Técnicas de procesado en array para realzado de voz en situaciones adversas

Autor: David Martínez Pérez

Director: José Javier López Monfort

Resumen

La voz es la principal forma de comunicación del ser humano. En los últimos años diversos avances tecnológicos como la telefonía móvil, los sistemas de videoconferencia o los dispositivos controlados mediante el habla, han propiciado un nuevo campo de investigación destinado al realzado de la señal de voz. Esto ha causado la proliferación de novedosas técnicas de tratamiento de señal aplicadas a tal fin.

En esta tesina, hemos llevado a cabo una revisión del estado del arte de los dos grupos principales de técnicas de realzado de voz: el beamforming y la separación ciega de fuentes de audio. Varios algoritmos han sido implementados y comprobados mediante diversos prototipos de arrays de micrófonos. Además, se ha propuesto un novedoso método de procesado en array para separación de fuentes en tiempo real. También se muestran resultados obtenidos mediante experimentos con grabaciones en salas reales.

Abstract

Speech is the main form of communication of the human being. In the last years, several technological advances, such as mobile communications, videoconferencing and speech-controlled systems, have brought about a new research field aimed at enhancing speech signals. For this purpose, several processing techniques have been recently developed.
In this work, we have carried out a state-of-the-art review of the two main speech enhancement techniques: beamforming and blind audio source separation. Several algorithms haven been implemented and tested over a set of microphone-array prototypes. Furthermore, we have proposed a novel array processing method for real-time audio separation. Experiments and results using recordings obtained in real rooms are discussed.

Author: Martínez Pérez, David, email: damarpe8@teleco.upv.es
Director: López Monfort, José Javier, email: jjlopez@dcom.upv.es
Fecha de entrega: 05-12-08

Índice general

Capítulo 1

Introducción

1.1. Procesado de señal en un array de micrófonos

Un *Array* es un conjunto ordenado de elementos. En nuestro caso, será un conjunto de transductores electroacústicos o micrófonos dispuestos con una determinada geometría, con el objetivo de mejorar la comunicación. Esto es posible ya que la capacidad de transducción del array depende de la posición espacial de la fuente que está recibiendo. Esta propiedad básica de todo array es conocida como directividad, y permite una captación máxima de señal en una determinada dirección atenuando las señales procedentes del resto de direcciones que serán consideradas como interferencias.

Esta tesina está enfocada a mejorar la calidad de la voz captada por un array de micrófonos. Existen muchas aplicaciones, (comunicaciones con manos libres, interfaces hombre-máquina, ayuda para personas con problemas auditivos, videoconferencia, etc) en las que se requiere el uso de micrófonos. La señal de interés normalmente es contaminada por ruido de fondo, reverberación e interferencias. El objetivo de esta tesina es comparar métodos del estado de arte en este campo así como proponer nuevas alternativas.

La señal de voz deseada y el ruido interferente normalmente ocupan bandas de frecuencia solapadas, por lo tanto, técnicas de realce de voz de solo un micrófono, como substracción espectral o filtrado de Kalman encuentran problemas para reducir el ruido de fondo, especialmente, sin introducir ruido musical audible o distorsión en la señal deseada. Si la fuente deseada y las fuentes interferentes están físicamente situadas en diferentes posiciones espaciales, es posible explotar esta diversidad espacial utilizando un array de micrófonos, de tal manera que podemos emplear juntas las características espaciales y las espectrales para construir algoritmos de realce de voz.

En concreto nos centraremos en las dos grandes técnicas que se vienen empleando. El beamforming o conformación de haz y la separación ciega de fuentes de audio.

1.2. Beamforming

El beamforming ha sido estudiado desde antiguo en muchas áreas, tales como radar, sonar, sismología o comunicaciones [1]. En el área de la mejora de voz, aplicaciones típicas son, manos libres para telefonía móvil, sistemas de ayuda a la escucha, y sistemas controlados por voz. Las señales grabadas generalmente están degradadas por una considerable cantidad de ruido de fondo [2]. Si queremos mejorar la SNR (Relación Señal a Ruido) de una señal sin deteriorar ésta, lo mejor es emplear algoritmos de beamforming.

El beamforming es una técnica de procesado de señal en la que se produce una mejora de la señal principal mediante combinaciones lineales de las diferentes salidas de cada uno de los micrófonos del array, de tal manera que las perturbaciones indeseadas se atenúan por el fenómeno de la directividad o selectividad espacial del array. Debido a que empleamos combinaciones lineales, no se introduce distorsión en la señal ni tampoco ruido musical, que normalmente es provocado por filtrados no lineales.

1.3. Separación ciega de fuentes

La mayoría de las señales de audio son mezcla de varias fuentes de audio, las cuales, están activas simultáneamente. Por ejemplo, los debates en directo son mezclas de varios locutores, los discos son mezclas de instrumentos musicales y cantantes, las bandas sonoras son mezclas de voces, música y sonidos naturales. La Separación Ciega de Fuentes de Audio (Blind Audio Source Separation, BASS), es el problema de recuperar cada señal fuente a partir de una señal mezclada [3].

1.4. Estructura de la tesina

Esta tesina de master se estructura de la siguiente manera:

En el Capítulo 2, hablaremos de la técnica de procesado en array más empleada, el beamforming (o también beam forming). Cuando la mejora de la señal principal se realiza mediante combinaciones lineales de las diferentes salidas de cada uno de los micrófonos del array, se habla de beamforming, de tal manera que las perturbaciones indeseadas se atenúan por el fenómeno de la directividad o selectividad espacial del array. En primer lugar, para comprender mejor como funciona el beamforming, se explicará la técnica más sencilla conocida como retardo y suma. A continuación se describen dos configuraciones de arrays muy empleadas, el array lineal uniforme y el array circular uniforme. Este último es el que hemos empleado en esta tesina para implementar los algoritmos. Para finalizar se muestran otros algoritmos más robustos para mejorar la supresión de ruido e interferencias.

El capítulo 3 se centra en otras técnicas conocidas como separación ciega de fuentes de audio. Estas técnicas son muy empleadas para intentar solventar el llamado problema "cocktail party effect". Este problema se basa en la capacidad de los humanos para centrarse en un sólo locutor cuando hay un número elevado de conversaciones (o interferencias) y un elevado ruido de fondo. Explicaremos una de las técnicas más empleadas de separación ciega de fuentes y extenderemos su uso al procesado en array, gracias a lo cual, podemos mejorar los resultados.

En el capítulo 4 se muestran las pruebas realizadas para evaluar los diferentes algoritmos propuestos.

Por último, en el capítulo 5 se encuentra un resumen de las conclusiones extraídas de la tesina.

Capítulo 2

Beamforming

Mediante el beamforming un receptor puede discriminar entre diferentes señales incidentes, dependiendo de cuál sea la localización espacial de las mismas. El beamforming es por tanto la manera más sencilla de mejorar una señal mediante un array [4]. Si apuntamos el eje de máxima captación de un array hacia la fuente deseada se estará atenuando la señal procedente de otras fuentes no situadas en dicho eje, y que se consideran como fuentes de ruido. Además, se atenuará el sonido procedente de las reflexiones en las paredes, el techo y el suelo de la sala reduciendo por lo tanto el nivel de reverberación.

Seguidamente se expone el modelo de señal que emplearemos cuando vayamos a referirnos al beamforming con un array de micrófonos.

2.1. Modelo de señal

Supongamos que solamente incide al array una fuente sonora, omnidireccional y de banda estrecha situada en el punto de coordenadas esféricas (r, θ, φ). Dicha fuente producirá en el origen de coordenadas una señal de presión acústica:

$$p(t) = p_0 e^{j(\omega t + \phi)} \tag{2.1}$$

con p_0 la presión acústica (valor eficaz), ω la pulsación de la vibración acústica producida por la fuente y ϕ su fase inicial. Considerando únicamente el camino directo de transmisión acústica, que une la fuente al micrófono, y que la fuente es omnidireccional, dicha fuente creará sobre el micrófono i-ésimo una onda de presión de la forma:

$$p_i(r_i^{'}, t) = p(t)\frac{r}{r_i^{'}} e^{-j\omega \frac{r_i^{'} - r}{c}} \tag{2.2}$$

donde, como se muestra en la figura 2.1 r es la distancia entre la fuente y el centro del array, y $r_i^{'}$ la distancia entre la fuente y el micrófono i-ésimo.

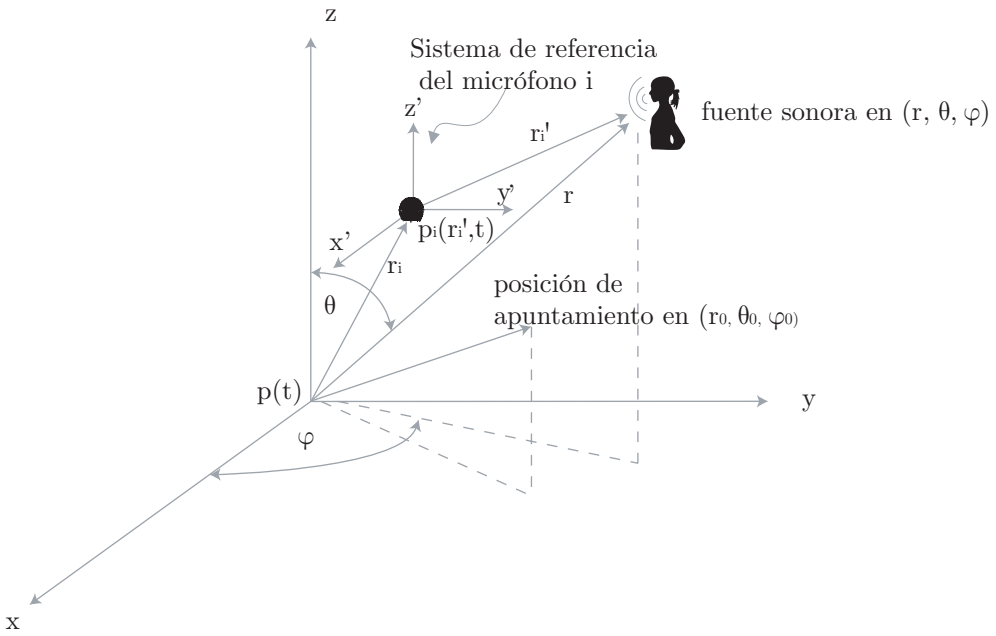


Figura 2.1: Sistema de referencia empleado para definir tanto las fuentes como el array de micrófonos.

Como herramienta para simplificar las expresiones, definiremos el micrófono de referencia. Este micrófono tiene índice cero (0) y está situado en el eje coordenado, es decir, las coordenadas de la fuente F cumplen $(r, \theta, \varphi) = (r_0^{'}, \theta_0^{'}, \varphi_0^{'})$. Dicho micrófono de referencia, puede ser uno de los integrantes del array, aunque no es necesario. Supondremos que el micrófono de referencia tiene una sensibilidad S_0 y es omnidireccional $D_0(\theta, \varphi) = 1$, por lo que si recibiese la onda de presión $p(t)$ daría como respuesta eléctrica:

$$x_0(t) = S_0 p(t) \tag{2.3}$$

Entonces, suponiendo todos los micrófonos con la misma sensibilidad y el mismo patrón de directividad, podemos expresar la señal eléctrica en el micrófono i-ésimo como:

$$x_i(t) = \frac{r}{r_i^{'}} e^{-j\omega \frac{r_i^{'} - r}{c}} x_0(t) \tag{2.4}$$

que se puede escribir también como:

$$x_i(t) = a_i(r_i^{'}, \theta_i^{'}, \varphi_i^{'}) x_0(t) \tag{2.5}$$

con,

$$a_i(r_i^{'}, \theta_i^{'}, \varphi_i^{'}) = \frac{r}{r_i^{'}} e^{-j\omega \frac{r_i^{'} - r}{c}} \tag{2.6}$$

que representa la respuesta particular del micrófono i ante una excitación procedente de la posición r, y relativa a la respuesta del micrófono de referencia. Si ahora consideramos la existencia de I micrófonos en el array, obtendremos un elemento de respuesta por cada micrófono, de tal manera que se hablará del vector de apuntamiento (*steering vector*):

$$\underline{a}(r,\theta,\varphi) = [a_1(r_1^{'},\theta_1^{'},\varphi_1^{'}), a_2(r_2^{'},\theta_2^{'},\varphi_i^{'}), \ldots, a_I(r_I^{'},\theta_I^{'},\varphi_I^{'})]^T \tag{2.7}$$

en el que se representan distintas atenuaciones y retardos, para cada uno de los micrófonos, en función de la posición de la fuente. De la misma manera, podemos expresar la salida eléctrica de cada micrófono en forma vectorial:

$$\underline{x}(t) = \underline{a}(r,\theta,\varphi)x_0(t) \tag{2.8}$$

Hasta ahora, hemos supuesto una única fuente, pero, en un entorno real, lo más habitual es tener una fuente de interés y varias interferentes, e incluso, en algunas aplicaciones, como pueden ser la grabación de debates en directo o la vigilancia, existen varias fuentes de interés, que además pueden estar activas en el mismo instante (figura 2.2). Si suponemos que existen M emisores diferentes, podemos definir la matriz de apuntamiento o enfoque (*steering matrix*) de dimensiones IxM, de la siguiente manera [5]:

$$\underline{\underline{A}}(r,\theta,\varphi) = [\underline{a}_1(r_1,\theta_1,\varphi_1), \ldots, \underline{a}_M(r_M,\theta_M,\varphi_M)]^T \tag{2.9}$$

con lo que la ecuación del array quedará como:

$$\underline{x}(t) = \underline{\underline{A}}(r,\theta,\varphi)\underline{x}_0(t) \tag{2.10}$$

donde $\underline{x}_0(t)$ representa el vector respuesta de referencia del array $\underline{x}_0(t) = [x_{01}(t), \ldots, x_{0M}(t)]^T$ Además, en un entorno real tendremos ruido. Si suponemos ruido blanco, gaussiano y aditivo, la ecuación del array quedará:

$$\underline{x}(t) = \underline{\underline{A}}(r,\theta,\varphi)\underline{x}_0(t) + \underline{n}(t) \tag{2.11}$$

La suposición de que el ruido aditivo es totalmente incoherente, es excesiva ya que una fuente ruidosa situada cerca del array puede ser captada por los micrófonos del array con elevada coherencia intercanal. No obstante, para simplificar el análisis seguiremos suponiendo ruido espacialmente blanco.

