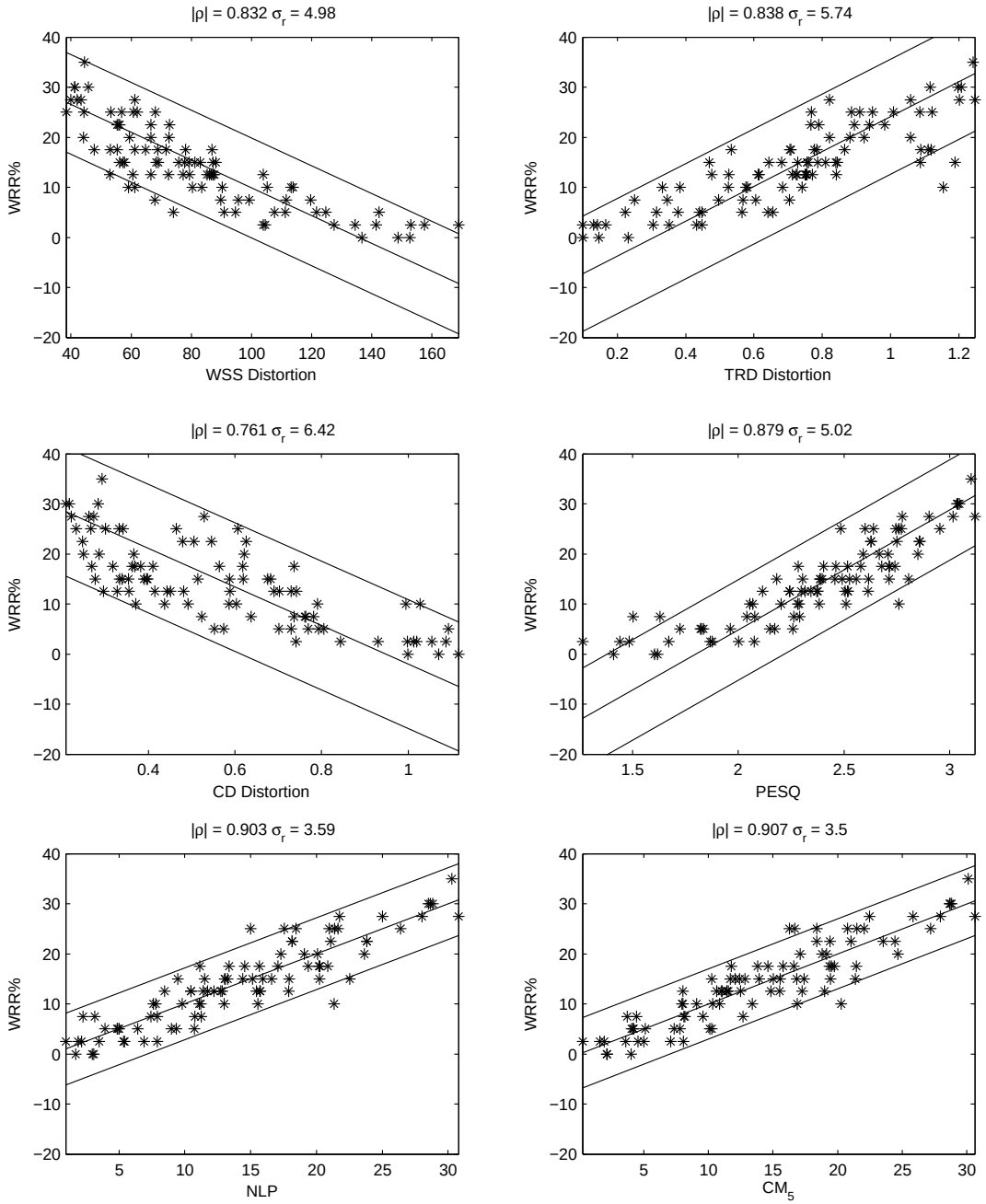


Figure 3.5. Regression analysis of quality measures for all the experimental cases.

measures seemed to be quite robust to variations of the experimental conditions.

It is interesting to note that the worst performance of the raw PESQ measure was obtained in the evaluation with competing speech as noise. This can be due to the fact that eliminating a competing voice is probably one of the hardest problems [Divenyi, 2004, chapter 20]. Since parts of the competing speech are present in the separated signal, the recognizer can confuse some phonemes. At the same time, being the desired source present in a good level, the output of the perceptual model will be not so different and the quality obtained would be good.

The performance of TRD is very competitive, particularly well correlated at medium reverberation times. This can be related to the fact that this measure was designed specifically to evaluate blind source separation algorithms. This could show an important relation (that was not necessarily obvious a priori) between the evaluation of the separation algorithm itself and its performance for speech recognition.

The relative good results of cepstral measures are not a surprise. Their quite uniform performance for all levels of reverberation can be related to the internal representation of signals in the recognizer, based in cepstral coefficients. Therefore, changes in these coefficients are reflected directly in recognition rate, giving some uniform behavior. Although one could expect a general better performance for MCD than for CD, the results not always agree. It would be interesting to perform the same experiments with a recognizer based in different feature extraction methods, like Rasta-PLP [Hermansky, 1990; Hermansky and Morgan, 1994; Koehler et al., 1994], to check whether the good performance is recognizer-related or it can actually be generalized. Comparing only the cepstral measures, MCD is very stable with reverberation, keeping high correlation levels in all cases. CD presents better correlation than MCD at higher reverberation times.

In opposition to the expected results, MNB performance was rather poor, compared to the previous ones. This measure was specially designed for speech vocoders, and perhaps the vocoder distortions are very different from the ones produced in blind source separation.

It is also interesting to verify that segmental SNR has been outperformed in all cases by almost all measures. The SNR evaluation neglects the filtering effects that arise in BSS. In this way, using this measure a perfectly separated but filtered signal, will be seen as heavily distorted, even if it keeps the main cues used by the recognition system to produce a good recognition rate. This must be considered in all works where improvement of algorithms is reported using this measure. According to these results, improvements reported in terms of SNR will not be reflected in recognition rates. We strongly recommend against the use of SNR for this kind of evaluations, something that is very common in practice.

The results presented here were obtained for this separation algorithm and this

specific recognizer, so strictly speaking they are only applicable for these cases. Nevertheless, we consider that as long as the separation algorithm uses similar processing principles (i.e. frequency domain BSS) and the speech recognizer uses the same paradigm (HMM with MFCC features) the results should not change qualitatively. On the other hand, all the experiments here presented were made with Japanese language. However, there are some studies (like [Voran, 1999a]) where it is shown that the results of objective quality measures for different languages are quite similar, and so we expect an objective measure not to change significantly when applied to different languages.

With relation to the composite measures, it can be seen that they do not provide a significant improvement in correlation with respect to NLP. NLP appears as a component in all the composite measures, and the addition of other measures provides only a marginal improvement. Given the increased cost related to the evaluation of several measures and combining them, it seems to be better to just use the NLP measure instead.

In brief, we strongly recommend the use of PESQ or NLP for the evaluation of BSS algorithms for speech recognition. It must be noted that the mapping from PESQ to NLP is monotonically increasing, and thus PESQ can be used instead to compare algorithms (if a method gives higher PESQ than the other one, it will give also higher NLP due to the monotonicity). These measures can be used alone or with other measures that evaluate other aspects of the separation algorithm as well, like SIR to account for the separation capabilities of the algorithms. Being PESQ a perceptual measure designed to "mimic" the auditory process, it is highly correlated with subjective quality evaluations [ITU, 2001]. This shows that evaluating the results of a BSS algorithm using PESQ (either in raw PESQ mode or the nonlinear version NLP) will be very powerful, because a good score would suggest both, a high recognition rate and a high perceptual quality.

3.6. Proposed experimental protocol

The results presented here showed the capabilities of PESQ as the most important objective quality measure to evaluate the performance of BSS algorithms for ASR systems. This is important when several methods of BSS are compared, but moreover, it is a valuable tool to compare the performance of the same method when varying some parameter. This method is easier and faster, for a general quality comparison, than using a whole speech recognition system. Therefore, we summarize here the proposed experimental protocol:

1. Design the BSS method.

2. Determine the key paramethers and the theoretical limitations of the method.

3. For a given room (that is, for fixed reverberation conditions) define the locations of sources and microphones. In this step, the theoretical limitations of the method should be taken into account.

4. Select the kind and number of competing sources and power ratios.

5. Obtention of mixtures. At this stage there are two possible methodologies: artificial mixtures and real mixtures.

   a) Artificial mixtures: in this case, what is measured in the real room is the IR from each source location to each microphone location. Using these IRs, the mixtures are obtained applying the convolutive mixture model.

   b) Real mixtures: for each source and noise, the sound sources are replayed with loudspeakers in the real room. The sound field generated is captured by the microphones.

6. Once the mixtures are obtained, the separation method is applied to each mixture.

7. At initial stages, the objective evaluation can be performed over a reduced set of sentences, to determine optimal parameters for the algorithms.

8. Once the best parameters are determined, the objective evaluation is performed on all the data. Additional measures can be also used to evaluate other dimensions of performance, like the average processing time to evaluate the potential for real-time usage.

9. The ultimate evaluation will be performed in a speech recognition task. It can be both from artificial mixtures or real mixtures.

In stage 3, the consideration of the theoretical limitations includes among others, the application of any restriction to number and/or locations of sources and/or microphones, restrictions on the room properties to low/moderate reverberation times, etc.

For stage 5, both methods for mixtures generation require the use of pre-recorded clean sources. This is important because the clean sources will be used as the comparison signal for the objective evaluation, and this will allow for repetibility of the experiments in other conditions. The advantage of the artificial mixtures approach in stage 5*a* is that the mixtures are easily generated for any number of sources, in a reduced time, allowing to evaluate the performance in larger databases, without the time requiered for real recordings. The main disadvantage is the reduced reality of the experiment, which makes the result less extrapolable to real situations.

On the other side, the usage of real mixtures in stage $5b$ allows for an extremely high realism for the experiment. The mixtures are real, there are no approximations in the mixture model. The main advantage is that results obtained in these conditions would be directly extrapolable to real applications. The main disadvantage is the time requiered to perform the recordings, moreover if the database needs to be large to have precise results (i.e. for ASR tests for example).

In stage 9, it is desirable to have real mixture results, although the problem of recording a large database can be a limitant in this aspect.

This framework for the evaluation of the methods will be followed for all the separation methods proposed in this document. The algorithm will be evaluated on diferent conditions of reverberation, with diferent kinds of competing noises, using both, real and artificial mixtures.

For the rest of the dissertation, the evaluation will be presented in the case of two mixtures and two sources. The algorithms that we will produced are valid for other number of sources and microphones, but the evaluation would be more difficult to interpretate in those cases.

3.7. Concluding remarks

In this chapter a review of the different approaches for BSS evaluation was presented. Based on this review, a specific framework for the evaluation of BSS algorithms for ASR was proposed, using measures that compare the output to the original clean measure. This kind of measures allows for very realistic experiments, without using knowledge of the mixing process. A set of 11 objective quality measures were selected and evaluated regarding their capability to predict the WRR of an ASR system. From this analysis the best measures were NLP and PESQ. Another important result is the demonstration of the very poor capabilities of SNR, a measure that is widely used to evaluate the performance of this kind of systems. Based on this results, in the rest of the dissertation the PESQ measure will be used as a valid index of performance for all stages of the development of new algorithms, keeping the evaluation with ASR systems as the last level of global evaluation. The experiments will follow the proposed framework to obtain results that could be extrapolated to real situations. The results presented in this chapter have been published in [Di Persia et al., 2007] and [Di Persia et al., 2008].

Chapter 4

Frequency-domain ICA with improved initialization, permutation correction and postfilter

As stated in Chapter 2, the standard approach to fd-ICA presents several drawbacks and it is sensitive to factors that impair its performance, particularly under the effects of reverberation. One of those analyzed factors was the sensitivity of the ICA algorithms to initialization, which is manifested as results of very different quality for the same signals used as input. A second aspect was the problem of permutations that need to be solved to obtain a consistent representation of each source, before time-domain reconstruction. And a third drawback was the fundamental limitation of the method as it is capable of rejecting the direct sound for competing sources but not the reflections of the desired or competing noises. This comes from the analysis of the beampattern generated by the separation matrices, that shows a behavior similar to that of null beamformers, with nulls in the competing sources directions. In the first algorithm to be proposed in this chapter, some alternatives to improve these aspects providing for better quality results will be explored. First, a technique for improvement of the initialization will be given. Next, a method to improve the permutation problem will be presented. Finally, the use of a Wiener postfilter to reduce the effect of the echoes and competing source, improving the output quality, will be introduced.

4.1. Robust initialization

In our preliminary works using the standard method [Handa et al., 2004; Tsutsumi et al., 2004], we explored two alternatives for the solution of the ICA problem. One was the complex JADE algorithm proposed in [Cardoso and Souloumiac, 1993], and the

Table 4.1. Quality and average processing times for standard JADE and FastICA based fd-ICA methods.

		JADE		FastICA	
Power	Noise kind	PESQ	Av. Time	PESQ	Av. Time
6 dB	Speech	2.47	5.18	2.57	14.67
	White	2.35	5.55	2.68	10.42
0 db	Speech	2.28	5.39	2.33	12.25
	White	2.13	5.58	2.22	19.77
Average		2.31	5.42	2.45	14.28

other was the complex FastICA algorithm of [Bingham and Hyvärinen, 2000]. Both algorithms are methods that exploit the nongaussianity to produce the separation, and in this way they are equivalent in their effects, although the results are different because they use different cost functions to measure such nongaussianity, and different approaches for the optimization. We have evaluated the PESQ score and the average time for processing using both approaches. For this task we used real data recorded in a room with reverberation time of about 200 ms. Five sentences uttered by four speakers were used, for a total of 20 utterances. Each of them were mixed with two kind of noises, white noise and speech noise, replayed by a different loudspeaker. For each kind of noises, two power ratios were used, 0 dB and 6 dB.

Table 4.1 shows the average results. In general, the FastICA algorithm gives better results, although it is more sensitive to initialization. We verified this as the JADE algorithm always converge, whereas for FastICA, it is usual to reach the maximum number of iterations without convergence in several frequency bins. At the same time, we found that JADE algorithm is stable and faster than FastICA. This suggests a natural strategy to improve the method: using JADE algorithm first, which will produce an initial estimate, and then refining the separation estimate by FastICA, which given a good initial estimation will improve the separation quality. We will present here a brief review of JADE algorithm (the FastICA method has been explained in Chapter 1).

JADE algorithm is an ICA method that uses explicit high order statistics information, introduced by means of the fourth-order circular cumulant tensor. For a random vector X with probability density function $f_X(\mathbf{x})$, the fourth order circular cumulant is given by

$$C^{jl}_{ik,X} = Cum\left\{X_i, X_j^*, X_k, X_l^*\right\} . \tag{4.1}$$

Cardoso and Souloumiac [1993] proposed to obtain the unitary matrix $\mathbf{W}$ of (1.9)

by means of maximizing the cost function

$$\mathcal{J}(\mathbf{W}) = \sum_{i,k,l=1}^{M} \left| C_{ik,Y}^{il} \right|^2 , \tag{4.2}$$

where $\mathbf{y}$ is the observed mixture vector. This optimization is equivalent to a join diagonalization of a set of eigen-matrices.

Given a matrix P, the fourth-order cumulant $C_{ik,X}^{jl}$ defines a linear transformation $\Omega(P)$ in such a way that

$$\Omega(P)_{i,j} = \sum_{k,l} C_{ik,X}^{jl} P_{k,l} . \tag{4.3}$$

This linear transformation has M^2 eigen-matrices $E_{p,q}$ that satisfy $\Omega(E_{p,q}) = \lambda_{p,q} E_{p,q}$. It was shown in [Cardoso and Souloumiac, 1993] that it is enough to find the M most significant eigenmatrices (i.e. those with larger associated eigenvalues) and perform approximate joint-diagonalizing of these eigen-matrices to obtain separation matrix $\mathbf{W}$ for signals in $\mathbf{x}$. This is the origin of the algorithm name, joint approximate diagonalizing of eigenmatrices. Before applying this algorithm, a whitening transformation is performed in order to eliminate second order correlations and simplify the algorithm convergence.

The JADE algorithm is very efficient as long as the number of sources is small. When the number is increased, the explicit calculation of the eigenmatrices slows down the algorithm, but for up to 5 or 6 sources it is quite fast.

Although the JADE algorithm has good separation capabilities, its performance can be improved by using the obtained separation matrix as initial value for another algorithm that refines it by optimizing some contrast function. In this case, FastICA algorithm has been used [Bingham and Hyvärinen, 2000].

As explained in Chapter 1, this is a Newton-like fixed-point iteration with quadratic convergence, and as such, it is very fast. As all fixed point methods, however, it depends on good initial conditions, and we have a good estimation using the output of JADE algorithm. In this way, we expect that the addition of JADE will speed up the convergence of FastICA and the total time will be less than the addition of the times needed for each method.

4.2. Indeterminacy solution

The permutation and scaling ambiguities need to be solved before recovering the time domain signals. Solving these problems does not mean that the resulting time domain signals will not have a permutation and a scaling, but that all the frequency

bins will have the same permutation (thus belonging to the same source) and scaling (thus they will not be subject to arbitrary colorations of the spectral contents).

4.2.1. Solution of the scaling problem

To solve scaling indeterminacy, a variant of the method proposed by [Murata et al., 2001] has been used. For the scaling ambiguity, the approach consists of recovering the filtered versions of the sources instead of the sources themselves. So, mixtures are modeled as $\mathbf{x} = \mathbf{v}_1, \ldots, \mathbf{v}_M$. Using the separation matrix $\mathbf{W}$ and its inverse $\mathbf{W}^{-1}$ (i.e. estimated mixing matrix), it can written

$$\begin{aligned} \mathbf{x} &= \mathbf{W}^{-1}\mathbf{y} \\ &= \mathbf{W}^{-1}\mathbf{W}\mathbf{x} \\ &= \mathbf{W}^{-1}\mathbf{I}\mathbf{W}\mathbf{x} \\ &= \mathbf{W}^{-1}\left(\mathbf{E}_1 + \cdots + \mathbf{E}_M\right)\mathbf{W}\mathbf{x} \\ &= \mathbf{W}^{-1}\mathbf{E}_1\mathbf{y} + \cdots + \mathbf{W}^{-1}\mathbf{E}_M\mathbf{y} \\ &= \mathbf{v}_1 + \ldots + \mathbf{v}_M\,, \end{aligned} \tag{4.4}$$

where $\mathbf{E}_i$ is a matrix with a one in the i-th diagonal element and zeros elsewhere. It is easy to prove that the representation of $\mathbf{v}_i$ is independent of the scaling in matrix $\mathbf{W}$. This method results of the application of the minimal distortion principle as defined in [Matsuoka, 2002]. The signals $\mathbf{v}_i$ represents the contributions of the i-th source in all microphones, and any of them can be used as a representative of the source. It must be noted that in this way the signal is recovered as received by the microphone, that is, including the filtering effect of the room.

4.2.2. Solution of the permutation problem

Now, for the permutation problem, the standard approach makes use of the fact that the envelopes of different sound signals must be different. Sometimes this property is called the amplitude modulation property. Also, if the signals are independent the correlation between envelopes of the separated sources in each bin must vanish or be small [Murata et al., 2001]. Furthermore, it can be expected that successive frequency bins of the same source should share the same or similar envelopes. This is the information used to solve permutation problem: starting from some frequency band, an estimation of the envelope based on previous classified bands is calculated. Then, for each separated signal in a new frequency bin, the correlation between its envelope and the estimated one for pre-classified bins is calculated, and the signal is assigned to that of maximum correlation value [Murata et al., 2001].

The envelope for the signal in a frequency bin is estimated in a two step approach. First, the magnitude of the complex signal is taken. Then, a low-pass moving average filter that promediates the last N_{env} values of the magnitude is used to smooth the amplitude. The envelopes for all the bins are calculated first. Then, an inner product is used to evaluate the similitude among the sources at each bin, and the bins are reordered accordingly. The bins which are clearly different are classified first, those that are more similar and confusable are left for the final stages.

In the original paper, the pre-classified envelopes are estimated as an average of all the previously classified envelopes in that class.

$$\epsilon\left(k\right)_j = \frac{1}{k}\sum_{i=1}^{k} E\left(i\right)_j \ , \tag{4.5}$$

where $\epsilon\left(\cdot\right)_j$ refers to the locally averaged envelope for source class j, and $E\left(\cdot\right)_j$ to the last classified envelope for this class. This equation can be also written as an autorregressive model

$$\epsilon\left(k\right)_j = \frac{k-1}{k}\epsilon\left(k-1\right)_j + \frac{1}{k}E\left(k\right)_j \ , \tag{4.6}$$

in which the weight used to introduce the information of the new envelope is variable, and is reduced with k.

In our work, instead of using this approach, we assume that in the averaging process, the last classified envelopes must have more weight since they will be more similar to the envelopes following for classification. Therefore instead of a simple averaging of envelopes, we update that value as

$$\epsilon\left(k\right)_j = (1-\alpha)\epsilon\left(k-1\right)_j + \alpha E\left(k\right)_j \ , \tag{4.7}$$

where $0 < \alpha < 1$ is a constant. That is, we use an autoregresive model for the estimation of the envelope, with a constant value for wheighting the new information. This model always keep a fraction of the previous estimation, but updates it by a constant percentage of the recently processed bin. In the simple average of all envelopes approach used in [Murata et al., 2001], this means that each new classified bin has a reduced effect on the global estimate of envelope. In that way, the k-th bin will contribute with a factor of $1/k$ to the envelope estimation. On the contrary, using our approach, the last classified bin will contribute with a factor of α to the average, being constant for all bins, thus giving more weight to the last classified bins.

Figure 4.1 shows this process for the bin at $k = 14$, in a 2-by-2 BSS case. The top row shows the amplitudes for the bin, the second row the envelopes calculated after the low pass filtering, and the third row shows the accumulated envelope using

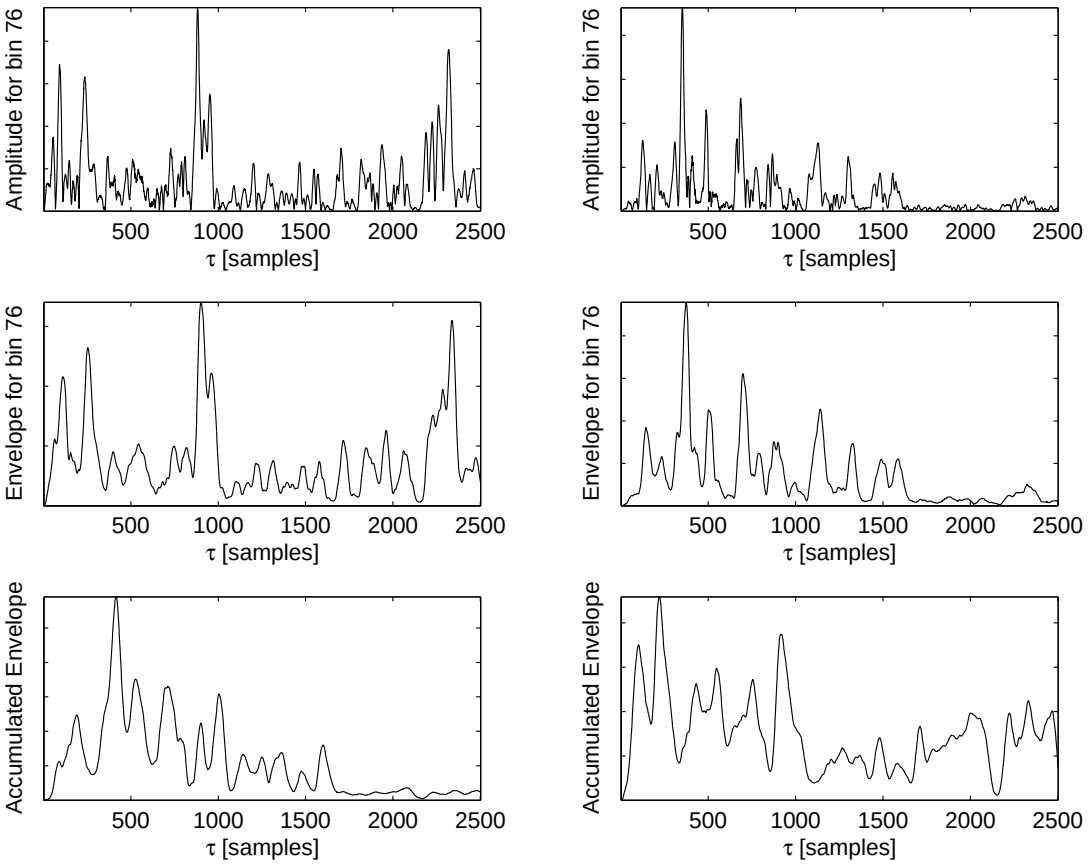


Figure 4.1. Permutation correction by the proposed correlation of envelopes. Top row: amplitudes of the two sources. Middle row: envelopes for the sources. Botton row: accumulated envelopes using our proposed method of (4.7).

our proposed modification, up to that bin. As it can be seen, the bin being processed presents a permutation. The correlation of the envelopes with the accumulated ones will allow to correct it. The figure shows that the accumulated envelopes are quite similar to the envelopes of the processed bin, which will produce a proper permutation identification and correction.

After this process we obtain time-frequency representations for each of the source components, in each sensor. That is, for each source we obtain N time-frequency representations (where N is the number of both, sensors and sources) each one corresponding to the effect of that source in one of the sensors, isolated from the other sources effects. Since usually only one representation for each source is needed, in this study we use the alternative of keeping the one with larger energy for each source to be used in the

following steps.

4.3. Post processing by Wiener filter

The separation result will not be perfect mainly due to the reverberation effect. When reverberation time increases, the performance of the algorithms tend to decrease, as already discussed. An important effect that contributes to this degradation is the behaviour of the separation matrix as null beamformers, as explained in Chapter 2. Each of these beamformers put a null towards the direction (blindly estimated by the ICA algorithm) of each undesired source, with a gain in the direction of the desired one. In this way, for each competing source, the method can reject only the main direct sound, but not its echoes coming from different directions. Also, for the desired source it let pass the direct source and all its echoes. For these reasons, the method is not capable of providing good separation when there are strong echoes, nor to produce reverberation reduction for the desired sources.

To improve this aspect we propose the use of a non-causal time-frequency Wiener filter as post-processing [Huang and Benesty, 2004]. Without losing generality, this will be explained for the two sources and two microphones case, with the generalization to more sources being straightforward.

The short-time Wiener filter $H_{\mathcal{W}}$ for a signal generated by the simple additive noise model is:

$$H_{\mathcal{W}}(\omega,\tau) = \frac{|\tilde{z}(\omega,\tau)|^2 - |\tilde{n}(\omega,\tau)|^2}{|\tilde{z}(\omega,\tau)|^2} \tag{4.8}$$

where $\tilde{n}$ represents the estimated additive noise. After the permutation and scaling problems correction in our algorithm we obtain two signals, $\tilde{v}_1(\omega,\tau)$ and $\tilde{v}_2(\omega,\tau)$, and provided that the separation process was successful they can be used as estimation of the clean sources.

