

UNIVERSIDAD POLITÉCNICA DE MADRID
ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN
DEPARTAMENTO DE SEÑALES, SISTEMAS Y RADIOCOMUNICACIONES



TESIS DOCTORAL

MEJORA DE SEÑAL DE VOZ EN CONDICIONES ACÚSTICAS ADVERSAS MEDIANTE ARRAYS DE MICRÓFONOS

Autor: José Luis Sánchez Bote

Director: Joaquín González Rodríguez



Madrid, 2004

TESIS DOCTORAL

MEJORA DE SEÑAL DE VOZ EN CONDICIONES ACÚSTICAS ADVERSAS
MEDIANTE ARRAYS DE MICRÓFONOS

AUTOR

JOSÉ LUIS SÁNCHEZ BOTE

DIRECTOR

JOAQUÍN GONZÁLEZ RODRÍGUEZ

El Tribunal nombrado para juzgar la Tesis Doctoral arriba citada, compuesto por:

Presidente:

LUIS ALFONSO HERNÁNDEZ GÓMEZ

Vocales:

FRANCISCO JAVIER HERNANDO PERICÁS

ALBERTO GONZÁLEZ SALVADOR

MANUEL A. SOBREIRA SEOANE

Secretario:

FRANCISCO JAVIER CASAJÚS QUIRÓS

Acuerda otorgar la calificación de: **SOBRESALIENTE CUM LAUDE**

En Madrid, a 23 de Junio de 2004

AGRADECIMIENTOS

Gracias a todos los que me han ayudado de forma directa o indirecta en la realización de esta Tesis Doctoral.

A mi director de Tesis, Joaquín González Rodríguez, por la colaboración científica y la revisión del texto.

A mis compañeros del DIAC, por su apoyo durante este largo periodo, especialmente a Danilo Simón Zorita, compañero de fatigas.

A Javier Alonso Valdesueiro, pieza fundamental para la implementación del prototipo y los experimentos finales.

A todos los miembros del ATVS por ayudarme siempre que se lo pedí.

A mi familia y amigos por soportarme.

Gracias a todos.

RESUMEN

El trabajo desarrollado en esta Tesis estudia el procesado de señal usando arrays de micrófonos, aplicado a la mejora de señal de habla (*Speech Enhancement*), cuando el locutor se encuentre en condiciones acústicas adversas, especialmente con altos niveles de ruido y reverberación. También se describe la implementación de un prototipo de array en tiempo real que utiliza las propuestas y soluciones aportadas en la Tesis.

El objetivo primordial perseguido es la propuesta y revisión de los métodos de procesado en array usando micrófonos que en algún sentido sean novedosos, y su implementación práctica en un prototipo que pueda funcionar de manera autónoma y que sirva para investigaciones futuras dentro de la misma línea de trabajo de procesado en array de señales acústicas.

El capítulo 1 es de introducción y en él se plantea la problemática del procesado con arrays microfónicos, sus antecedentes y el marco de trabajo en el que se ha realizado la Tesis, así como los objetivos y estructura de la misma. Después sigue el cuerpo principal de la Tesis, dividido en tres partes, cuyo contenido se expresa resumidamente a continuación.

La “Parte 1: Mejora de habla con arrays de micrófonos” contiene una revisión del estado del arte sobre el procesado multicanal de señal de voz en condiciones de ruido y reverberación. Está a su vez dividida en cuatro capítulos, del 2 al 5.

El capítulo 2 trata la conformación de haz (*beamforming*) y su aplicación a los arrays microfónicos. Se estudian los métodos más frecuentes de conformación de haz y su relación con la reducción de ruido y reverberación en el habla. Con relación a las propuestas de la Tesis, tiene especial relevancia lo concerniente a arrays anidados y a conformación superdirectiva.

El capítulo 3 está dedicado a la localización de fuente usando un array de micrófonos y se describen los métodos más conocidos para evaluar la posición de la fuente de señal acústica a partir de la información multicanal proporcionada por el array. De entre los métodos más habituales se describe la maximización de la potencia de salida del array y la estimación de retardos por correlación cruzada. Como métodos más novedosos se tratan los basados en la descomposición de la señal multicanal en subespacios vectoriales.

El capítulo 4 se dedica a las técnicas de mejora de habla mediante postfiltrado. Se describe el filtrado de Wiener multicanal y la sustracción espectral, así como otras variantes de filtrado de especial interés aquí, como la técnica de supresión auditiva de ruido (método ANS). También se estudian las técnicas de estimación de los espectros de potencia de voz y de ruido. Adicionalmente, en el capítulo 4 se describe la problemática de la derreverberación ciega, en especial la técnica de la derreverberación por descomposición cepstral, utilizada en las propuestas del autor.

Finalmente, el capítulo 5, último de esta primera parte, trata las técnicas de estimación objetiva de la calidad de la señal de habla utilizadas para evaluar los resultados a la salida del

procesador en array. Se estudian los estimadores objetivos de mejora de habla basados en la relación señal a ruido y los que utilizan el índice de inteligibilidad STI (*Speech Transmission Index*).