2.2. Beamformer de retardo y suma

El objetivo fundamental de todo beamforming es el de obtener a la salida de nuestro procesador una señal $y(t)$, que represente lo más fielmente posible la excitación acústica $p(t)$ que produce la fuente de interés en el centro del array, eliminando o atenuando en

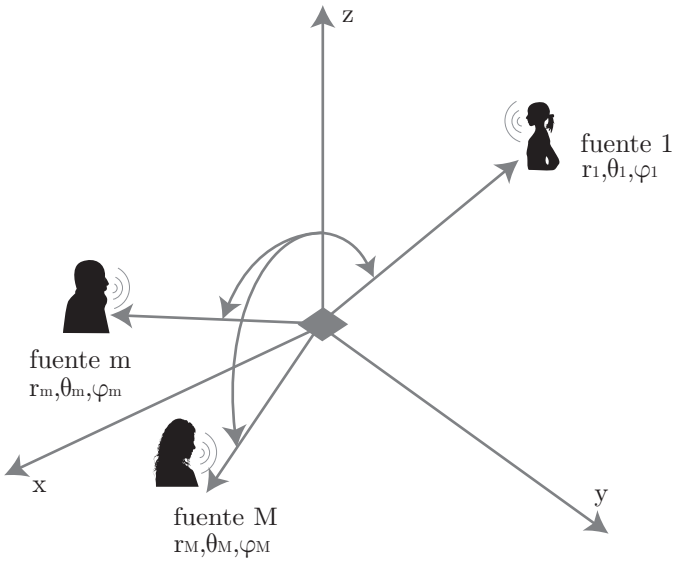


Figura 2.2: Array de micrófonos que capta M fuentes.

mayor medida las fuentes de perturbación ajenas a la principal (ruido, reverberación o fuentes secundarias). El beamforming se basa en combinaciones lineales de las diferentes salidas de cada uno de los micrófonos del array. Su estructura más sencilla es la de retardo y suma (Delay and Sum, DS). Con esta estructura, todas las salidas de los micrófonos del array son multiplicadas por un coeficiente, constante o variable con la frecuencia, para ser posteriormente sumadas y producir la salida única del array $y(t)$. Para ello, cada canal microfónico es filtrado con un filtro w_i.

Inicialmente, supondremos que estamos procesando una señal de banda estrecha, (una sinusoide de una sola frecuencia). El beamformer de retardo y suma, consiste en dos etapas básicas de procesado. La primera etapa consiste en desplazar la señal recibida por cada sensor, para compensar el tiempo que tarda, una onda incidente al array, en recorrer la diferencia de caminos entre dicho sensor y el de referencia. Podemos hablar entonces del vector de coeficientes de filtrado $\underline{w} = [w_1, w_2, \ldots, w_I]^T$, que dependerá del ángulo de llegada de la fuente de interés al array. Si el módulo de los coeficientes no depende de la frecuencia, y su fase es proporcional a dicha frecuencia, el proceso de filtrado, que equivale a retardar cada una de las señales del array queda:

$$y_i = w_i^* x_i(t) \tag{2.12}$$

donde el $*$ representa el operador conjugado. El segundo paso, consiste en sumar las señales desplazadas, dando como salida del beamformer DS:

$$y(t) = \sum_{i=1}^{I} w_i^* x_i(t) = \underline{w}^H \underline{x}(t) = \underline{w}^H \underline{a}(r, \theta, \varphi) x_0(t) \tag{2.13}$$

siendo H el operador Hermitiano (transpuesto y conjugado). Si ahora consideramos señales de banda ancha, tal y como es el caso de la voz, los coeficientes variarán con la frecuencia, por lo que en el dominio del tiempo habrá que considerar la convolución de la señal con la respuesta impulsiva de los pesos. De esta manera, obtenemos:

$$y(t) = \sum_{i=1}^{I} w_i^*(t) * x_i(t) \tag{2.14}$$

El proceso de filtrado puede definirse tanto en el dominio del tiempo, como en el de la frecuencia, como se muestra en la figura 2.3. Sin embargo, la implementación práctica es muy diferente. El filtrado temporal requiere filtros FIR o IIR, mientras que el filtrado frecuencial modifica directamente el espectro de la señal. En este último caso, es necesario realizar la transformada de Fourier de la señal temporal como fase previa al filtrado, pero aprovechándonos de las implementaciones basadas en FFT, podemos conseguir filtrados computacionalmente más eficientes en el dominio de la frecuencia. Podemos reescribir la ecuación 2.8, en el dominio frecuencial como:

$$\underline{X}(f) = \underline{A}(f, r, \theta, \varphi)\underline{U}(f, r, \theta, \varphi)X_0(f) + \underline{N}(f) \tag{2.15}$$

donde el término $\underline{U}(f, r, \theta, \varphi)$ representa la directividad de los micrófonos, $X_0(f)$ es la señal incidente en el dominio frecuencial. Y finalmente, como se muestra en la figura 2.3 la respuesta del array en el dominio frecuencial es:

$$Y(f) = \sum_{i=1}^{I} W_i^*(f)X_i(f) = \underline{W}^H\underline{X}(f) \tag{2.16}$$

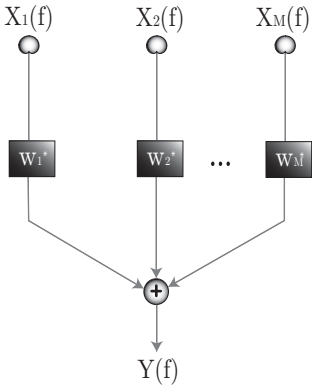


Figura 2.3: Esquema frecuencial del beamformer DS.

En sistemas reales, el conjunto de vectores, $\underline{W}(f)$, es una matriz compleja de tamaño $L \times I$, donde L es el número de frecuencias discretas (o bins) e I es el número de micrófonos. A cada conjunto de pesos $\underline{W}(f)$, le corresponde una determinada forma de haz (o

Beampattern) $B(f, \theta, \varphi)$, que representa la directividad del beamformer y que podemos calcular como:

$$B(f, r, \theta, \varphi) = \sum_{i=1}^{I} W_i^*(f) A_i(f, r, \theta, \varphi) U_i(f, r, \theta, \varphi) \tag{2.17}$$

Ahora nos queda la cuestión de cómo calcular los pesos. El retardo y suma se basa en maximizar la potencia del array, por lo que el vector de pesos, tal y como se demuestra en [5], viene dado por:

$$\underline{w}(r_0, \theta_0, \varphi_0) = \frac{\underline{a}(r_0, \theta_0, \varphi_0)}{\underline{a}^H(r_0, \theta_0, \varphi_0)\underline{a}(r_0, \theta_0, \varphi_0)} \tag{2.18}$$

donde, $(r_0, \theta_0, \varphi_0)$ son las coordenadas espaciales de la fuente principal o DOA (directon of Arrival). El denominador es un factor de normalización.

La fase de los pesos controla la dirección de apuntamiento del array, y su módulo controla la forma del diagrama de directividad (anchura del lóbulo principal y nivel de lóbulo principal a secundario).

2.3. Array lineal

El array más sencillo en el que podemos pensar es el array lineal uniforme (Uniform Linear Array, ULA). Suponiendo que la fuente está en campo lejano y que los micrófonos son omnidireccionales y de la misma sensibilidad, tal y como se muestra en la figura 2.4, el steering vector quedaría de la siguiente manera:

$$\underline{A}(f,\theta) = e^{j\frac{\omega}{c}\underline{x}cos(\theta)} \tag{2.19}$$

siendo $\underline{x}$ el vector de posiciones de los micrófonos del array y c la velocidad del sonido ($\simeq 340m/s$). Empleando, 2.18, obtenemos el siguiente vector de pesos:

$$\underline{W}(f,\theta_0) = \frac{1}{I}e^{j\frac{\omega}{c}\underline{x}cos(\theta_0)} \tag{2.20}$$

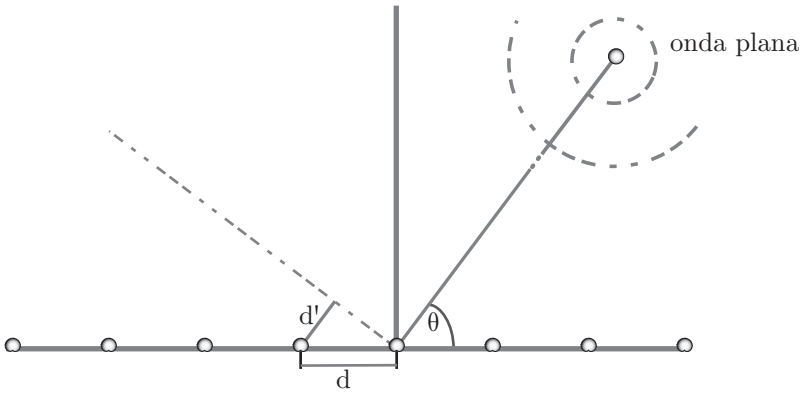
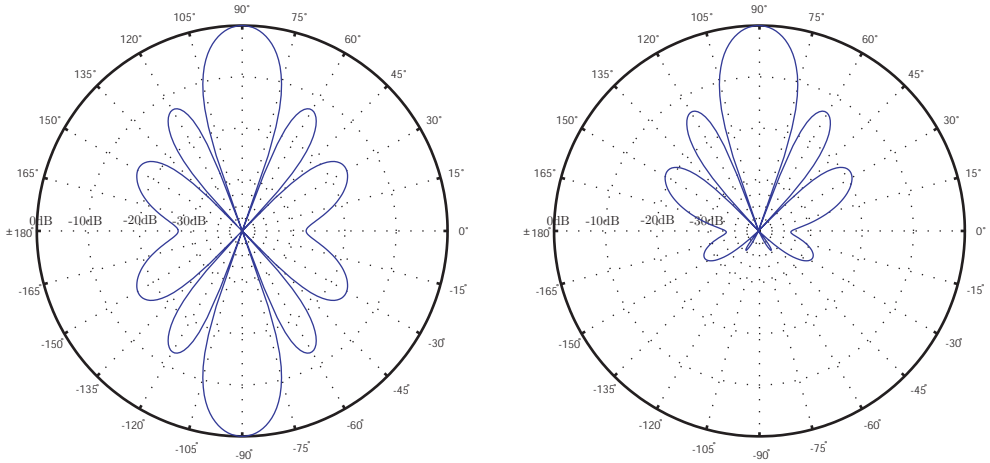


Figura 2.4: Array lineal Uniforme de 8 micrófonos.