In this way, in order to eliminate residual information from source v_2 on source v_1 we can use the short-time power spectrum of $\tilde{v}_1$ as numerator (estimation of clean source) and add the short-time power spectrums of $\tilde{v}_1$ and $\tilde{v}_2$ as the estimation of noisy power spectrum in denominator. Moreover, as we know that each resulting signal contains some residual information from the other one, and that this sharing of information would be not uniform over the whole time-frequency plane, time-frequency weights can be used to reduce the effect of the filter at specific time-frequency slots, as expressed in the following equation:

$$H_{\mathcal{W},1}(\omega,\tau) = \frac{|\tilde{v}_1(\omega,\tau)|^2}{|\tilde{v}_1(\omega,\tau)|^2 + C(\omega,\tau)\,|\tilde{v}_2(\omega,\tau)|^2}\,. \tag{4.9}$$

where the weighting matrix $C(\omega,\tau)$ will have real values between 0 and 1.

If the time-frequency contents of $\tilde{v}_1$ and $\tilde{v}_2$ are very similar (which would imply that for such time-frequency coordinate the separation was not well done), the weights in C must be close to zero, otherwise they must be near to one. There are several ways to set these weights. One is to ignore them (that is, setting $C(\omega, \tau) = 1$ for all ω and τ). Another simple way may include dot products to determine time and frequency similitude of power spectrums.

For a given bin k corresponding to frequency ω_k, the correlation at lag zero $\mathcal{E}\{\tilde{v}_1(\omega_k, \tau)\tilde{v}_2(\omega_k, \tau)\}_\tau$ taken along the time index τ will be zero, as a consequence of the independence obtained by application of ICA at that bin. Nevertheless, if the expectation is taken for a given τ, along the frequency bins k, it will be different to zero and will quantify the similitude among the fourier transforms at that time instant. The matrix C is then defined as a function of τ only, and calculated using the correlation normalized by the norms of the two signals, that is

$$C(\omega, \tau) = C(\tau) = 1 - \left| \frac{\mathcal{E}\{\tilde{v}_1(\omega_k, \tau)^* \tilde{v}_2(\omega_k, \tau)\}_k}{\left(\sum_k \tilde{v}_1(\omega_k, \tau)^* \tilde{v}_1(\omega_k, \tau) \sum_k \tilde{v}_2(\omega_k, \tau)^* \tilde{v}_2(\omega_k, \tau)\right)^{1/2}} \right| . \quad (4.10)$$

As it can be seen, if the separated signals for that time slot are very similar, the quotient in the right side will be near to one, and C will be near zero for that time, thus the Wiener filter will have no effect. If the two signals are too similar, it means that the separation was not succesfull and thus if the Wiener filter is applied it will introduce too much distortion.

The short-time Wiener filter to improve source v_2, $H_{\mathcal{W},2}(\omega, \tau)$ is calculated in a similar way to (4.9), with the roles of v_1 and v_2 interchanged. After calculation of Wiener filter, the output signals are calculated as:

$$\widehat{y}_1(\omega, \tau) = H^1_{\mathcal{W}}(\omega, \tau)\widehat{v}_1(\omega, \tau) , \quad (4.11)$$

with a similar equation for the other source. Finally the time signals are obtained by inverse STFT.

To show the capabilities of this Wiener filter, assume that there is a sound field produced by white and stationary signals, with equal power from all directions. That is, suppose that the microphone array receives equal power from all angles and for all frequencies and times. In this case, the behaviour of the combined separation and Wiener filter process can be analized using the beampatterns, as the beampattern output will be the real power at the output of the separation, as a function of the arrival angle. Figure 4.2 shows the beampatterns obtained from the separation matrix in the bin corresponding to 2000 Hz for an example of 2-by-2 mixtures sampled at 8 kHz, with sources located at ± 26 degrees (for other frequencies the analysis is equivalent). The top row shows the beampatterns obtained from the separation matrix. For each beampattern, it can be seen that in the direction of each source, the gain is unitary

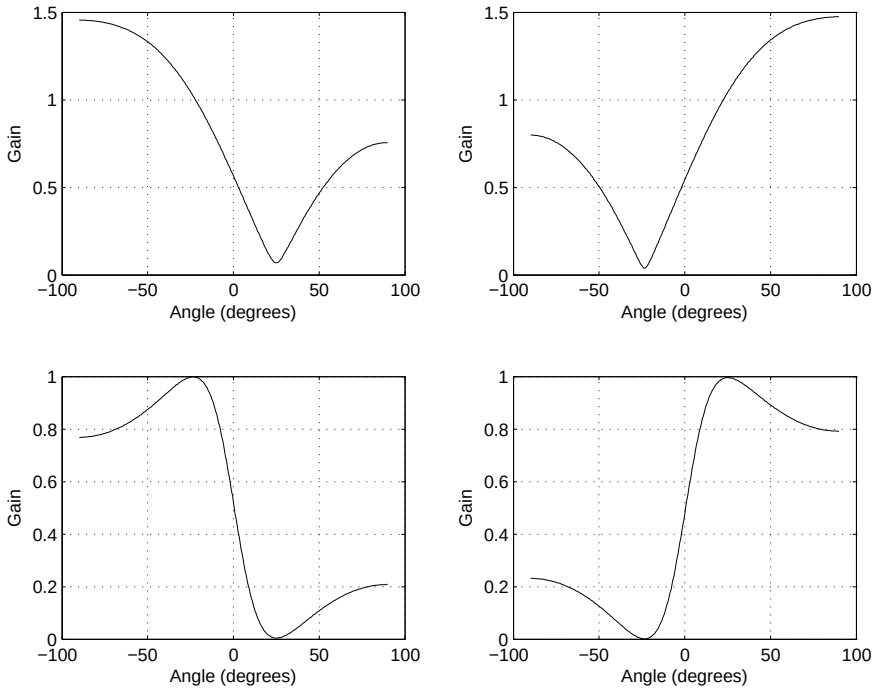


Figure 4.2. Effect of the Wiener postfilter on the beampatterns. a) the beampatterns generated from the separation matrix. b) the beampatterns after application of the Wiener filter.

(which is a consequence of the minimal distortion principle), and in the direction of the other source the gain tends to zero. In the bottom row, we have applied the equation of the Wiener filter to these patterns. That is, if the gains are $G_1(\theta)$ and $G_2(\theta)$, and as they are also the output amplitudes as a function of the angle, the first Wiener filter will be $G_1(\theta)^2/(G_1(\theta)^2+G_2(\theta)^2)$, and the same for the other filter. This is a way to visualize the approximate global effect of the whole processing. As it can be seen, the Wiener filter maintains unitary gain in the desired directions and nulls in the interference directions, but also produces attenuation in all other directions, which mitigates the effect of all echoes including both, those from the undesired noise (which improves separation) and these from the desired source (which reduces the reverberation).

Clearly, in real situations the input signals will be neither of the same power for all directions as assumed, nor white and stationary. Nevertheless, the signal with stronger component will in general come from the detected directions, with the echoes of lower power arriving from different directions, and thus the resulting effect would be even better than the depicted one. That is, the case shown in Figure 4.2 represents the worst scenery of possible inputs, and thus for more realistic cases an even better behaviour can be expected.

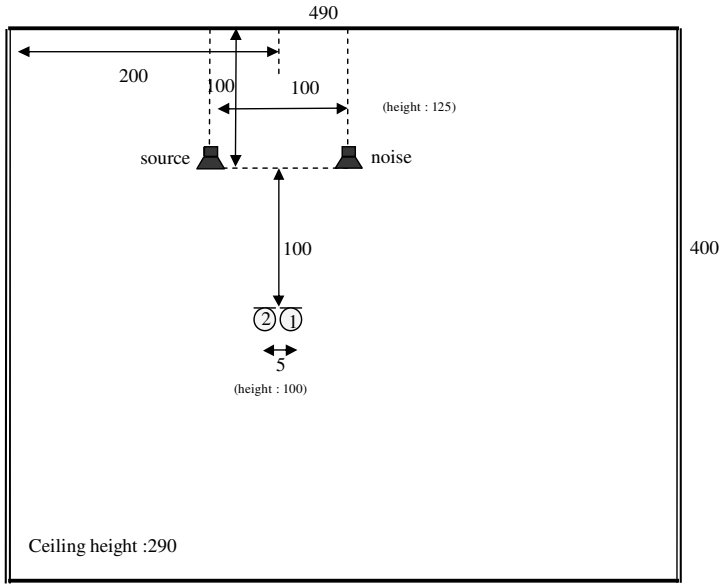


Figure 4.3. Experimental setup (all dimensions in cm).

4.4. Results and discussion

4.4.1. Experimental setup

In the precedent sections we proposed several modifications to the standard fd-ICA approach of [Murata et al., 2001] to improve the separation quality, aiming at three of the identified problems of the methods (initialization, permutation ambiguity solution and effects of reverberation). We have tested the capabilities of this algorithm using sentences from Albayzin Spanish speech corpus [Moreno et al., 1993]. From this large database, a subset of 20 sentences were selected. There are 5 sentences spoken by 4 speakers. Selected sentences were 3001, 3006, 3010, 3018, 3022, from speakers aagp and algp (female), and mogp and nyge (male). Those 20 sentences were recorded in a room according to Figure 4.3.

Two loudspeakers were used, one to reproduce desired speech source and the other to reproduce some kind of noise. The resulting sound field was recorded with two Ono Sokki MI 1233 omnidirectional measurement microphones, with flat frequency response between 20 Hz to 20 kHz and with preamplifiers Ono Sokki model MI 3110.

The interfering sources were of two kinds, speech and white noise. For speech noise, sentence 3110 and speakers aagp and nyge where selected. To contaminate female

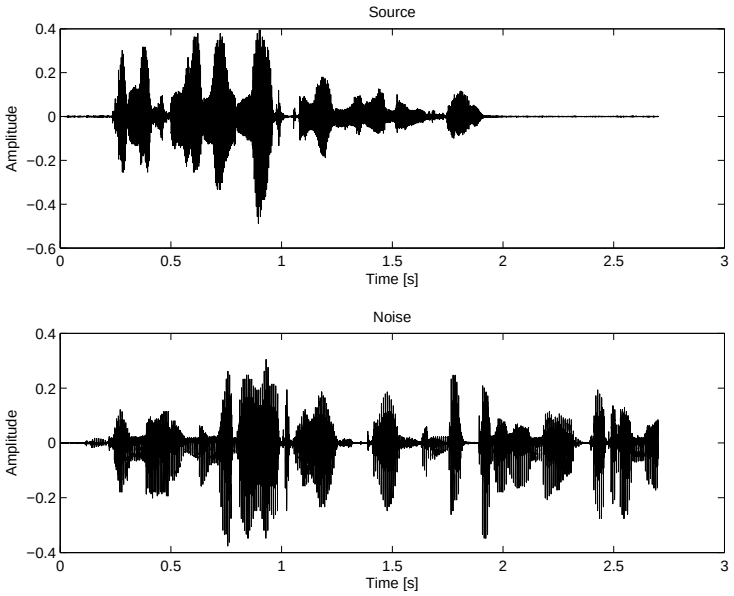


Figure 4.4. Example - Source signals. Top: aagp3002, desired source, female; bottom: nyge3110, noise, male.

utterances, male speech (nyge) was used, and vice versa. White noise from Noisex database [Varga and Steeneken, 1993] was used. For both noise kinds, two different signal to noise output power ratios were selected, 0 dB and 6 dB. A 0 dB output power ratio means that at speakerphones, both signals were replayed with equal powers. More data on this recording set can be found in Appendix A.2.

All recordings where made at 8000 Hz of sampling frequency with 16 bits quantization. The room used was a sound proof room, with additional plywood reverberation boards in two of the walls to increase reverberation times up to about 200 ms. The algorithm in all its variants was coded in Matlab language, and all tests were run in a Pentium 4 processor of 3 GHz of frequency, with 1 GB of RAM. For the JADE and FastICA methods, the original codes available at the respective authors web pages were used.

4.4.2. Separation example

To show an example of the algorithm output, the signal aagp 3002 mixed with speech noise at 0 dB was processed with the algorithm. Figure 4.4 shows the source signals (clean, before mixing), Figure 4.5 shows the resulting mixed signals as measured by microphones and Figure 4.6 shows resulting separated signals. This example shows a good separation with large noise reduction, even in a mixture with very strong

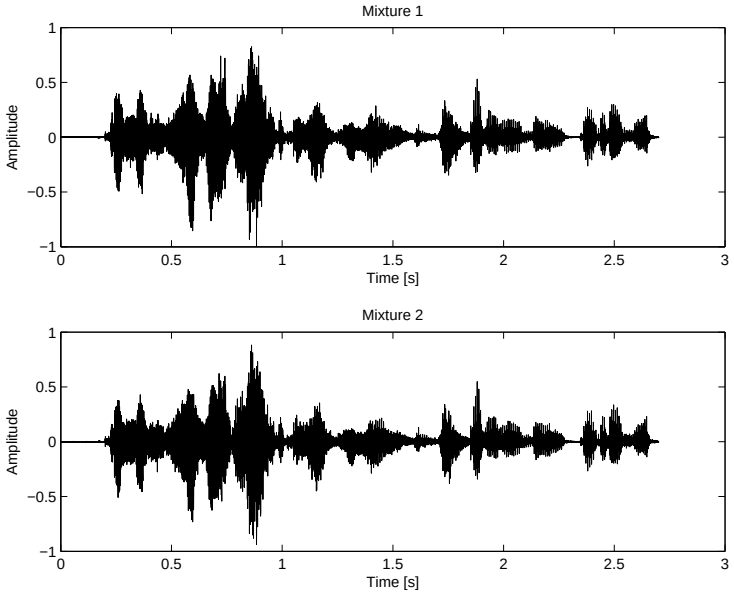


Figure 4.5. Example - Mixed signals. Top: microphone 1; bottom: microphone 2. Signal to noise power ratio: 0 dB.

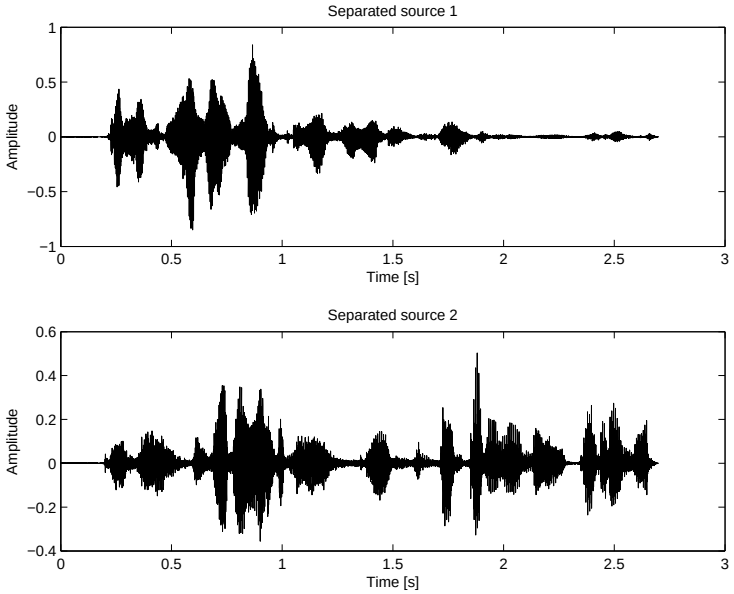


Figure 4.6. Example - Separated signals. Top: separated source 1; bottom: separated source 2.

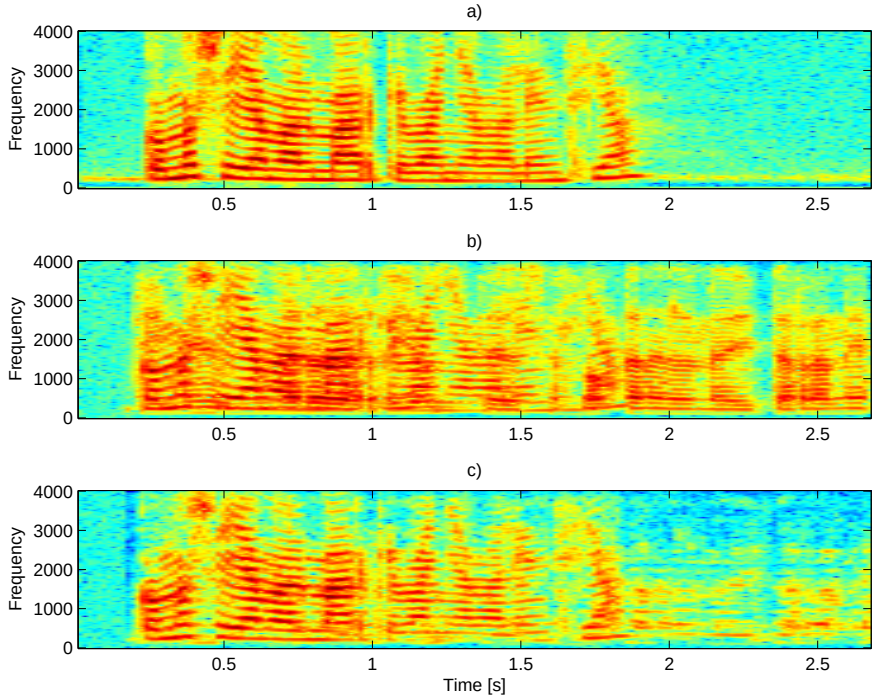


Figure 4.7. Spectrograms of a) source signal; b) mixed signal and c) separated signal, for a mixture of speech with speech interference emitted with equal power (power ratio of 0 dB).

noise. When listening to the output, the utterance by female speaker can be clearly distinguished, while male noise is heard as a very low volume background.

In Figure 4.7, spectrograms (in dB scale) for original clean signal, mixed signal and separated signal are shown. A large improvement can be seen, specially in the area where desired signal has low amplitude (towards the end), and also it can be seen how the main structures of the original signal are present in an enhanced way in the separated signal.

4.4.3. Beampatterns

A second interesting aspect to be evaluated is the kind of beampatterns generated by the method, and how the permutation and scaling solution affect them. To provide a general idea of this process beampatterns were generated at different stages of the method. Figure 4.8 presents in the top row the beampaterns generated for the two outputs before correcting any of the problems, and in the second row, the same but after correction of the scaling problem. It can be seen that the indeterminacy in scaling produces a rather bad pattern, with the high frequencies being overamplified. When

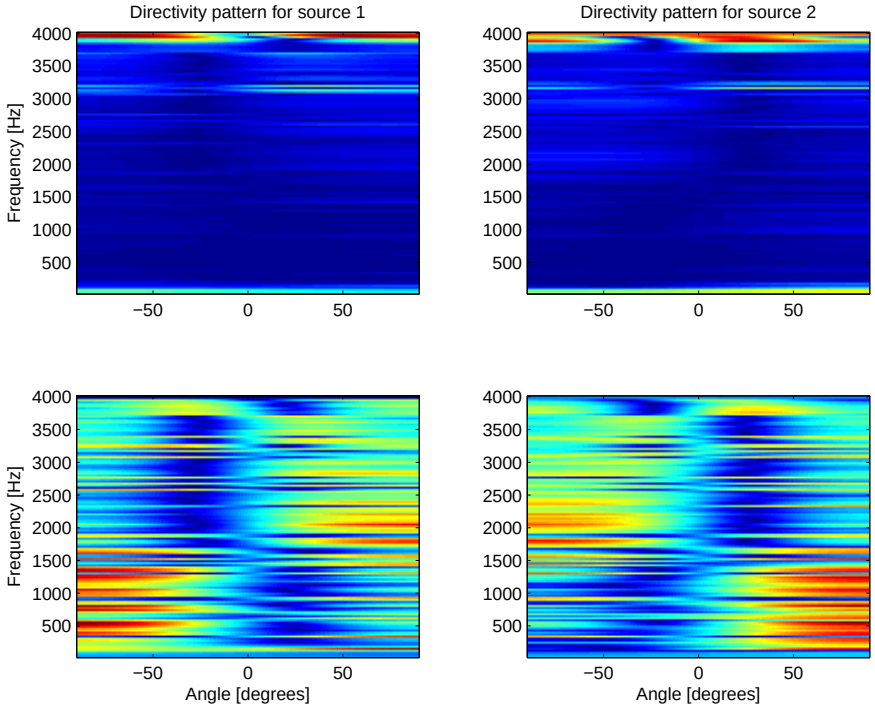


Figure 4.8. Beampattern for each source, before correction of permutations. Top row, beampatterns before correction of the scaling problem. Botton row, beampatterns after correction of the scaling problem.

the amplitude problem is corrected, the beampattern gain seems to be more uniform. Also note that the bins which suffer from the permutation problem can be clearly seen.

Figure 4.9 presents in the top row the beampatterns generated after correction of the permutation problem by the standard method of Murata et al. [2001]. As it can be seen, some of the previous existing permutations have been corrected, but now there are some additional permutations. That is, in this case the method was not able to solve succesfully the permutations. The second row shows the beampatterns generated after the correction of the permutation using the here proposed method. Clearly, most of the permutations have been corrected.

Note also two particularities of the beampatterns. One is that the null is not always in the same angle. According to the location of the sources, nulls should appear at about ± 26 degrees, but this angle fluctuates as a consequence of the reverberation and the nonuniform convergence of the ICA method. This is one of the disadvantages of the bin wise fd-ICA method, as explained in Chapter 2. In second place, there are some bins for which the beampattern does not have a well defined null (for example, in

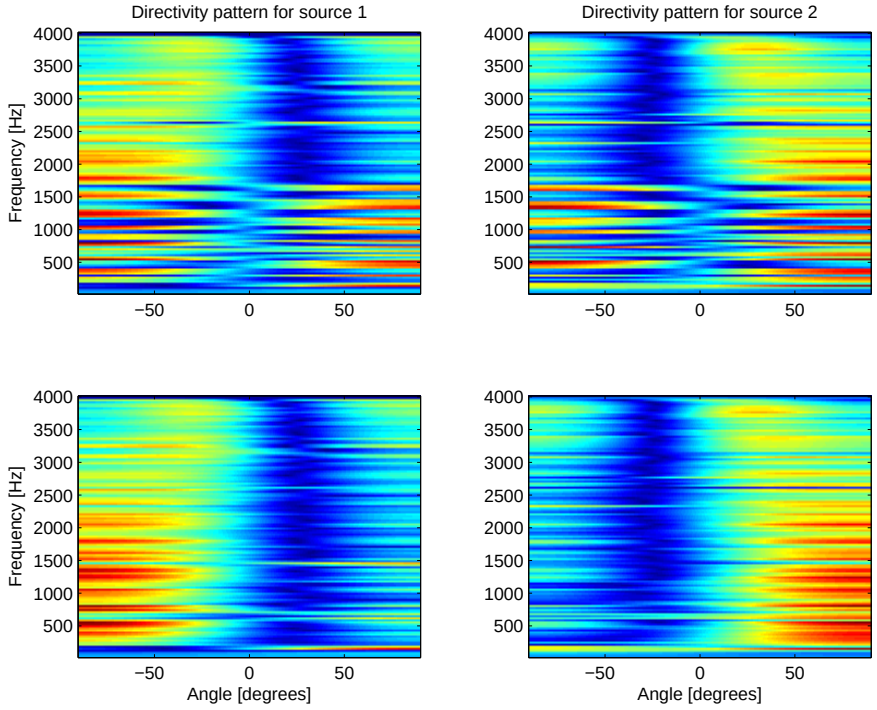


Figure 4.9. Beampattern for each source, after correction of permutations. Top row, permutations corrected by the standard method. Botton row, permutations corrected by the proposed method.

the bin around 2600 Hz), showing that the ICA method has failed for these bins. This is also another structural limitation of the method.

4.4.4. Determination of the algorithm parameters

In first place it was necessary to determine the best value to use for α, which controls the AR approximation of the accumulated envelope for the permutation solving method. The proposed algorithm was employed with five different values of α, and also compared to the use of the standard method for permutation solution [Murata et al., 2001](denoted as Std in the table), by means of the PESQ quality measure[1]. Table 4.2 shows the results when using $C(\omega, \tau) = 1 \, \forall \, \omega, \tau$ in the Wiener postfilter. As it can be seen, the best quality is obtained for a value of $\alpha = 0.2$. Also, it is shown that our method outperforms the standard one.

[1]Note that the only difference in the Std. method is in the permutation problem solution, but we use our proposed method for all the other stages, with JADE, FastICA, and Wiener postfilter

Table 4.2. Quality measured as PESQ scores when varying the α parameter in the permutation correction method. In this case, $C(\omega, \tau) = 1$.

Power	Noise kind	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	Std.
6 dB	Speech	2.74	2.89	2.89	2.89	2.77	2.73
	White	2.70	2.77	2.81	2.84	2.84	2.86
0 db	Speech	2.47	2.57	2.58	2.54	2.51	2.41
	White	2.40	2.37	2.43	2.47	2.45	2.47
Ave.		2.58	2.65	2.68	2.69	2.65	2.62

Table 4.3. Quality measured as PESQ scores when varying the α parameter in the permutation correction method. In this case, $C(\omega, \tau) = C(\tau)$.

Power	Noise kind	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	Std.
6 dB	Speech	2.74	2.87	2.88	2.88	2.76	2.73
	White	2.71	2.76	2.81	2.84	2.84	2.84
0 db	Speech	2.40	2.49	2.50	2.48	2.44	2.36
	White	2.39	2.36	2.42	2.48	2.46	2.50
Ave.		2.56	2.62	2.65	2.67	2.62	2.61

We also tested the use of the alternative calculation of $C(\omega, \tau)$ given in (4.10) instead of the constant values. Table 4.3 presents the results. It is evident that there is no significant difference among the two options and that $\alpha = 0.2$ is still the best option, although the variable C produces a little degradation of quality. This is caused mainly by the time periods where there is only one active signal (particularly, at the beginning and the end of the signals). Although the weigths try to evaluate when the separation failed, whenener there is only the undesired signal the value of C will be near to zero (as for that time both signals will contain only contributions of the undesired signal) and the Wiener postfilter will have little effect, degrading the performance of the method. Given these results, in the following experiments the value of $C(\omega, \tau) = 1$ will be used.

To evaluate whether these results are independent of the Wiener postfilter usage, the experiment was repeated using $C(\omega, \tau) = 1$, but without the Wiener postfilter. Table 4.4 shows the results. It can be seen again that our method outperforms the standard permutation solution, and also that the value of $\alpha = 0.02$ is again the optimum. The difference with the standard method for permutation solution is not too large, however. This seems to show that the combination of our permutation solution

Table 4.4. Quality measured as PESQ scores when varying the α parameter in the permutation correction method. In this case, $C(\omega, \tau) = 1$ and we do not use the Wiener postfilter.

Power	Noise kind	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	Std.
6 dB	Speech	2.59	2.67	2.68	2.67	2.59	2.66
	White	2.60	2.63	2.65	2.68	2.68	2.69
0 db	Speech	2.36	2.42	2.40	2.40	2.37	2.33
	White	2.24	2.20	2.25	2.30	2.30	2.32
Ave.		2.45	2.48	2.49	2.51	2.49	2.50

with the Wiener postfilter has a potentiating effect on the separation quality.