La “Parte 2: Propuestas, experimentos y resultados preliminares” presenta las propuestas realizadas por el autor de forma preliminar, enfocadas a la implementación posterior del prototipo de array. Contiene dos capítulos que se sintetizan seguidamente.

En el capítulo 6 se describe el comportamiento teórico del array microfónico en la configuración que va a ser usada en las pruebas iniciales y también las bases de datos de señal de voz multicanal que van a ser evaluadas en esos experimentos iniciales.

El capítulo 7 contiene las principales aportaciones algorítmicas de la Tesis, así como las pruebas y experimentos realizados sobre las bases de datos descritas en el capítulo 6. Las propuestas realizadas en el capítulo 7 consisten en dos procesadores multicanal combinados (procesador MW-MPAP y procesador ANS-MW) y un método de evaluación objetiva de la calidad de habla (índice E-RASTI), basado en el cálculo de los índices de inteligibilidad que se usan actualmente en evaluaciones de tipo electroacústico.

A continuación, la “Parte 3: Implementación sobre DSP de un prototipo en tiempo real” se dedica a los aspectos relacionados con el diseño y las pruebas del prototipo de array. Contiene a su vez otros dos capítulos.

El capítulo 8 describe los elementos *hardware* utilizados y el *software* diseñado, que componen el prototipo final implementado.

En el capítulo 9 se desarrolla la propuesta final de mejora de señal de voz usando el prototipo de array en tiempo real. Esta propuesta contiene las ideas planteadas en las pruebas preliminares del capítulo 7, añadiendo como novedad la utilización de un conformador superdirective en la banda de baja frecuencia. Se ofrecen pruebas electroacústicas de directividad que corroboran el paralelismo existente entre las predicciones teóricas destacadas por el autor y los resultados de dichas medidas. En el capítulo 9 además se describen las pruebas y experimentos con el prototipo de array sobre una base de datos multicanal en español, generada por el autor con el empleo del procesador implementado.

Por último se ha considerado una parte a modo de epílogo que contiene el capítulo 10 con las conclusiones y aportaciones destacables del trabajo de investigación y las líneas propuestas de trabajo futuro y un capítulo de bibliografía, con la relación de referencias usadas a lo largo de la Tesis.

A la vista de los contenidos que son tratados en este trabajo de investigación, queda patente el esfuerzo que se ha hecho en fijar las bases del procesado en array de señales acústicas, con especial énfasis en la implementación de un prototipo en tiempo real, de tal manera que la labor realizada pueda ser utilizada para mejoras y desarrollos futuros que complementen y continúen la línea de investigación seguida a lo largo de esta Tesis.

ABSTRACT

In this Ph. D. Thesis it has been studied the application of microphone arrays to speech signal enhancement when the speaker is placed in adverse acoustic conditions, specially into rooms with high noise and reverberation levels. Additionally the implementation of a real time array prototype is described, using the proposals and solutions exposed in the Ph. D. Thesis.

The primary goal of this work is the proposition and revision of those array processing methods using microphones which are relatively new, and its implementation on a prototype able to work autonomously and which can be useful for future research that continues the same work guidelines about acoustic signal array processing followed here.

Chapter 1 is an introduction where signal processing using microphone arrays is presented, and also the antecedents and background associated with this Ph. D. Thesis work. In this chapter the goals and Ph. D. Thesis structure are described as well. In the following, the main body of the text is organized in three parts, whose contents are summarized below.

“Section 1: Speech enhancement with microphone arrays” contains a state of the art revision about multichannel speech signal processing in noise and reverberation, and it is organized in four chapters, from 2 to 5.

Chapter 2 describes beamforming and its application to microphone arrays. The most frequent beamforming methods are studied in detail, and how they can be applied to enhance noisy and reverberant speech. In this chapter the theory of nested arrays and superdirective beamforming is shown, which is of special interest for its relation with the Ph. D. Thesis proposals.

Chapter 3 deals with source localization using microphone arrays, and the most popular methods to search the position of the acoustic signal source by means of the multichannel information provided by the array are described. In the category of conventional methods of source localization, those based on the maximization of the array power output has been considered and also the methods that use the interchannel cross correlation for estimating the time aligning delay related to the source position. The newest techniques for source localization, known as subspace methods, apply a decomposition of the multichannel signal into vectorial subspaces, and they have been also reviewed in chapter 3.

Chapter 4 studies speech enhancement using postfiltering techniques. Multichannel Wiener filtering and spectral subtraction are described as well as another specially relevant filtering alternatives like the audible noise suppression technique (ANS method) that has been used along this research work. Different procedures of speech and noise power estimation are also analysed. Additionally, chapter 4 offers a revision of blind dereverberation techniques, and focuses specially in the dereverberation algorithms based on cepstral decomposition that are present in the author’s proposals.

Finally, chapter 5, the last one in this first section, deals with methods for speech quality objective estimation which have been used to evaluate the results at the array output in the author's experiments. Different objective estimators based on the signal to noise ratio are studied, and also those that apply the STI (Speech Transmission Index) intelligibility index.