Si representamos el diagrama de directividad de un ULA (ver figura 2.5 (a)), podemos ver como el array tiene simetría en la semicircunferencia $\theta \in [0^o - 180^o]$ y la semicircunferencia $\theta \in [0^o - (-180^o)]$, por lo que, si queremos apuntar hacia una fuente situada en $\theta = 90^o$, inevitablemente, amplificaremos el ruido o interferencia situado en $\theta = 270^o$, lo cual, obviamente no deseamos. Una primera apoximación para solventar este problema puede ser emplear micrófonos cardioides apuntando hacia la semicircunferencia de interés. Como se muestra en la figura 2.5 (b), el nulo en la parte trasera del diagrama de radiación de los micrófonos cardioides solventa el problema, sin embargo, esta situación impide que podamos apuntar satisfactoriamente hacia fuentes situadas en esa semiesfera. Otra forma de hacer frente a este tipo de situaciones, y así tener cobertura en los 360^o, es recurrir a los arrays circulares.

Existen otros dos problemas todavía más limitantes, que se desprenden de la propia naturaleza de la señal de voz. Dicha señal se extiende, aproximadamente, desde los



(a) (b)

Figura 2.5: Diagrama de directividad de un ULA de 8 micrófonos cuando $\theta = 90^{o}$, $d = 5$cm y $f = 2.5$kHz (a) Con micrófonos omnidireccionales y (b) con micrófonos cardioides.

125 Hz hasta unos 8 Khz, abarcando unas seis octavas. Debido a esto, aparecen los problemas de aliasing espacial y escasa directividad a baja frecuencia.

Podemos interpretar un array como un sistema que realiza un muestreo espacial de la señal a procesar. Es como si tuviéramos una señal discreta espacial, cuyo período de muestreo espacial sería $\Omega_s = \frac{2\pi}{\Delta x}$ siendo Δx la separación entre micrófonos. Por lo tanto, empleando el criterio de Nyquist, podemos concluir que la máxima separación entre micrófonos, vendrá impuesta por la máxima frecuencia a captar según:

$$\Delta x \leq \frac{c}{2f_{max}} \tag{2.21}$$

Para solventar el problema del aliasing, podemos disminuir la separación de los micrófonos, pero esto implicaría tener un array más pequeño para el mismo número de elementos, lo que agrava el problema de la escasa directividad a baja frecuencia. Si suponemos un ULA de I micrófonos con ponderación rectangular y queremos que el array tenga al menos un cero en el diagrama de radiación, según el criterio de resolución de Rayleigh ha de cumplirse que:

$$I\Delta x \geq \frac{c}{f_{min}} \tag{2.22}$$

Lo que viene a significar que para que un array fuera aceptablemente directivo a $f = 150$ Hz necesitaríamos que tuviera un tamaño $D = 2.27$ m, lo cual, para las aplicaciones habituales es totalmente inviable.

2.4. Array circular

Como hemos visto, con un array lineal es imposible distinguir entre una fuente que incida por un determinado ángulo, y otra que incida con el mismo ángulo, pero de signo contrario. Es lo que se conoce como ambigüedad delante-detrás. Un array circular no presenta este problema de discriminación delante-detrás [6]. A partir de ahora nos centraremos en el estudio del Array Circular Uniforme (Uniform Circular Array, UCA). En la figura 2.6 se muestra un ejemplo de esta configuración que es con la que se han realizado todos los algoritmos de conformación de haz descritos en la tesina.

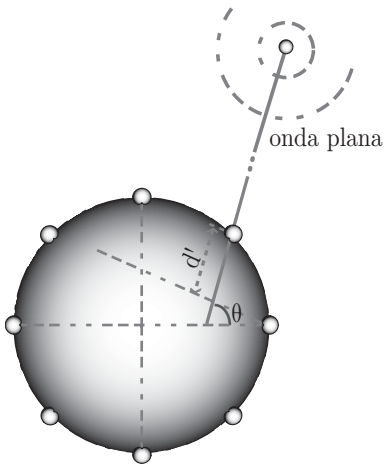


Figura 2.6: Array Circular Uniforme de 8 micrófonos.

El steering vector de un array circular de I elementos y radio R, con el punto de referencia de fase en el centro del array, y la DOA (θ), medida respecto a la línea que une el centro del array con el primer elemento es dado por [7]:

$$\underline{a}(\theta) = \left[e^{j\frac{w}{c}R\cos\theta}, e^{j\frac{w}{c}R\cos(\theta-\frac{2\pi}{I})}, \ldots, e^{j\frac{w}{c}R\cos(\theta-\frac{2\pi(I-1)}{I})}\right]^T \tag{2.23}$$

En el array lineal vimos como podíamos mejorar los resultados de directividad empleando micrófonos cardioides en lugar de omnidireccionales. Sin embargo, si observamos la figura 2.7 (a), vemos como el efecto de emplear elementos directivos en arrays circulares no parece ofrecer una particular mejora, con la notable excepción del caso de excitaciones de arco restringidas [8]. El uso más conocido de arrays circulares probablemente sea el array Wullenweber, el cual emplea un arco de alrededor de 120º. En 2.7 (b) se representa el diagrama de directividad a $2.5kHz$, de un CUA de 8 micrófonos omnidireccionales frente al obtenido empleando sólo 3 micrófonos cardioides para simular una configuración Wullenweber.

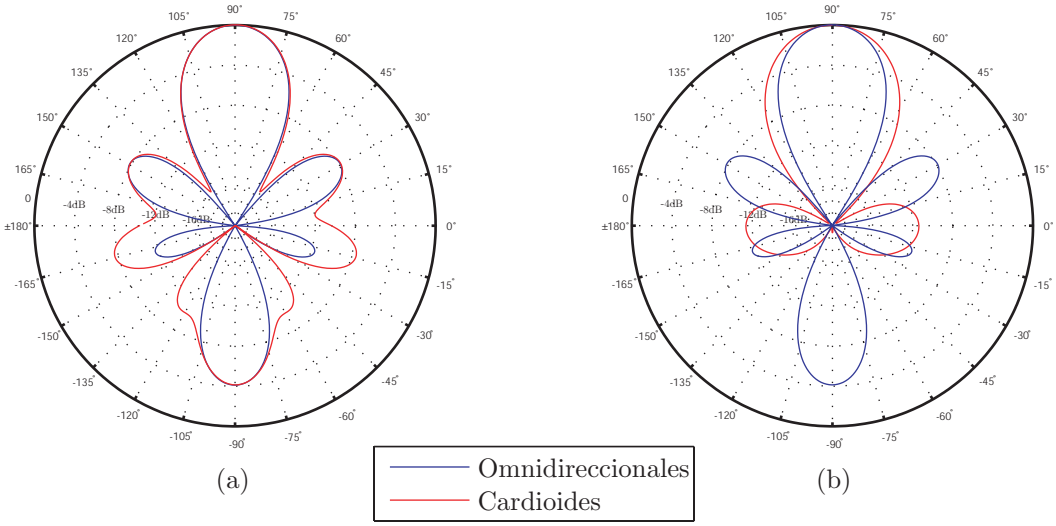


Figura 2.7: Diagrama de directividad a $f = 2.5kHz$, de un CUA de $r = 0.085m$. (a) Para configuraciones de 8 micrófonos, (b) comparativa entre 8 micrófonos omnidireccionales y 3 cardioides.

Debido a que la señal de voz es de banda ancha, en la figura 2.8 se muestra el diagrama de directividad, para todas las frecuencias, de un array circular de $r = 0.085m$ compuesto por 8 micrófonos cardioides, que es el que emplearemos para evaluar los resultados. En esta gráfica podemos apreciar los dos problemas comentados en la sección 2.3, la menor directividad a baja frecuencia, y el efecto del aliasing. En (a) hemos empleado todos los micrófonos, y en (b) una configuración Wullenweber. Gracias a esta configuración conseguimos una mayor directividad además de reducir los problemas anteriormente citados.

2.5. Beamformer en tiempo real con solución suboptima

Los algoritmos para supresión óptima de ruido se basaban en encontrar soluciones paramétricas dada la geometría del array. Para reducir la complejidad, posteriores diseños se basaron en encontrar soluciones suboptimas bajo diferentes criterios u objetivos. Algunos de estos algoritmos son el beamformer de directividad constante, el MVDR (minimum-variance distortionless response), o incluso los canceladores de lóbulos laterales [1]. En la práctica, los mejores resultados se alcanzan con algoritmos adaptativos, los cuales, compensan los cambios en las posiciones de la fuente objetivo y de las interferencias. Sin embargo, en escenarios como pequeñas salas de conferencias, las fuentes sonoras varían muy rápidamente sin que los algoritmos adaptativos tengan tiempo de converger.

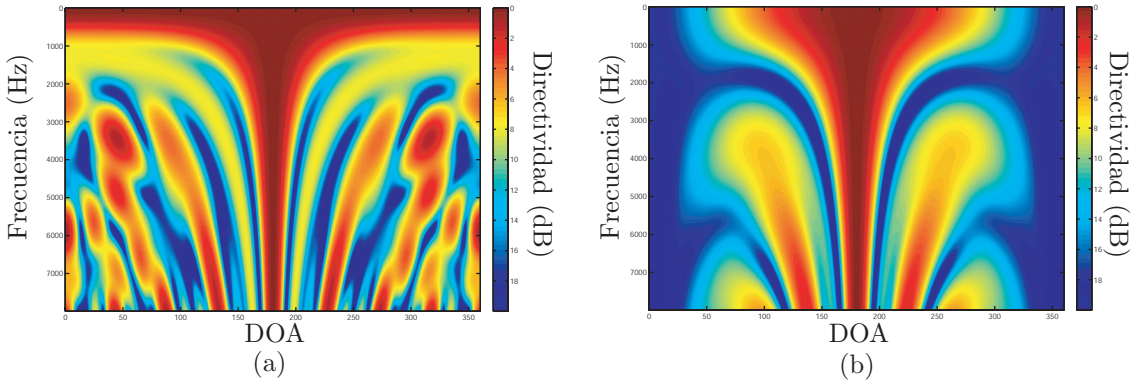


Figura 2.8: Diagrama de directividad de un array circular apuntando a 180^o, (a) para 8 micrófonos cardioides (b) empleando solo los 3 micrófonos más cercanos a la fuente de interés.

En [9], se propone un algoritmo donde los haces óptimos para cada dirección espacial son calculados off-line, para facilitar el procesado en tiempo real de las señales.

Para cada frame de N muestras, aplicamos el vector de pesos óptimos a la expresión 2.16 con el fin de calcular la energía de la señal para cada posición espacial. Finalmente, nos quedaremos con los pesos de la dirección que maximiza la potencia de salida. El vector de pesos óptimos correspondiente a cada dirección espacial ha sido previamente calculado mediante un algoritmo off-line que se explica más adelante.

Como ya hemos visto, los arrays de micrófonos mejoran la relación señal a ruido (SNR) gracias a su selectividad espacial. Hasta ahora, hemos considerado que sólo había ruido espacialmente blanco. Sin embargo, ahora consideraremos una situación más real, en la cual, las señales capturadas contienen dos fuentes de ruido, ruido instrumental y ruido acústico isotrópico (o ruido de ambiente). El ruido instrumental, $N_I(f)$, suele tener un espectro aproximadamente blanco. Es debido al micrófono, el preamplificador y el conversor analógico digital, por lo tanto, está incorrelado para los diferentes canales. El ruido isotrópico, $N_A(f)$, está correlado para los diferentes canales.

Podemos definir la ganancia de ruido de ambiente como el volumen del haz del array de micrófonos:

$$G_A(f) = \frac{1}{V} \oint B(f, \theta, \varphi) dc \tag{2.24}$$

donde V es el volumen de trabajo del array, es decir, el conjunto de las posiciones espaciales que queremos cubrir. La Ganancia de ruido instrumental viene dada por:

$$G_I(f) = \sqrt{\sum_{i=1}^{I} W_i(f)^2} \tag{2.25}$$

Por último, podemos definir el valor cuadrático medio del ruido a la salida del beamformer como:

$$E^2 = \int_0^{\frac{fs}{2}} [(G_A(f)N_A(f))^2 + (G_I(f)N_I(f))^2]df \tag{2.26}$$

Diseñar el vector de pesos óptimo, significa encontrar los pesos que maximiza la supresión de ruido para una determinada posición de apuntamiento, lo que equivale a minimizar la expresión 2.26. Además, suele ser conveniente establecer una serie de restricciones para evitar distorsión y no linealidades:

$$\begin{array}{c} |B(f,\theta,\varphi)| = 1 \\ \arg(B(f,\theta,\varphi)) = 0 \end{array} \quad \forall f \epsilon [f_B, f_E] \tag{2.27}$$

donde f_B y f_E son los límites de la banda de frecuencia de trabajo, en la cual, la forma de haz debería de ser igual que la diseñada por el algoritmo de beamforming. La minimización de E_N bajo estas restricciones, puede ser resuelta con métodos tradicionales de optimización como los basados en gradientes conjugados [10]. Sin embargo, debido a la hipersuperficie multimodal de la expresión 2.26, encontrar el mínimo por métodos de descenso del gradiente es una tarea difícil ya que suelen caer en mínimos locales. Para solventar este problema en [9], se propone sustituir la minimización directa con restricciones de E_N por una búsqueda unidimensional por mínimos cuadrados de un patrón de error con normalización. El parámetro a optimizar será el ancho del haz, y la normalización es para asegurar que los pesos calculados cumplen las restricciones.