Another interesting aspect to note from the results in Tables 4.2, 4.3 and 4.4 is that, in the case of white noise, the results for all versions of permutation solutions, including the standard one, are quite similar showing a very little variation. This is due to the fact that the white noise spectrum envelope will always be very flat, and thus the clasification of envelopes will be easier no matter how the accumulated envelope is determined.

Once determined the best α to use, we wanted to evaluate the effect of the JADE inicialization approach. For this, we tested the method, using our proposed permutation solution, with respect to the initialization and also to the use or not of the postfilter. Two indexes were used to evaluate the quality: the PESQ score as presented in Chapter 3, and the average processing time. We compared the results for the standard version using either only JADE (J) or only FastICA (F) or JADE as initialization followed by FastICA (JF) or our whole approach including the Wiener Postfilter (JFW).

For all the algorithms our proposed method was employed to solve the permutations. For the permutation correction algorithm we used the best value of $\alpha = 0.2$ and $C(\omega, \tau) = 1$, and the order of the lowpass filter was fixed to $N_{env} = 40$ (we found that this number has not a large effect on the quality of the permutation solution, the important thing is to produce some smoothing of the envelopes). We use a lowpass filter with all the coefficients with the same weight (that is, a simple averaging).

Table 4.5 presents the PESQ scores for the different approaches. Note that the combined Jade-FastICA method produces an improvement with respect to using either one of the approaches alone. On the other hand, the Wiener postfilter produces an important increment of the separation quality. This is related to the enhancement in separation and some reverberation reduction, due to the elimination of the residual echoes from directions different that the desired one, as discussed in Section 4.3 and Figure 4.2.

Table 4.5. Quality measured as PESQ scores for the different methods.

Power	Noise kind	J	F	JF	JFW
6 dB	Speech	2.49	2.69	2.67	2.89
	White	2.35	2.67	2.69	2.84
0 db	Speech	2.35	2.35	2.40	2.54
	White	2.24	2.28	2.30	2.47
Average		2.36	2.50	2.51	2.69

Table 4.6. Average processing times for the different methods.

Power	Noise kind	J	F	JF	JFW
6 dB	Speech	5.18	11.41	11.82	12.79
	White	5.55	10.15	10.33	10.71
0 db	Speech	5.39	11.63	11.91	11.96
	White	5.58	17.55	17.19	17.51
Average		5.42	12.68	12.81	13.24

Table 4.6 presents the average processing time for the different cases. As it can be seen, the increase in separation quality is obtained at a cost of increased processing time. The average duration of the utterances was 3.55 seconds, showing that the processing speed is about 4 times slower than requiered for real-time separation. Also note that the time used in JF is just a little longer than the one of FastICA alone. This shows that using Jade indeed provides a good initialization that stabilizes the FastICA method, which then converges faster to produce a total time equivalent to that of FastICA only but with higher quality.

4.4.5. Speech recognition evaluation

The performance of the algorithm for ASR was tested by means of a speech recognition system to estimate the improvement on word recognition rate before and after separation. For this test, a continuous speech recognizer based on tied Gaussian-mixtures Hidden Markov Models (HMM) was used. This recognizer was trained with 585 sentences from Minigeo subset of Albayzin database (the training set did not include any of the sentences used in the test).

After training, the recognition system was tested with the clean sentences of our test set. To study how the mixing process degrades the recognition output we also evaluated the recognition performance over the mixtures. Then we applied the separation

Table 4.7. Word recognition rate in robust speech recognition. For reference, with clean sources $WRR\% = 93\%$.

Power	Noise kind	Mixtures	J	F	JF	JFW
6 dB	Speech	44.50	77.00	85.19	82.50	84.00
	White	19.50	48.00	77.50	78.00	84.50
0 db	Speech	30.00	66.50	63.50	63.50	79.50
	White	7.20	30.50	42.00	47.50	69.00
Average		25.30	55.50	67.05	67.88	79.25

algorithms of the previous evaluation, using either only Jade (J) or only FastICA (F) or both methods (JF) or using both methods combined with the Wiener postfilter (JFW), in all cases using our proposed method for permutation solution with $\alpha = 0.2$. We also tested the WRR of the mixtures, that is when the recognizer is fed with the raw data without any separation. Table 4.7 shows the results of these tests. As it can be seen from these results, recognition rate improvements are in the order of 30 to 60%. Also it should be noted that the recognition rates in some cases are only 9% below those of the clean sources.

4.5. Concluding remarks

The separation method presented provides an improvement in quality respect to the basic standard methods of fd-ICA. The use of FastICA produces good quality results, and the initialization by JADE yields a more stable separation method. The modification in the permutation alignment method provides a good improvement of quality, due to its better capabilities to solve the permutations. The Wiener postfilter provides an important increase of separation quality, while at the same time produces some reduction of reverberation, an aspect that standard fd-ICA algorithms cannot do. The good quality of the resulting separation was verified using an objective quality measure (PESQ) and the application of interest (ASR), showing an important improvement of the recognition rate. The main drawback of this approach is its computational cost: the processing time was several times that one of the duration of the signal to process. If real-time or almost real-time operation is desired, more efficient (in terms of processing speed) methods are needed. The advances of this algorithm were presented in two publications, [Di Persia et al., 2006a] and [Handa et al., 2006].

Chapter 5

Multi-resolution BSS

In the previous chapter several improvements to the standard fd-ICA methodology were introduced, in order to improve the separation quality of the resulting signals. Altough this objective was achieved, the main drawback of the presented approach was the computational cost. Even as the computation time is relatively small, it is several times the duration of the signals to process. This makes impossible the application of the method for real-time processing. Also, there was the problem of residual permutations.

In order to improve the computational cost and the permutation aspects, in this chapter a solution is proposed that explores a simplified alternative mixing model. The simplifications introduced make the algorithm very fast, and moreover, they eliminate the need of permutation correction since the method is inherently permutation-free. Moreover, the separation stage is followed by a Wiener filtering stage which improves the separation and dereverberation. This process is iterated a number of times using different window lengths in the STFT, which produces an improvement at multiple resolutions. In the following sections a detailed presentation of this method will be made. First the general idea of the multirresolution algorithm will be presented. Then, the fast separation stage based in a simplified mixture model will be introduced. This stage is the key to producing a fast algorithm. Then a postfiltering stage will be explained which improves the separation. Next, some implementation isuued are discussed. The chapter ends with results and discussion of them.

5.1. Introduction

The separation methods discussed up to now have the dissadvantage of a high computational cost. Also, there was the problem of permutations. Although the solution proposed in Chapter 3 allows for a good permutation correction that performs better

that previous approaches, for longer reverberation times the number of bins with residual permutations is increased. This is due to the smearing effect of the reverberation, which smooths the envelopes and makes more difficult its characterization. The ideal method of separation would be to completely avoid the permutations to occur. There have been some attempts to do so, as explained in Chapter 2. In [Parra and Spence, 2000], the separation filters are transformed to the time domain, then the coefficients after some index are made zero, and transformed back to frequency. This method is performed in each iteration, and the amount of nonzero coefficients usually is much smaller than the length of the filters (which is the window length used in the STFT). This constrain produces a kind of coupling on the separation matrices for all bins, which avoids the existence of permutations. Another approach [Lee et al., 2007] uses an explicit multivariate pdf for all frequency bins, which are directly coupled and thus the separation matrices need to be updated for all bins at the same time. Nevertheless all these approaches also have a high computational cost.

The methods analyzed up to now considered only one fixed possibility for the window length for the STFT. It must be clear that different selections of window lengths would produce different input data for the separation methods. The key idea of this chapter is how can one exploit automatically the window length to produce a better separation. A simple idea would be to use the method proposed in the previous chapter for separation, but with different options of window length, and keep the result that produces better separation according to some measurement. The main disadvantage of this approach would be that the separation method is quite slow, thus performing separation with let us say, five different window lengths, would produce an algorithm five times slower. It is clear that this approach can be usefull only if there is an enough fast separation method available.

A second idea to improve the separation would be to iterate this process. Each time-frequency representation using different window lengths (and thus, different time-frequency resolution) has a different set of data over which the separation can work. In this way, if some information was not successfully separated at a given resolution, perhaps at a different one it can be. This leads us to the proposition of an iterative approach based in the steps presented in the block diagram of Figure 5.1.

5.2. Multi-resolution BSS algorithm

The key point in this method will be to have some fast way to determine, at each iteration, which window length (and thus, which time-frequency resolution) will yield the best results. This will be done using a simplified separation method, that is fast enough to allow the scheme deligned before. That is, to perform separation over different time-frequency representations and to select the one that yields best results.

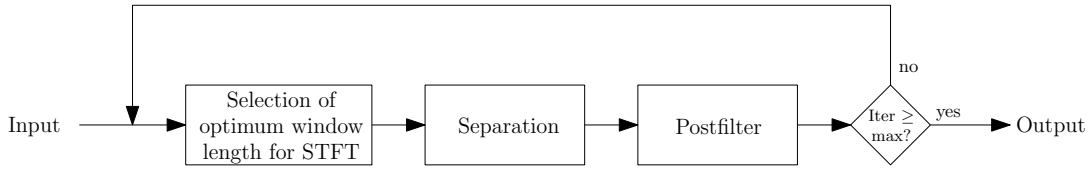


Figure 5.1. Block diagram of the proposed iterative method.

The selection of the best window length will be made by evaluation of the correlation coefficient among the two outputs. A low value of this coefficient will indicate a better separation.

After the best window length was selected, the separation at that time-frequency resolution is performed, and a Wiener postfilter is applied. In this Wiener postfilter, the correlation coefficient is used as a weigth to reduce the effect of a deficient separation (as will be seen later). The algorithm can be expressed as in figure 5.2.

The name of this method should be now clear: as we use a different window length at each iteration, each STFT uses a different resolution in the time-frequency plane. Small window length implies a better time resolution but a worst frequency resolution, and vice versa.

5.2.1. Simplified separation stage

For the development of this algorithm we start by stating the hypothesis used. We assume, as usual, an equal number of sources and microphones ($M = N$) and the statistical independence of the original sources. As already explained in Chapters 1 and 2, assuming that the room behaves like a LTI system, the mixture can be written as:

$$x_j(t) = \sum_{i=1}^{M} h_{ji}(t) * s_i(t) \,, \tag{5.1}$$

where $x_j(t)$ is the j-th microphone signal, $s_i(t)$ is the i-th source, $h_{ji}(t)$ is the impulse response of the room from source i to microphone j, and $*$ stands for convolution. Taking a short-time Fourier transform (STFT) of the previous equation, the convolution becomes a multiplication as explained in the previous chapters, thus we have the mixture model in the time-frequency domain:

$$\mathbf{x}(\omega, \tau) = H(\omega)\mathbf{s}(\omega, \tau) \tag{5.2}$$

This equation shows that for each frequency bin the mixture can be considered instantaneous, in such a way that a simple mixing matrix can be estimated. If one

1. *Select a set $\mathcal{W}_L$ of window length candidates with N_L possible values.*
2. *For each window length in the set $\mathcal{W}_L$,*
   - *Perform separation by the fast method.*
   - *Reconstruct the time domain signals using the overlap-and-add method.*
   - *Calculate the correlation coefficient among the time-domain separated signals.*
3. *Select the window length that produced the lower correlation coefficient.*
4. *Perform separation again, using this optimum window length.*
5. *Apply the Wiener postfilter.*
6. *Reconstruct the time domain signals.*
7. *Remove the window length used from the list of candidates $\mathcal{W}_L$.*
8. *Return to step 2, until the desired number of iteration is reached.*

Figure 5.2. Multiresolution BSS algorithm.

makes another simplification on this equation, assuming that the mixing matrix is independent of frequency, then the problem can be reduced to find only one mixing system, as expressed in the following equation:

$$\mathbf{x}(\omega, \tau) = H\mathbf{s}(\omega, \tau) \tag{5.3}$$

Each component of this vector equation can be thought as a mixture using mixing matrix H of several two-dimensional signals, and for that it is equivalent to the instantaneous mixture of images, with the exception that the signals are complex-valued. We will present and analyze the method for the case of two sources and two mixtures, the generalization being straightforward.

To perform separation, the problem is to find a separation matrix W such that the separated signals result as independent as possible. Since in this case the signals used are two-dimensional, they need to be scanned to one dimension. There are several possible scanning alternatives, but if the sources correspond to iid random signals, the scanning should not alter the results. We have performed a scanning in which, for each

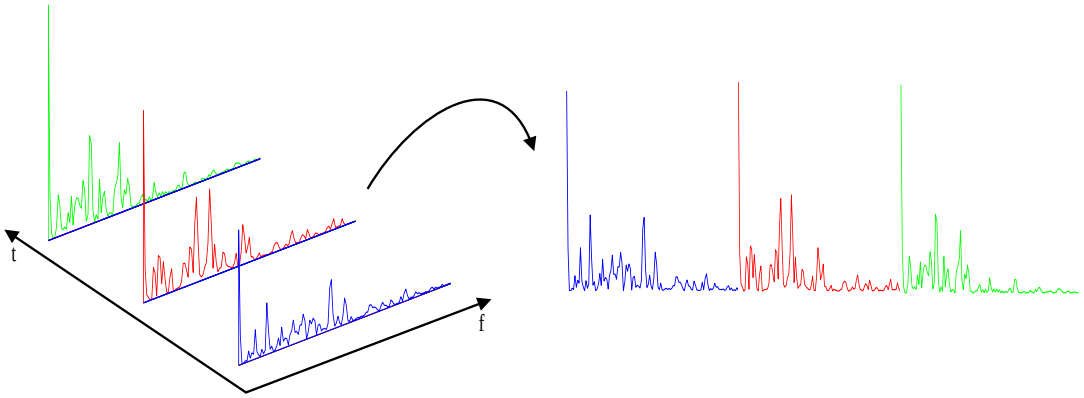


Figure 5.3. Stacking of the information of succesive windows to form a long data vector.

time, we stack all the frequency values of the signal, and proceed to the successive time index to form a long vector for each input and signal. Figure 5.3 shows schematically this process for the time-frequency representation of one signal.

The size of this vector will be variable, depending on the overlap factor used in the STFT. If the transform is used with a step size of half the window length, this will produce a doubling in the amount of data, but as it is only necessary to perform separation for the values corresponding to positive frequencies (the values for negative frequencies are the complex conjugates of the positive ones), in this case the amount of data will be equal to the length of the original signal. If more overlapping is used, a larger amount of data will be produced, most of it very redundant.

For the separation, the JADE algorithm as presentend in Capter 2 was used. This method was selected because it can perform the separation quite fast, even if the data vector is very long, as will result in our method. Calling W the separation matrix estimated by JADE, the separation is expressed by the equation:

$$\mathbf{z}(\omega, \tau) = W\mathbf{x}(\omega, \tau) \tag{5.4}$$

Alternatively, FastICA algorithm can be used to estimate W, altough it will make the whole algorithm slower, because FastICA is computationally more expensive than JADE, although we can expect a better separation quality when using this method. We can also use a combination of both algorithms, to give more robustness to the method, as was done in the previous chapter.

The separated signals will still have permutation and amplitude indeterminacies. The amplitude indeterminacy is solved as proposed by [Murata et al., 2001], that is, application of the minimal distortion principle [Matsuoka, 2002], as was done in Chapter

2. The permutation needs not to be corrected at this level: as the indeterminacy applies to the whole signals, its only effect is that one cannot know to which source belongs each final separated signal, but there will not be permutations among the frequency bins, yielding a consistent time-frequency representation for each source. Furthermore, note that this method will have a large amount of data to learn the separation matrices, and thus the step size used in the STFT can be relatively larger than in the standard fd-ICA methods, where usually a small step size is employed to increase the amount of data available at each bin [Araki et al., 2003].

5.2.2. Postprocessing stage

After applying separation, two STFT that will in the ideal case correspond to the two independent sources are obtained. However, as this separation was performed on a simplified model, it will not be perfect and some of the information from the other source will remain, although attenuated. To further reduce the cross-channel contents, we use a time-frequency Wiener filter [Huang and Benesty, 2004] as stated in the previous chapter . In the present case two signals, $z_1(\omega,\tau)$ and $z_2(\omega,\tau)$ will be obtained and if the separation process was good enough they could be used as estimation of the clean sources. So in order to eliminate residual information from source z_2 on source z_1 we can use the short-time power spectrum of z_1 as numerator (estimation of clean source) and add short-time power spectrum of z_1 and z_2 as the estimation of noisy power spectrum in denominator of the Wiener filter. In the same way as in the previous Chapter (Section 4.3) we will use a weight in the denominator to compensate for the imperfect separation, as shown in the following equation

$$H_{\mathcal{W},1}(\omega,\tau) = \frac{|z_1(\omega,\tau)|^2}{|z_1(\omega,\tau)|^2 + (1-\alpha)\,|z_2(\omega,\tau)|^2}\,, \tag{5.5}$$

where the weighting constant α measures the similarity among the separated signals, with values between 0 and 1. The purpose of this constant is to reduce the effect of the Wiener filter if the separation stage was defective. If the separated signals $z_1(t)$ and $z_2(t)$ are very similar, the weight $1-\alpha$ must be near to zero, otherwise it must be near to one. In this way, when the reconstructed signals are too similar, the Wiener filter has little effect, reducing the distortions.

There are several ways to set those weights, the easier one is to assume that the separation will be good enough and setting α to zero. In this dissertation we use the square root of the normalized correlation coefficient

$$\alpha = \sqrt{\frac{\langle z_1(t), z_2(t)\rangle}{\|z_1(t)\|\|z_2(t)\|}}\,, \tag{5.6}$$

where $z_i(t)$ are the time-domain signals. Note that this coefficient was already calculated as a measure of similarity, to select the optimum window length, and thus it is not neccesary to calculate it again.

The Equation 5.5 is the time-frequency Wiener filter to improve source z_1. A similar equation is used for the other source, with the roles of z_1 and z_2 interchanged. After calculation of Wiener filter, the output signals are calculated as

$$y_1(\omega, \tau) = H_{\mathcal{W},1}(\omega, \tau) z_1(\omega, \tau) \,, \tag{5.7}$$

with a similar equation for the other source. Finally the time signals are obtained by inverse STFT.

5.2.3. Implementation issues

Some aspects of this algorithm need to be clarified. In first place, the algorithm can select the optimum widow length to use. But this capability produces an increment of the computational cost. Instead of just one optimization for each iteration, there are a number N_L of optimizations performed. As each step requires transforming and antitransforming with different windows lengths, this requires also the additional calculation of N_L extra STFT and ISTFT.

Also, note that during the selection of the window length, the separation result in the time-frequency plane can be saved to be used in the final separation stage. This, however, would require an amount of additional memory equal to N_L times the amount needed to save one separation result, as the cases for all window sizes need to be saved. As an alternative, the separation in time-frequency for the optimum window length can be recalculated, without saving the intermediate results. This would require less memory, but will add one whole separation stage.

The other parameter that can be controlled is the window overlapping. Some overlapping is needed to assure that inverse STFT exists. At the same time, using too much overlapping will increase redundancy in the transform, which is not good for separation where the sources are assumed to have no temporal correlations. So a moderate amount of overlapping must be used. Through some preliminary experiments we have determined that this parameter has not a strong influence on the method, and a 50% overlapping is a good compromise that allows for reconstruction and at the same time provides good separation.

5.3. Results and discussion

There are some aspects of this algorithm that need to be addressed. We need to check the behaviour of this separation algorithm for the different parameters it

depends on. There are basically two aspects that need to be considered: the number of iterations needed for good separation, and the kind of separation algorithm that works better (JADE, FastICA or a combination of them).

For the tests we used the framework proposed in Chapter 3, with real mixtures. We selected 5 sentences from Albayzin database [Moreno et al., 1993], uttered by two male and two female speakers, for a total of 20 utterances. This set was mixed with two kind of noises, white noise and competing speech, using the room setup presented in Figure 5.4. The room used had a reverberation time of $\tau_{30} = 195$ milliseconds. Two power ratios were used, 0 dB and 6 dB. More data on this recording set can be found in Appendix A.2. The algorithm was programmed in Matlab. The signals used for test were sampled at 8 kHz.

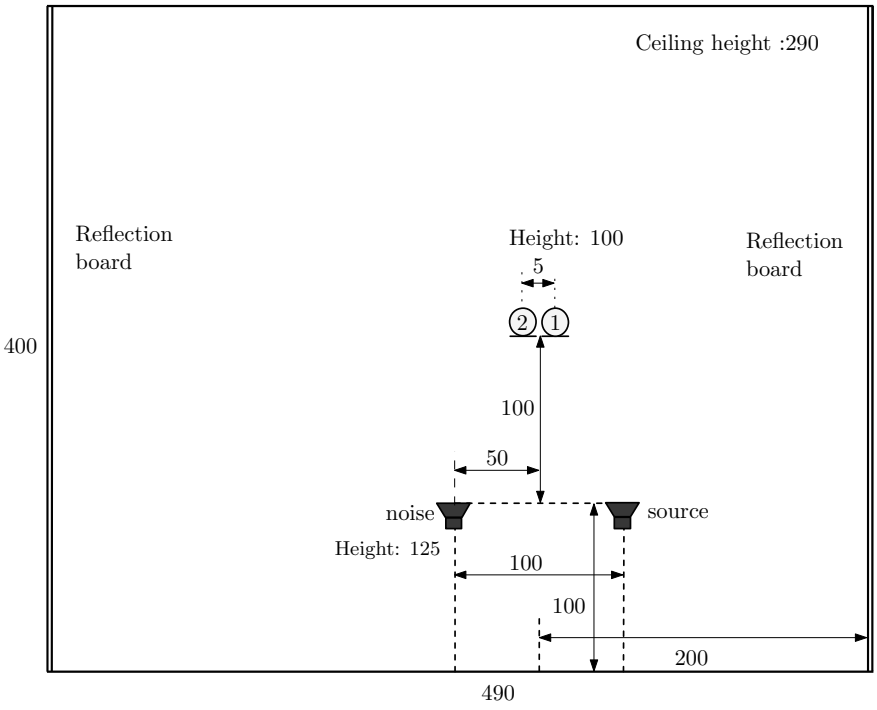


Figure 5.4. Experimental setup used in experiment.

5.3.1. Determination of the set of window lengths $\mathcal{W}_L$

First we want to determine the best set of windows lengths to use. As already explained, each window length in the set increases the computational cost (and also the

Table 5.1. Selection of optimal window length. The value shown is the number of times each length was chosen as the optimal.

Power ratio	Noise	4096	2048	1024	512	256	128	64	32
6 dB	Speech	0	1	0	7	5	3	2	2
	White	0	2	1	2	2	6	4	3
0 dB	Speech	0	0	0	4	5	5	3	3
	White	0	3	4	5	4	4	0	0
Total		0	6	5	18	16	18	9	8

memory requirements, if the separation result is saved for each tested length). In this way N_L should be kept as small as possible. To evaluate this, we used the proposed algorithm with JADE, only one iteration, and fixed the candidate windows length as $\mathcal{W}_L = \{4096, 2048, 1024, 512, 256, 128, 64, 32\}$ samples. Table 5.3.1 shows how many times each window length was chosen as the optimum length, for each kind of noise and power ratio.

This table shows that, in general, for speech noise a shorter window length yields better results, meanwhile for white noise the algorithm performs better with a larger window length. We wanted to keep N_L as low as possible so we reduced the set to half its size, selecting the four values of window length that were chosen the larger number of times as optimum lengths. In this way, the selected set is $\mathcal{W}_L = \{512, 256, 128, 64\}$, whose members were selected as optimum the 76.25% of times.

5.3.2. Determination of the iteration number and ICA algorithm to use

The next test was aimed to detect the best number of iterations to use. We have evaluated the quality measured by PESQ and also the average separation time for each case. As each iteration includes a Wiener filter, it is not good to have a high number of iterations because each Wiener filter application will produce some amount of distortion, by reduction of desired componets that were not successfully separated. We tried the three different methods with JADE (MR-J), with FastICA (MR-F) and with combined JADE and FastICA (MR-JF), and three values for the number of iterations (1, 2 and 3 iterations). We also added the results for a standard fd-ICA method using FastICA as presented in table 4.1 in Chapter 4 (Std-F). Table 5.2 presents the PESQ evaluation results for these cases, and Table 5.3 the average processing time.

Table 5.2. Quality measured as PESQ scores for different options of ICA method and number of iterations.

		MR-J			MR-F			MR-JF			Std-F
Power Ratio	Noise	1	2	3	1	2	3	1	2	3	
6 dB	Speech	2.43	2.37	2.27	2.45	2.42	2.23	2.45	2.44	**2.48**	2.51
	White	2.56	2.56	2.45	**2.57**	2.56	2.42	**2.57**	**2.57**	2.42	2.57
0 dB	Speech	**2.37**	2.35	2.28	2.18	2.13	2.03	2.18	2.17	2.11	2.35
	White	2.28	**2.33**	2.19	2.25	2.28	2.18	2.25	2.28	2.18	2.31
Average		**2.41**	2.40	2.30	2.37	2.35	2.21	2.37	2.36	2.30	2.43

Table 5.3. Average processing times in seconds, for different options of ICA method and number of iterations.

		MR-J			MR-F			MR-JF			Std-F
Power Ratio	Noise	1	2	3	1	2	3	1	2	3	
6 dB	Speech	0.70	1.14	1.52	7.61	11.42	11.32	8.15	8.99	9.76	7.60
	White	0.74	1.22	1.58	1.98	6.93	7.12	2.44	3.53	4.40	6.00
0 dB	Speech	0.73	1.17	1.52	7.46	10.84	11.35	10.49	10.45	11.12	5.88
	White	0.69	1.24	1.67	2.16	8.13	12.35	2.75	3.91	4.91	9.14
Average		0.72	1.19	1.57	4.80	9.33	10.53	5.96	6.72	7.55	7.16

Regarding the ICA method, JADE seems to outperform the FastICA and the combined JADE+FastICA methods, both with respect to resulting PESQ score and processing time. In particular, JADE algorithm is very fast, taking in average only 0.72 seconds to separate signals that last for an average of 3.55 seconds, that is, the algorithm can produce results in a fraction of the real time. The better behaviour of JADE in front of FastICA seems to be related to the robustness of each method. The simplification we made in the mixing model is that the mixing matrix was only one, but in reality there are one for each frequency bin, and they are not necessarily similar. Due to this, by packing all data in one vector we are mixing data generated with many different mixing conditions. JADE algorithm seems to be more robust to this subject and it produces a rather better general estimation, while the FastICA estimation seems to be not so good. In comparisson with the results for a standard fd-ICA approach, the PESQ obtained by our method with JADE is equivalent, but our method can produce this result in only a fraction of the time of the standard method.

With respect to the number of iterations, only one iteration seems to produce a better PESQ, and of course it produces a reduction in processing time. When heard and compared, the increase on the number of iterations produces an improvement in separation, being the competing noise less heard, but at the same time it introduces more distortions and musical noises, and for this reason the PESQ score produces a worst result. From a subjective point of view, it produces a better separation which allows a better understandig of the speech (that is, it seems to improve the intelligibility), although at the same time it introduces distortions. Nevertheless, it can be usefull to use two iterations in applications where the important aspect is intelligibility. In that case, the reduction of competing noise will be a gain and some musical noise can be tolerated.