"Section 2: Proposals, experiments and preliminary results" contains the author's contributions considered as preliminary work, which will be used in the subsequent implementation of the array prototype. It comprises two chapters as is shown below.

In chapter 6 the theoretic behaviour of the array in the configuration to be used in the initial experiments is described. The multichannel speech databases in use in those preliminary experiments are also described.

Chapter 7 contains the main algorithmic contributions of this Ph. D. Thesis, and also the experiments with the multichannel databases described in chapter 6. The proposals offered in chapter 7 consist in two combined multichannel processors (MW-MPAP processor and ANS-MW processor) and a speech quality objective evaluation method (E-RASTI index), based on the well-known RASTI intelligibility index, frequently used at present in electroacoustical-type evaluations.

Later, "Section 3: Implementation on DSP of a real time prototype", shows those aspects related with the design and tests of the array prototype. This section contains also two different chapters, which are described below.

Chapter 8 shows the prototype hardware elements in use and the developed software elements as well, which are the main parts of the final array implementation.

In chapter 9 the speech enhancement final proposal using the real time array prototype is described. This proposal contains the ideas presented in the preliminary studies of chapter 7, adding as innovation a superdirective beamformer in the low frequency band. Several directivity measurements are shown here, demonstrating the good correlation between the theoretical predictions and the array behaviour. Lots of experiments using a multichannel Spanish database, generated by the author using the array prototype, are also described in chapter 9. Unlike the experiments described in chapter 7, these have been done entirely with the array prototype and in real conditions, established by the author.

Finally, the epilogue contains chapter 10 with the most outstanding contributions and conclusions obtained in the research work of this Ph. D. Thesis. It includes also the guidelines for future work, proposed to continue the development of multichannel speech enhancement using microphone arrays. In this epilogue a bibliography chapter is included, with the list of references used through the text.

RELACIÓN DE ACRÓNIMOS Y SÍMBOLOS UTILIZADOS

ACRÓNIMOS

ADC	<i>Analog-Digital Converter</i> (conversor analógico-digital).
AI	<i>Articulation Index</i> (índice de articulación).
Alcons	<i>Articulation loss of consonants</i> (pérdida de articulación de consonantes).
ALU	<i>Arithmetic-Logic Unit</i> (unidad aritmético-lógica).
AMT	<i>Auditory Masking Threshold</i> (umbral de enmascaramiento auditivo).
ANS	<i>Audible Noise Suppression</i> (supresión de ruido audible).
ANS-MW	<i>Audible Noise Suppressor-Modified Wiener</i> (supresor de ruido audible-Wiener modificado).
ATVS	Área de Tratamiento de Voz y Señales.
B&K	<i>Brüel and Kjaer</i> .
BWS	<i>Blue Wave Systems</i> .
CDB	<i>Constant Directivity Beamformer</i> (conformador de directividad constante).
CMU	<i>Carnegie Mellon University</i> . Base de datos real.
CSSM	<i>Coherent Signal Subspace Method</i> (método de subespacio de señal coherente).
DE	<i>Dual Excitation</i> (excitación dual).
DFT	<i>Discrete Fourier Transform</i> (transformada discreta de Fourier).
DI	<i>Directivity Index</i> (índice de directividad).
DIAC	Departamento de Ingeniería Audiovisual y Comunicaciones.
dLAR	distancia euclídea de parámetros LAR.
DOA	<i>Direction Of Arrival</i> (dirección de llegada). Posición del campo acústico donde se sitúa una fuente sonora.
dRCEP	distancia euclídea de parámetros RCEP.
DRT	<i>Diagnostics Rhyme Test</i> .
DSP	<i>Digital Signal Processor</i> (procesador de señal digital).
ESPRIT	<i>Estimation of Signal Parameters via Rotation Invariance Technique</i> (estimación de los parámetros de la señal por la técnica de invarianza de rotación).
E-RASTI	<i>Emulated-RASTI</i> (RASTI emulado).
FFT	<i>Fast Fourier Transform</i> (transformada rápida de Fourier).
FIFO	<i>First Input First Output</i> (lo primero en entrar es lo primero en salir).
FLOPS	<i>Floating-point Operations Per Second</i> (operaciones en punto flotante por segundo).
GAI	ganancia en AI para una señal procesada por el array.
GCC	<i>Generalized Cross-Correlation</i> (correlación cruzada generalizada).
GdLAR	ganancia en dLAR en una señal de habla procesada por el array.
GdRCEP	ganancia en dRCEP en una señal de habla procesada por el array.
GJBF	<i>Griffith-Jim BeamFormer</i> (conformador de Griffith-Jim).
GNMR	ganancia en NMR al aplicar un procesador en array.