Si nos fijamos en 2.26, vemos que al estrechar el ancho de haz, aumentamos la directividad, lo que provoca una reducción del ruido ambiental. Sin embargo, la solución para aumentar la directividad pasa por explotar pequeñas diferencias entre las señales, que tras la normalización realza el ruido instrumental (no correlado). Si ensanchamos el haz ocurre lo contrario, tenemos más ruido ambiental pero menos instrumental, por lo tanto, existe un ancho de haz que proporciona una solución casi óptima.

Para crearnos el patrón de error, en primer lugar tenemos que definir la función objetivo o haces óptimos que deberán de depender del ancho de haz, que como ya hemos dicho, será el parámetro a optimizar. Debido a que áreas rectangulares causarían rizados en la forma de haz recurriremos a funciones coseno suavizadas:

$$T(r,\theta,\varphi,\delta) = \cos\left(\frac{\pi(r_T - r)}{k\delta}\right)\cos\left(\frac{\pi(\theta_T - \theta)}{\delta}\right)\cos\left(\frac{\pi(\varphi_T - \varphi)}{\delta}\right) \tag{2.28}$$

donde (r,θ,φ) es el punto de enfoque objetivo, δ es el ancho de haz, y K es sólo un conversor dimensional. Una vez definida esta función objetivo podemos encontrar el conjunto de pesos que mejor ajusta el haz real al deseado mediante:

$$\xi = |\underline{T} - \underline{B}|^2 \tag{2.29}$$

donde $\underline{B}$ es el vector de ganancias actuales del haz, por lo tanto, la meta es alcanzar $\underline{B} \approx \underline{T}$. Para alcanzar una solución mediante mínimos cuadrados necesitamos un sistema lineal de ecuaciones sobredeterminado, por lo que elegiremos L puntos uniformemente espaciados en el área de trabajo, siendo $L > I$. Por lo tanto, los vectores $\underline{T}$ y $\underline{B}$ tendrán dimensiones $L \times 1$. Para calcular $\underline{B}$ recurriremos a la expresión 2.17, particularizada para cada posición espacial.

Llamaremos vector de pesos óptimos $\underline{W}_{OPT}$ a los pesos que minimizan el error cuadrático medio (MMSE) definido en 2.29. $\underline{W}_{OPT}$ es el vector de pesos que mejor ajusta el haz conseguido con nuestro beamformer al deseado, pero no cumple las restricciones definidas en 2.27. Para asegurar ganancia unidad y desplazamiento de fase cero para las señales originadas en el punto a optimizar, recurriremos a la siguiente normalización:

$$\underline{W}_N(f) = \frac{\underline{W}_{opt}(f)}{B(f, \theta_T, \varphi_T)} F(f) \tag{2.30}$$

donde $F(f)$ es la respuesta en frecuencia deseada, la cual normalmente será plana entre f_B y f_E, con pendientes de caída suavizadas.

Hasta ahora, nos hemos limitado a encontrar los pesos óptimos normalizados para una determinada función objetivo, sin embargo, nuestra meta es encontrar esos pesos para la función objetivo cuyo ancho de haz minimiza el ruido a la salida del beamformer. Para ello haremos una búsqueda en el intervalo $[\delta_{min}, \delta_{max}]$. Valores típicos son $\delta_{min} = 10^{\text{o}}$, si exploramos los 360^{o} con 36 haces, y $\delta_{max} = 250^{\text{o}}$. Por lo tanto, nuestro problema de optimización será encontrar el δ que minimiza E_N.

Finalmente, para obtener la matriz de pesos $\underline{\underline{W}}$ completa, repetiremos los pasos anteriores para cada bin de frecuencia.

2.6. Algoritmos adaptativos

Aunque sabemos que el algoritmo de DS es óptimo en cuanto a reducción de ruido totalmente incorrelado, no es el algoritmo óptimo para un escenario real. Con dicho algoritmo la posición y el nivel de los lóbulos laterales viene fijada por la configuración del array, por lo que, si una fuente interferente se sitúa en uno de estos lóbulos, no será suficientemente atenuada.

Ahora, además de buscar una respuesta máxima en la dirección de apuntamiento, buscaremos atenuar las interferencias procedentes de otras direcciones. Es decir, queremos un algoritmo que se adapte al escenario de ruido e interferencias para lograr una solución que maximice la SINR. Dicha solución se alcanza mediante:

$$\underline{W}(f) = \alpha \underline{\underline{R}}^{-1} \underline{a}(f, \theta_0) \tag{2.31}$$

donde α es una constante que se impone para que la señal cumpla ciertas restricciones. Emplearemos el caso particular de mínima varianza con restricciones lineales (LCMV), más conocido como Frost Beamformer, el cual proporciona una salida sin distorsión de la fuente de interés. Los pesos óptimos son calculados como:

$$\underline{W} = \left(\frac{\underline{\underline{R}}^{-1} \underline{a}(f)}{\underline{a}(f)^H \underline{\underline{R}}^{-1} \underline{a}(f)} \right)^* \tag{2.32}$$

donde $\underline{\underline{R}} = E[\underline{X}(\tau), \underline{X}(\tau)^H]$ es la matriz de correlación del vector de observación $\underline{X}(\tau)$. Dicha matriz de correlación se puede reescribir como:

$$\underline{\underline{R}} = \begin{bmatrix} R_{11} & R_{12} & \dots & R_{1I} \\ R_{21} & R_{22} & \dots & R_{2I} \\ \vdots & \vdots & \ddots & \vdots \\ R_{I1} & R_{I2} & \dots & R_{II} \end{bmatrix} \tag{2.33}$$

donde $R_{ij}(f)$ representa la correlación cruzada entre las señales captadas por el micrófono i y j para el bin de frecuencia f. La función de correlación cruzada puede ser aproximada en el dominio de la frecuencia como:

$$R_{ij}(\tau) \approx \sum_{k=0}^{L-1} X_i(k) X_j(k)^* e^{\frac{j2\pi k\tau}{L}} \tag{2.34}$$

siendo τ el retardo entre los micrófonos i y j y L el número de bins de frecuencia. Con esta estima de la correlación no se distingue suficientemente las fuentes interferentes ya que los picos tienden a solaparse, y por lo tanto, no podremos anular las interferencias suficientemente. Para obtener mejores resultados recurriremos a calcular la correlación

cruzada blanqueada, en la cual, cada bin de frecuencia del espectro contribuye la misma cantidad a la correlación final.

$$R_{ij}^{w}(\tau) \approx \sum_{k=0}^{L-1} \frac{X_i(k)X_j(k)^*}{|X_i(k)||X_j(k)|} e^{\frac{j2\pi k\tau}{L}} \tag{2.35}$$

En la implementación adaptativa, los pesos óptimos solo son calculados durante periodos de ruido, es decir, cuando la voz no está activa. La adaptación durante periodos de voz podría proporcionar un incorrecta solución y la consiguiente cancelación de la señal deseada. Estimar los períodos de inactividad de voz no es una tarea trivial. Para ello hemos recurrido a un algoritmo de detección de voz (Voice Activity Detection, VAD) basado en un modelo estadístico descrito en [11].

En los micrófonos direccionales, la respuesta de fase varía significativamente con la frecuencia, la dirección e incluso de micrófono a micrófono. Para micrófonos cardioides típicos la respuesta de fase es razonablemente plana para ángulos pequeños, pero se hace altamente variable conforme aumenta el ángulo de llegada [12]. Esta variación de fase no es necesariamente un inconveniente en arrays lineales, ya que el ángulo de llegada es prácticamente el mismo en todas las direcciones. Sin embargo, en arrays circulares uniformes se convierte en un serio problema, ya que el sonido llega de direcciones uniformemente distribuidas en los 360^{o} y por lo tanto, la estima de la matriz de correlación es incorrecta. Para solucionar este problema, empleamos solamente los 3 micrófonos más cercanos a la fuente de interés. Por una parte, estamos perdiendo capacidad para atenuar interferencias, ya que ahora sólo podemos colocar dos ceros en el diagrama de directividad. Sin embargo, también tiene una ventaja, y es que ahora estamos reduciendo sensiblemente el coste computacional.

Capítulo 3

Separación Ciega de Fuentes

La Separación Ciega de Fuentes de Audio (Blind Audio Source Separation, BASS), es el problema de recuperar cada fuente de audio a partir de una señal mezclada. En este caso, la señal captada por el array de micrófonos.

En primer lugar explicaremos una técnica que se basa en la suposición de que para un determinado punto tiempo-frecuencia sólo hay una fuente activa. Dicho método, conocido como DUET (Degenerate Unmixing Estimation Technique) es uno de los más famosos y extendidos, ya que obtiene muy buenos resultados para el caso de mezclas infradeterminadas.

3.1. Suposiciones

Sin las apropiadas restricciones en las fuentes mezcladas, el problema de la separación ciega de fuentes tiene un número infinito de soluciones. Por lo tanto, para obtener una única solución debemos hacer suposiciones acerca de las fuentes o los filtros de mezcla [3].

3.1.1. Mezclas anecoicas

Supongamos un par de micrófonos al que le llegan las señales provenientes de N fuentes, $s_j(t),\ j = 1, \ldots, N$. La señal capturada por los micrófonos puede expresarse como:

$$x_i(t) = \sum_{j=1}^{N} h_{ij} s_j(t) \qquad i = 1, 2, \tag{3.1}$$

donde $h_{ij}(t)$ es la respuesta al impulso entre la fuente j y el micrófono i. Si sólo consideramos el camino directo, las dos mezclas anecoicas resultantes pueden expresarse como:

$$x_1(t) = \sum_{j=1}^{N} s_j(t) \tag{3.2}$$

$$x_2(t) = \sum_{j=1}^{N} a_j s_j(t - \delta_j) \tag{3.3}$$

siendo δ_j el retardo de llegada entre los dos micrófonos, y a_j es un factor de atenuación relativo correspondiente a la relación de las atenuaciones de los caminos entre las fuentes y los micrófonos. Emplearemos Δ para definir el máximo retardo posible entre los micrófonos (la separación de los micrófonos dividido por la velocidad del sonido), por lo tanto $|\delta_j| \leq \Delta, \forall j$.

3.1.2. Fuentes ortogonales W-disjuntas

Diremos que dos funciones $s_j(t)$ y $s_k(t)$, son ortogonales W-disjuntas (W-disjoint orthogonal, WDO), si para una función de enventanado $W(t)$, las correspondientes transformadas de Fourier enventanadas $\hat{s}_j(\tau, \omega)$ y $\hat{s}_k(\tau, \omega)$ son disjuntas:

$$\hat{s}_j(\tau, \omega)\hat{s}_k(\tau, \omega) = 0, \qquad \forall \tau, \omega, \ \forall j \neq k. \tag{3.4}$$

Esta suposición, es crucial, ya que permite separar las fuentes presentes en una mezcla empleando máscaras binarias. Por ejemplo, para la fuente j, obtendremos la siguiente máscara:

$$M_j(\tau, \omega) := \begin{cases} 1 & \text{si } \hat{s}_j(\tau, \omega) \neq 0 \\ 0 & \text{otro caso} \end{cases} \tag{3.5}$$

y entonces, para separar $\hat{s}_j$ de la mezcla:

$$\hat{s}_j(\tau, \omega) = M_j(\tau, \omega)\hat{x}_1(\tau, \omega), \qquad \forall \tau, \omega. \tag{3.6}$$

3.1.3. Estacionariedad local

A la hora de trabajar con señales de voz en el dominio de la frecuencia recurrimos a la Transformada de Fourier dependiente del tiempo, o de tiempo corto (STFT). A cada punto de esta transformación lo llamaremos punto tiempo-frecuencia (TF), ya que, de alguna forma, aporta información de un determinado intervalo de tiempo para un determinado intervalo de frecuencias.

Como sabemos, un desplazamiento en el tiempo equivale a un cambio de fase en frecuencia, sin embargo, cuando empleamos ventanas temporales de duración finita, esto

puede no llegar a cumplirse. Hay que elegir una duración de ventana tal que asegure que la señal de voz puede considerarse estacionaria en esa ventana, y por tanto no perdemos resolución temporal, pero que a su vez cumpla que $\forall\delta, |\delta| \leq \Delta$.