Figure 5.5 shows this phenomena. In the first row, a clean speech signal and its spectrogram are shown. In the second row, the result of the separation, with only one iteration, is presented. The third row shows the separation with two iterations. In the time domain signals it can be seen that with two iterations, most of the competing source at the end of the signal have disappeared, which produces a better separation as less interference is heard. But in the spectrogram, it can be seen that the cost is a deterioration of the frequency contents, specially at low frequencies, that produces distortions negatively affecting the PESQ and heard as musical noise.

The main cause of this effect is the constant directivity matrix. Analyzing this effect from the beamformer interpretation, the beampatterns generated by this approach are not uniform with frequency, that is, instead of having a null pointing at a constant angle, the null location varies with frequency. Figure 5.6 shows the beampattern for the algorithm after one iteration, with JADE method. The desired sources are located at about -26 and 26 degrees, and so the optimum beamformer pair should put a

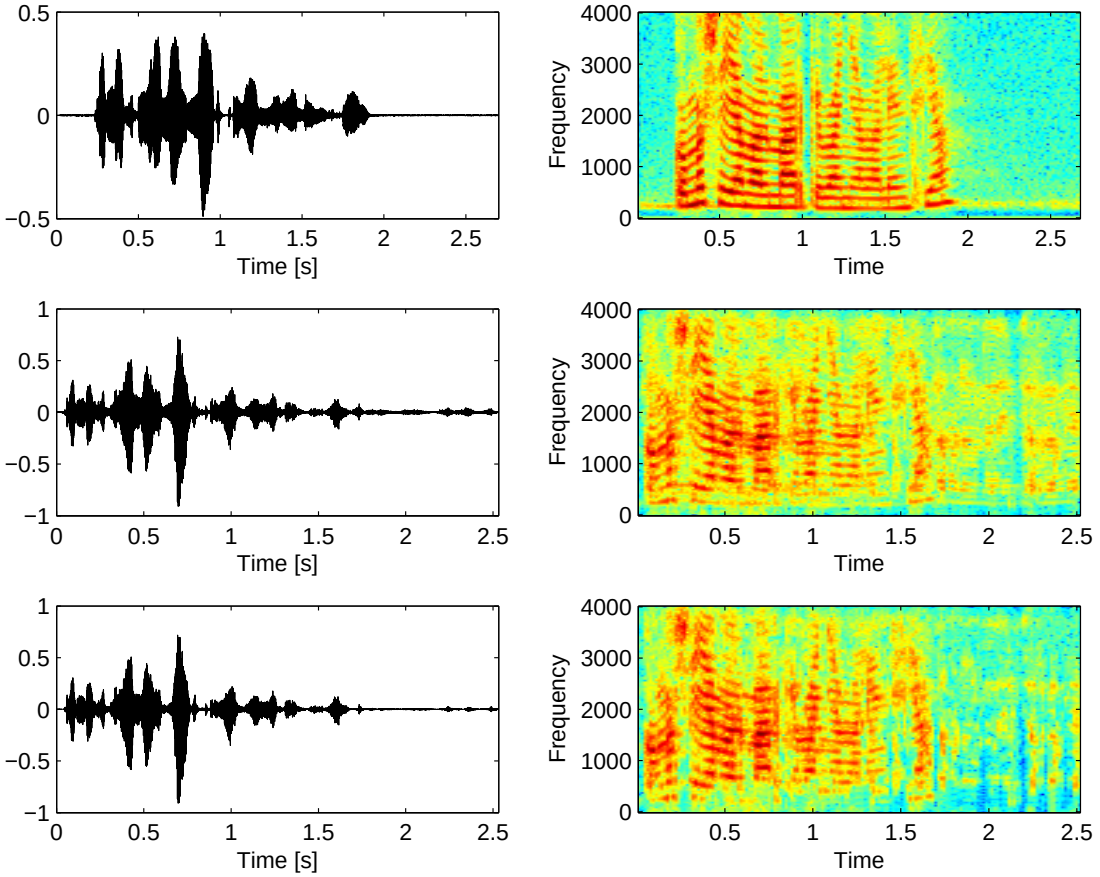


Figure 5.5. Effect of the iteration number on the separation.

null in these directions. But as the separation matrix is kept constant with frequency, the beampatterns show the null at different locations. In the figure, the color scale represents the gain of the beamformer.

The results for the combined JADE and FastICA and for FastICA alone are very similar, but the combined one takes less time. This is consistent with the idea that a good initialization by JADE will speed up the FastICA convergence, although due to the less robustness of FastICA in the present case, this will not produce an improvement of the separation quality.

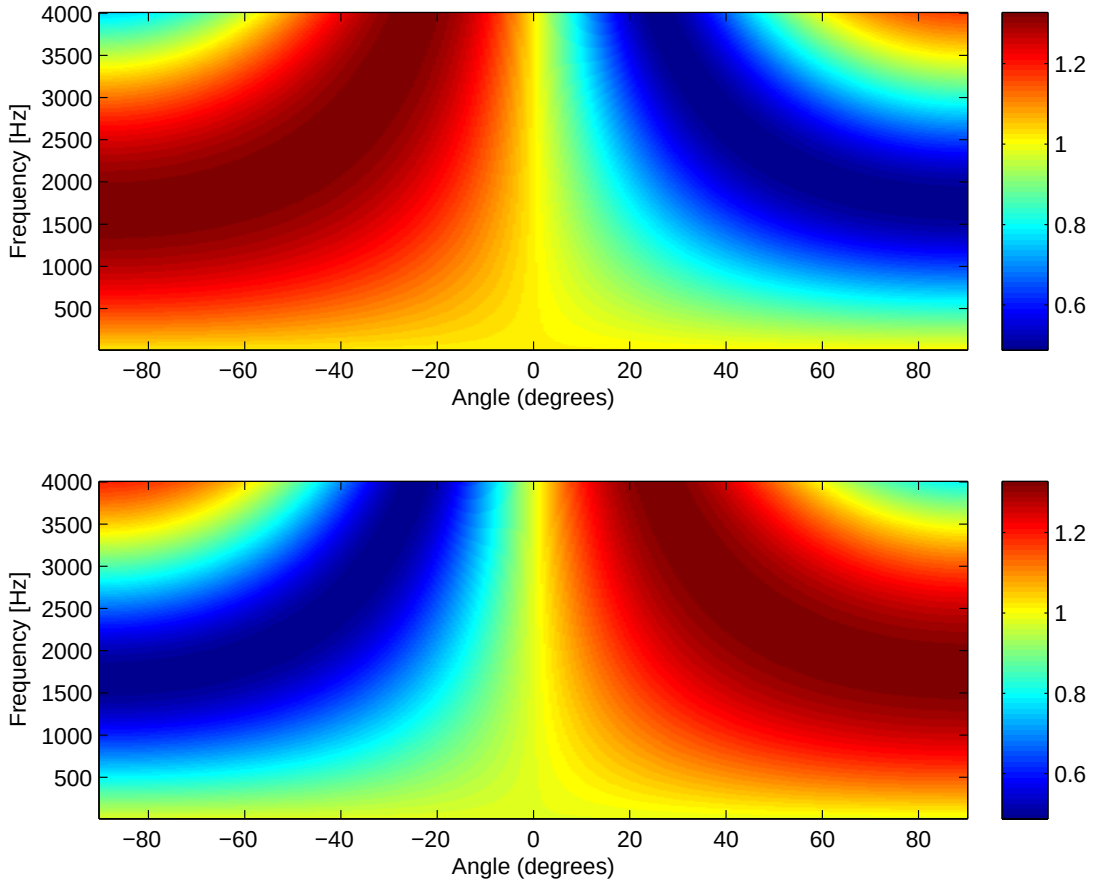


Figure 5.6. Beampatterns generated with fixed separation matrices.

5.3.3. Speech recognition results

Finally we present an evaluation of the speech recognition for the evaluated algorithms. We used a recognizer trained with the minigeo subset of Albayzin database [Moreno et al., 1993]. The recognizer used was the same as in Chapter 4, Section 4.4.1: an HMM based continuous speech recognition system based on monophones, with Gaussian tied mixtures, using 12 MFCC coefficients as features with delta and acceleration coefficients. We have selected for each ICA method (JADE, FastICA and combined), the number of iterations that provides the best output quality as measured by PESQ in Table 5.2, that is with only one iteration for each ICA method. Table 5.4 presents the results for the evaluated algorithms, and also the recognition rate of the

mixtures.

Table 5.4. Word recognition rates for the different alternatives. For reference, with the clean sources $WRR\% = 93\%$.

Power Ratio	Noise	Mixtures	J	F	JF
6 dB	Speech	44.50	70.50	74.50	74.50
	White	19.50	73.00	71.50	71.50
0 dB	Speech	30.00	70.50	53.50	53.50
	White	7.20	36.00	35.00	35.00
Average		25.30	62.50	58.63	58.63

As it can be seen the proposed method of multiresolution BSS using JADE and only one iteration produces the best recognition rate over the alternatives using FastICA, comfirming the PESQ predictions. Also, the method produces an improvement of about 30% in the recognition rates with respect to the mixtures. It must be noted that, although the improvement is not so important, the processing time is considerably reduced, allowing for almost real-time processing.

5.4. Concluding remarks

In this chapter, a new method for separation, called the multiresolution BSS (MRBSS) method was proposed. This method uses a simplification considering a mixture with only one mixing matrix, to separate the signals with a very fast algorithm. The method automatically selects the optimum window length for the STFT, and for that reason we named it a multiresolution approach, as at each iteration a different time-frequency resolution is used. In the experimental section it was shown that JADE algorithm performs better in this case than FastICA, having a higher robustness in the presence of data from different mixing conditions. Although for ASR the best results were obtained with only one iteration in the separation algorithm, the method seems to produce better separation with more iterations, although also introduces a larger amount of distortions that can be heard as musical noise. The results of these algorithms were published in [Di Persia et al., 2006c,d,e].

Chapter 6

Pseudoanechoic Model BSS

The algorithm presented in Chapter 4 showed a good separation quality, but with an increase in computational cost that makes it less atractive for real time applications. The main disadvantage of this approach is the permutation indeterminacy among source identification, as the separated sources can have some residual permutations for each frequency bin. Considering this, some means to avoid the permutation problem are required. The algorithm proposed in Chapter 5 is permutation-free. This algorithm results in a very fast separation, but it presents as a drawback a lowering in the separation quality.

In this chapter we propose a new method of fd-ICA based on a simplified mixture model, which assumes a high similarity between impulse responses from a given source to all the microphones. This method can be used to generate a separation matrix for each frequency bin, having no indeterminacy among bins, and with high processing speed. As a consequence, in the directivity patterns the direction of the null is always the same, which improves the separation quality. Also a time-frequency Wiener postfilter is applied to enhance the output by reducing the residual noise. The chapter starts with a description of the problems to solve and the presentation of a new pseudoanechoic mixture model. Next, a method to improve the convergence to the ICA algorithm and produce more robust results will be presented. The separation algorithm based on the pseudoanechoic model is next introduced. Then, an extensive evaluation of the algorithm capabilities and parameters is presented in the results and discussion section.

6.1. Introduction

As already mentioned, one of the main disadvantages of the standard fd-ICA method is the presence of permutations. Although the method introduced in Chapter 4 has better capabilities to solve the permutations, it still produces residual permutations.

It must be noted that this problem is a direct result of the indeterminacy that arises in ICA from the lack of information about the sources. Thus, to obtain a consistent time-frequency representation for each source, this approach requieres to correct the permutations and the arbitrary scaling among frequency bins. Although there are some approaches to solve the permutation problem, based either on the correlation between frequency bands [Murata et al., 2001] or on the estimation of directivity patterns of the separation matrix [Sawada et al., 2004], these approaches tend to fail when the reverberation time of the environment increases.

A detailed analysis of the working principles and limitations of fd-ICA for reverberant environments is presented in [Araki et al., 2003]. In this approach, in order to capture the impulse response effect, a large frame size is required for the STFT analysis. This reduces the amount of data available in each frequency bin and produces a deficient estimation of the separation matrix. This was shown in Chapter 4 as some frequencies for which the ICA method failed (see Figure 4.9). As a consequence, there is a tradeoff between the need of long frames to deal with reverberation and the need of short frames to properly estimate the separation matrix. Furthermore, in the same work the BSS processing is compared to a set of null beamformers, and it is shown that for longer reverberation times, the directivity pattern produced by fd-ICA is increasingly deteriorated, mainly in low frequencies, due to wrong estimation of the mixing matrices. This increases the rate of permutation misalignments, producing poorer results, mainly due to the non-uniform directivity pattern that produces a degradation in some frequency bands. To overcome these drawback a new mixing model will be introduced.

6.2. Pseudoanechoic mixture model

To obtain a robust method of separation, a simplified mixture model is proposed. The mixing model and the separation algorithm derived from it will be explained for the case of two sources and two microphones. The generalization to more sources and microphones is straightforward, and will be sketched after presenting the algorithm. In a 2-by-2 configuration, there are four impulse responses (IR) that characterize the transmission path from each source to each microphone. We assume the usage of omnidirectional microphones, both pointing in the same direction to avoid phase inversions. The IR from source i to microphone j will be denoted as $h_{ji}(t)$. The source vector is $\mathbf{s}\,(t) = [s_1(t)\; s_2(t)]^T$ and the mixture vector is denoted $\mathbf{x}\,(t) = [x_1(t)\; x_2(t)]^T$, as usual. Figure 6.1 shows these variables. Using this setting, the mixtures can be expressed as

$$\begin{aligned} x_1\,(t) &= s_1\,(t) * h_{11}\,(t) + s_2\,(t) * h_{12}\,(t) \\ x_2\,(t) &= s_1\,(t) * h_{21}\,(t) + s_2\,(t) * h_{22}\,(t)\ . \end{aligned} \tag{6.1}$$

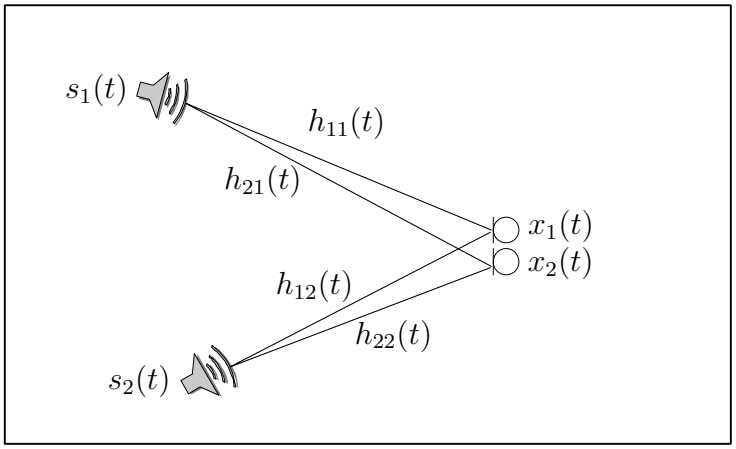


Figure 6.1. Environment description and notation for a two sources and two microphones case.

where $*$ stands for convolution. As it can be seen, each signal x_i is produced by the addition of two terms, generated by each source s_i after convolution with different IRs. In the general case of arbitrary microphone locations, the IRs from a source, say s_1, to both microphones, h_{11} and h_{21}, will be quite different, and thus after convolution with them the results will have very different waveforms. This means that the contributions of the same source on each microphone would behave as completely different signals. Thus the problem behaves as under-complete, as it is like a 4 sources and 2 microphones mixture. For that reason instantaneous independent component analysis fails to solve the problem.

Now assume that the microphones, instead of having arbitrary positions, are restricted to remain "near" to each other. This idea is not unusual, in fact if one think in the separation of the ears in animals it can be quite small. Moreover, this can be even desirable, for the production of multimedia devices of small size. The sound generated by some source corresponds to local changes in pressure from a steady, stable value. Thus, what we are interested in, is the evolution along time and space of the pressure, relative to the steady value. This pressure variation, denoted by $p(\mathbf{v})$, where $\mathbf{v} = [x, y, z, t]$ represents the concatenation of space coordinates and the time evolution, can be modeled using the classic fluid mechanics theory, as presented in Chapter 1. For a flow at small velocity (which is the usual case for sound at normal power levels), the sound field is characterized by the classical wave equation [Kuttruff, 2000]. As this equation includes second order partial derivatives in time and space, the pressure function that solves the equation must be a C^2 class function, that is, a twice continuously-differentiable function of space and time coordinates. Therefore, both the pressure and its first derivative must be continuous. This continuity implies that the limit of the pressure must exist in all space-time coordinates of the domain. In

other words, at two "near enough" points of the space-time coordinates, the pressure cannot be too different. In this phrase, the terms "near enough" refer to the euclidean norm $||\mathbf{v}_1 - \mathbf{v}_2||$ being small. The meaning of this, is that if the microphones are near enough, the IR measured from the same source at both microphones will have a similar waveform, possibly affected by some delay and scaling. This observation motivates the following assumption that we will use to simplify the mixture model: Given enough near located microphones, the impulse responses from one source to all the microphones are similar in shape, and are only modified by a delay and a scaling factor. That is,

$$\begin{aligned} h_{21}(t) &\simeq \alpha h_{11}(t - d_1) \\ h_{12}(t) &\simeq \beta h_{22}(t - d_2) \ . \end{aligned} \tag{6.2}$$

To simplify the notation let $h_1(t)$ and $h_2(t)$ denote $h_{11}(t)$ and $h_{22}(t)$, respectively. Denoting $z_1 = s_1 * h_1$ and $z_2 = s_2 * h_2$, we can rewrite (6.1) as

$$\begin{aligned} x_1(t) &= z_1(t) + \beta z_2(t - d_2) \\ x_2(t) &= \alpha z_1(t - d_1) + z_2(t) \ . \end{aligned} \tag{6.3}$$

After a STFT, and assuming the time invariance of the impulse responses (as usual for static or short duration sources), this can be written as

$$\mathbf{X}(\omega, \tau) = A(\omega)\, \mathbf{Z}(\omega, \tau) \ , \tag{6.4}$$

where the mixing matrix A has the form

$$A(\omega) = \begin{bmatrix} 1 & \beta e^{-jd_2\omega} \\ \alpha e^{-jd_1\omega} & 1 \end{bmatrix}. \tag{6.5}$$

In this model, the parameters α, β, d_1 and d_2 are the relative attenuations and delays of the impulse responses arriving at different microphones, and the effect of the room is included in $\mathbf{Z}(\omega, \tau)$. The separated sources (convolved by the impulse responses h_1 and h_2) can be obtained using $W(\omega)$, the inverse of the mixing matrix $A(\omega)$ for each frequency bin. In this way, we have a specific mixing matrix for each frequency bin, and thus a specific separation matrix, which will produce the desired specific directivity patterns.

In the standard fd-ICA formulation, the problem consists of estimating a 2-by-2 complex separation matrix for each frequency bin (that can be hundreds). Using this new model, the problem has been reduced to the estimation of four parameters, named α, β, d_1 and d_2. If one can obtain a reliable estimate of these parameters for some frequency bin, then they can be used to build the mixing matrix $A(\omega)$ for other frequency bins. Given $A(\omega)$, the separation matrix $W(\omega)$ is obtained as its inverse, and

the separation is realized by applying it to each mixed frequency bin. This matrix works in a similar way than that one of standard fd-ICA methods, and can be interpreted as a pair of null beamformers.

Two aspects must be noted: by using this model, there are no amplitude nor permutation ambiguities among bins, thus there is no need for permutation correction stages after the separation. Also, an algorithm based on this approach is expected to have a low computational cost, as only one optimization (to estimate the parameters) would be needed. This is the opposite for standard fd-ICA approaches, that need one ICA optimization for each frequency bin and to solve the permutation and amplitude ambiguities afterwards.

The main assumption of this pseudoanechoic model, that is, the similar waveforms of the impulse responses from the same source in all the microphones, has also been observed in other works. In a recent work [Melia and Rickard, 2007], in the context of underdetermined BSS methods for echoic environments, the authors analyze the IR for closely spaced microphones (2.5 cm). They present some graphics that show very similar impulse responses for 4 consecutive microphones, and then state that these suggest that the impulse responses are merely delayed and scaled versions of each other. Moreover, in [Katayama et al., 2004] a pair of microphones are located with their tips almost coincident, and therefore the authors simplify the mixture model because they consider the IR to be identical. In this case directional microphones are used, and directionality is what allows the separation. Although the above mentioned works propose the usage of closely spaced microphones, they do not explore the theoretical bases for their use.

This approach is quite different from the anechoic model used by example in [Saruwatari et al., 2003]. In the anechoic model, the effect of reflections is disregarded, and no restrictions on the microphones location are imposed. Therefore the anechoic model only works for rooms with very small reverberation. On the contrary, our model takes reflections into account, and considers that their effect can be grouped into some latent variables z_i, which are obtained as outputs. This is clearly explained in Figure 6.2. In part a) of this figure, the anechoic model is shown in the left, and the block diagram in the right side shows the usual transformation to relative parameters. Both models are completely equivalent, and they can be applied only if the mixture was anechoic. In addition, if a far field simplification is used, $\alpha = \beta = 1$, as in [Saruwatari et al., 2003]. On the other side, in part b) a fully echoic model is shown in the left, which is valid no matter how long the filters are. The transformation in the right, which yields relative parameters, is possible for near enough microphones. In that case, both models are completely equivalent, and so the one in the right models a fully echoic mixture. Although the right sides of both models are similar in structure, they clearly differ in their principles and conditions of applicability. Given the similitude to the anechoic model we called this “pseudoanechoic” model.

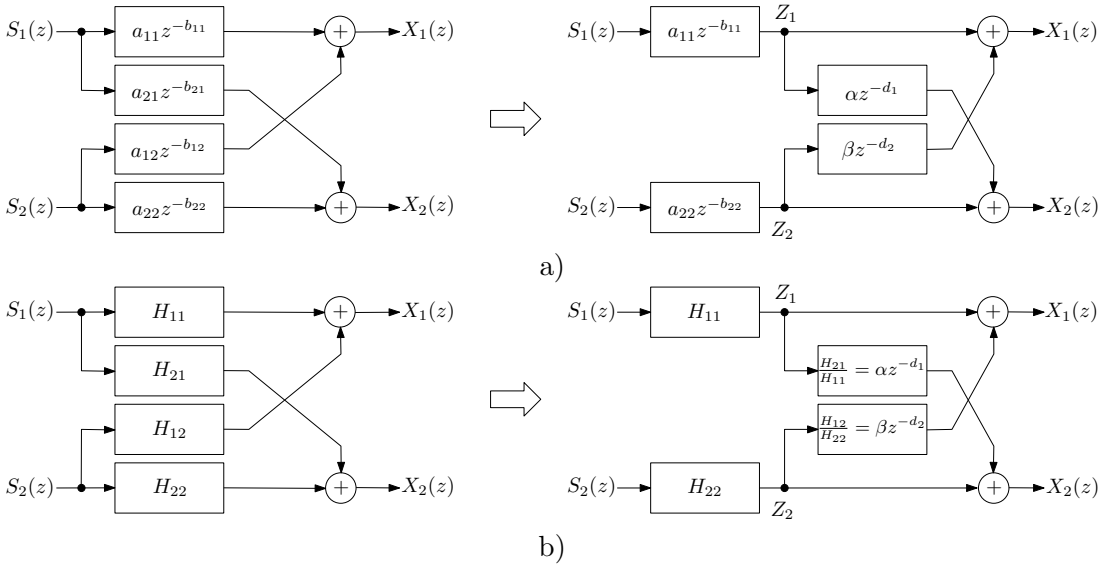


Figure 6.2. Block diagrams comparing a) the anechoic, and b) the pseudoanechoic models. For both cases, in the left is the general case, and in the right is the equivalent model using relative parameters.

In the pseudoanechoic model the reverberation time is not a limiting aspect, because as it can be seen in Figure 6.2, the transformation to relative parameters only depends on the validity of the assumption of similar waveforms of the impulse responses. In [Saruwatari et al., 2003], the anechoic model is used to synthesize null beamformers that do not take into account the amplitude attenuations, using some closed formulation. On the contrary, our method takes reflections into account, considers both delays and attenuation factors, synthesize the mixing matrix for each frequency bin using the estimated parameters, and calculates the separation matrices by direct inversion of the estimated mixing matrix. This yields very different equations for the separation matrix coefficients with respect to those obtained by synthesizing null beamformers with constant attenuations.

6.3. Convergence of the ICA algorithm

The standard approach of fd-ICA has, besides the permutation and scaling ambiguitites, two other important drawbacks, as analized in Chapter 2. One is the large amount of data needed to provide proper convergence to the ICA algorithm in each frequency bin. This can be difficult to achieve, due to the variable length of the mixtures, that can be as short as hundred milliseconds for short utterances. Another important

problem is the non-uniform convergence of the ICA algorithm for all frequency bins. In some bins the algorithm can give poor separation, while in others no separation at all. This depend on the characteristics of the mixtures at those bins (for example, it can present a low SNR, or even some of the signals not having any information in that bin).

Now, analyzing the simplified model we have proposed, it is clear that what is needed is a good estimation of the separation matrix for one specific bin, and this estimation will allow the construction of the separation matrix for all bins. However, if one uses the data of one arbitrary bin, the two above mentioned problems can cause a bad estimation. If this happens, the constructed separation matrices will be all wrong and the method will completely fail.

This shows the need to provide robustness to the estimation. One way to do this can be the usage of long signals only, which can help with the amount of data issue. However, this approach cannot solve the problem of a bad estimation due to the intrinsic characteristics of the signals in the selected bin.

We propose an alternative method for robust estimation of the mixing parameters, based in the idea of using several frequency bins instead of just one. Assuming that the microphones are near enough, the difference among the separation matrices for succesive bins comes from the angle of complex values of the mixing matrix. Let us illustrate this with an example. Suppose that we consider the element $A_{1,2}$ of the mixing matrix, which contains the delay d_2. The modulus is fixed, and the angle changes with frequency:

$$\angle(A_{1,2}(\omega_0) = -\omega_0 d_2 = \frac{-2\pi f_s}{N} k d_2 = -\varrho k d_2 \,, \tag{6.6}$$

where ϱ is a constant, f_s is the sampling frequency, N is the length of the frame used in the FFT, and k the bin index. As it can be seen, the angle changes linearly with the bin index. So, if three successive angles (for bins $k-1$, k and $k+1$) are averaged, the resulting angle should be the same as the one of the central bin (k). This is the idea used in our bins packaging: the contributions in excess due to the higher frequency bins will compensate the contribution of the lower frequency bins, giving a result that is closer to that of the central bin. By adding more bins, the ICA algorithm will converge better (because it has more data to estimate the statistics), and the discrepancies from the bins at both sides of the central one will be cancelled to produce a central estimation. This gives robustness, because for example, if the ICA algorithm does not converge using the central bin alone (as it is usual for isolated bins in standard fd-ICA), the addition of the lateral ones will make it converge. Of course that if we add too much bins, we expect the discrepancies among mixing matrices at both extremes of the bin range to have more weight and to start producing wrong results. Furthermore, the time used by the ICA algorithm will be increased due to the larger amount of data to process and the degradation produced by inconsistent data.