GSC	<i>Generalized Sidelobe Canceller</i> (cancelador generalizado del lóbulo lateral).
GSNR	ganancia en SNR al aplicar un procesador en array.
GSNR _A	ganancia en SNR _A al aplicar un procesador en array.
GSVD	<i>Generalized Singular Value Decomposition</i> (descomposición generalizada en valores singulares).
IFT	<i>Inverse Fourier Transform</i> (transformada inversa de Fourier).
IPS	<i>Instructions Per Second</i> (instrucciones por segundo)).
IQML	<i>Iterative Quadratic Maximun Likelihood</i> (máxima verosimilitud cuadrático-iterativa).
LAR	<i>Log Area Ratio</i> .
l _{DML}	<i>logarithmic Deterministic ML</i> (ML determinista logarítmica).
LFIFO	número de muestras utilizadas de la FIFO del DSP.
LPC	<i>Linear Predictive Coding</i> (codificación por predicción lineal).
MIN-NORM	<i>MINimum NORM</i> (norma mínima).
ML	<i>Maximum Likelikood</i> (máxima verosimilitud).
MMSE	<i>Minimum Mean Squared Error</i> (mínimo error cuadrático medio).
MOS	<i>Mean Opinion Score</i> (baremo de la opinión media).
MPAP	<i>Minimum Phase-All Pass</i> (fase mínima-paso todo).
MS	<i>Mean Squared</i> (media cuadrática).
MSE	<i>Mean Squared Error</i> (error cuadrático medio).
MTF	<i>Modulation Transfer Function</i> (función de transferencia de modulación).
MUSIC	<i>MUltiple Signal Classification</i> (clasificación de señales múltiples).
MVDR	<i>Minimum Variance Distortionless Response</i> (respuesta de mínima varianza sin distorsión).
MW	<i>Modified Wiener</i> (Wiener modificado).
MW-MPAP.....	<i>Modified Wiener-Minimum Phase All Pass</i> (Wiener modificado-fase mínima paso todo).
NMR	<i>Noise to Masking Ratio</i> (relación de ruido a umbral de enmascaramiento).
NSS.....	<i>Non-linear Spectral Subtraction</i> (Sustracción espectral no lineal).
PAMS	<i>Perceptual Analysis Measurement System</i> (sistema de medida de análisis perceptual).
PHAT	<i>PHASE Transform</i> (transformada de fase).
PMC	<i>PCI Mezzanine Card</i> .
PSQM	<i>Perceptual Speech Quality Measurement</i> (medida de la calidad de habla perceptual).
RASTI.....	<i>RApid STI</i> (índice STI rápido).
RCEP	<i>Real CEPstrum</i> (cepstrum real).
RMS.....	<i>Root Mean Squared</i> (raíz de la media cuadrática).
SD-ANS-MW	<i>Super Directive-ANS-MW</i> (ANS-MW superdirectivo).
SF.....	<i>Spreading Function</i> (función de ensanchamiento).
SFM	<i>Spectral Flatness Measure</i> (medida de la suavidad espectral).
simCMU-1.....	base de datos multicanal simulada a partir de CMU. Considera reverberación pero no ruido.
simCMU-1.1	base de datos multicanal simulada a partir de simCMU-1 mediante la adición de ruido. No considera la reverberación del ruido.
simCMU-2.....	base de datos multicanal simulada a partir de CMU. Considera reverberación y ruido.
SML	<i>Stochastic ML</i> (ML estocástica).
SNR	<i>Signal to Noise Ratio</i> (relación señal a ruido), medida normalmente a priori.
SNR _A	relación SNR con ponderación A.

SNR_{ap}	relación SNR aparente para el cálculo del STI.
SNR_{post}	relación señal a ruido a posteriori.
SNR_{T}	relación SNR truncada.
SPL	<i>Sound Pressure Level</i> (nivel de presión sonora).
SRAM	<i>Shared RAM</i> (RAM compartida).
SRP	<i>Steered Response Power</i> (potencia de respuesta conformada).
STFT	<i>Short Time Fourier Transform</i> (transformada de Fourier a corto plazo).
STI	<i>Speech Transmission Index</i> (índice de transmisión del habla).
SUS	<i>Semantically Unpredictable Sentences</i> (frases impredecibles semánticamente).
TDOA	<i>Time Difference Of Arrival</i> (diferencia de tiempo de llegada).
UPM	Universidad Politécnica de Madrid. Base de datos multicanal real obtenida con el prototipo implementado.
VAD	<i>Voice Activity Detection</i> (detección de actividad de voz).
WSF	<i>Weighted Subspace Fitting</i> (adecuación por subespacios ponderados).