Además, al igual que ocurría con los arrays lineales, hemos de tener en cuenta que la separación máxima entre micrófonos debe de cumplir la ecuación 2.22 para evitar el fenómeno del aliasing.

3.2. DUET

Las suposiciones de mezcla anecoica y estacionariedad local, nos permite reescribir 3.2 y 3.3 en el dominio de la frecuencia como:

$$\begin{bmatrix} \hat{x}_1(\tau,\omega) \\ \hat{x}_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ a_1 e^{-i\omega\delta_1} & \dots & a_N e^{-i\omega\delta_N} \end{bmatrix} \begin{bmatrix} \hat{s}_1(\tau,\omega) \\ \vdots \\ \hat{s}_N(\tau,\omega) \end{bmatrix} \tag{3.7}$$

Además, teniendo en cuenta la suposición de fuentes ortogonales W-disjuntas, sólo habrá una fuente activa en cada punto TF (τ,ω), con lo que podemos describir el proceso de mezcla como:

$$\text{para cada } (\tau,\omega), \quad \begin{bmatrix} \hat{x}_1(\tau,\omega) \\ \hat{x}_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 \\ a_j e^{-i\omega\delta_j} \end{bmatrix} \hat{s}_j(\tau,\omega), \quad \text{para algún } j. \tag{3.8}$$

donde j se refiere a la fuente activa en el punto tiempo-frecuencia (τ,ω).

Los parámetros de mezcla asociados con cada punto tiempo-frecuencia, pueden ser calculados como:

$$\tilde{a}(\tau,\omega) := \left| \frac{\hat{x}_2(\tau,\omega)}{\hat{x}_1(\tau,\omega)} \right| \tag{3.9}$$

$$\tilde{\delta}(\tau,\omega) := \frac{-1}{\omega} \angle \frac{\hat{x}_2(\tau,\omega)}{\hat{x}_1(\tau,\omega)} \tag{3.10}$$

Ahora, ya conocemos las parejas de parámetros de mezcla estimados para cada punto tiempo-frecuencia, por lo tanto, podemos calcular las máscaras de la siguiente manera:

$$M_j(\tau,\omega) := \begin{cases} 1 & \text{si } (\tilde{a}(\tau,\omega), \tilde{\delta}(\tau,\omega)) = (a_j, \delta_j) \\ 0 & \text{otro caso} \end{cases} \tag{3.11}$$

siendo, a_j, δ_j, los parámetros de mezcla para la fuente j. Sin embargo, estos parámetros a priori no los conocemos, al igual que tampoco conoceremos el número de fuentes. Además, en la práctica, no todas las suposiciones se cumplen estrictamente, y por lo tanto, los parámetros de mezcla estimados no son exactamente los parámetros de mezcla reales.

Para solventar estos problemas en [13] se propone el cálculo de Histogramas Bidimensionales Ponderados y Suavizados. Otras opciones pueden ser algoritmos de detección de fuentes o DOA, como [14, 15].

3.2.1. Histogramas bidimensionales ponderados y suavizados

En lugar de estimar a_j, ahora calcularemos la llamada atenuación simétrica α_j, de tal manera que si la señal llega más fuerte al micrófono 1 $\alpha_j < 0$, y si la señal llega más fuerte al micrófono 2 $\alpha_j > 0$:

$$\tilde{\alpha}(\tau,\omega) := \left|\frac{\hat{x}_2(\tau,\omega)}{\hat{x}_1(\tau,\omega)}\right| - \left|\frac{\hat{x}_1(\tau,\omega)}{\hat{x}_2(\tau,\omega)}\right| \tag{3.12}$$

Además, los puntos que contribuyan a la elaboración del histograma suavizado han de cumplir que:

$$I(\alpha,\delta) := \{(\tau,\omega) : |\tilde{\alpha}(\tau,\omega) - \alpha| < \Delta_\alpha, |\tilde{\delta}(\tau,\omega) - \delta| < \Delta_{delta}\}, \tag{3.13}$$

donde, Δ_α y Δ_δ, son los anchos de resolución suavizada. Finalmente, el histograma es construido:

$$H(\alpha,\delta) := \int\int_{(\tau,\omega)\in I(\alpha,\delta)} |\tilde{x}_1(\tau,\omega)\tilde{x}_2(\tau,\omega)|^p \omega^q d\tau d\omega \tag{3.14}$$

siendo, p y q parámetros de ponderación de los estimadores. El parámetro p afecta a la amplitud simétrica y el parámetro q al retardo relativo. Valores típicos son los propuestos en el algoritmo original del DUET [16], $p = 0$ y $q = 0$ o los propuestos en [13], $p = 1$ y $q = 0$.

Como se observa en la figura 3.1, los N picos del histograma corresponden a las N fuentes, y la localización de los picos serán los parámetros de mezclas de dichas fuentes.

3.2.2. Separación de las fuentes

Una vez que hemos identificado los picos, nuestra meta es determinar las máscaras tiempo-frecuencia que separarán las fuentes de la mezcla. Para lograrlo asignaremos cada punto tiempo-frecuencia con el pico más cercano a los parámetros locales estimados para dicho punto. El primer paso es convertir la atenuación simétrica de los picos localizados en los histogramas al parámetro de atenuación. Para ello:

$$\tilde{a}_j = \frac{\tilde{a}_j + \sqrt{\tilde{a}_j^2 + 4}}{2} \tag{3.15}$$

Ahora, asignamos un pico a cada punto tiempo frecuencia:

$$J(\tau,\omega) := \arg\min_k \frac{\left|\tilde{a}_k e^{-i\tilde{\delta}_k\omega}\hat{x}_1(\tau,\omega) - \hat{x}_2(\tau,\omega)\right|^2}{1+\tilde{a}_k^2} \tag{3.16}$$

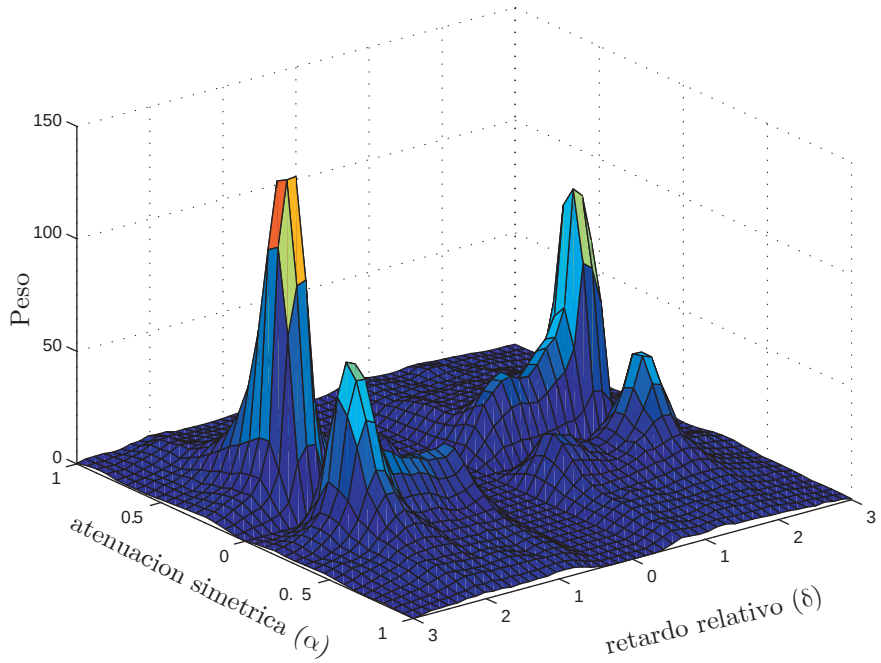


Figura 3.1: Histograma bidimensional de atenuación simétrica y retardo relativo para una mezcla de cuatro fuentes.

y entonces, calculamos las máscaras tiempo frecuencia como:

$$\tilde{M}_j(\tau,\omega) := \begin{cases} 1 & \text{si } J(\tau,\omega) = 1 \\ 0 & \text{otro caso.} \end{cases} \tag{3.17}$$

El último paso es calcular la señal separada (y posteriormente devolverla al dominio temporal):

$$\hat{s}_j(\tau,\omega) = \tilde{M}_j(\tau,\omega) \left(\frac{\hat{x}_1(\tau,\omega) + \tilde{a}_j e^{-i\tilde{\delta}_j \omega} \hat{x}_2(\tau,\omega)}{1+\tilde{a}_j^2} \right). \tag{3.18}$$

3.3. Separación ciega de fuentes con arrays de micrófonos

En la sección anterior estudiamos uno de los métodos de separación ciega de fuentes más famosos y eficaces, el DUET. Si suponemos fuentes estáticas en los histogramas deberían aparecer deltas perfectas, pero si observamos otra vez la figura 3.1, vemos que estas deltas en realidad son distribuciones supergaussianas. Esto es debido a que las fuentes, que habíamos supuesto disjuntas, en realidad no lo son para todos los puntos tiempo-frecuencia, lo que hace que que se vayan ensanchando estas deltas. Como consecuencia, la separación obtenida por métodos como DUET, conservará parte de las interferencias e introducirá artefactos de ruido musical. Además, en escenarios reales, debido a la reverberación de la sala, estas distribuciones se ensancharán más todavía, lo que conlleva peores resultados en la separación.

Para minimizar estos problemas, se presenta un método que consiste en aplicar las ventajas del tratamiento en array para solventar los problemas e incrementar la eficiencia de un método ya de por sí robusto como es el DUET.

La idea básica del DUET es separar las fuentes a partir de la información que nos proporcionan dos características de las mezclas. Estas características son las diferencias de amplitud, y las diferencias de fase. En primer lugar, vamos a intentar conseguir una mayor separación entre las clases o fuentes, y así poder clasificar o separar con menor error. El DUET, como la mayoría de métodos de separación de BASS, puede considerarse como una técnica de separación no paramétrica. En concreto, una basada en la regla del vecino más cercano. Cuando la mínima probabilidad de error es pequeña, la probabilidad de error del vecino más cercano también es pequeña, pero conforme la mínima probabilidad de error para las distintas clases se va haciendo similar, la probabilidad de error del vecino más cercano también aumenta hasta $P_e = 1/N^{o}clases$ [17]. Por tanto, si conseguimos separar o ampliar los parámetros de las fuentes, de algún modo, estamos disminuyendo la probabilidad de error al separar las clases.

Si empleamos dos micrófonos omnidireccionales muy próximos las diferencias de amplitud son muy pequeñas. Si aumentamos la separación entre ellos para aumentar estas diferencias de amplitud, se produce aliasing en las altas frecuencias que provoca datos erróneos en la estima del retardo relativo. Otra posible solución podría ser emplear micrófonos direccionales apuntando a diferentes posiciones del espacio para aumentar la diferencia de amplitud. Sin embargo, estos micrófonos consiguen dicha diferencia de amplitud gracias a los desfases producidos en su cavidad, y como se muestra en [18], la información de fase que se obtiene con micrófonos de gradiente es muy sensible a desviaciones en amplitud entre los micrófonos, que pueden deberse, tanto a las tolerancias de fabricación como a pequeñas imperfecciones a la hora de posicionar los micrófonos en el array, por lo que estos micrófonos no son apropiados para tal fin.

La solución propuesta consiste en combinar la información de amplitud, extraída a

partir de micrófonos de gradiente, con la información de fase obtenida mediante dos micrófonos omnidireccionales posicionados muy juntos. Ahora podemos ampliar la separación entre los micrófonos cardioides. Existe un límite en la separación entre dichos micrófonos establecido por la suposición de estacionariedad local, pero con el tamaño de las ventanas empleado ($\approx 64ms$), esta distancia es del orden de metros, por lo que no supone ninguna limitación.