6.4. Separation algorithm

According to the previous sections, the key point to achieve separation using the pseudoanechoic model is to be able to find a good estimate for matrix $A(\omega)$ for a given frequency bin. This allows the estimation of the mixing parameters, and thus we can build a separation matrix for each frequency bin. To realize this idea, an algorithm was designed as shown in Figure 6.3. In this algorithm, there are several subjects that need to be clarified. We will detail all the steps in the following.

Step 1) Transformation to the time-frequency domain: This transformation is done by means of a standard STFT using a Hanning window [Deller et al., 1993]. This transformation is used to obtain the time-frequency representations for all the mixtures, as explained in Chapter 2.

The two relevant parameters in this transformation are the window length N and the frame shifting interval R. As in this method the impulse response is considered as a part of the signal to obtain, these parameters are not so critical. In usual fd-ICA, a large window length is used to capture the impulse response characteristics. This increases the number of frequency bins to be processed. As the reverberation is included in the model in this new approach, a relatively small window length can be used without significant degradation of separation. This will speed-up the algorithm as less frequency bins need to be processed. Regarding the frame shifting interval, in standard fd-ICA a small value is used mainly to increase the amount of available data. This increase of the data length does not necessarily imply a better separation, because the data is highly redundant and the convergence may be slow. On the contrary, in this new algorithm the amount of data is decided by other aspects (see Step 2), and so the frame shifting interval can be increased (even to half of the window length) to reduce computational costs.

Step 2) Selection of frequency bin: This selection is not trivial, the ideal frequency bin would be one which presents a good signal to noise ratio, and for which the ICA algorithm will produce good directivity patterns. We have selected the frequency bin ω_ℓ based on knowledge of the source characteristics, but some better designed automatic decision algorithms can be developed. To give robustness to the method we use not only one frequency bin, but we select a number Δ of frequencies to each side of the chosen one, and pack all in one long vector, as already discussed. The use of lateral bins not only makes the estimation for the central bin more robust, but also increases the amount of available data (which improves the convergence properties of ICA). In this way we can fix a minimal number of K training samples for the ICA algorithm. If each frequency bin has a length of L samples, we set $\Delta = \max(3, \lceil (K/L - 1)/2 \rceil)$, where $\lceil \cdot \rceil$ means rounding to the nearest higher integer. That is,we use a number of bins enough to have, combined, K samples for ICA training, and if this number is less than 3, we fix it to 3. The separation matrix obtained from this process will correspond

1. *Apply a STFT to switch to the time-frequency domain.*
2. *Choose some frequency bin* ω_ℓ.
3. *Estimate the separation and the mixing matrix for* ω_ℓ *by ICA.*
4. *Convert the mixing matrix to the normalized form of (6.5).*
5. *Use the obtained matrix to calculate the four parameters:* α, β, d_1 *and* d_2.
6. *Separate each frequency bin. For each* ω*:*
   a) *Calculate* $A(\omega)$ *according to (6.5).*
   b) *Calculate separation matrix* $W(\omega)$ *by inversion of* $A(\omega)$.
   c) *Obtain the estimated source contributions* $\tilde{\mathbf{Z}}(\omega,\tau) = W(\omega)\,\mathbf{X}(\omega,\tau)$.
7. *Apply the time-frequency Wiener filters to enhance the output signals.*
8. *Reconstruct the temporal signals by inverting the STFT.*

Figure 6.3. Separation algorithm based on the proposed mixture model.

to the frequency ω_ℓ of the central bin. The selection of central bin will be discussed in more depth in Section 6.5 and Subsection 6.5.4

Step 3) Estimation of mixing and separation matrices: We use the complex version of FastICA algorithm, as proposed in [Bingham and Hyvärinen, 2000]. This algorithm uses a deflationary approach where each source is extracted sequentially. After finding the separation matrix, the mixing one is calculated as its inverse.

Step 4) Conversion of mixing matrix to normalized form: The normalized form consists of ones in the main diagonal, and in general this will not be the case with the estimated mixing matrix. To obtain the normalized form of (6.5), all elements in column i must be divided by the i-th element of the diagonal. This step is responsible for the elimination of the amplitude ambiguity because all scaling effects of the mixing matrix are absorbed into $\mathbf{z}$.

Step 5) Estimation of the mixing parameters: Once the mixing matrix in nor-

malized form is obtained, the parameters can be calculated as:

$$\alpha = |a_{21}|\,, \qquad d_1 = -\frac{N}{2\pi\ell f_s}\Im m(\ln(a_{21}))\,,$$
$$\beta = |a_{12}|\,, \qquad d_2 = -\frac{N}{2\pi\ell f_s}\Im m(\ln(a_{12}))\,, \tag{6.7}$$

where ℓ is the index of the central frequency bin used in step 2, f_s denotes the sampling frequency, and $\Im m(\cdot)$ is the imaginary part function. It must be noted that the delay estimations will be valid only if $\frac{2\pi\ell f_s}{N} d_i < \pi$, which follows from the periodicity of the complex exponentials. This requirement will be satisfied if the microphone spacing is small enough to avoid spatial aliasing, and of course, if the mixing matrix is successfully estimated.

Note that this robust estimation of the parameters is quite different from the direction of arrival (DOA) estimation used in the field of fd-ICA [Saruwatari et al., 2003; Sawada et al., 2004]. For ICA-based DOA estimation, an ICA problem is solved in each frequency bin, and after solving the permutation problem, each global DOA is estimated by averaging the DOAs estimated on each frequency bin, all under an anechoic model (as in eq. (13) of [Saruwatari et al., 2003]). The estimations for each frequency bin are affected by many disturbances like different noise powers, bad convergence of the ICA algorithm, and residual permutations, as seen in Chapter 4 (Figure 4.9). All of these noise sources affect the estimation of each bin and thus the resulting average estimates will lack robustness. The robustness of our approach is a consequence of the use of several frequency bins in step 2 and the absence of permutations.

Step 6) Separation: In this step, a specific mixing matrix for each frequency bin is calculated using the estimated parameters. A separation matrix is obtained as its inverse, and the separated sources are calculated using it. It must be noted that the structure of the mixing matrix with ones in the diagonal can be exploited to speed-up the calculation of the separation matrix by reducing the amount of multiplications. After this step we obtain an estimation $\tilde{\mathbf{Z}}(\omega_k, \tau) = [\tilde{\mathcal{Z}}_1(\omega_k, \tau)\ \tilde{\mathcal{Z}}_2(\omega_k, \tau)]^T$ of the sources $\mathbf{Z}(\omega_k, \tau)$.

Step 7) Wiener filtering: Due to the behavior of the separation algorithm as a pair of null beamformers [Araki et al., 2003], the estimated sources will still have some residual noise, as already discussed in the previous chapters. To reduce this residual we propose to use a pair of non-causal time-frequency Wiener filters as post-processing [Huang and Benesty, 2004, chapter4]. The short-time Wiener filter $F_{\mathcal{W},1}$ to enhance source 1 is

$$F_{\mathcal{W},1}(\omega_k, \tau) = \frac{\left|\tilde{\mathcal{Z}}_1(\omega_k, \tau)\right|^2}{\left|\tilde{\mathcal{Z}}_1(\omega_k, \tau)\right|^2 + \left|\tilde{\mathcal{Z}}_2(\omega_k, \tau)\right|^2}\,. \tag{6.8}$$

where $\tilde{\mathcal{Z}}_2(\omega_k, \tau)$ in the denominator is used as an estimation of the residual noise. The short-time Wiener filter to improve source $\tilde{\mathcal{Z}}_2$, $F_{\mathcal{W},2}(\omega_k, \tau)$ is calculated in a similar way to (6.8), with the roles of $\tilde{\mathcal{Z}}_1$ and $\tilde{\mathcal{Z}}_2$ interchanged.

The behavior of this Wiener filter will depend of the capability of having a good estimation of source and noise, that is, it depends on a good result for the previous separation stage. This is equivalent to say that, if after this filtering the result is of good quality, it was a consecuence of a suficiently good previous separation.

Step 8) Signal reconstructions: We use the overlap-add inverse STFT to reconstruct the source signals [Allen and Rabiner, 1977]. The reconstruction formula presented in Chapter 2 is applied to each estimated source $\mathcal{Z}_i(\omega, \tau)$ to obtain its time-domain representation $z_i(n)$.

As this new algorithm is based in the proposed pseudoanechoic model, we will call it pseudoanechoic model blind source separation (PMBSS).

It must be noted that as we are searching for $\mathbf{z}$ but not for $\mathbf{s}$, the algorithm will achieve separation but not reverberation reduction (with the exception of the reduction of reverberation introduced by the Wiener filter). As the reverberation is considered as part of the target signals, the algorithm will be less sensitive to it, and thus will achieve better separation for the cases where standard fd-ICA methods fail. If reverberation reduction is also desired, a second processing stage should be employed. In this way we split the problem into two simpler ones. Also note that this method is "indeterminacy free", in the sense of the elimination of the need to solve the usual fd-ICA permutation and amplitude ambiguities among different bins, but the time-domain reconstructed signals will still contain an arbitrary scaling (and moreover, a filtering) and permutation. So our method has the advantage over other fd-ICA approaches of eliminating the between-bins permutations and scalings, which makes it more robust, even as it still has a global permutation ambiguity among the resulting signals.

This algorithm can easily be generalized for the q-by-q mixture case. Up to step 4, no modifications are needed. For step 5, in the presented algorithm the amplitudes and delays are relative to the reference signal (the one in the main diagonal). For the general case, as in the normalized form the main diagonal has always ones, the parameters to estimate are the amplitude of the off-diagonal elements, and the pairwise relative delays in the exponents, that is, two parameters for each off-diagonal element. So, there are $2q(q-1)$ parameters to estimate, and their estimation is straightforward from the normalized form, using the corresponding analogous to (6.7). Step 6 does not need any change. For step 7, the Wiener filter formulation needs to be modified, as the noise spectrum estimate will be the sum of all the other estimated source spectrums. Step 8 remains unchanged. Although this generalization is straightforward, we will restrict our experiments to the 2-by-2 case, leaving for future works the analysis of the general case behavior.

6.5. Results and discussion

The pseudoanechoic model was built assuming that the impulse responses from one source to all microphones should have approximately the same waveform. We need to investigate how the spacing between microphones will affect the algorithm performance. Furthermore, there are two parameters: window length and frame shifting interval, that need to be calibrated. These aspects will be studied in the first and second part of the current section.

We are interested in the application of this algorithm to automatic speech recognition (ASR), and also in applications where the processed speech is presented to a human listener, like in hearing aids. Therefore we need to evaluate our algorithm regarding both, speech recognition tests and perceptual quality. In the following subsections, all these issues will be explored in detail.

As in the previous chapters, for all the experiments speech sentences extracted from Albayzin Spanish database [Moreno et al., 1993] ave been used. Also we have used white noise source from Noisex-92 database [Varga and Steeneken, 1993]. All the signals were resampled to 8000 Hz. These sources were mixed in different conditions, using both speech and white noise as competing sources, to generate appropriate data sets for the experiments.

The mixtures were separated by the algorithm of Figure 6.3. In all experiments, for this algorithm we used a different central frequency ω_c in the case of speech noise and white noise. For speech-speech mixtures, a high frequency bin was used, as in general the separation matrix is better estimated in high frequencies, as shown in [Araki et al., 2003]. For speech signals with telephony quality bandwidth, the maximum frequency of interest present in the signals is 4000 Hz. A frequency band located at 5/8 of the maximum frequency was selected in this case. On the other hand, for speech-white mixtures a low frequency bin was used. This is due to the fact that white noise presents equal power in all frequencies, whereas speech tends to have more power in low frequencies due to the low-pass characteristics of the glottal response [Deller et al., 1993]. In this way, the signal to noise ratio is degraded with increasing frequencies. A bin located at 3/8 of the maximum frequency was selected in this case. The desired number of data samples to use with FastICA was empirically fixed at $K = 5000$. We have performed a lot of experiments with FastICA and for simple noise free examples, 2000-3000 samples are enough for the algorithm to converge to a good solution. As in this case the real mixtures are subject to many noise sources and disturbancies, we used a larger number to assure better convergence.

For the evaluation of the method in robust speech recognition, we performed the tests with a state-of-the-art continuous speech recognition system. For this task, a three-state semi-continuous hidden Markov model for context-independent phonemes and silences was used [Huang et al., 1991]. Gaussian mixtures were utilized as observa-

tion probability densities. After four reestimations using the Baum-Welch algorithm, tying was applied in the model to reduce the number of parameters. For this, a pool of 200 Gaussians for each model state was selected. After tying, another 12 reestimations were performed. A bigram language model was used for recognition, estimated from transcriptions of the training sentences [Jelinek, 1998]. For the front-end, we performed parametrization with standard mel frequency cepstral coefficients (MFCC), including energy and the first derivative of these features [Rabiner and Juang, 1993]. This recognition system was built using the HTK toolkit [Young et al., 2005]. The results were evaluated using the word recognition rate (WRR). To evaluate the separation quality from an objective point of view we have used the PESQ raw score in narrow band mode as defined in [ITU, 2001]. For all these evaluations, we have used the framework proposed in Chapter 3.

Given the large amount of parameters to explore, if all of them were varied in an exhaustive search, the number of required experiments would grown exponentially. To avoid this problem we have sorted the parameters according to their influence on the algorithm, and then explored the variation of each one independently, while the other parameters remained fixed. Although this would produce a sub-optimal set of parameters, it would allows us to explore a larger area of the parameter space, with a reasonable number of experiments and time.

6.5.1. Effects of microphone spacing

The proposed algorithm uses a physically plausible assumption to simplify the mixture model. The key question to be analyzed in this section is how plausible is that hypothesis in real cases.

As already discussed, the motivation for the assumption comes from the physics of sound propagation, taking into account the continuity of the sound field. Intuitively, if the microphones are "near enough", then they should measure similar variations of the sound field, and thus the IR measured at those points should have approximately the same shape, but affected by some delay and scaling. This produces two main aspects that needed to be determined: one is how much near the microphones must be for the hypothesis to be applicable, and the second one is how sensitive the algorithm is concerning to a poor adjustment to the hypothesis.

As an example to illustrate the first point, Figure 6.4 shows two impulse responses recorded in the room of Figure 6.5. The impulse responses were measured from source 1 to microphones 1 and 5, which were spaced by 4 cm. The distance from the source to the microphones was about 113 cm, with an angle respect to the array center of 26 degrees, in a room with 349 ms of reverberation time.

The top part of Figure 6.4 shows the impulse responses. In a general view the IRs seems to have similar global characteristics, although due to the scale it is difficult

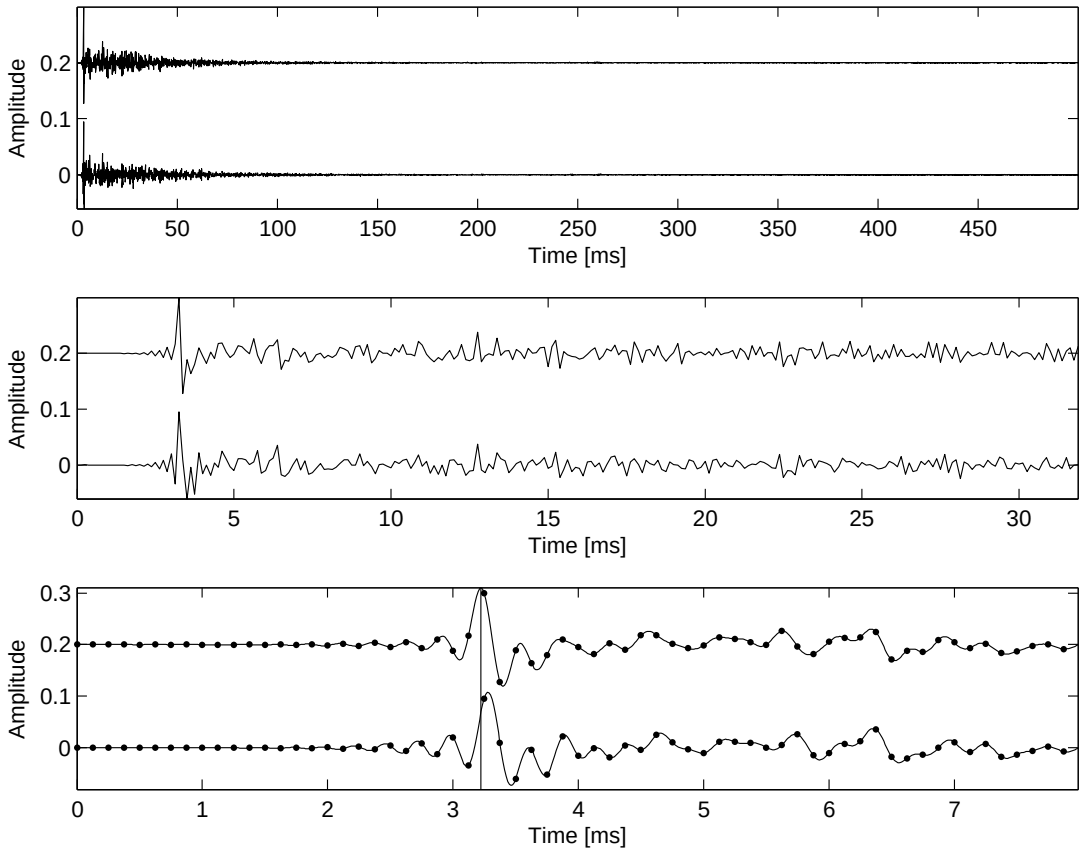


Figure 6.4. Impulse response characteristics for 4 cm spacing, recorded as in Figure 6.5. Top: first 0.5 seconds. Center: zoom showing the first 256 samples. Bottom: first 64 samples, resampled at 10 times the original sampling frequency. In all panels, one of the signals has a constant of 0.2 added to separate the two plots.

to realize how similar they really are. In the central panel, a zoom of the initial 256 samples of the IRs is shown. In this image it is easier to see the similitude between the two impulse responses. Although there are some parts showing small differences, most of them can be attributable to the combined effect of the fractional delay and the sampling. This can be seen in the bottom panel, where a zoom of the first 64 samples is shown. To generate this plot, a resampling using bandlimited interpolation was used, to elevate the sampling frequency to 10 times the original (i.e., from 8 kHz to 80 kHz). Also, the original samples are shown with dark dots. In this plot, the fractional delay can be clearly seen. The delay corresponds to about 2/5 of the original sampling time, which agrees with the spatial setup. The bottom panel also shows that most of the local

differences in the waveforms have disappeared, which confirms that the morphological differences were artifacts produced by the sampling. Considering this example, the assumption about the similarity of the IR waveforms seems to be very plausible. It must be noted that in this example the microphone spacing is of 4 cm, and even with such a "wide" separation, the similitudes are evident. This is supported also by the results in [Melia and Rickard, 2007], in which four impulse responses measured with a 2.5 cm uniform spacing are found to be very similar in shape. The authors conclude that the IR can be possibly considered as delayed and scaled copies.

Considering the second aspect (the sensibility of the algorithm with respect to the hypothesis), it is necessary to evaluate how the separation performance is modified by dissimilar impulse responses, thus one needs to explore the effect of microphone spacing. If the spacing is too large, the impulse responses from one source to the microphones will be too different and the hypothesis will become invalid. Also, spatial aliasing can be produced for large microphone distance. The maximum allowable distance to avoid spatial aliasing for a sampling frequency of 8000 Hz is about 4.3 cm, as was shown in Chapter 2 [Brandstein and Ward, 2001]. On the other hand, too small spacing will cause the relative amplitudes to be very near to one and the relative delays to be very small. An accurate estimation of these parameters will thus be difficult to obtain, mainly because of the limited measurement precision and the microphone noise.

We have explored this issue using synthetic mixtures of speech. To produce the mixtures we have selected five sentences from Albayzin database spoken each one by two male and two female speakers, for a total of 20 utterances. Also, one sentence spoken by a male and a female speakers were selected to be used as competing noise. This sentence was used because it was longer than any of the other sources, and so the same sentence could be used to interfere with all the target sources. To compete with male speech, female speech was used, and vice versa. The utterance duration ranged from 2.26 s to 4.65 s, with an average duration of 3.55 s. We also used white noise from Noisex database.

The mixtures were generated by convolving each source with an impulse response measured in a real room and adding the results to generate each microphone signal. The impulse responses were measured using the method of time stretched pulses (TSP) [Suzuki et al., 1995]. A condenser desktop microphone with omnidirectional frequency response from 20 Hz to 20 kHz was used. The measurements were made in the room depicted in Figure 6.5. We used 5 microphone locations with a spacing of 1 cm, with a careful synchronization to preserve relative amplitude and delays between impulse responses. The average reverberation time, measured by the Schroeder backintegration method [Schroeder, 1965], was of $\tau_{60} = 349$ ms. The impulse responses measured from pairs of microphones with spacings of 1, 2, 3 and 4 cm were used (longer spacings would introduce spatial aliasing). According to the naming convention of Figure 6.5, the pairs

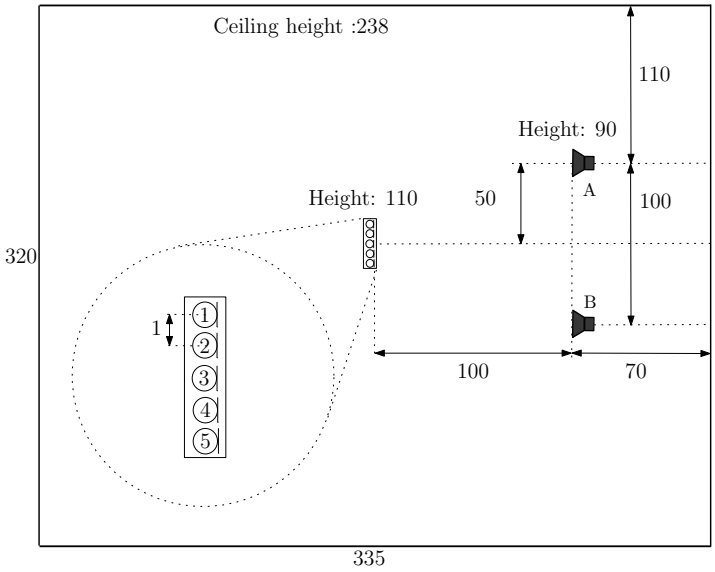


Figure 6.5. Experimental setup for two sources and five locations of the microphone. All dimensions are in cm.

of microphones were 2-3 and 3-4, 1-3 and 3-5, 1-4 and 2-5, and 1-5 for 1, 2, 3 and 4 cm spacing, respectively.

The effect of different noise powers was also explored. For each noise kind (speech or white) we used two different power ratios (0 dB and 6 dB) by properly scaling the source signals. Thus, we performed tests for a total of 16 mixing conditions.

First, we wanted to investigate the optimal spacing, so we performed separation for the different spacings. We have repeated the separation using 3 window lengths (128, 256 and 512 samples) and two different frame shifting intervals for each window length (one quarter and one half of the window length). Figure 6.6 shows the average PESQ scores over the 20 sentences. In this figure, we have averaged the results for the four noise conditions to produce a single value for each spacing, window length and shift interval. It can be seen that the optimal spacing is 4 cm in all cases. A too short spacing makes difficult to accurately estimate the parameters and thus the algorithm fails. Spacings longer that 4.3 cm will cause spatial aliasing. This behavior is repeated if the analysis is discriminated for each noise condition, showing always the best separation at 4 cm. According to this result, the spacing was fixed at 4 cm for the following experiments.

Figure 6.6. Effects of microphone spacing on the PMBSS method. Gray bar: 1 cm spacing, dashed bar: 2 cm spacing, white bar: 3 cm spacing, black bar: 4 cm spacing.

6.5.2. Effect of window parameters

Once fixed the optimum spacing, the effect of window length and frame shifting interval was explored. For this evaluation we proceeded in a similar way to the previous experiment, but only for the case of 4 cm spacing. We used five window lengths (128, 256, 512, 1024 and 2048 samples) with a frame shifting size fixed on half of the window length. For the evaluation we used PESQ and also the average processing time. As a fast separation algorithm with a good quality performance is required, we used the ratio of time to quality as a cost-to-benefit or tradeoff function to determine the optimum window length.

It could be argued that the processing time is not a good index of complexity because different implementations of the same algorithm would yield different times. Nevertheless, the time requirements of our algorithm are not caused by high complexity tasks that could be performed with different implementations of algorithms, but by simpler task that are repeated many times. The FastICA algorithm used is the same, and is performed only once, with the same amount of data samples, so its influence on the calculation time for different frame lengths is equivalent. Thus, the two processes that have a strong effect on processing time are the calculation of the separation matrices (which involves a matrix inversion for each frequency bin) and the separation of the data itself, which implies a matrix-matrix multiplication for each time-frequency tile.

With increasing frame length, the number of bins to process is increased, which

Table 6.1. Effects of the window length. Tradeoff is the ratio of Time to PESQ score.

Window	PESQ	Time [s]	Tradeoff
128	2.166	0.550	0.254
256	2.215	0.558	0.251
512	2.234	0.647	0.290
1024	2.177	0.616	0.283
2048	2.094	0.716	0.342

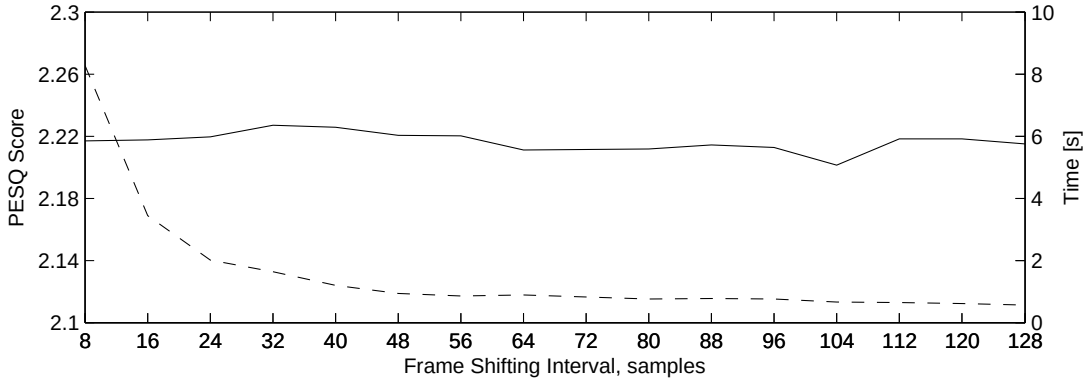


Figure 6.7. Effects of the frame shifting interval. Solid line: PESQ score, Dashed line: average processing time.

means that more matrix inversions need to be calculated, but the amount of data to separate is the same. So the change of computation time is mainly due to the increased number of matrix inversions needed. Thus, a larger window will have to perform more matrix inversions, so its processing time would be directly increased. Table 6.1 shows the results. It can be seen that the optimum (the minimum of the Time/PESQ ratio) is obtained for a window of 256 samples.

Finally the effect of the frame shifting interval was studied. Fixing the microphone space in 4 cm and the window length in 256 samples, we explored the shifting parameter from 8 to 128 samples with increments of 8 samples. The maximum shifting interval was fixed to 128 samples because at least a half window redundancy is necessary to obtain a proper reconstruction of time domain signals. Figure 6.7 shows the PESQ score and the average processing time for this analysis. As it can be noted from the figure, there is no significant quality change for different shifting intervals. Nevertheless, with small shifting values the processing times grow very fast. With small shifts, the number of frames in each frequency bin increases. Given that we fixed a minimum value of at least $\Delta = 3$ for the lateral bins, as the shift interval diminishes the amount of data

used in FastICA increases and the algorithm is slowed down. On the other hand, when the shift interval increases, as the amount of data is fixed at $K = 5000$, the redundancy is reduced and the convergence is faster. According to this, we selected 128 samples as the best frame shifting interval.