SÍMBOLOS

1	vector columna de unos.
a, a	vector de apuntamiento (<i>steering vector</i>), elemento del vector de apuntamiento.
A, A	vector de apuntamiento (<i>steering vector</i>), o elemento del vector de apuntamiento, ambos en el dominio de la frecuencia.
A	filtro de ponderación en frecuencia de tipo A.
\mathbf{am}	matriz de apuntamiento (<i>steering matrix</i>).
b	índice de banda crítica.
b	respuesta al impulso eléctrica del micrófono ante una excitación acústica, excluyendo los efectos del retardo acústico.
B	banda o intervalo de frecuencias.
B	filtro de ponderación en frecuencia de tipo B.
B	función de transferencia eléctrica del micrófono ante una excitación acústica, excluyendo los efectos del retardo acústico.
B	número de bandas críticas considerados en el filtrado ANS.
\mathbf{B}	matriz de bloqueo.
B_1, B_2, B_3	cada una de las subbandas de frecuencia consideradas para el array anidado prototipo.
B_T	banda de frecuencias total considerada para el procesado en banda única.
c	coeficiente de ponderación utilizados en el cálculo del AI.
c	coeficiente de ponderación utilizados en el cálculo del STI.
c	función de correlación cruzada generalizada.
c	velocidad del sonido.
C	espectro ensanchado por SF para el procesado auditivo.
C	factor de ponderación para el array CDB.
C	filtro de ponderación en frecuencia de tipo C.
C	función coherencia para el procesador MW.
C	pesos complejos que representan la conformación de haz.
\mathbf{C}	vector de coeficientes de calibración del array.
\mathbf{C}'	vector de coeficientes de calibración del array corregidos.

\mathbf{c}_X	vector de cepstrum real correspondiente a la señal X_0 o de referencia del array.
\mathbf{c}_Y	vector de cepstrum real correspondiente a la señal de entrada al array Y.
D	coeficiente de renormalización para el procesado auditivo.
D	directividad de un micrófono.
D_{FA}	directividad del factor de array.
DI_{MAX}	índice de directividad máximo.
D_μ	directividad de los micrófonos del array.
E	función de error.
f	frecuencia.
f	secuencia discreta que representa la posición de los micrófonos del array en el eje z.
F	frecuencia de modulación para el cálculo del STI.
F	transformada de Fourier de la secuencia del array f.
f_1, f_2	frecuencias de división del espectro en las tres subbandas B_1, B_2 y B_3 , del array anidado prototipo.
$F_{20}, F_{21}, \dots, F_{25}$	frecuencias de modulación en la banda de 2kHz para el cálculo del E-RASTI.
$F_{50}, F_{51}, \dots, F_{54}$	frecuencias de modulación en la banda de 500Hz para el cálculo del E-RASTI.
f_a	frecuencia de <i>aliasing</i> espacial.
f_{nulo}	frecuencia en la que aparece el primer nulo de directividad.
f_{R1}, f_{R2}	frecuencias límite entre las que se realiza la detección VAD recursiva en el procesador propuesto.
f_s	frecuencia de muestreo.
G	número de parámetros LAR considerados.
\mathbf{g}_X	vector de parámetros LAR de la señal con espectro X.
h	respuesta al impulso eléctrica del micrófono ante una excitación acústica.
\hat{h}	cepstrum de h.
\mathbf{h}, h	vector de respuestas al impulso (o un elemento de \mathbf{h}) de los filtros aplicados a la señal del array.
H	función de transferencia eléctrica del micrófono ante una excitación acústica.
\mathbf{H}, H	vector de funciones de transferencia en frecuencia (o elemento de \mathbf{H}) asociadas a cada canal del array para la conformación de haz o para el postfiltrado de mejora.
h_{all}	parte paso todo de la respuesta al impulso h.
H_{all}	parte paso todo de la función de transferencia H.
H_{ANS}	filtro ANS.
H_{LBF}	filtro LBF (Jeannes R. le Bouquin y G. Faucon).
h_{min}	parte fase mínima de la respuesta al impulso h.
\hat{h}_{min}	parte fase mínima del cepstrum de h.
H_{min}	parte fase mínima de la función de transferencia H.
H_{MW}	postfiltro para el procesador MW.
\hat{h}_r	cepstrum real de h.
\mathbf{H}_S, H_S	vector filtro de sustracción espectral multicanal (o elemento de \mathbf{H}_S).
H_w	filtro óptimo de Wiener.
\mathbf{H}_w	vector filtro óptimo multicanal de Wiener.
i	índice del cada micrófono dentro del array.
I	intensidad acústica.