En la figura 3.3 (a) se presenta una comparación, con la que nos podemos hacer una idea intuitiva de las diferencias entre este procedimiento y el DUET clásico. La representación corresponde a una situación real con dos fuentes incidiendo a un array lineal. Las fuentes inciden por $\theta_{F1} = 45^o$ y $\theta_{F2} = 90^o$, mientras que el array consta de dos micrófonos omnidireccionales a $1cm$ cada uno del centro del array y dos micrófonos cardioides a $9cm$ cada uno del centro apuntando uno hacia $\theta_{M1} = 0^o$ y el otro hacia $\theta_{M2} = 180^o$. En (a) debido a que las fuentes están relativamente cercanas, y que la energía de una de las fuentes es superior a la otra, no es fácil apreciar los dos picos del histograma para aplicar correctamente DUET. Sin embargo, en (b), al haber aumentado la diferencia en la amplitud captada por los micrófonos, es más fácil obtener dichos picos. No obstante, la mejora

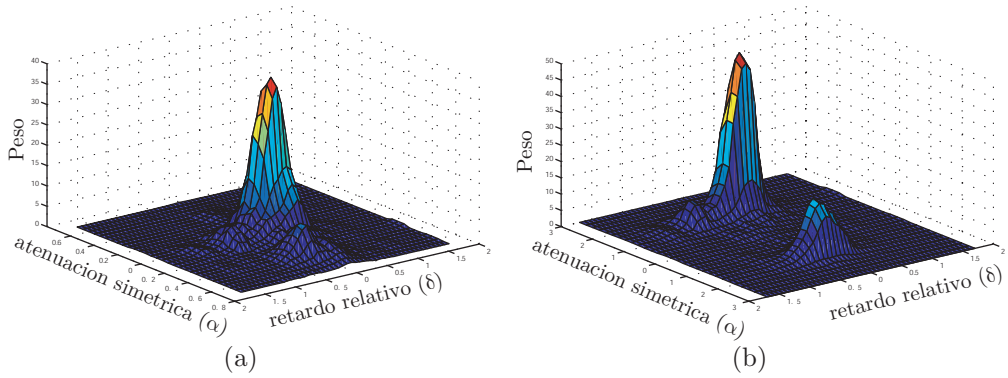


Figura 3.2: Histogramas de dos fuentes, una en $\theta_{F1} = 45^o$ y otra en $\theta_{F2} = 90^o$,(a) por el método del DUET, (b) calculando la amplitud a partir de 2 cardioides separados 5 cm y apuntando uno a $\theta = 0^o$ y el otro a $\theta = 180^o$.

más importante se produce a la hora de separar las fuentes, ya que ahora, una vez calculadas las respectivas máscaras a partir del histograma, separaremos la señal de la mezcla multiplicando su máscara correspondiente por el micrófono cardioide más cercano a esta señal. Como en dicho micrófono ya se ha producido una pequeña separación espacial de las fuentes, la SIR obtenida es mayor.

Al igual que ocurría con arrays lineales, con DUET tenemos ambigüedad delante-detrás, ya que realmente, estamos captando la señal con un array de dos micrófonos. Además, los parámetros que necesitamos estimar son la amplitud y la fase de las ondas

incidentes respecto al array de micrófonos. Por tanto, tenemos exactamente el mismo problema que antes. Para solventar este problema podemos recurrir a un array bidimensional que resuelve este tipo de ambigüedad.

Imaginemos un escenario con cuatro fuentes situadas en $\underline{\theta} = [0^o\ 90^o\ 180^o\ 270^o]$. Como hemos dicho, con DUET, sería imposible separar las cuatro fuentes. Para superar esta limitación proponemos una nueva configuración que se muestra en la figura 3.3. El array propuesto consta de 8 micrófonos, pero como la información la procesaremos en grupos de 4 micrófonos, definiremos subarrays, a los que nos referiremos con la siguiente nomenclatura:

$$S_k = [i - j,\ m - n] \tag{3.19}$$

donde, los índices $i - j$, hacen referencia a la pareja de micrófonos destinados a estimar la información de fase y los índices $m - n$, se corresponden con la pareja de micrófonos que se encargará de estimar la amplitud simétrica. S_k es la señal separada para la fuente k. Dada esta configuración, podemos emplear el subarray $[1 - 3,\ 6 - 8]$, para intentar

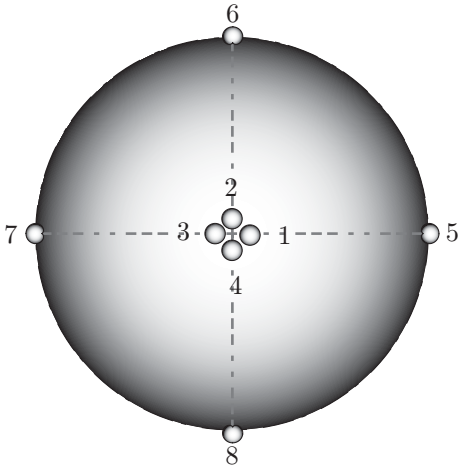


Figura 3.3: Configuración compuesta por un CUA de $R = 1cm$ de 4 micrófonos omnidireccionales y otro CUA de $R = 9cm$ de 8 micrófonos cardioides.

separar las cuatro fuentes antes comentadas. En la figura 3.4, se observa como, con este procedimiento, es posible distinguir los 4 picos del histograma que se corresponden con las 4 fuentes, por lo que calculando sus respectivas máscaras podremos obtener las fuentes separadas a partir de la mezcla.

A pesar de todo, cuando hay más de dos fuentes en un entorno reverberante, el número de reflexiones empieza a ser considerable, por lo que las distribuciones en el histograma se ensanchan y tienden a solaparse. Si a esto le unimos el hecho de que el nivel de energía con el que inciden las diversas fuentes al array suele ser diferente, y por lo tanto,

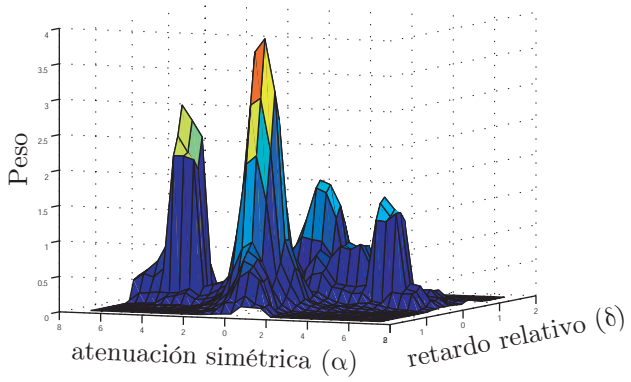


Figura 3.4: Histograma calculado con el subarray $[1-3,\ 6-8]$, cuando inciden 4 fuentes por las direcciones $\underline{\theta} = [0^o\ 90^o\ 180^o\ 270^o]$.

el pico de la fuente con mayor energía suele enmascarar al resto, se hace realmente complicado extraer los centroides de las fuentes a partir de los histogramas. Estos problemas hacen que sea prácticamente imposible aplicar un método ciego de separación de fuentes, sobre todo si queremos que sea un método automático y en tiempo real.

Una forma de solventar este problema puede ser emplear algoritmos de localización y seguimiento de fuentes como los propuestos en [14, 15] como fase previa a la separación, con lo que propiamente no sería una separación ciega pero si que conseguiríamos cumplir los objetivos. Sin embargo, estos métodos, suelen necesitar conocer de antemano el número de fuentes además de emplear también arrays de más de dos elementos. En esta tesina se propone otro método, en el que en principio supondremos que las fuentes que queremos separar están situadas angularmente próximas a los micrófonos cardioides del array.

Sigamos considerando el ejemplo de cuatro fuentes situadas frente a los micrófonos cardioides. Supongamos que de las cuatro fuentes, nos interesa la situada en $\theta = 0^o$, considerando las demás como interferencias. Si empleamos el subarray $[1-3,\ 5-7]$, obtendremos tres picos en el histograma, uno para la fuente de interés $(\alpha > 0\ y\ \delta > 0)$, otro para la fuente situada en $\theta = 180^o (\alpha < 0\ y\ \delta < 0)$, y el último para las otras dos fuentes $(\alpha = 0\ y\ \delta = 0)$. Si ahora aplicamos DUET, podemos extraer la fuente de interés (y la situada en $\theta = 180^o$). Si fuera necesario obtener las cuatro fuentes, bastaría con aplicar el mismo proceso empleando ahora el subarray $[2-4,\ 6-8]$.

A priori, no conocemos la DOA exacta de la fuente, ni tampoco si está o no enmascarada por el resto. Sin embargo, como hemos visto, al emplear este subarray, la fuente de interés siempre queda situada en el cuadrante $alpha > 0, delta > 0$, por lo que, para localizar las coordenadas del pico basta con aplicar:

$$P_0(\alpha, \delta) = \text{máx}[Hist(\alpha, \delta)]|_{(\alpha>0\ \&\ \delta>0)} \tag{3.20}$$

$$P_{i1}(\alpha,\delta) = \text{máx}[Hist(\alpha,\delta)]|_{(\alpha<0\ \&\ \delta<0)} \tag{3.21}$$

$$P_{i2}(\alpha,\delta) = \text{máx}[Hist(\alpha,\delta)]|_{(\alpha<\alpha_{min}\ \&\ \delta<\delta_{min})} \tag{3.22}$$

donde α_{min} y δ_{min} son dos umbrales próximos a cero que nos permiten hacer una búsqueda cercana a cero para las otras dos fuentes interferentes. Hasta ahora hemos aprovechado la geometría del array para eliminar la ambigüedad delante-detrás y solventar el problema de la identificación de fuentes en el histograma.

Por otra parte, podemos aprovechar que hemos pasado de un sistema infradeterminado (más fuentes que sensores) a otro sobredeterminado (más sensores que fuentes) para seguir mejorando la calidad de la señal separada. Para ello aplicaremos iterativamente un mecanismo de separación similar al que acabamos de comentar. La diferencia es, que en cada iteración sólo identificaremos dos picos, el de la fuente de interés y el de la fuente interferente 1, con lo que simplificaremos el proceso. Podemos ver el proceso como una especie de cancelador sucesivo de interferencias espaciales (ver figura 3.5), en el que para cada iteración empleamos el subarray óptimo para reducir tanto la interferencia como el ruido provenientes de un determinado sector espacial. Esta cancelación la hacemos identificando el pico del histograma correspondiente a la fuente objetivo y a la interferencia para posteriormente calcular las máscaras de separación en función de la distancia de cada punto tiempo-frecuencia al centroide de la fuente.

El proceso completo para el anterior ejemplo sería el siguiente:

1. $S_1 = [1-3,\ 5-7]$, cancelamos la interferencia cuasada por la fuente $\theta = 180^{\text{o}}$. La fuente $\theta = 90^{\text{o}}$ y $\theta = 270^{\text{o}}$ siguen estando muy presentes.

2. $S_2 = [1-2,\ S1-6]$, cancelamos la interferencia provocada por la fuente $\theta = 90^{\text{o}}$.

3. $S_3 = [1-4,\ S2-8]$, finalmente, cancelamos también la interferencia provocada por la fuente $\theta = 270^{\text{o}}$. Obteniendo la señal separada.

Como hemos elegido el subarray óptimo para cada interferencia, siempre tenemos la fuente de interés en el cuadrante $(\alpha > 0\ \&\ \delta > 0)$, y la interferente en el cuadrante $(\alpha < 0\ \&\ \delta < 0)$, por lo que podemos emplear esta información como prior (probabilidad a priori) para mejorar el mecanismo de clasificación [17]. Ahora la expresión para asignar un pico a cada punto tiempo-frecuencia queda como:

$$J(\tau,\omega) := \arg\min_{k} \left[P_k(\tau,\omega) \frac{\left| \tilde{a}_k e^{-i\tilde{\delta}_k \omega} \hat{x}_1(\tau,\omega) - \hat{x}_2(\tau,\omega) \right|^2}{1+\tilde{a}_k^2} \right] \tag{3.23}$$

donde $P_k(\tau,\omega)$ es la probabilidad a priori de que el punto TF (τ,ω) pertenezca a la fuente (k). Experimentalmente comprobamos que, cuando las fuentes están espacialmente distanciadas por un ángulo similar o mayor al ángulo de separación de los micrófonos en el

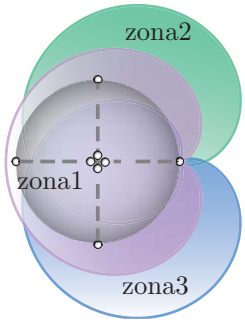


Figura 3.5: Cancelación iterativa de interferencias para la señal captada por un micrófono del array.

array, la mejor separación se obtiene para valores $P_k \approx 1$. Lo que quiere decir, que asignamos a la máscara de la fuente que pretendemos separar todos los puntos TF $\alpha > 0, \delta > 0$ y ninguno de los puntos TF $\alpha < 0, \delta < 0$. Dependiendo de lo que hagamos con el resto de puntos podemos hablar de separación Hard o separación Soft, aplicando las siguientes máscaras:

$$\tilde{M}_j^H(\tau,\omega) := \begin{cases} 1 & \text{si } \alpha > 0 \ \& \ \delta > 0 \\ 0 & \text{otro caso.} \end{cases} \tag{3.24}$$

$$\tilde{M}_j^S(\tau,\omega) := \begin{cases} 0 & \text{si } \alpha < 0 \ \& \ \delta < 0 \\ 1 & \text{otro caso.} \end{cases} \tag{3.25}$$

La elección entre una u otra dependerá de la aplicación, con la máscaras hard, se consigue una mayor separación a costa de aumentar los artefactos de ruido musical, mientras que con las máscaras soft [1] estos artefactos son menores que con las hard pero aumenta el nivel de interferencias. Una buena elección puede ser emplear máscaras hard para aplicaciones de reconocimiento automático del habla (RAH), y máscaras soft en aplicaciones destinadas al ser humano como mejora de señales de voz en videoconferencia o en sistemas de ayuda a la escucha (hearing aids).