6.5.3. Evaluation on artificial mixtures

Once obtained the optimum values for spacing, windows length and frame shifting size, we wanted to check the performance of the separation algorithm in a larger database. In this test, the separation algorithm was used as a pre-processing stage for an automatic speech recognizer. Also, the objective quality evaluation by means of PESQ score was carried out. For this task we used a larger subset of Albayzin database. This subset consists of 600 sentences spoken by 6 male and 6 female speakers, with a vocabulary of 200 words, from the central area of Spain. The average duration of sentences was 3.95 s, with a minimum of 1.88 s and a maximum of 8.69 s.

The sentences were mixed artificially with impulse responses for microphones 1 and 5 in Figure 6.5. As noise sources we used competing speech and white noise. For speech, we selected two Albayzin sentences (different from the 600 used as sources), one from a female speaker to interfere with male speakers and one from a male speaker to contaminate female speech. The noise powers were adjusted to produce a power ratio of 0 dB and 6 dB at the simulated speakers. In this way, 4 sets of mixtures (2 noise kinds with 2 noise powers) were generated. After mixture, the separation algorithm with the optimum window length and frame shifting size obtained in previous experiments were used.

For the speech recognition task we used an ASR system like the one described in Section 6.5. The leave-k-out cross validation method with 10 partitions of the 600 sentences was used to test the acoustic model of the ASR system. For each partition 20% of the sentences were selected randomly to form a test set and the other 80% used as train sets. The results of the 10 partitions were averaged.

It is known that reverberation reduces the automatic recognition rates [Kinsbury and Morgan, 1997], even if the recognition system is trained with speech recorded in the same reverberant room [Benesty et al., 2005]. According to this, as our algorithm is not aimed to reduce reverberation, we cannot expect the recognition rate to be equivalent to that of clean speech. The maximum obtainable performance would be near to that of reverberant speech, without interference. We have used the artificially reverberated signals to evaluate this maximum performance. Table 6.2 shows the resulting PESQ scores and WRR, for the different noise kinds and powers. Besides the results of the separation algorithm, we present the results for the mixtures, the sources and the reverberant (but clean) sources.

As it can be appreciated, an average improvement of more that 35% in WRR and

Table 6.2. Recognition rate and PESQ score for synthetic mixtures.

		Mixtures		Separated	
Power ratio	Noise	WRR%	PESQ	WRR%	PESQ
6 dB	Speech	32.44	1.89	63.94	2.33
	White	16.73	1.74	63.36	2.30
0 dB	Speech	20.96	1.56	51.35	2.15
	White	12.26	1.50	43.09	2.05
Average		20.60	1.67	55.44	2.21
Reverberant				70.22	2.33
Clean				93.98	4.50

more than 0.5 in PESQ score is obtained by pre-processing the speech with the proposed algorithm. Also, it can be seen that for the case of 6 dB mixtures, the separated signal achieves a WRR that is near to the maximum attainable for our kind of algorithms. The average PESQ score for processed signals is also similar to that of reverberant ones.

Something interesting can be seen in the case of speech noise with 6 dB of power ratio. For this case, PESQ is the same as for reverberant speech. Nevertheless, speech recognition is lower than that one of reverberant speech. We know that the separated signals still have some residual interference, and so the quality should be degraded. However the PESQ is higher than expected because some reverberation has also been removed (this can be noted by inspection of spectrograms and by listening carefully). This reverberation reduction is an effect of the Wiener filter, since the signal used as estimation of the noise has some echoes of the desired source arriving from different directions. The Wiener filter will thus reduce in some amount the reverberant echoes. In this way, a small quality improvement in reverberation compensates the reduction of quality due to residual noise and PESQ is the same, although the remaining noise degrades speech recognition.

6.5.4. Robustness of the ICA estimation

There are two subjects that remain to be explored. One is how the quality and the robustness of the method are affected by the number of lateral bins selected, and the other one is how sensitive the algorithm is to the central bin selection.

To answer the first question an experiment using the real mixtures data of Chapters 4 and 5 was performed. For this experiment, a bin located in the center of the frequency range (in order to be able to add bins symmetrically at both sides) was selected. Then, the separation algorithm was applied using no lateral bins, one lateral bin to each side,

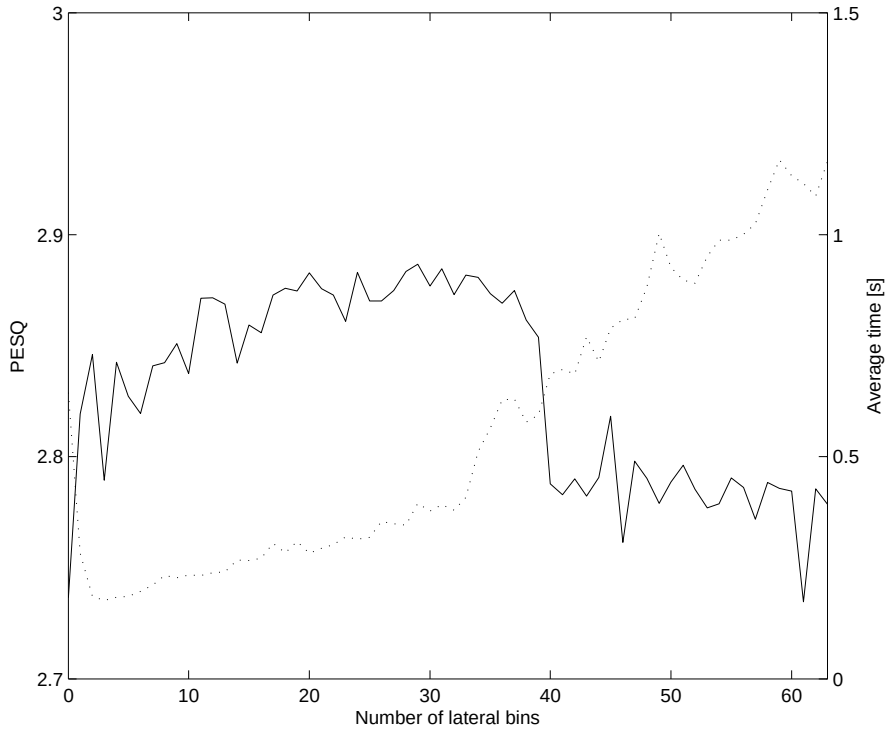


Figure 6.8. Effect of the number of lateral bins used in the ICA algorithm convergence. Left axis and solid line: PESQ score; right axis and dotted line: average processing time.

two lateral bins to each side, and so on. We performed the separation and averaged the PESQ and processing times for the 20 test sentences, for the case of white noise and 6 dB of power ratio (similar results were obtained for other noise and powers). A frame length of 256 samples was used, so the bin index goes from 0 to 128. In this way, we used the index 64 as central bin, and added lateral bins to each side from 0 to 63 bins. In the x axis of Figure 6.8 the number of lateral bins used is shown. Two y axes were used: the average PESQ obtained at the left side (in solid line), and the average processing time used for separation at the right side (dotted line).

As it can be seen in this figure, the addition of lateral bins is beneficial, as the quality is always improved compared to the case when using only the central bin. It can also be seen that adding more bins tends to increase the quality, up to a limit, where it starts decreasing. This agrees with our assumption that to some extent, by adding lateral bins, the effect of the upper an lower bins discrepancies is cancelled, and the quality is augmented. However, if too much bins are added, the discrepancies have more weight and the quality is lowered.

Also, it can be seen that the processing time is decreased initially, proving that

the addition of lateral bins produced faster convergence than using only the central bin, even when the ICA algorithm has to process more data. When more and more bins are added, the processing time is increased, so it would be desirable to keep the lateral bins as low as possible for reducing the processing time, but large enough to provide a good quality. This shows that the adition of lateral bins indeed provides an improvement in the convergence properties of ICA, both in terms of convergence speed and estimation quality.

Now for the second problem, in all the experiments up to now we have selected a fixed central bin to be used in ICA. This bin was chosen using some a priori knowledge about the source characteristics. Clearly, the best case would be to have some method that could determine which would be the bin that will produce the best separation quality. Although more research is needed before producing such a method (and this will be the subject of future researchs), it is important to know how sensitive is the method to the bin selection. To verify this robustness, we performed an experiment using a real mixture from the database described in Appendix A.2, contaminated with white noise at 0 dB of power ratio. The frame length was fixed at 256 samples with 128 samples of step size for the STFT. This means that we had the bin index changing from 0 to 128. From this range we excluded the extreme indexes (because using symmetrical lateral bins is not possible), and performed separation using each bin as the central one. That is, we set the central bin at 1, performed separation, and measure the resulting PESQ score, then changed the central bin to 2, and so on. We repeated this process, for three cases: using no lateral bins, using 5 lateral bins, and using 10 lateral bins [1]. The results are presented in Figure 6.9.

As it can be seen in this figure, for 0 lateral bins there are a lot of deep valleys in the PESQ score. These valleys correspond to bins where the ICA algorithm failed to converge. This behaviour is typical of standard fd-ICA approaches, which estimate the separation matrix in each frequency bin. For these cases the quality of the convergence of ICA is not uniform for all bins, so even if the other problems (i.e. permutations and scalings) are solved, the quality will be variable for different bins. That is, if we were using a standard approach with an ICA problem in each bin, each valley would correspond to a wrongly separated bin. What can be seen is that using 5 lateral bins, most of the valleys have been eliminated, and using 10 bins (solid line), the quality in function of the central bin is quite smooth. With 10 lateral bins, even if the quality has some changes when varying the central bin, these changes are small and will not produce an important degradation in quality.

This experiment shows, on one side, that the addition of lateral bins provides robustness, as it produces a good estimation for cases where using only the central bin

[1] The extreme cases where the central bin was lower than the number of lateral bins was excluded, because the desired number of lateral bands cannot be used symmetrically

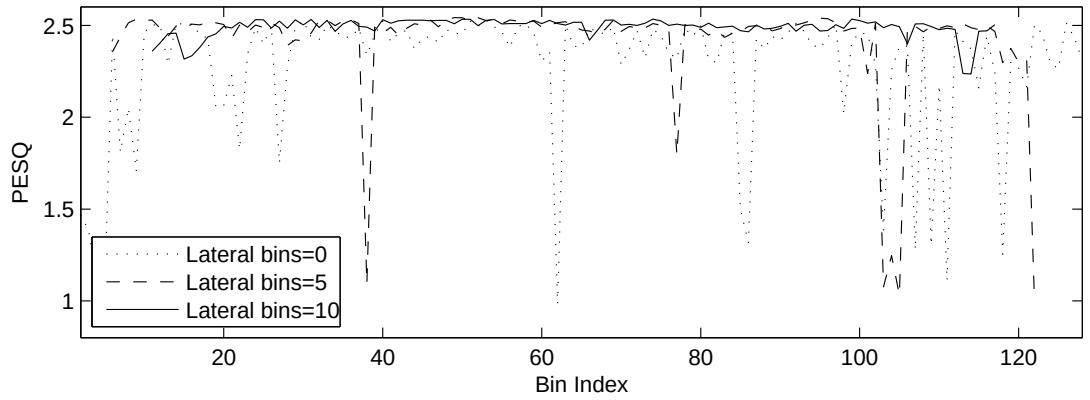


Figure 6.9. Effect of the central bin selection on the quality of separation, for different numbers of lateral bins.

the ICA algorithm fails. On the other side, it shows that the method is quite robust to a wrong selection of central bin, which makes the central bin selection a not so critical aspect.

It is also interesting to analyze the beampatterns generated by this separation method. Figure 6.10 shows these patterns for an example of a real mixture of two speech sources with 0 dB of power ratios. The recording was done as in Figure 6.11, which has the sources at about ± 26 degrees from the central axis perpendicular to the array axis. As it can be seen in the figure, the beampatterns determined for both sources detect correctly the right direction, and also it can be noted that the null location is the same for all frequencies (compare this figure with Figures 5.6 and 4.9).

6.5.5. Evaluation on real mixtures

For this experiment we recorded the same 20 sentences used in Sections 6.5.1 and 6.5.2, contaminated with the same noises, but in a real room as shown in Figure 6.11. The environment is an acoustically isolated room that naturally has a reverberation time of less than $\tau_{60} = 77$ ms. To increase the reverberation time, plywood reflection boards were added in two of the walls. The average reverberation time for this case was about $\tau_{60} = 195$ ms. More details on this database can be found in Appendix A.2.

After mixture, separation was performed using the proposed separation algorithm, the one proposed in [Parra and Spence, 2000] (we will call this Parra), and the one proposed in [Murata et al., 2001] (we will call this Murata). Both algorithms are fd-ICA methods, that obtain independence by exploiting the nonstationary structure of the speech signals, using second-order statistics. Murata algorithm uses the correlation among envelopes of the frequency bins to solve the permutation problem, and Parra

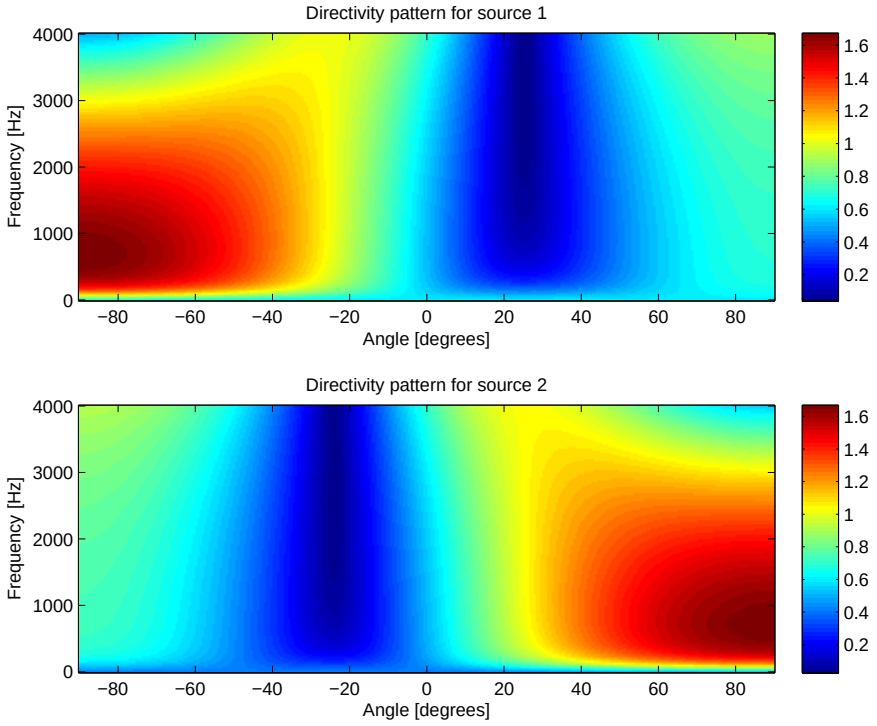


Figure 6.10. Beampaterns generated by the PMBSS method.

algorithm avoids permutation by imposing constrains in the structure of the time-domain separation filters. For Murata algorithm, we used a window length of 256 samples (32 ms) and a frame shifting interval of 10 samples (1.25 ms) and we selected 40 correlation matrices to diagonalize, as suggested by the authors. For Parra algorithm, as the signals used here were of very different durations than the ones reported by the author, we tried several combinations of filter and window lengths (128/1024, 256/512, 256/2048 and 512/3072). The best results were obtained for a filter length of 256 samples (32 ms) with a window length of 512 samples (64 ms).

An implementation of the Parra algorithm was obtained from the author web page[2], and the implementation of Murata algorithm was obtained from Shiro Ikeda web page[3]. All the algorithms were programmed in Matlab language, and the separation tests were ran in a Pentium 4 EMT64 of 3 GHz, with 1GB of RAM. For the proposed algorithm we used two variants, without including the time-frequency Wiener postfilter

[2] `http://newton.bme.columbia.edu/~lparra/publish/`
[3] `http://www.ism.ac.jp/~shiro/research/index.html`

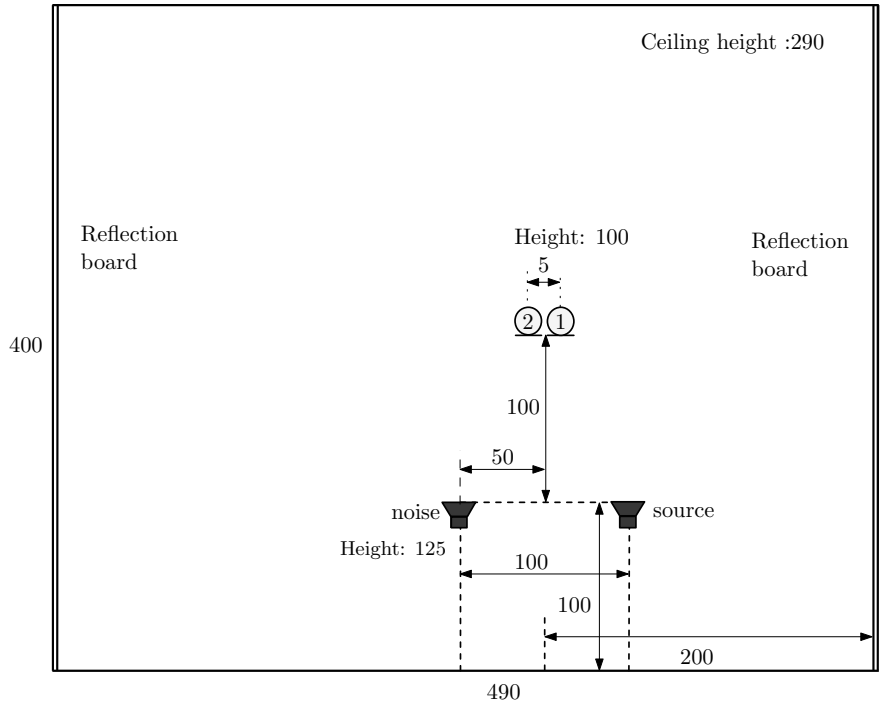


Figure 6.11. Experimental setup for a two sources-two microphones case.

of step 7 (PMBSS I) and with the Wiener filter (PMBSS II).

To test the performance of the algorithm, we used a speech recognition system similar to that described in Section 6.5. The acoustic model was trained with 585 sentences from a subset of Albayzin database (the training set does not include neither any of the sentences used in the test, nor the ones used as interfering voices).

Also, we performed two additional evaluations. One using the PESQ to have a perceptual objective quality evaluation, and another using the average processing time for each of the algorithms. Table 6.3 shows the WRR for this experiment, and Tables 6.4 and 6.5 the PESQ scores, and the average processing time, respectively.

As can be seen, Murata algorithm produces some degradation of both, WRR and PESQ. This is due to the fact that the algorithm cannot handle the reverberation times involved on these tests. The only effect of the processing is to introduce distortions that degrade the performance. For Parra algorithm, some improvement is noted, although it is not enough. The here proposed algorithm can handle the separation in this environment and produces a large improvement in WRR and PESQ. Even if we do not use the Wiener postfilter, the quality of the output of our algorithm outperforms

Table 6.3. Word recognition rate for the evaluated algorithms and the mixtures. PR is the power ratio in the loudspeakers.

PR dB	Noise	Mixed	Murata	Parra	PMBSS I	PMBSS II
6	Speech	44.50	25.00	49.50	83.07	85.50
	White	19.54	15.00	27.50	61.00	85.50
0	Speech	30.00	27.00	49.00	62.50	83.00
	White	7.20	11.00	20.00	24.00	67.50
Ave.		25.31	19.50	36.50	57.64	80.38

Table 6.4. Quality measured as PESQ scores for the evaluated algorithms and mixtures. PR is the power ratio in the loudspeakers.

PR dB	Noise	Mixed	Murata	Parra	PMBSS I	PMBSS II
6	Speech	2.11	1.97	2.22	2.51	2.83
	White	1.98	1.86	2.37	2.57	2.83
0	Speech	1.73	1.71	2.19	2.26	2.59
	White	1.64	1.67	2.16	2.25	2.54
Ave.		1.86	1.80	2.23	2.40	2.70

Table 6.5. Average processing time in seconds for the evaluated algorithms and mixtures. PR is the power ratio in the loudspeakers.

PR dB	Noise	Murata	Parra	PMBSS I	PMBSS II
6	Speech	9.49	6.48	0.36	0.43
	White	8.79	7.01	0.26	0.27
0	Speech	9.56	6.60	0.31	0.42
	White	8.98	6.48	0.24	0.28
Ave.		9.21	6.64	0.29	0.35

the other evaluated alternatives, showing that the separation stage indeed works better than the previous approaches.

For comparisson, we have also evaluated the use of a different postfilter. In other works in the area, a binary mask postfilter was used after a first stage of fd-ICA separation, to improve the results. The main assumption for binary masks is that each time-frequency sample has information of only one source. But for reverberant signals this assumption collapses, and so a continuous mask should produce better results. We have implemented the binary mask postfilter presented in [Mori et al., 2005] and used it instead of our Wiener postfilter (with the same first stage), evaluating the PESQ scores. For speech noise, the PESQ scores obtained were 2.69 and 2.49 for 6 dB and 0

dB of power ratios, respectively. For white noise, the PESQ scores were 2.73 and 2.45 for 6 dB and 0 dB, respectively. These results represent in average only 52.2% of the improvement obtained with our Wiener postfilter.

Regarding the processing times, the proposed algorithm is more than 26 times faster than Murata and more than 18 times faster than Parra algorithms. It should be noted that no special optimization was made in the implementation to make a faster code.

6.6. Concluding remarks

In this chapter a new mixture model was introduced, exploiting the use of closely spaced microphones, which is a usual and even a desirable situation in many applications. This mixture model allows to produce a consistent separation matrix for all frequency bins, which avoids the existence of arbitrary permutations among bins, and also eliminates the scaling ambiguity. By a proper selection of a group of bins instead of isolated ones, the ICA algorithm is improved with a better convergence speed and a better estimation of the mixing matrix. The time-frequency Wiener filter improves even more the quality of the separated signals, providing for increased undesired sources rejection, and also to some degree, reverberation reduction.

The proposed algorithm is very fast, as it needs only one optimization step. The experiments show that with both artificial and real mixtures, it is able to produce a good quality result measured by PESQ, in a fraction of the time needed for the other methods. The recognition rates are better than with any of the considered alternatives, including some standard fd-ICA methods. The processing times are about 1/5 of the real time processing, allowing the use of this method for practical applications. The results presented in this chapter have been published in [Di Persia et al., 2009a] and in japanese conferences [Di Persia et al., 2009b][Di Persia et al., 2009c].

Chapter 7

General discussion and conclusions

In the previous chapters, several contributions to the state of the art in BSS for speech sources have been presented. The contributions ranged from basic methodological results, like the establishment of a consistent experimental protocol for the study of BSS algorithms, to the study of objective methodologies to evaluate the performance of the algorithms, to the development of new separation algorithms based on new mixing models. Although each chapter has a special section with the partial results, it is important to have a comparisson and summarization of all the important findings of this research. The following sections will be devoted to this important subject.

7.1. Comparative results

This doctoral dissertation was devoted to developments in two main areas in the BSS research. One area was methodological, from the point of view of the stablishment of a consistent experimental protocol for the evaluation of BSS algorithms. In this area, a particularly important part was the evaluation and determination of appropriate objective quality measures to evaluate the quality of the results after application of BSS algorithms, as reported in [Di Persia et al., 2008, 2007]. The second area was the development of new separation algorithms. Based on a thoroughly study of the properties and limitations of the standard fd-ICA approach [Handa et al., 2004; Tsutsumi et al., 2004], three new methods for fd-ICA were proposed. The first one introduced a better permutation correction scheme, an improved initialization stage based on combined JADE and FastICA methods, and a postfiltering stage using a time-frequency Wiener filter. This method produced a better separation quality, although with increased computational cost [Di Persia et al., 2006a; Handa et al., 2006]. The second one introduced a strong simplification in the mixing model which makes the method permutation-free, and operates selecting the best time-frequency resolution to use, in an iterative separation scheme [Di Persia et al., 2006c,d,e]. This alternative produced

Table 7.1. Quality measured as PESQ scores for all the methods explored and for the mixtures.

Power ratio	Noise	Mixtures	Murata	Parra	JFW	MRBSS	PMBSS
0 dB	Speech	2.11	1.97	2.22	2.89	2.43	2.83
	White	1.98	1.86	2.37	2.84	2.56	2.83
6 dB	Speech	1.73	1.71	2.19	2.54	2.37	2.59
	White	1.64	1.67	2.16	2.47	2.28	2.54
Average		1.86	1.80	2.23	2.69	2.41	2.70

a faster method with relatively good quality, but has some distortions introduced by a beampattern that varies with frequency. The third method uses a simplification of the mixture model based on close spaced microphones, that can produce the desired beampatterns, based on a stabilized convergence of FastICA, without indeterminecies, and that performs even better in terms of processing speed [Di Persia et al., 2009a,b,c]. In this section we will unify the obtained results, comparing the three aspects according to our proposed experimental protocol: objective quality measured by PESQ, average processing time and word recognition rate.

In order to contrast the performance of the developed methods with other approaches, they will be compared to two state-of-the-art fd-ICA approaches that have been and are widely used to compare performances. These are the algorithms by Parra [Parra and Spence, 2000] and Murata [Murata et al., 2001]. The data used for this comparison consists of the 20 utterances mixtures in a real room, the same used in each chapter to report the recognition rate results of the proposed methods (a subset of the Spanish database described in Appendix A).

Table 7.1 presents the PESQ scores [ITU, 2001] for the different methods. This score has a range from -0.5 to 4.5, and for example, a G.726 ADPCM vocoder at 16 kbps gives a PESQ score for Spanish speech signals of 2.56 [ITU, 2005]. As can be seen in this table, our methods outperform the evaluated alternatives. With respect to the mixtures, the quality improvement is significant, ranging from scores of 0.55 for MRBSS, to 0.82 for JFW and to 0.84 for PMBSS.

Table 7.2 presents the word recognition rates for the same methods. As a reference point, the clean sources produced a WRR of 93%. All the proposed methods behave better than the alternative state-of-the-art ones, confirming the prediction of PESQ scores as analyzed in Chapter 3. The best method is clearly the PMBSS.

Finally, Table 7.3 presents the average processing time for the alternatives evaluated. All the methods were programed in Matlab and evaluated on the same computer to provide consistent results. The methods from other authors were obtained from their web pages and were not altered in any way. Also the table presents the average time

Table 7.2. Word recognition rate % for all the methods explored and for the mixtures.

Power ratio	Noise	Mixtures	Murata	Parra	JFW	MRBSS	PMBSS
0 dB	Speech	44.50	25.00	49.50	84.00	70.50	85.50
	White	19.54	15.00	27.50	84.50	73.00	85.50
6 dB	Speech	30.00	27.00	49.00	79.50	70.50	83.00
	White	7.20	11.00	20.00	69.00	36.00	67.50
Average		25.31	19.50	36.50	79.25	62.50	80.38

Table 7.3. Average processing time in seconds for all the methods explored and average recognition time for the mixtures.