I	número de micrófonos del array.
I	matriz identidad.
I ₂	intensidad acústica filtrada en la octava de 2kHz para el cálculo del E-RASTI.
I ₅	intensidad acústica filtrada en la octava de 500Hz para el cálculo del E-RASTI.
I _{B1} , I _{B2} , I _{B3}	número de micrófonos respectivamente en las subbandas B ₁ , B ₂ y B ₃ , del array anidado prototipo.
I _T	número total de micrófonos del array anidado prototipo.
k	índice de trama.
k	número de onda de la transmisión acústica en el aire.
K	número de tramas temporales bajo análisis.
L	número de muestras temporales de cada trama de análisis.
L ₁	número de componentes cepstrales usados para la evaluación de GdRCEP.
m	índice de fuente.
m	índice de modulación del habla.
M	número de fuentes presentes con el array.
m ₂₁ , ..., m ₂₅	índices de modulación m comparativos en la banda de 2kHz para el cálculo del E-RASTI.
m ₅₁ , ..., m ₅₄	índices de modulación m comparativos en la banda de 500Hz para el cálculo del E-RASTI.
mx ₂₁ , ..., mx ₂₅	índices de modulación m de entrada en la banda de 2kHz para el cálculo del E-RASTI.
mx ₅₁ , ..., mx ₅₄	índices de modulación m de entrada en la banda de 500Hz para el cálculo del E-RASTI.
my ₂₁ , ..., my ₂₅	índices de modulación m de salida en la banda de 2kHz para el cálculo del E-RASTI.
my ₅₁ , ..., my ₅₄	índices de modulación m de salida en la banda de 500Hz para el cálculo del E-RASTI.
n	índice de banda de filtrado en el análisis temporal de la señal del array.
n	índice del micrófono para un array CDB.
n, n	vector de ruido a la salida del array, ruido de un micrófono.
N	número de puntos en frecuencia calculados por ventana, para el procesado por tramas.
N	número total de bandas de filtrado en el análisis temporal de la señal del array.
N	vector espectro del ruido (perturbación acústica) a la salida del array.
NB	número de subbandas para el array anidado.
n _m	vector de perturbación a la salida del array debida a las fuentes ajenas.
n _n	vector de perturbación a la salida del array debida al ruido aditivo.
n _r	vector de perturbación a la salida del array debida a la reverberación.
n _v	número de ventanas temporales de voz que caben en LFIFO.
O	parámetro de <i>offset</i> para el procesado auditivo.
p	presión acústica debida a una fuente acústica, medida en el centro de coordenadas.
P	potencia a la salida del array actuando como conformador.
P ₁ , P ₂ , ..., P ₈	puntos origen de las señales de voz y de ruido para generar la base de datos simCMU-2.
P _{MU}	espectro MUSIC.
q	índice del camino acústico para un sistema multirayecto.

- Q factor de apertura de un array CDB.
 Q factor de directividad.
 Q número total de caminos acústicos considerados para un sistema multirayos.
 Q_{MAX} factor de directividad máximo.
 r función de ponderación para el cálculo de \hat{h}_{min} .
 \mathbf{r} vector de posición de un punto genérico del campo acústico.
 r, θ, φ coordenadas esféricas de un punto genérico del campo acústico.
 r, θ, φ coordenadas esféricas de un punto genérico del campo acústico, referidas a uno de los micrófonos del array.
 r_0, θ_0, φ_0 coordenadas esféricas de apuntamiento del array.
 \mathbf{r}_i vector de posición del micrófono i -ésimo del array.
 \mathbf{r}_i vector de posición de un punto genérico del campo acústico respecto al micrófono i -ésimo del array.
 \mathbf{R}_{nn} igual que la matriz \mathbf{R}_{xx} pero calculada sólo con la perturbación acústica o ruido genérico.
 R_{xx} autocorrelación de x .
 \mathbf{R}_{xx} matriz de correlaciones cruzadas medida entre las señales de salida de los micrófonos del array, cuando no existe perturbación acústica.
 \mathbf{R}_{yy} igual que \mathbf{R}_{xx} pero considerándose incluida la perturbación acústica o ruido genérico.
 \mathbf{S} sensibilidad de un micrófono.
 S solapamiento interventana para el procesado por tramas.
 S_1 desplazamiento interventana para el procesado por tramas.
 \mathbf{T} matriz de transformación para el método CSSM.
 T umbral de enmascaramiento auditivo o AMT.
 T_{60} tiempo de reverberación.
 t_F tasa de FFT's por segundo.
 T_L duración temporal de la trama de voz captada.
 T_n lapso temporal donde predomina el ruido.
 ton factor de tonalidad de la señal de habla que entra al procesador.
 T_S duración temporal del solapamiento interventana para el procesado por tramas.
 T_s periodo de muestreo.
 T_{S_i} duración temporal del desplazamiento interventana para el procesado por tramas.
 T_y lapso temporal donde predomina la señal de voz.
 t_Δ tiempo de caída del estimador recursivo.
 U número de tramas temporales anteriores que se consideran para la calcular la media móvil.
 UC umbral de coherencia para el procesador MW.
 \mathbf{u}_n autovectores generadores del subespacio ruido.
 \mathbf{U}_n matriz de autovectores generadores del subespacio ruido.
 \mathbf{U}_x matriz de autovectores generadores del subespacio señal.
 \mathbf{U}_y matriz de autovectores generadores del espacio vectorial originado por la señal de salida del array.
 v función de ponderación temporal o ventana.
 \mathbf{w}, w vector de filtrado del array y elemento de \mathbf{w} , para realizar la conformación de haz en el dominio del tiempo o con una sola frecuencia.