Emplear este tipo de máscaras tiene otra gran ventaja, y es que no se basan en la información del histograma, por lo tanto, no necesitamos calcularlos y por consiguiente buscar los picos, lo que hace que el algoritmo sea computacionalmente más eficiente. Además, no tener que calcular los histogramas deriva en no tener que hacer estadísticas de la señal para obtener sus parámetros, o lo que es lo mismo, podemos tomar decisiones en cada ventana. Por lo tanto es un algoritmo de separación de fuentes apto para aplicaciones en tiempo real.

[1] Aunque hemos llamado a estas mácaras soft, no hay que confundirlas con las empleadas en otros métodos de BASS, ya que en este caso siguen siendo máscaras binarias.

Para situaciones como la descrita, con una fuente situada en cada micrófono cardioide, el algoritmo, como se verá en el capítulo de resultados, ha demostrado ser muy robusto, logrando la separación de todas las fuentes con una gran SIR, y una muy buena SAR. En aplicaciones típicas de videoconferencia, o ayuda a la escucha, habrá solo una fuente de interés y varias fuentes interferentes sin localización fija. Para este tipo de aplicaciones, quizás las más frecuentes e importantes, el algoritmo sigue demostrándose muy robusto, siempre y cuando las fuentes interferentes más cercanas no estén situadas a menos de $30^o - 40^o$ de la fuente de interés. Para superar esta limitación podemos recurrir a las siguientes estrategias:

- **Volver a emplear la información de los histogramas:** Estas técnicas las habíamos desechado por ser computacionalmente menos efectivas, y en la mayoría de aplicaciones los picos de los histogramas deberán ser obtenidos por inspección visual como se indica en [13], o bien mediante un robusto algoritmo de DOA (aunque esto sólo nos da la posición estimada para el retardo relativo). Sin embargo, este tipo de aplicaciones rara vez exigirán procesamiento en tiempo real, por lo tanto, podemos aplicar el modelo basado en probabilidades a priori explicado con anterioridad.

- **Aplicar técnicas de Beamforming:** El problema se produce cuando tenemos que separar dos fuentes que angularmente están muy cercanas, ya que en ninguna iteración del algoritmo, logramos que la fuente de interés se sitúe en el cuadrante $\alpha > 0, \delta > 0$, y la interferente en $\alpha < 0, \delta < 0$. Podemos aplicar un beamformer sencillo como el DS hacia cada una de las fuentes presentes en la escena y emplear estas señales como entradas del algoritmo. Con esto ampliamos las diferencias de amplitud y conseguimos la fuente de interés en el plano $\alpha > 0$ y la interferente en $\alpha < 0$.

- **Aumentar el número de micrófonos:** Esta es la mejor solución. Aumentando el número de micrófonos direccionales, como se muestra en la figura 3.6 aumentamos la resolución angular. Además, esta mejora es compatible con las dos anteriores. Sobre todo con la del Beamforming, ya que al aumentar el número de micrófonos del array, aumentamos también la efectividad del beamforming.

Obviamente, no por aumentar el número de micrófonos podemos seguir separando más y más fuentes, ya que estas separaciones cada vez son de menor calidad, llegando un momento en que es imposible mejorar.

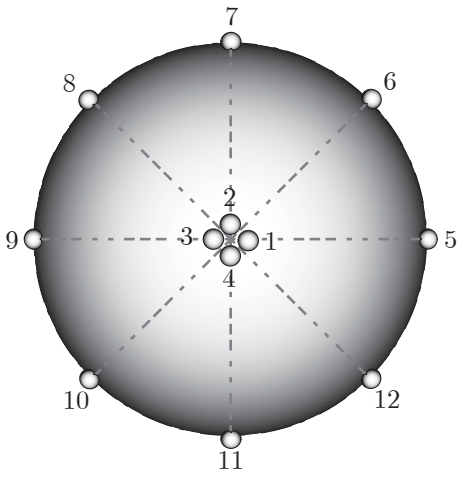


Figura 3.6: Configuración compuesta por un CUA de $R = 1cm$ de 4 micrófonos omnidireccionales y otro CUA de $R = 9cm$ de 8 micrófonos cardioides.

Capítulo 4

Medidas y Experimentos

En este capítulo se presentan las medidas y los experimentos realizados tanto para el caso del Beamforming como para el de la separación ciega de fuentes.

En el caso del Beamforming, la DOA de las fuentes ha sido calculada mediante dos procedimientos. En el caso de una sola fuente hemos empleado el proceso descrito en la sección 2.5. Mientras que cuando teníamos más de una fuente hemos recurrido a una modificación (debida a los micrófonos cardioides y a considerar fuentes estáticas) del algoritmo descrito en [15].

Para la separación ciega de fuentes, como en todas las configuraciones de array empleadas teníamos micrófonos omnidireccionales muy cercanos, hemos empleado una extensión a los 360º del método descrito en [19].

4.1. Beamforming

Como vimos en teoría, para el caso de ruido espacialmente blanco y una única señal incidiendo en el array (es decir, sin interferencias), el beamforming de retardo y suma es el óptimo ya que es el que consigue la mejor SNR de salida. Por ello, en primer lugar analizaremos los resultados del Beamforming de Retardo y Suma para un CUA de $R = 0.085$ m. El array consta de 8 micrófonos cardioides de tipo electret de muy bajo coste. En la tabla 4.1, se muestra un resumen de los resultados obtenidos. En dichos resultados, podemos observar, como al alejarnos de la dirección de apuntamiento de un micrófono, la ganancia obtenida disminuye ligeramente, alcanzando su punto mínimo en la intersección entre dos micrófonos. Este efecto es similar, al que ocurre en los arrays lineales, cuando apuntamos a la dirección broadside respecto a la dirección endfire, pero con esta configuración, el efecto es mucho menor, lo que da como resultado un ancho de haz prácticamente constante en cualquier dirección de apuntamiento.

Como ya sabemos, podemos mejorar las prestaciones del array empleando configu-

DOA	SNR_{IN}	SNR_{OUT}	G
9º	12.2	20.0	7.8
45º	11.6	19.7	8.1
64º	5.6	12.2	6.6
93º	12.5	21.5	9.0
124º	10.6	17.7	7.1

Tabla 4.1: Relaciones SNR (dB) para un CUA de 8 micrófonos cardioides.

DOA	G_{1mic}	G_{2mic}	G_{3mic}	G_{4mic}	G_{5mic}
9º	3.9	4.5	4.6	3.5	2.3
45º	0.2	1.6	1.2	1.3	0.0
64º	1.5	2.0	1.2	1.1	0.4
123º	-0.2	0.3	0.6	0.2	0.0
270º	1.9	2.4	2.3	1.9	1.1

Tabla 4.2: Ganancias en (dB) entre la configuración inicial y una Wullenweber de N micrófonos.

raciones Wullenweber. Estas configuraciones son concebidas para un gran número de elementos, pero en nuestro caso, el array se reduce a tan solo 8 micrófonos, que suele ser el número de canales de entrada de las tarjetas de sonido comerciales. Supongamos que la fuente de interés incide directamente sobre la posición de uno de los micrófonos, p. ej. 90º. Si solo utilizamos los tres micrófonos más cercanos tendremos un arco de 90º, pero si empleamos los cinco micrófonos más cercanos, el arco se nos amplía hasta 180º. Para ver que configuración es la óptima se ha comparado la SNR obtenida para diverso número de micrófonos respecto a cogerlos todos como hacíamos hasta ahora. Ver tabla 4.2.

Tras observar los resultados, puede verse como la mejor configuración suele ser la de emplear tan sólo los dos micrófonos más cercanos a la fuente de interés, siendo el resultado muy parecido al de emplear los tres micrófonos más cercanos. Sin embargo, lo que más llama la atención, es que la diferencia entre emplear estas configuraciones y sólo un micrófono, son mínimas. Es decir, en una configuración de tan pocos micrófonos en círculo, y al ser estos direccionales, no hacer beamforming y quedarnos sólo con el micrófono que apunta hacia la fuente de interés, ofrece prácticamente el mismo resultado.

En el capítulo 2 dijimos que era posible obtener mejoras considerando el ruido instrumental de los micrófonos para hacer la estimación de la ganancia de ruido del beamformer. Para ello explicamos un método en el que se calculaban los haces casi óptimos teniendo en cuenta el ruido ambiental y el instrumental. En la tabla 4.3 se muestra una

DOA	**SNR_{IN}**	**SNR_{1mic}**	**SNR_{DS}**	**SNR_{NI}**	**G_{1mic}**	**G_{DS}**	**G_{NI}**
100º	12.4	20.5	22.7	26.0	8.1	10.3	13.6
130º	10.5	18.2	19.8	22.1	7.7	9.3	11.6
260º	9.2	18.3	23.6	26.0	9.1	14.4	16.8
350º	12.2	23.9	24.5	26.9	11.7	12.3	14.7

Tabla 4.3: Comparativa entre el DS y el algoritmo que contempla el ruido instrumental.

Grabación	**DOA**	**$SINR_{IN}$**	**$SINR_{DS}$**	**$SINR_{NI}$**	**$SINR_{Frost}$**	**G_{DS}**	**G_{NI}**	**G_{Frost}**
2 fuentes	80º	-5.3	-6.1	-4.7	3.1	-0.8	0.6	8.4
	130º	5.3	7.5	7.9	6.8	2.2	2.6	1.5
4 fuentes	350º	-2.7	3.0	3.1	8.6	5.7	5.5	11.3
	100º	-3.1	2.4	1.3	7.9	5.5	4.4	11.0
	170º	-3.9	-0.1	-0.5	1.6	3.8	3.4	5.5
	260º	-9.7	-3.1	-2.3	-1.6	6.6	7.4	8.1

Tabla 4.4: Ganancias en la SINR (dB) entre el DS, el algoritmo que contempla el ruido instrumental y el Frost Beamformer.

comparativa entre el DS y dicho método. Como podemos ver, los resultados mejoran sensiblemente.

En muchos escenarios, además de tener ruido ambiental, tendremos interferencias localizadas en una determinada posición espacial. Por lo que, un parámetro que nos interesa medir es la relación Señal a Interferencia y Ruido (SINR). Como ya sabemos, el beamformer de retardo y suma establece los lóbulos y ceros en posiciones fijas dada la geometría del array. Por su parte, el segundo algoritmo descrito, establece el ancho óptimo para cada frecuencia, pero este ancho, es calculado off-line, por lo tanto no tiene en cuenta la señal interferente. Para intentar mejorar este aspecto se introdujo un algoritmo adaptado a la interferencia, el Frost Beamformer. En la tabla 4.4 se muestran resultados de SINR en diversas grabaciones para los tres métodos descritos:

Tras observar los resultados podemos ver como el Frost Beamformer obtiene los mejores resultados en cuanto a SINR. Sin embargo, mientras que los beamformers invariantes en el tiempo son diseñados bajo la suposición de ruido isotrópico, los algoritmos adaptativos funcionan mejor considerando fuentes de ruido puntuales [20]. Aunque este tipo de procesamiento adaptativo es lineal y no introduce artefactos ni ruido musical, los algoritmos adaptativos pueden tener algunos residuos audibles y distorsiones, por lo que no podemos decir que exista un algoritmo óptimo para todas las aplicaciones.

Método	Parámetro	Grabación 1		Grabación 2		Grabación 3	
		F1 = 2º	F2 = 163º	F1 = 2º	F2 = 111º	F1 = 99º	F2 = 126º
DUET	**SDR**	2.1	4.3	0.2	2.6	-4.8	-3.1
	SIR	19.2	11.2	17.2	6.5	-3.2	11.5
	SNR	28.8	28.5	32.6	38.0	35.3	26.7
	SAR	2.2	5.6	0.4	5.7	5.1	-2.7
SOFT	**SDR**	12.0	9.7	9.1	1.7	-2.5	-0.4
	SIR	38.9	41.2	23.7	21.2	6.1	16.5
	SNR	47.3	45.8	45.0	40.7	37.2	38.9
	SAR	12.0	9.8	9.2	1.8	-1.0	-0.2
HARD	**SDR**	6.6	7.9	6.0	1.2	-12.0	-0.3
	SIR	50.1	44.3	35.7	34.2	3.4	19.2
	SNR	53.3	65.0	42.5	43.2	32.5	44.2
	SAR	6.5	7.8	6.0	1.2	-10.2	-0.2

Tabla 4.5: Comparación entre métodos para diversas configuraciones de 2 fuentes.