Power ratio	Noise	Mixtures	Murata	Parra	JFW	MRBSS	PMBSS
0 dB	Speech	0.34	9.49	6.48	11.41	0.70	0.43
	White	0.33	8.79	7.01	10.15	0.74	0.27
6 dB	Speech	0.37	9.56	6.60	11.63	0.73	0.42
	White	0.66	8.98	6.48	17.55	0.69	0.28
Average		0.42	9.21	6.64	12.68	0.72	0.35

used by the recognizer to perform recognition on the mixtures (in the column labelled Mixtures). This parameter is usefull to compare the processing time for separation, because for a practical application the total time needed for recognition (including the separation preprocessing and the recognition itself) would be a critical point for real-time applications. As can be seen, the PMBSS method uses about the same time as the needed for recognition. Given that the average duration of the sentences used was 3.55 seconds, it seems that the use of PMBSS as preprocessing for the ASR system can be done in almost real time.

As it can be seen, the JFW algorithm achieved a higher quality than Murata and Parra ones, but at the cost of increased processing time. By contrast, MRBSS takes only 5.7% of the time of JFW, although this speed comes at the price of a reduction of 16.75% in recognition rate relative to JFW (but it still performs better than Murata and Parra). Finally, PMBSS produces a better recognition rate than JFW, and moreover, using only 2.8% of its processing time.

7.2. Main findings

During the development of this research, several improvements were introduced, that resulted in a high speed and high quality method for fd-ICA based BSS. Each method was developed aiming at the solution of the main problems of standard fd-ICA methods. The main findings of this study are:

- In the area of BSS experimentation and evaluation:
  1. A wide evaluation of objective quality measures as predictors of the recognition rate has been performed. As a result, the best measures were those related to perceptual evaluation, like PESQ and its nonlinear variant.
  2. The very bad performance of the segmental SNR has been demostrated. This is an important aspect because of its widespread use in the scientific comunity.
  3. A consistent framework for the experimentation on BSS applied to ASR has been developed, both using real and artificial mixtures.
- In the area of separation algorithms themselves:
  1. A more robust initialization for standard fd-ICA methods has been proposed.
  2. A way to improve the permutation alignment method for standard fd-ICA methods has been developed.
  3. The use of a simplified model with only one mixing matrix for all frequency bins has been proposed. It has resulted in a fast and relatively good separation method, using multiple resolutions in an iterative approach. The main advantage of the method is the absence of ambiguities among frequency bins. We have studied its limitations and the reason for them, mainly the defective beampattern generated by the hypothesis.
  4. A new pseudoanechoic mixture model for close spaced microphones has been proposed. Under this model the separation can be accomplished by a very efficient algorithm which avoids the ambiguities, and thus provides high quality results.
  5. The use of lateral bins to provide robustness to the ICA method was proposed, and we have shown that this alternative provides faster results and robustness. Although this methodology was used in combination with the pseudoanechoic model to produce the PMBSS method, it can be also used with other approaches.

6. For all kind of algorithms, the use of Wiener time-frequency filters has been shown to produce an important quality improvement by reduction of the competing sources and their echoes, and also by partially reducing the auto-echoes.
7. Significant and important improvements of the recognition rates have been obtained. For our best algorithm the WRR improvement with respect to using the mixtures was of 55.07%, and the processing time was only 10% of the duration of the signals.

This list of advances shows that all the proposed objectives of this doctoral research have been accomplished.

7.3. Future work

From these findings, there is much work left for the future. Each one of the proposed algorithms has many parameters that can be improved, and that have the potential to even increase their performance. A list of them includes the following aspects:

1. In relation to the objective quality measures, it would also be desirable to evaluate the performance of the measures as predictors of subjective quality. This would require the design of subjective listening test with a large number of listeners to provide significant results. Having a complete characterization of the quality measures, this would provide for an improvement of the framework to include applications for human listeners.

2. The reverberation still has a degrading effect. Although the proposed algorithms can handle a wider range of reverberation conditions that standard fd-ICA methods, they are not immune to it. Moreover, MRBSS and PMBSS provide results where the reverberation is still present, although the separation was achieved. As reverberation is a degrading aspect for many applications, including ASR, some methods to eliminate the reverberation effect of the separated sources may be developed. During this research some advances were done in that direction [Ohta et al., 2005], but further work is needed.

3. The Wiener postfilter used was very basic and there is space for many improvements. For example, the 2-norm was used as an estimation of power, but in many works related to Wiener filters, other exponents have been reported to produce better results. Moreover, the filter for each time-frequency slot was estimated using information of only that slot, but due to the reverberation, it is likely that

previous values of time could be used to improve the estimation. These aspects show a wide field for research and experimentation of improvements, in combination with any of the proposed methods.

4. Also the Wiener filter can be seen as a continuous masking related to the binary masking methods used in CASA and sparsity based approaches. This research line can be expanded with knowledge from those areas to produce new methods for both separation and postfiltering.

5. The method of packaging of lateral bins that was used to stabilize the ICA algorithm in PMBSS is very interesting and opens many possibilities. The first one is researching on the automatic selection of the best central bin and the best number of lateral bins. Some ideas in this direction include the usage of correlation measurements among the signals of a bin, or their mutual information as indexes of separability, and selecting the more separable one as the central bin. Also, this method can be extended to provide robustness to standard fd-ICA approaches. Some research was already done in that direction [Di Persia et al., 2006b; Noguchi et al., 2006; Ohta et al., 2008; Yasui et al., 2006], although an important number of alternatives are left for further research.

7.4. Conclusion

The main objectives of this doctoral research were to produce methods based on machine learning techniques aimed at improving the recognition rate of automatic speech recognition systems. Three alternative algorithms were presented, based on the fd-ICA approach. All of them produced considerably higher recognition rates than standard alternatives. Each one of the methods had their own advantages and disadvantages. The first one, called JFW, allows for a high quality separation, but the computational cost is higher than the standard methods. The second proposed method, called MRBSS, sacrifices a little quality with respect to JFW (although it is still better than the standard approaches) but produces a large improvement in the computation speed. The third proposed method, PMBSS, produces results with higher quality than JFW, and with higher processing speed than MRBSS, being the best method obtained. All of them are more robust to reverberation than the state-of-the-art methods evaluated, thus fulfilling other of our main objectives. Also, a consistent and complete experimental framework for designing the experiments and for the objective evaluation of the separation algorithms has been established. We can affirm that all the proposed objectives have been accomplished. Moreover, the development of the methods has opened a large number of opportunities for future improvements and many research lines can be started from this point.

Appendix A

Speech and impulse responses databases

At several moments during the development of this research, it became necessary to obtain databases of clean, reverberant and mixed signals in diverse environments. These recordings were made with the help of some special equipment and with software specially programmed for those tasks. For all the environments used, special care was taken in the evaluation of the reverberation times measurements as well as other acoustic characteristics, also with the help of some specially produced software. In the following sections, the different data set used will be described in detail[1]

A.1. Japanese TV-commands database

This database was designed to study the capabilities of a TV-set remote controlling system. The underlying idea in such a system is to be able of use voice commands to control the TV status and produce a high level navigation through the programs. In this way, a person can, for instance, request the list of all channels in where some program related to the keyword "sports" is being broadcasted or programmed. This task was accomplished as part of a group of researchers working on this application in the Intelligent Mechanisms Laboratory, jointly with Omron Corporation.

This kind of system proposes a hard to solve problem, for many reasons. One of them is the need to work in a remote and location-independent way. Also this task must be performed under a wide variety of noise and competing sound sources (for example, even the same sound being emitted by the TV-set itself acts as a competing source). A database using several standard commands in Japanese language was recorded.

[1]Most of the software needed for those tasks was written with the help of the Speech Processing Library, specially programmed for this study and which is described in Appendix B

Table A.1. Japanese commands for a TV set (phonetic transcription and its approximate meaning).

Command	Meaning
kikoenai	"I can't hear"
urusai	"Its too loud"
cancel	Cancel the previous command
NHK	The name of a Japanese broadcasting station
yon channel	"Channel four"
roku	"Six"
Kansai terebi	"Kansai TV", a broadcasting station
BS	"Channels via Broadcasting Satellite"
spohrts news	News programs on sports
eega	"Movies"
News station	The name of a news program
taiga dorama	A genre of TV programs
Hanshin taigaas	The name of a baseball team
Waarudo kappu	"World cup"
Hosino kantoku	The name of a baseball team
Bekkamu	"Beckham", the name of a soccer player
Ohsaka	"Osaka", name of a city
Okinawa	Name of a Japanese perfecture
Amerika	"America"
Europa	"Europe"

The database consisted of two kinds of materials, the desired commands, and several kind of noises used to interfer with them. The first material consisted in the clean database, composed by 20 typical commands that could be issued to such a system. These commands were recorded in a sound-proof room, with close to mouth microphones to avoid the incidence of noise. The reverberation time in this room is lower than 100 ms, and as close to mouth microphones were used, this avoided the effect of echoes. The microphones used were Onno-Sokki, model MI 1233 with flat response from 20 Hz to 20 kHz and preamplifiers Onno Sokki MI 3110. To produce this corpus, 10 male speakers and 10 female speakers were selected. The sampling frequency used was 44000 Hz, downsampled later to 16000 Hz. The command list is presented in table A.1, where the phonetic transcription of the Japanese commands has been written, together with its approximate meaning in English.

The clean speech corpus was complemented with three kind of noises, used to

contaminate the clean sources, representing the typical kinds of noise that can happen in the system working conditions. One of the noises was speech, uttered by speakers different than those used to record the commands. Two long sentences were recorded, with contents that did not overlap with the commands. The spoken sentences were pronounced by one male and one female speaker.

To evaluate the effect of self induced noise by the TV set, a long recording of a TV program, during broadcast of an advertisement including speech and background music, was also selected. Finally, the last kind of noise was the one produced by several computers in a cluster room. This noise presents the effect of several coolers, fans and diverse sounds generated by computers.

All the speech material was produced in one recording session, and later it was postprocessed with bandpass filtering to eliminate low frequency coupling and line noise. The bandpass filter was a butterworth of order 4, with passband from 70 to 7900 Hz.

This database was recorded with the assistance of the members of the Intelligent Mechanisms Laboratory, and was employed in the experiments of Chapter 3, in which the clean signals were used to generate several real room mixtures with different combinations of localizations, sound powers and in different reverberation conditions.

A.2. Spanish database

We wanted to produce a database of Spanish speech recordings, mixed in a real reverberant room, in different conditions of reverberation, power and interfering noise. Three reverberation conditions were used, in two different environments. A sound proof room (that is, an acoustically isolated room to avoid the interference of exterior sound sources) with a small reverberation time was used. Also, to produce a longer reverberation time, we added plywood reflection boards covering all the surface of two of the sound proof room walls. Finally, we utilized a living room simulator, constructed in a similar size and with similar materials than usual Japanese living rooms.

For these rooms, we selected some positions for the microphone and sources, as shown in Figures A.1 and A.2. We replayed the desired source from one speakerphone, while at the same time reproducing the competing noise in the other one, and recorded the sound field through 4 measurement microphones spaced by 5 cm among each. Microphone marked as 1 is nearer to the desired source. These recordings were performed following the protocol presented in Chapter 3. Figure A.3 shows a photograph of the four microphones used in the recordings, and Figure A.4, the experimental setup in the living room simulator.

After recording the mixtures, each sound source was individually replayed, in order to have a record of the sources only affected by the room (but without competing

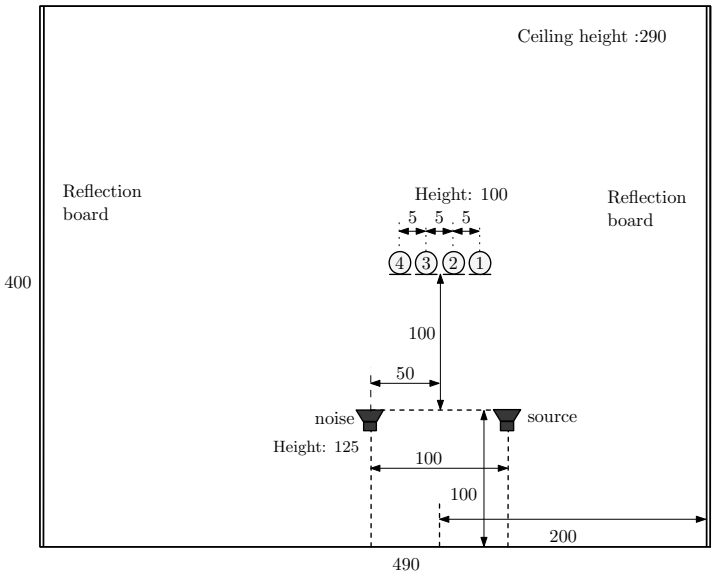


Figure A.1. Sources and microphones locations for the Spanish sound proof recordings. All dimensions are in cm.

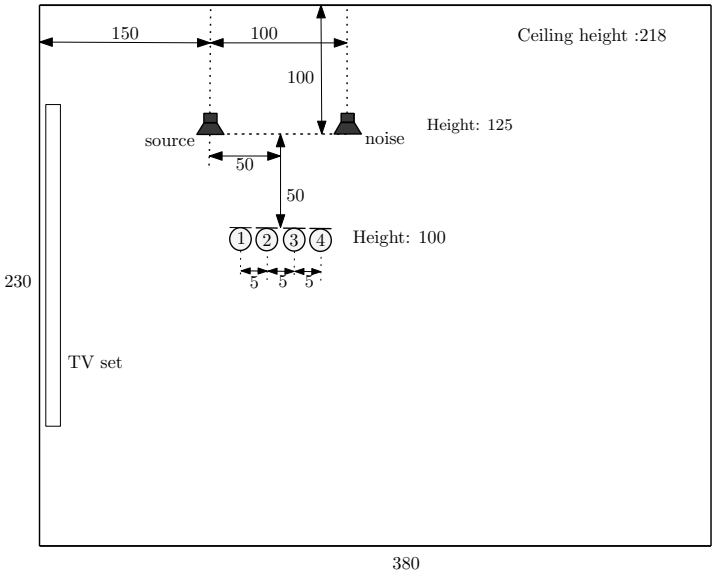


Figure A.2. Sources and microphones locations for the Spanish living room simulator recordings. All dimensions are in cm.

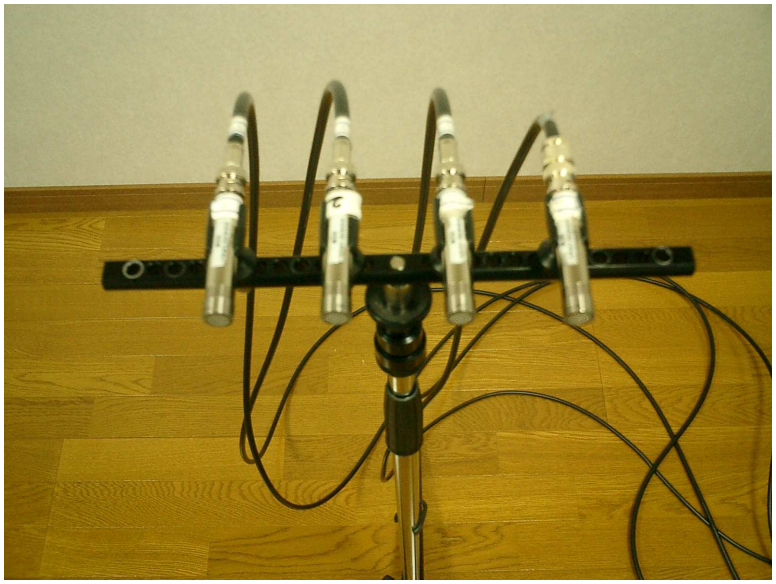


Figure A.3. Microphones used in the Spanish recordings.

Figure A.4. The experimental setup in the living room simulator.

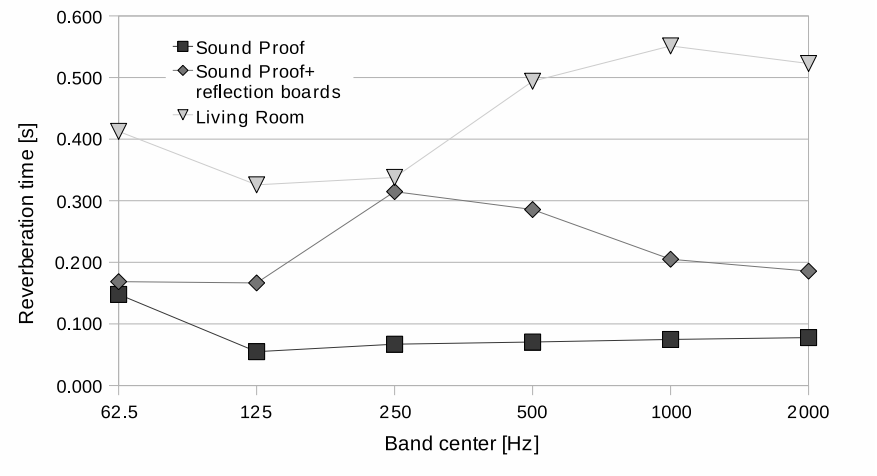


Figure A.5. Average reverberation time τ_{30} in octave bands, for the three environments used in the Spanish database.

sources). This data could be valuable to develop reverberation reduction methods.

For this database, 5 sentences from Albayzin database [Moreno et al., 1993] were selected; those sentences are: 3002, 3006, 3010, 3018 and 3022, according to the syntax of the Albayzin corpus. Four human speakers (2 male, 2 female) were used, their Albayzin codes are: aagp and algp (female); mogp and nyge (male). The Albayzin name convention includes a speaker code followed by a sentence number, so the sentences names are of the kind: aagp3002.au. The signals are sampled at 8000 Hz, with 8 bits resolution (with μ-law encoding, equivalent to about 14 bits of linear coding resolution). The average duration of the 20 sentences used was 3.55 s.

Two kind of interfering signals were used. One is white noise, obtained from Noisex database [Varga and Steeneken, 1993], which was subsampled to 8000 Hz and only the first 5 s were used. The other kind of noise was speech. For this, sentence 3110 was used. To interfere with male spoken sentences, female speech was used and for interfering with female spoken sentences, male speech was used. For those, aagp3110 and nyge3110 were used as female and male speech noise respectively.

As mentioned previously, for each environment two different power ratios of source and noises were used, 0 and 6 dB. For this, the signals were generated in the form of 2 channels wav files, being channel 1 the source and channel 2 the noise. The amplitude of noise in each case was modified to provide a power ratio for that pair of signals according to the desired one.

For all cases, the impulse responses from all sources to all microphones were measured using the TSP methodology presented in Chapter 1 and a software produced

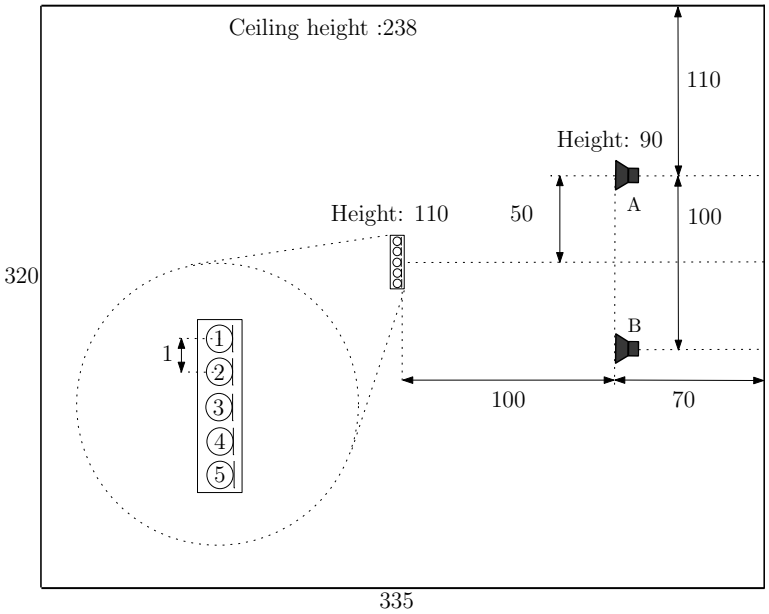


Figure A.6. Spatial setup in the room used for the impulse responses recording. All dimensions are in cm.

with the library described in Appendix B. For each room we evaluated the global τ_{30} reverberation time as an average for all the source-microphones locations, and also we evaluated it in octaves. For the sound-proof room, the average reverberation time is 77 ms, for the sound proof room with reflection boards is 195 ms, and for the living room simulator is 530 ms. Figure A.5 shows the average reverberation time in octaves, for the three rooms. For this evaluation, a software programmed using the library presented in Appendix B was also used.

This database was used to evaluate the methods presented in Chapters 4, 5 and 6. In those chapters, we used the data from the central microphones numbered 2 and 3, recorded in the sound proof room with two reflection boards, for the evaluation.

A.3. Impulse responses for artificial mixtures

This set of recordings was designed to have a bank of well known impulse responses, in order to produce artificial convolutive mixtures using them. A room with moderate reverberation time was selected. The room was a bedroom, with carpet in the floor and a curtain covering a window. It had some furniture, a desktop, a chair and two beds.

The measurement of the impulse responses was done using a standard desktop

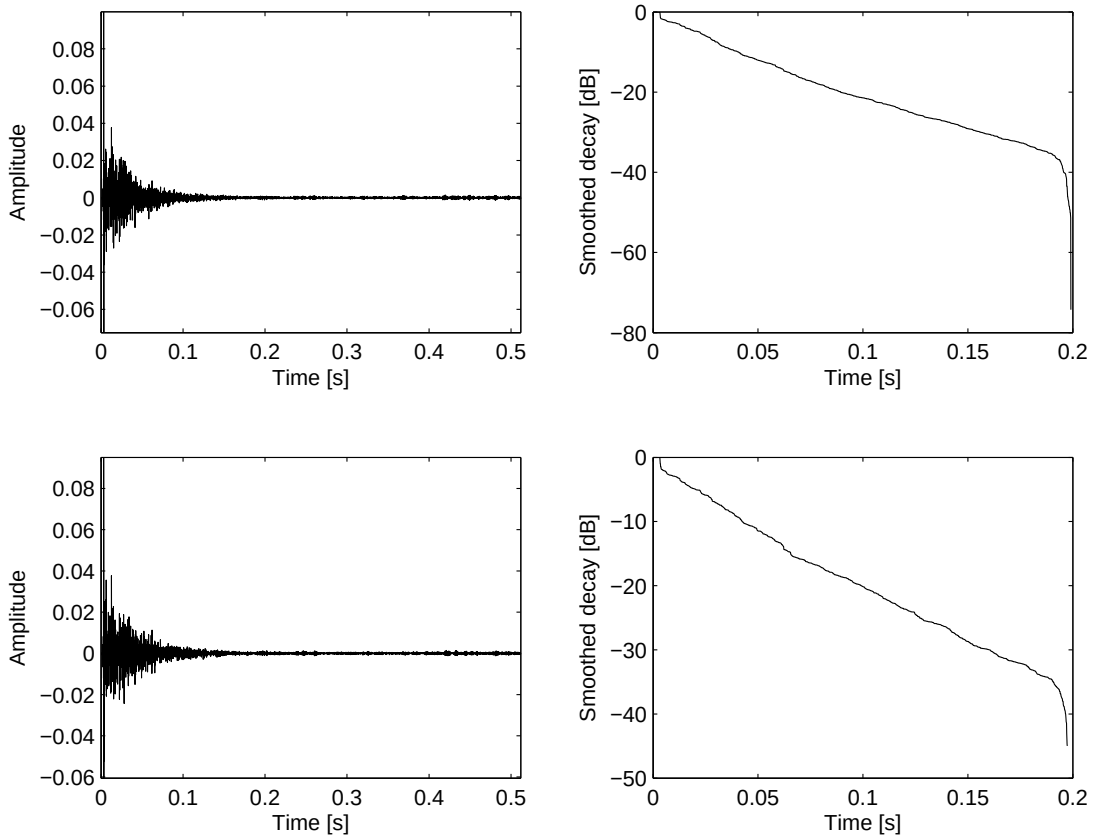


Figure A.7. Two of the impulse responses and its corresponding Schroeder backintegration curve.

microphone with plain frequency response from 20 Hz to 20 KHz. The recording was done in a spatial setup as depicted in Figure A.6. Five microphone locations were used, spaced 1 cm among them. Two loudspeakers with flat frequency response from 20 Hz to 20 kHz, model Edifier R18, with 2 watt of power, were used as sound sources to replay the excitation sounds. The measurements were done using the TSP technique as described in Chapter 1 [Suzuki et al., 1995]. This method allows for a low noise (the noise floor was measured at -59 dB).

The sampling frequency was fixed at 8000 Hz. For the TSP method, stimulus signals of one second were generated. These stimulus were replayed twice and the responses were synchronously averaged to improve the signal to noise ratio of the measurement. The method used a circular deconvolution with the inverse stimulus to produce the impulse response.

Synchronization is an important subject. We wanted the recordings to allow for a precise estimation of the delay introduced by the sound propagation in its arrival to two microphones located next to each other. If we wanted this delay to be preserved in the recording it would be necessary to know exactly both the starting time of reproduction, and the latency of the sound recording and reproducing system. It must be noted that the mentioned delay is fractional, that is, it is a fraction of the sampling time. After the recordings, using an upsampling to ten times the sampling frequency we verified that the measured delays were in coincidence with the predicted ones according to the spatial setup.

The resulting recording then consists of the impulse responses from each one of the two sources locations (A and B in Figure A.6) to each one of the microphone locations (labelled 1 to 5 in the figure) for a total of ten impulse responses. The reverberation time τ_{30} was measured for each one of these impulse responses, using the Schroeder backintegration method [Schroeder, 1979] to smooth the power decay curve. The average global reverberation time for the room (averaged over the ten impulse responses) was 349 ms.

Figure A.7 presents, as an example, two of the measured impulse responses and their corresponding power decay curves. In the first row, the impulse response from the source located at the point labelled A to the microphone 1 is presented, while in the botton row, the impulse response from source A to microphone 5 is shown.

To have an idea of the frequency dependent variations of the impulse responses, we also have studied for each of the impulse responses, the reverberation time τ_{30} but after bandpass filering with octave band filters (as usually recommended for evaluation of room acoustics). Figures A.8 and A.9 present this analysis for the impulse responses measured from source A and B respectively. It can be seen that although there are small variations in the reverberation times, all the impulse responses from one source to the five microphones have similar characteristics.

All the software used for recording this impulse responses as well as the one for analyzing the impulse response characteristics was programmed by the author using the library presented in Appendix B. The impulse responses of this database were used in the artificial mixture experiments presented in Chapter 6.