W, W	vector de filtrado del array y elemento de W , para realizar la conformación de haz en el dominio de la frecuencia.
W_C	número de puntos considerados para el <i>liftering</i> cepstral paso bajo del procesador MPAP.
\hat{W}_{low}	<i>liftering</i> paso bajo en el dominio cepstral.
W_{RS}	vector de filtrado del array para el caso de retardo y suma, en el dominio del tiempo o considerada una sola frecuencia.
W_S	vector de coeficientes en frecuencia para implementar la suma de todos los canales del array.
W_{SD}	vector de filtrado en frecuencia para el array superdirectivo.
W'_{SD}	vector de filtrado en frecuencia para el array superdirectivo, si se aplica después de la alineación temporal de los canales.
x, x	vector de respuesta del array, respuesta de un micrófono. Representa la salida eléctrica del array.
X	vector espectro de la señal limpia a la salida del array, cuando no existe perturbación acústica.
X, Y, Z	vectores unitarios que definen el sistema cartesiano de coordenadas en el que se sitúa el array.
X', Y', Z'	vectores unitarios que definen el sistema cartesiano de coordenadas particular, con respecto a uno de los micrófonos del array.
x₀	respuesta del micrófono de referencia o señal de referencia. Representa la señal de voz limpia.
x₀	vector de respuesta de referencia del array cuando capta varias fuentes.
X₀	espectro del micrófono de referencia o espectro de referencia. Representa la señal de voz limpia.
x₂, y₂	señal temporal filtrada en la octava de 2kHz para el cálculo del E-RASTI.
x₅, y₅	señal temporal filtrada en la octava de 500Hz para el cálculo del E-RASTI.
y	salida temporal de un conformador genérico.
y	vector de respuesta del array si se incluyen las perturbaciones acústicas (ruido, reverberación o fuentes ajenas).
Y	espectro de salida de un conformador genérico.
Y	vector espectro de la señal sucia a la salida del array, incluyendo la perturbación acústica.
y_{all}	parte paso todo de la señal y.
Y_{all}	parte paso todo del espectro Y.
Y_{ANS}	espectro de salida del procesador ANS-MW (o sólo el procesador ANS).
y_{ANS}	señal de salida en tiempo para el procesador ANS-MW (o sólo el procesador ANS).
Y_B	vector de salida bloqueada del array, en frecuencia.
Y_{beam}	espectro paso todo conformado a partir de los espectros Y _{all} de salida del array.
Y_{LBF}	espectro de salida del postfiltro H _{LBF} .
ŷ_{low}	cepstrum fase mínima conformado después de aplicar un <i>liftering</i> cepstral.
Y_{low}	cepstrum inverso de ŷ _{low} .
y_{min}	parte fase mínima de la señal y.
Y_{min}	parte fase mínima del espectro Y.
Y_{MPAP}	espectro de salida del conformador MPAP.
Y_{MW}	espectro de salida del procesador MW.
y_{MW-MPAP}	señal de salida en tiempo para el procesador MW-MPAP.
Y_{MW-MPAP}	espectro de salida del conformador MW-MPAP.

\mathbf{Y}_R	vector de salida en frecuencia del array una vez aplicada la alineación temporal.
y_{RS}	señal de salida en tiempo para el conformador de retardo y suma.
\mathbf{Y}_{RS}	espectro de salida del conformador de retardo y suma.
y_S	señal de salida en tiempo de un sustractor espectral multicanal.
\mathbf{Y}_S	espectro de salida de un sustractor espectral multicanal.
y_{SD}	señal de salida en tiempo del conformador superdirectivo.
\mathbf{Y}_{SD}	espectro de salida del conformador superdirectivo.
y_W	señal de salida en tiempo del filtro de Wiener multicanal.
\mathbf{y}_W	vector de señales temporales filtradas por Wiener, antes del conformador.
\mathbf{Y}_W	espectro de salida del filtro de Wiener multicanal.
\mathbf{Y}_w	vector de señales en frecuencia, filtradas por Wiener, antes del conformador.
z	coordenada de posición de un micrófono sobre el eje Z para un array lineal.
α	factor de atenuación de una fuente sonora con respecto a otra presente.
α	factor de forma para la estimación de coherencia en el procesador MW.
α	factor de sobresustracción.
β	coeficiente de reflexión de una pared del recinto.
β	factor de suelo espectral (<i>spectral flooring</i>).
β	respuesta acústica al impulso del fenómeno de reflexión de la onda acústica con una pared del recinto.
δ	delta de Dirac.
δ	parámetro de control de ganancia del filtro ANS, dependiente de NMR.
Δf	intervalo de Schroeder.
Δt	intervalo de actualización de trama para análisis por ventanas temporales.
Δz	periodo de muestreo espacial (separación intermicrofónica) para un array lineal uniforme.
$\Delta z_{B1}, \Delta z_{B2}, \Delta z_{B3}$	periodos de muestreo espacial para las subbandas B_1 , B_2 y B_3 , del array anidado prototípico.
$\Delta\theta$	distancia angular entre los dos nulos de captación más próximos a la dirección de apuntamiento.
ϵ	factor de forma para el filtrado auditivo ANS.
ϕ	fase.
Φ_{NN}	igual que Φ_{XX} pero calculada sólo con la perturbación acústica o ruido genérico.
Φ_{XX}	autoespectro de X.
Φ_{XX}	matriz de espectros cruzados medida entre las señales de salida de los micrófonos del array, cuando no existe perturbación acústica.
Φ_{YX}	vector de espectros cruzados entre el vector Y y la señal X.
Φ_{YY}	igual que Φ_{XX} pero considerándose incluida la perturbación acústica o ruido genérico.
γ	factor de forma para la sustracción espectral.
Γ_{NN}	matriz coherencia de ruido.
Γ'_{NN}	matriz coherencia de ruido calculada después de la alineación temporal de los canales.
η	factor de sobresupresión para el procesador ANS propuesto.
λ	longitud de onda del sonido en el aire.
λ	parámetro de actualización del estimador recursivo.
λ_a	parámetro de actualización (ataque) del estimador recursivo.