4.2. Separación ciega de fuentes

La separación de fuentes de audio suele tener como último fin la escucha de las propias fuentes, por lo tanto, la calidad de una determinada separación estará relacionada con la distorsión percibida entre las señales separadas y las originales (normalmente desconocidas). Estas distorsiones serán interferencias de otras fuentes, artefactos de ruido musical, distorsiones de timbre y distorsiones espaciales respecto a la fuente objetivo. Los artefactos de ruido musical, suelen ser causados por filtrados no lineales de los datos. Estos ruidos son más molestos que el resto de distorsiones, sobre todo para aplicaciones musicales que requieren alta calidad. Criterios objetivos como Relación Distorsión a Fuente (SDR), Relación Interferencias a Fuente (SIR) y la Relación Artefactos a Fuente (SAR), han sido empleados en este trabajo para hacer comparaciones entre métodos. Para ello hemos empleado las librerías publicadas en la *First Stereo Audo Source Separation Evaluation Campaign,* [21].

En la tabla 4.5 se presentan los resultados correspondientes a la separación de varias grabaciones con 2 fuentes, mediante DUET y mediante nuestro algoritmo tanto para máscaras HARD como SOFT.

A la vista de los resultados podemos ver como el método propuesto funciona mejor que el DUET. Cuando las fuentes están suficientemente separadas se consiguen ganancias en SIR superiores a 20 dB. Sin embargo, al ir disminuyendo esta distancia no podemos aprovechar las ventajas del procesamiento en array y esta ganancia va disminuyendo.

Método	**Parámetro**	**F1 = 2º**	**F2 = 111º**	**F3 = 163º**	**F4 = 264º**
DUET	**SDR**	1.3	-15.3	-3.6	-6.3
	SIR	10.7	-8.8	0.0	5.3
	SNR	29.0	23.0	24.8	1.9
	SAR	2.2	-4.8	1.9	-4.8
SOFT	**SDR**	5.8	5.4	3.1	1.5
	SIR	21.0	23.1	19.7	25.3
	SNR	45.5	40.7	48.0	35.7
	SAR	6.0	6.5	3.3	1.5
HARD	**SDR**	3.9	6.3	3.2	0.6
	SIR	27.2	37.1	31.6	34.6
	SNR	43.7	41.2	47.0	38.0
	SAR	3.9	6.3	3.2	0.6

Tabla 4.6: Comparación entre métodos para una configuración de 4 fuentes.

De entre nuestras dos implementaciones, vemos también como el HARD consigue una mayor SIR a costa de una menor SAR. En audio, las medidas subjetivas suelen ser más apropiadas que las objetivas, Para saber que es mejor, si una mayor SIR o una menor SAR, hemos recurrido a analizarlas subjetivamente. Para situaciones en las que las fuentes están suficientemente separadas, los resultados obtenidos con las máscaras HARD, son mejores que con la SOFT, pero conforme la separación angular va disminuyendo, el número de artefactos musicales aumenta y es más conveniente emplear máscaras SOFT, que además distorsionan menos la señal de interés (mayor SDR).

Como ya sabemos, con DUET solo podemos separar fuentes comprendidas en un arco de 180^{o}, pero en muchas aplicaciones necesitaremos cubrir los 360^{o}. Una de las formas de conseguirlo es aumentando el número de micrófonos del array para formar geometrías que no presentan dicha ambigüedad. En la tabla 4.6, se muestran los resultados obtenidos con una grabación de cuatro fuentes distribuidas por todo el espacio. Como podemos apreciar, el método de separación iterativa con máscaras HARD, obtiene unos resultados del entorno de 30 dB de SIR, manteniendo unos niveles de SDR y de SAR aceptables.

Una de las aplicaciones más típicas y en las que más repercusión puede tener este método es la mejora de la calidad de voz en videoconferencias. En este tipo de aplicaciones, lo más habitual es tener una fuente de voz que será la que nos interese y una o varias fuentes interferentes distribuidas por el resto del espacio y que irán variando su posición. Para ver como se comporta el método ante este tipo de situaciones hemos hecho un experimento (ver tabla 4.7), en el cual, tenemos una fuente en $\theta = 270^{\text{o}}$ y hasta 7 fuentes interferentes distribuidas entre $\theta \epsilon [0^{\text{o}}, 180^{\text{o}}]$. Los resultados obtenidos superan con

Parámetro	**DUET**	**SOFT**	**HARD**
SDR	-7.4	2.9	0.9
SIR	3.0	11.2	23.1
SNR	19.9	38.8	51.0
SAR	-5.2	3.8	0.9

Tabla 4.7: Comparación para sistema de videoconferencia con la fuente de interés en $\theta = 270^{o}$ y 7 fuentes interferentes distribuidas entre $\theta\epsilon[0^{o}, 180^{o}]$.

Método	**Parámetro**	**F1 = 7º**	**F2 = 60º**	**F3 = 124º**	**F4 = 144º**	**F5 = 152º**
NORMAL	**SDR**	-23.2	-18.2	-29.8	-28.2	-31.3
	SIR	-11.9	0.2	-16.0	0.1	-3.5
	SNR	25.8	18.0	19.6	11.2	14.2
	SAR	-10.6	-15.4	-13.4	-24.8	-26.0
DS	**SDR**	-14.5	-4.0	-10.4	-16.2	-19.2
	SIR	-5.1	19.7	3.7	11.6	5.8
	SNR	35.0	31.3	24.6	19.5	22.0
	SAR	-7.7	-3.9	-8.6	-15.9	-18.1

Tabla 4.8: Parámetros de separación con 5 fuentes muy cercanas con el método normal y con beamforming DS antes de aplicar el método normal.

creces al DUET. Además, con estos métodos, conseguimos una SIR y una SNR muy por encima de la lograda con cualquier beamformer, aunque a costa de introducir artefactos y algo de distorsión en la señal original. Para paliar este inconveniente, una solución típica es añadir un poco de la señal original, de tal manera que camuflamos los artefactos con un fondo de ruido que corresponde a la señal original más la interferente. Es decir, eliminamos la percepción psicoacústica de los artefactos introduciendo más interferencia (disminuyendo la SIR). Aún así, los resultados siguen superando a los obtenidos con las técnicas de beamforming.

El método propuesto para separación de fuentes funciona bien siempre que las fuentes interferentes tengan una gran separación angular. En caso contrario, los resultados empiezan a empeorar, ya que la suposición de que la fuente de interés está en $\alpha > 0\, y\, delta > 0$ no siempre se cumple. Para minimizar estos problemas se ha empleado el array de la figura 3.6. Además se ha empleado un sencillo beamformer (DS), apuntando hacia cada una de las fuentes, con el fin de aumentar la probabilidad de que se cumpla suposición anteriormente comentada. Los resultados se muestran en la tabla 4.8.

Capítulo 5

Resumen y Conclusiones

A lo largo de este trabajo, se ha llevado a cabo una revisión de diferentes algoritmos para la mejora de la señal de voz. En concreto, nos hemos centrados en algoritmos de procesado en array, ya que con ellos, podemos aprovechar las ventajas de la distribución espacial de las fuentes y el ruido.

En primer lugar hemos planteado la problemática de una señal de banda ancha como es la de la voz para el tratamiento en array. En especial, al restringirnos a aspectos prácticos tales como un número moderado de elementos y un tamaño de array relativamente pequeño.

Hemos revisado distintos algoritmos de beamforming, una de las técnicas tradicionalmente más empleadas. En concreto hemos presentado tres algoritmos, el retardo y suma, que es el algoritmo más sencillo y típico, un algoritmo suboptimo de reducción de ruido con implementación en tiempo real, y uno de los algoritmos más empleados para supresión de interferencias con restricciones (LCMV) al que le hemos añadido un robusto detector de actividad para evitar problemas de pérdida de la señal de interés.

Todos los algoritmos se han implementado en un sistema real consistente en un array circular uniforme de radio $r = 0.085m$ compuesto por 8 micrófonos cardioides de tipo electret. Se ha elegido esta configuración por que por su reducido tamaño, y su bajo coste, puede ser implementada en sistemas reales. Estos algoritmos han demostrado producir una mejora sensible de la calidad de voz respecto a emplear únicamente un micrófono omnidireccional. Sin embargo, la mejora introducida está todavía muy lejos de ser la ideal.

En una búsqueda por mejorar los resultados se ha recurrido a técnicas de separación ciega de fuentes de audio, ya que ha sido demostrado experimentalmente que las máscaras tiempo-frecuencia son más poderosas que el beamforming para la separación de mezclas de audio en entornos reales reverberantes [3].

Dentro de este tipo de procesamiento, hemos analizado la que probablemente sea la técnica más popular, DUET. A partir de ella, hemos construido un array que nos permita explotar las ventajas de la información espacial para suprimir las limitaciones presentes.

Para finalizar, se ha construido otro array más complejo en el que se ha aplicado un procesado que combina el beamforming con la separación ciega para mejorar los resultados en situaciones más adversas.

Como línea futura, se plantea seguir el camino de aprovechar lo mejor de las dos técnicas para lograr mejores resultados.

Agradecimientos

En primer lugar, me gustaría darle las gracias a todos los miembros del Grupo de Tratamiento de Audio y Comunicaciones (GTAC), por su gran acogida y compañerismo. En especial, al Prof. José J. López y a Máximo Cobos, por su continua ayuda en mi trabajo de investigación.

También quiero aprovechar dar gracias a mi familia, y en especial a mis padres, por apoyarme en todas mis decisiones y animarme siempre hasta el final. Para finalizar, no me quiero olvidar de mis amigos y mis compañeros de clase por haber hecho esta etapa de mi vida más amena y llevadera.

Este trabajo ha sido financiado por el Ministerio de Educación y Ciencia bajo el proyecto TEC2006-13883-C04-01 y fondos FEDER.

Bibliografía

[1] M. Brandstein and D. Ward, *Microphone Arrays*. Springer Verlag, 2001.

[2] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Springer, 2005.

[3] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Blind audio source separation," technical report, University of London, centre for digital music, November 2005.

[4] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoustics, Speech, and Signal Processing Magazine*, pp. 4–24, April 1998.

[5] Naidu and P. S., *Sensor array signal processing*. Boca Raton, CRC Press, 2001.

[6] A. Karbasi and A. Sugiyama, "A new doa estimation method using a circular microphone array," *European Association for Signal and Image Processing, EURASIP*, 2007.

[7] M. Wax and J. Sheinvald, "Direction finding of coherent signals via spatial smoothing for uniform circular arrays," *IEEE Transactions on antennas and propagation*, 1994.

[8] A. Rudge, K. Milne, A. Olver, and P. Knight, *The Handbook of Antenna Design*, vol. 2. London, IEE Electromagnetic Waves Series, 1983.

[9] I. Tashev and H. S.Malvar, "A new beamformer design algorithm for microphone arrays," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2005.

[10] R. Fletcher, *Practical Methods of Optimization*. John Wiley Sons, 1987.

[11] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, January 1999.

[12] Y. Rui, D. Florêncio, W. Lam, and J. Su, "Sound source localization for circular arrays of directional microphones," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2005.

[13] S. Makino and S. Rickard, *Blind Speech Separation*. Springer, 2007.

[14] A. Karbasi and A. Sugiyama, "A new doa estimation method using a circular microphone array," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007.

[15] J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3d localization and tracking of sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, 2007.

[16] O. Y. A. Jourjine, S. Rickard, "Blind separation of disjoint orthogonal signal: Demixing n sources from 2 mixtures," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, June 2000.

[17] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. WILEY, 2nd edition ed., 2000.

[18] T. V. den Bogaert, J. Wouters, T. J. Klasen, and M. Moomen, "Blind separation of disjoint orthogonal signal: Demixing n sources from 2 mixtures," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2005.

[19] M. Cobos, J. J. López, and D. Martínez, "Two-microphone multi-speaker localization based on a laplacian mixture model," *IEEE Signal Processing Letters*, accepted for publication.

[20] I. Tashev and A. Acero, "Microphone array post-processing using instantaneous direction of arrival," *International Workshop on Acoustic Echo and Noise Control*, September 2006.

[21] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," *International Conference on Independent Component Analysis and Signal Separation (ICA)*, Septiembre.