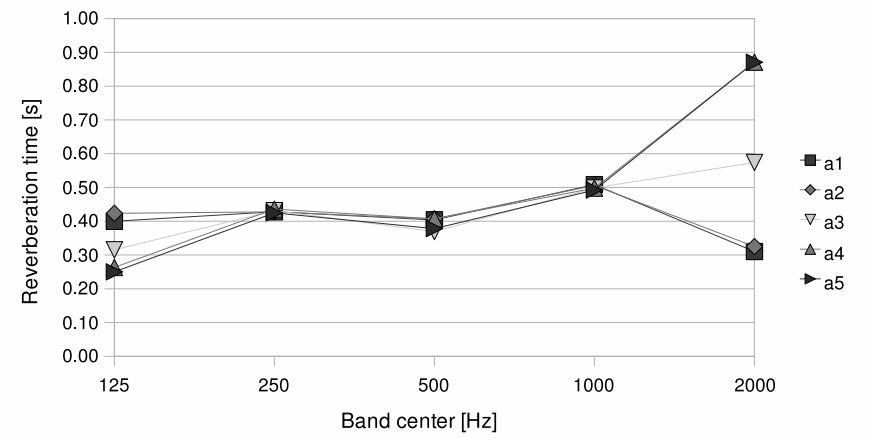


Figure A.8. Reverberation times by octave for the impulse responses measured from source A to all the microphones in the IR database.

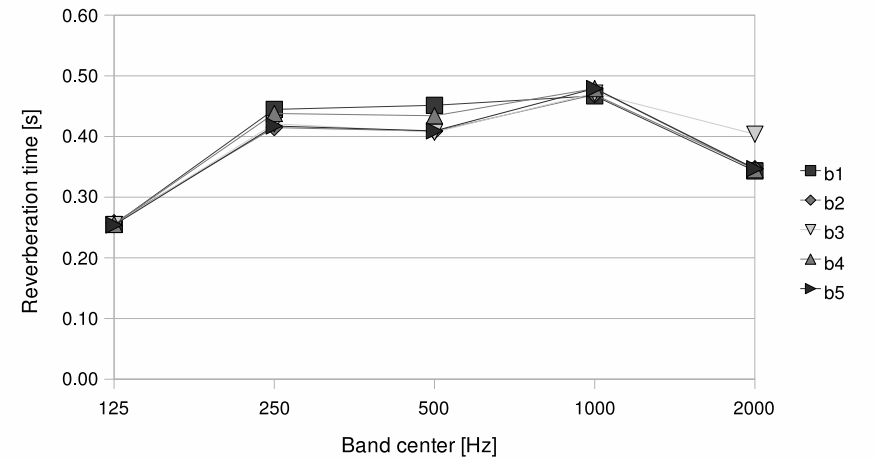


Figure A.9. Reverberation times by octave for the impulse responses measured from source B to all the microphones in the IR database.

Appendix B

C++ library for speech processing

During the development of this doctoral research, there was a need for the implementation of advanced methods and algorithms in C++ language. Although many of the methods were initially developed in Matlab, mainly because of graphics capabilities that allow for a fast verification of parameter effects, this language is interpreted and cannot compete with the speed of execution of compiled code. Due to this, in parallel with the algorithms development, a library of routines in C++ to support the research was developed. In its actual form, it includes more that 700 functions. Being a general speech processing library, it needs functions to perform all kind of operations, from creation and manipulation of vectors and matrices, to arithmetic functions over these entities, to statistical functions, algebra, signal processing operations, transformations, etc. These kind of operations were implemented to work both with real and complex valued elements. It includes low-level operations like matrix-vector multiplication or linear filtering, to high level functions like pitch contour determination, acoustic measurement functions, ICA algorithms, etc. This library will be available at the author's website to help other researchers in their work[1]. Its name is SPeech Processing library (SPP). In the following a brief review of its main capabilities and some functions specially related to this research will be presented.

B.1. General description

This is a library for speech analysis and processing, mainly oriented to speech quality improvement, noise reduction and sources separation. It is strongly based on Matrix and Vector types of both real and complex values. The aim is to provide a framework for quick development of new algorithms. An important part of this is compatibility with Matlab/Octave as they are widely used tools to implement new

[1] http://ldipersia.wikidot.com/software

algorithms. So the task flow usually consists of: developing new algorithm in Matlab, debuging it, refining it, and when it works well, translating it to C++ using this library. The code is mainly a C code because it almost does not make use of advanced programing techniques like classes and templates, though it uses some of C++ upgrades for versatility (for example, operator overloading). This has some reason: the main intention is to make this library cross-platform and compiler-independent compatible. Therefore, a special effort in making the lesser usage of capabilities that are (potentially) not supported for most common versions of compilers was made.

This library is intended for research purposes. It is not specially optimized either for speed nor memory usage, although some care has been taken to reduce overhead and increase speed. It must be at least as fast as Matlab in most of the code and usually faster.

All the library functions are documented by means of comments in each .h file. During the installation process, the documentation is compiled[2], generating a set of html files describing all the library capabilities. The library is released as open source software, under a very simple MIT license. To use the library, the programmer only needs to include the header SPP.h in his/her program main file.

B.2. Library structure

The library is built around some basic data types. It includes a type for double precision complex scalars, with the use of operator overloading to provide the usual set of operations on them. It also provides a complete set of mathematical functions over the complex scalars.

Since the signals are represented as real and complex vectors, and most of the signal processing operations can be defined in terms of vector operations or matrix-vector operations, the library has types to define both, matrices and vectors of several types, like double precision real, complex, integer, unsigned integer values. It also provides a wide set of functions to perform simple operations over these data types, like creating them, resizing, obtaining their size, taking norms, adding them, etc.

The central idea of the library was to provide fast prototyping of algorithms and also to be function-compatible with Matlab. In this way, many of the usual functions available in Matlab to manipulate vectors and matrices have been implemented in the library.

There are many functions to operate with matrices and vectors. No functions to open and save files of known formats have been included because there are many free libraries to do that, with many kinds of standard formats (like wav, au, snd) and for

[2]Using the Doxygen documentation generation program.

several platforms. So it was assumed that the potential user has his/her own functions to get the speech input.

In file *sppfiles.h* there are very simple functions to save data in text files and in a binary format. They are intended just to save temporal data or allow the import of processed files from other software.

In file *sppsig.h* there are several functions for common signal processing operations like convolution, filtering, filter design, etc.

In file *spptran.h* there are several functions to take transformations of signals. This includes only one dimensional Fourier, cosine and wavelet transform and some time-frequency analysis.

In the files *sppralg.h* and *sppcalg.h* there are functions for linear algebra operations.

In the file *sppbss.h* there are functions to perform PCA, ICA and BSS of speech signals.

In the file *spprmath.h* there are functions to operate with real numbers that are not implemented in the programming language, like random number generation with normal distribution.

File *spproom.h* has functions for room acoustic measurements and simulations.

File *sppconve.h* has some auxiliary functions to convert between SPP types and some other standard C++ types.

B.3. Functions used in this study

During this research, many of the needed processing was done using this library. We will provide a short description of the most relevant functions used.
Functions used for impulse response measurements and reverberation times determination:

- TSP_Stimulus: This function generates the TSP stimulus signals used for impulse response measurements.
- Impulse_TSP: This function takes the recorded room response to the TSP stimulus and the inverse stimulus as inputs, and determines the impulse response of the room.
- Octave_filter: This function designs an octave filter, used for the measurement of octave-band reverberation times.
- Schroeder: Implements the reverse Schroeder integration used to determine the smoothed decay of impulse responses.

- RevTime: This function calculates the global reverberation times and the reverberation times by octave. It uses the Schroeder backintegration method and produces estimations of τ_{20}, τ_{30} and the early decay time.

Functions for short-time Fourier transform and filtering:

- STFT: Calculates the STFT of a real-valued signal.
- ISTFT: Calculates the inverse STFT of a real-valued signal.
- spl: Calculates number of frames in STFT, given the input signal, the window length and the window step size.
- ispl: Returns the length of a reconstructed signal by ISTFT, given the matrix of the STFT, the window length and the window step size.
- FastConvOLA: Fast convolution using FFT and overlap-and-add method. This is usefull and very fast when the filter length is much smaller than that one of the signal to be filtered. Used for example to produce artificial mixtures.

Functions for whitening, ICA and Blind Source separation:

- Whitening: Performs whitening of a real and complex-valued matrices.
- FastICA: Performs the FastICA algorithm for real and complex-valued data.
- JADE: Performs the JADE algorithm proposed by Cardoso for real and complex-valued signals.
- SimpleBSS: This function implements the PMBSS method presented in Chapter 6, for 2 by 2 mixtures.

Bibliography

Allen, J. B. and Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950.

Allen, J. B. and Rabiner, L. R. (1977). A unified approach to short-time Fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558–1564.

Amari, S.-I., Douglas, S. C., Cichocki, A., and Yang, H. H. (1997). Multichannel blind deconvolution and equalization using the natural gradient. In *Proceedings of the IEEE Workshop on Signal Processing Advances in Wireless Communications*, pages 101–104.

Aoshima, N. (1981). Computer-generated pulse signal applied for sound measurement. *The Journal of the Acoustical Society of America*, 69(5):1484–1488.

Araki, S., Makino, S., Sawada, H., and Mukai, R. (2004). *Independent Component Analysis and Blind Signal Separation*, chapter Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA, pages 898–905. Lecture Notes in Computer Science. Springer.

Araki, S., Mukai, R., Makino, S., Nishikawa, T., and Saruwatari, H. (2003). The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Transactions on Speech and Audio Processing*, 11(2):109–116.

Araki, S., Sawada, H., Mukai, R., and Makino, S. (2007). Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. *Signal Processing*, 87(8):1833–1847.

Asano, F., Ikeda, S., Ogawa, M., Asoh, H., and Kitawaki, N. (2001). Blind source separation in reflective sound fields. In *Proceedings of the International workshop on hands-free speech communication*, pages 51–54.

Asano, F., Ikeda, S., Ogawa, M., Asoh, H., and Kitawaki, N. (2003). Combined approach of array processing and independent component analysis for blind separation of acoustic signals. *IEEE Transactions on Speech and Audio Processing*, 11(3):204–215.

Barnwell, T. P. (1980). Correlation analysis of subjective and objective measures for speech quality. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 706–709.

Basseville, M. (1989). Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349–369.

Beerends, J. G. and Stemerdink, J. A. (1994). A perceptual speech quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 42(3):115–123.

Benesty, J., Makino, S., and Chen, J., editors (2005). *Speech Enhancement*. Signals and Communication Technology. Springer.

Beranek, L. (2004). *Concert Halls and Opera Houses: Music, Acoustics, and Architecture*. Springer, 2nd ed. edition.

Beranek, L. L. (1996). *Acoustics*. The Acoustical Society of America.

Berkhout, A. J., de Vries, D., and Boone, M. M. (1980). A new method to acquire impulse responses in concert halls. *The Journal of the Acoustical Society of America*, 68(1):179–183.

Bingham, E. and Hyvärinen, A. (2000). A fast fixed-point algorithm for independent component analysis of complex valued signals. *International Journal of Neural Systems*, 10(1):1–8.

Blauert, J. (1997). *Spatial Hearing - the Psychophysics of Human Sound Localization*. MIT Press.

Brandstein, M. and Ward, D., editors (2001). *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 1 edition.

Cardoso, J.-F. and Souloumiac, A. (1993). Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370.

Cichocki, A. and Amari, S. (2002). *Adaptive Blind Signal and Image Processing. Learning Algorithms and applications*. John Wiley & Sons.

Crandell, C. and Smaldino, J. (2000). Classroom acoustics for children with normal hearing and with hearing impairment. *Language, Speech, and Hearing Services in Schools*, 31(4):362–370.

Deller, J., Proakis, J., and Hansen, J. (1993). *Discrete Time Processing of Speech Signals*. Macmillan Publishing, New York.

Di Persia, L., Milone, D., Rufiner, H. L., and Yanagida, M. (2008). Perceptual evaluation of blind source separation for robust speech recognition. *Signal Processing*, 88(10):2578–2583.

Di Persia, L., Milone, D., and Yanagida, M. (2006a). Convolutive blind source separation with Wiener post-filtering for robust speech recognition. In *Proceedings of the Argentinian Symposium of Tecnology*. Sociedad Argentina de Informática.

Di Persia, L., Milone, D., and Yanagida, M. (2006b). Multi-bin independent component analysis and blind sound source separation device using the same. Japanese Patent Application 2006-199420, Universidad Nacional de Entre Ríos, Universidad Nacional del Litoral, Doshisha University.

Di Persia, L., Milone, D., and Yanagida, M. (2006c). Permutation-free blind source separating method and device. Japanese Patent Application 2006-199420, Universidad Nacional de Entre Ríos, Universidad Nacional del Litoral, Doshisha University. En Tramite.

Di Persia, L., Milone, D., and Yanagida, M. (2009a). Indeterminacy free frequency-domain blind separation of reverberant audio sources. *IEEE Transactions on Audio, Speech and Language Processing*, 17(2):299–311.

Di Persia, L., Morimoto, Y., and Yanagida, M. (2009b). Blind Source Separation based on a Simplified Mixing Model. *IEICE Technical Report*, 108(411):109–114.

Di Persia, L., Morimoto, Y., and Yanagida, M. (2009c). Permutation-Free Blind Source Sseparation based on a Simplified Receiving Model. In *Spring Meeting of Acoustical Society of Japan*.

Di Persia, L., Noguchi, T., Ohta, K., and Yanagida, M. (2006d). Performance of permutation-free ica. *IEICE technical report. Speech*, 106(78):1–6.

Di Persia, L., Ohta, K., and Yanagida, M. (2006e). A method for solving the permutation problem in ica. *IEICE technical report. Speech*, 105(686):53–58.

Di Persia, L., Yanagida, M., Rufiner, H. L., and Milone, D. (2007). Objective quality evaluation in blind source separation for speech recognition in a real room. *Signal Processing*, 87(8):1951–1965.

Divenyi, P., editor (2004). *Speech Separation by Humans and Machines*. Springer, 1 edition.

Doclo, S. and Moonen, M. (2003). Design of far-field and near-field broadband beamformers using eigenfilters. *Signal Processing*, 83(12):2641–2673.

Douglas, S. C., Gupta, M., Sawada, H., and Makino, S. (2007). Spatio-temporal fastica algorithms for the blind separation of convolutive mixtures. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1511–1520.

Douglas, S. C. and Sun, X. (2003). Convolutive blind separation of speech mixtures using the natural gradient. *Speech Communication*, 39(1-2):65–78.

Everest, F. A. (2001). *Master Handbook of Acoustics*. McGraw-Hill, 4 edition.

Finitzo-Hieber, T. and Tillmann, T. (1978). Room acoustics effects on monosyllabic word discrimination ability by normal and hearing impaired children. *Journal of Speech and Hearing Research*, 21(3):440–458.

Fox, B., Sabin, A., Pardo, B., and Zopf, A. (2007). Modeling perceptual similarity of audio signals for blind source separation evaluation. In *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation*, pages 454–461.

Furui, S. (1989). *Digital Speech Processing, Synthesis, and Recognition*. Marcel Dekker Inc.

Gales, M. (1998). Predictive model-based compensation schemes for robust speech recognition. *Speech Communication*, 25(1-3):49–74.

Gilkey, R. and Anderson, T., editors (1997). *Binaural and Spatial Hearing in Real and Virtual Environments*. Lawrence Erlbaum Associates.

Gong, Y. (1995). Speech recognition in noisy environments: A survey. *Speech Communication*, 16(3):261–291.

Gotanda, H., K. Nobu, Koya, T., Kaneda, K., Ishibashi, T., and Haratani, N. (2003). Permutation correction and speech extraction based on split spectrum through fastica. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 379–384.

Gray, A. and Markel, J. (1976). Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5):380–391.

Gray, R. M., Buzo, A., Gray, A. H., and Matswama, Y. (1980). Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):367–376.

Gribonval, R., Benaroya, L., Vincent, E., and Févotte, C. (2003). Proposals for performance measurement in source separation. In *Proceedings of the 4th International*

Symposium on Independent Component Analysis and Blind Signal Separation, pages 763–768.

Grimm, M. and Kroschel, K., editors (2007). *Robust Speech Recognition and Understanding*. I-Tech Education and Publishing.

Handa, A., **Di Persia**, L., Ohta, K., and Yanagida, M. (2006). Separation of mixed speech signals of short duration using wiener filter as postprocessing for frequency-domain ica. *Information Processing Society of Japan SIG Notes*, 2006(12):1–6.

Handa, A., Tsutsumi, K., **Di Persia**, L., and Yanagida, M. (2004). Separation of mixed speech in reverberant environment using independent component analysis. Spring Meeting of Japanese Society of Information Processing.

Hansen, J. H. and Arslan, L. M. (1995). Robust feature estimation and objective quality assessment for noisy speech recognition using the credit card corpus. *IEEE Transactions on Speech and Audio Processing*, 3(3):169–184.

Hansen, J. H. and Pellom, B. (1998). An effective quality evaluation protocol for speech enhancement algorithms. In *Proceedings of the Interternational Conference on Spoken Language Processing*, volume 7, pages 2819–2822.

Hawley, M. L., Litovsky, R. Y., and Colburn, H. S. (1999). Speech intelligibility and localization in a multi-source environment. *The Journal of the Acoustical Society of America*, 105(6):3436–3448.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752.

Hermansky, H. and Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589.

Hu, Y. and Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238.

Huang, X. D., Ariki, Y., and Jack, M. A. (1991). *Hidden Markov Models for Speech Recognition*. Edinburgh Universitary Press.

Huang, Y. A. and Benesty, J., editors (2004). *Audio Signal Processing for next-generation multimedia communication systems.* Kluwer Academic Press.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons, Inc.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430.

Ikeda, S. and Murata, N. (1998). An approach to blind source separation of speech signals. In *Proceedings of the 8th International Conference on Artificial Neural Networks*, volume 2, pages 761–766.

Ikram, M. and Morgan, D. (2002). A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.*, volume 1, pages I–881–I–884.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):67–72.

ITU (2001). Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *ITU-T Recommendation P.862*.

ITU (2005). Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2. *ITU-T Recommendation P.862.3*.

Jelinek, F. (1998). *Statistical Methods for Speech Recognition*. The MIT Press.

Juang, B.-H. and Rabiner, L. (2005). Automatic Speech Recognition - A Brief History of the Technology. In *Encyclopedia of Language and Linguistics*. Elsevier.

Jutten, C. (1987). *Calcul neuromimétique et traitement du signal, analyse en composantes indépendantes*. These d'etat, INPG.

Jutten, C. and Herault, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10.

Kahrs, M. and Brandenburg, K., editors (2002). *Applications of Digital Signal Processing to Audio and Acoustics*. The Kluwer International Series In Engineering And Computer Science. Kluwer Academic Publishers.

Katayama, Y., Ito, M., Barros, A. K., Takeuchi, Y., Matsumoto, T., Kudo, H., Ohnishi, N., and Mukai, T. (2004). Closely arranged directional microphone for source separation. In *Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation*, pages 129–135.

Kawahara, T., Lee, A., Kobayashi, T., Takeda, K., Minematsu, N., Sagayama, S., Itou, K., Ito, A., Yamamoto, M., Yamada, A., Utsuro, T., and Shikano, K. (2000). Free software toolkit for Japanese large vocabulary continuous speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*, volume 4, pages 476–479.

Kawamoto, M., Matsuoka, K., and Ohnishi, N. (1998). A method of blind separation for convolved non-stationary signals. *Neurocomputing*, 22(1-3):157–171.

Kinsbury, B. and Morgan, N. (1997). Recognizing reverberant speech with RASTA-PLP. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1259–1262.

Klatt, D. H. (1982). Prediction of perceived phonetic distance from critical-band spectra: a first step. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1278–1281.

Koehler, J., Morgan, N., Hermansky, H., Hirsch, H. G., , and Tong, G. (1994). Integrating RASTA-PLP into speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 421–424.

Kuttruff, H. (2000). *Room Acoustics*. Taylor & Francis, fourth edition.

Lambert, R. H. (1996). *Multichannel blind deconvolution: Fir matrix algebra and separation of multipath mixtures*. PhD thesis, University of Southern California, Department of Electrical Engineering.

Lavandier, M. and Culling, J. F. (2008). Speech segregation in rooms: Monaural, binaural, and interacting effects of reverberation on target and interferer. *The Journal of the Acoustical Society of America*, 123(4):2237–2248.

Lee, A., Kawahara, T., and Shikano, K. (2001). Julius — an open source real-time large vocabulary recognition engine. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1691–1694.

Lee, I., Kim, T., and Lee, T.-W. (2007). Fast fixed-point independent vector analysis algorithms for convolutive blind source separation. *Signal Processing*, 87(8):1859–1871.

Lee, T.-W. (1998). *Independent Component Analysis - Theory and Applications*. Springer, 1st edition.

Lee, T.-W., Bell, A. J., and Orglmeister, R. (1997). Blind source separation of real world signals. In *Proceedings of the International Conference on Neural Networks*, volume 4, pages 2129–2134.

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, 19(22):1–15.

Manolakis, D. G., Ingle, V. K., and Kogon, S. M. (2005). *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*. Artech House Publishers.

Mansour, D. and Juang, B. H. (1989). A family of distortion measures based upon projection operation for robust speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(2):1659–1671.

Matsuoka, K. (2002). Minimal distortion principle for blind source separation. In *Proceedings of the SICE Annual Conference*, volume 4, pages 2138–2143.

Melia, T. and Rickard, S. (2007). Underdetermined Blind Source Separation in Echoic Environments Using DESPRIT. *EURASIP Journal on Advances in Signal Processing*, 2007:19 pages.

Mitianoudis, N. and Davies, M. (2001). New fixed-point ica algorithms for convolved mixtures. In *Proceedins of the 3rd International Conference on Independent Component Analysis and Signal Separation*, pages 633–638.

Miyoshi, M. and Kaneda, Y. (1988). Inverse filtering of room acoustics. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(2):145–152.

Montgomery, D. C. and Runger, G. C. (2003). *Applied Statistics and Probability for Engineers*. Third Edition, third edition.

Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J., and C. Nadeu (1993). Albayzin speech database design of the phonetic corpus. Technical report, Universitat Politècnica de Catalunya (UPC), Dpto. DTSC.

Mori, Y., Saruwatari, H., Takatani, T., Ukai, S., Shikano, K., Hiekata, T., and Morita, T. (2005). Real-time implementation of two-stage blind source separation combining SIMO-ICA and binary masking. In *Proceedings of the 2005 International Workshop on Acoustic Echo and Noise Control*, pages 229–232.

Murata, N., Ikeda, S., and Ziehe, A. (2001). An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41:1–24.

N. Nocerino, Soong, F. K., Rabiner, L. R., and Klatt, D. H. (1985). Comparative study of several distortion measures for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 25–28.

Noguchi, T., Ohta, K., Yanagida, M., and **Di Persia**, L. (2006). Frequency-domain independent component analysis by overlap piecewise integration of separation processing. In *Proceedings of the of the Fourth Joint Meeting of the ASA and ASJ*, volume 120, pages 3048–3048. JASA.

Ohta, K., **Di Persia**, L., and Yanagida, M. (2005). Removing reflected waves using temporal and spectral subtraction without a priori knowledge. In *International Workshop on Nonlinear Signal and Image Processing (NSIP 2005)*, page 16. IEEE.

Ohta, K., Noguchi, T., Yasui, K., **Di Persia**, L., and Yanagida, M. (2008). Frequency Domain ICA Connecting of Adjacent Frequency Bins. *IEICE Transactions on Information and Systems (Japanese Edition)*, J91-D(1):130–135.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15:267–273.

Oja, E. (1989). Neural networks, principal components, and subespaces. *International Journal of Neural Systems*, (1):61–68.

O'Shaughnessy, D. (2008). Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, 41(10):2965–2979.

Park, H.-M. and Stern, R. (2006). Spatial separation of speech signals using continuously-variable masks estimated from comparisons of zero crossings. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 1165–1168.

Parra, L. and Alvino, C. (2002). Geometric source separation: merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, 10(6):352–362.

Parra, L. and Spence, C. (2000). Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, 8(3):320–327.

Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall.

Rix, A., Beerends, J., Hollier, M., and Hekstra, A. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 749–752.

Rufiner, H. L., Torres, M. E., Gamero, L., and Milone, D. H. (2004). Introducing complexity measures in nonlinear physiological signals: application to robust speech recognition. *Physica A: Statistical Mechanics and its Applications*, 332(1):496–508.

Sakamoto, S., Seimiya, T., and Tachibana, H. (2002). Visualization of sound reflection and diffraction using finite difference time domain method. *Acoustical Science and Technology*, 23(1):34–39.

Saruwatari, H., Kurita, S., Takeda, K., Itakura, F., Nishikawa, T., and Shikano, K. (2003). Blind source separation combining independent component analysis and beamforming. *EURASIP Journal on Applied Signal Processing*, 2003(11):1135–1146.

Sawada, H., Mukai, R., Araki, S., and Makino, S. (2004). A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Transactions on Speech and Audio Processing*, 12(5):530–538.

Schobben, D., Torkkola, K., and Smaragdis, P. (1999). Evaluation of blind signal separation methods. In *Proceedings of the 1st International Workshop on Independent Component Analysis and Blind Signal Separation*, pages 261–266.

Schroeder, M. R. (1965). New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(3):409–412.

Schroeder, M. R. (1979). Integrated-impulse method measuring sound decay without using impulses. *The Journal of the Acoustical Society of America*, 66(2):497–500.

Srinivasan, S., Roman, N., and Wang, D. (2006). Binary and ratio time-frequency masks for robust speech recognition. *Speech Communication*, 48(11):1486–1501.

Sun, H., Shue, L., and Chen, J. (2004). Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume I, pages 865–868.

Suzuki, Y., Asano, F., Kim, H.-Y., and Sone, T. (1995). An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses. *The Journal of the Acoustical Society of America*, 97(2):1119–1123.

Tsutsumi, K., Handa, A., **Di Persia**, L., and Yanagida, M. (2004). Validity studies of blind source separation using independent component analysis in real environment. *Technical Report of IEICE. EA*, 103(609):1–6.

Van Veen, B. and Buckley, K. (1988). Beamforming: a versatile approach to spatial filtering. *ASSP Magazine, IEEE*, 5(2):4–24.

Varga, A. and Steeneken, H. (1993). Assessment for automatic speech recognition II NOISEX- 92: A database and experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251.

Vincent, E., Gribonval, R., and Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech and Language Processing*, 14(4):1462– 1469.

Voiers, W. D. (1977). Diagnostic acceptability measure for speech communication systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 204–207.

Voran, S. (1999a). Objective estimation of perceived speech quality. I. Development of the measuring normalizing block technique. *IEEE Transactions on Speech and Audio Processing*, 7(4):371–382.

Voran, S. (1999b). Objective estimation of perceived speech quality .II. Evaluation of the measuring normalizing block technique. *IEEE Transactions on Speech and Audio Processing*, 7(4):383–390.

Wang, D. and Brown, G. J., editors (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press.

Yamada, T., Kumakura, M., and Kitawaki, N. (2006). Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):2006–2013.

Yasui, K., Noguchi, T., Ohta, K., **Di Persia**, L., and Yanagida, M. (2006). Source separation by frequency-domain ica on locally connected frequency bins. *IEICE Technical Report. Natural language understanding and models of communication*, 106(441):59–64.

Yilmaz, O. and Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2005). *The HTK book (for HTK Version 3.3)*. Cambridge University Engineering Department, Cambridge.

This thesis was written in LaTeX, compiled with pdfTeX and edited with Kile.