λ_c	parámetro de actualización (caída) del estimador recursivo.
λ_n	autovalores del espacio vectorial del ruido.
λ_x	autovalores del subespacio vectorial de la señal.
λ_y	autovalores del espacio vectorial generado por el array.
μ	constante de restricción para el array superdirectivo.
θ_{nulo}	ángulo de llegada para el primer nulo de captación del array a partir del lóbulo principal.
σ^2	potencia de ruido captado.
τ	retardo.
ω	pulsación.
Ω	pulsación o frecuencia angular.
Ω_{nulo}	pulsación angular para el primer nulo de captación del array a partir del lóbulo principal.
Ω_s	pulsación de muestreo angular.
Ψ	función de filtrado o de peso para la GCC.

ÍNDICE

<i>RESUMEN</i>	VII
<i>ABSTRACT</i>	IX
<i>RELACIÓN DE ACRÓNIMOS Y SÍMBOLOS UTILIZADOS</i>	XI
<i>ÍNDICE</i>	XXI
<i>PREÁMBULO</i>	1
1 INTRODUCCIÓN	3
1.1 El array de micrófonos como dispositivo de mejora de voz.....	4
1.2 Objetivos de la Tesis.....	6
1.3 Precedentes y marco de trabajo	8
1.4 Estructura de la Tesis.....	9
<i>PARTE 1: MEJORA DE HABLA CON ARRAYS DE MICRÓFONOS</i>	13
2 CONFORMACIÓN DE HAZ (<i>BEAMFORMING</i>)	15
2.1 Array de micrófonos	15
2.1.1 Respuesta del array ante una fuente única.....	15
2.1.2 Respuesta del array ante varias fuentes simultáneas	18
2.1.3 Efecto de la reverberación en la respuesta del array	19
2.1.4 Efecto del ruido aditivo	22
2.1.5 Momentos del array	23
2.1.6 Conformador de haz convencional. Estructura de filtrado y suma.....	24
2.1.7 Algunos tópicos sobre directividad	28
2.1.8 Array lineal.....	29
2.2 Array de banda ancha	38
2.2.1 Array de directividad constante (CDB o <i>Constant Directivity Beamformer</i>)	39
2.2.2 Array anidado	42
2.2.3 Array superdirectivo	45
2.3 Conformación adaptativa.....	54
3 LOCALIZACIÓN DE FUENTE	59
3.1 Maximización de la potencia de salida.....	61
3.2 Localización basada en estimación de retardos (TDOA)	63
3.2.1 TDOA (<i>Time Difference Of Arrival</i>).....	64
3.2.2 Método de la Correlación Cruzada Generalizada (GCC o <i>Generalized Cross-Correlation</i>)	65
3.3 Localización de alta resolución basada en estimación espectral	67
3.3.1 Métodos de banda estrecha.....	67
3.3.2 Métodos paramétricos	72
3.3.3 Métodos de banda ancha	73

4 MEJORA DE HABLA MEDIANTE POSTFILTRADO	75
4.1 Reducción de ruido mediante postfiltrado	76
4.1.1 Filtrado de Wiener multicanal	76
4.1.2 Sustracción espectral	83
4.1.3 Filtrado perceptual. Método ANS	89
4.1.4 Técnicas de estimación de potencia de ruido y de señal de habla	97
4.1.5 Otros esquemas de reducción de ruido	102
4.2 Derreverberación	102
4.2.1 Derreverberación mediante descomposición en componentes fase mínima-paso todo	107
4.2.2 Otros métodos de derreverberación mediante postfiltrado	111
5 EVALUACIÓN OBJETIVA DE CALIDAD DE LA SEÑAL DE HABLA	113
5.1 Medidas de la relación señal a ruido. SNR, SNRA y NMR	115
5.2 Medidas basadas en predicción lineal (LP)	120
5.3 Índice de articulación AI	123
5.4 Índices de inteligibilidad STI y RASTI	124
5.4.1 Índice de modulación del habla, m	125
5.4.2 Cálculo del STI	127
5.4.3 El índice RASTI	128
<i>PARTE 2: PROPUESTAS, EXPERIMENTOS Y RESULTADOS PRELIMINARES</i>	131
6 ELEMENTOS PRELIMINARES DE TRABAJO	133
6.1 Array anidado de 15 canales: características	134
6.2 Base de datos real CMU	142
6.3 Base de datos simulada: simCMU-1 y simCMU-2	144
7 PROPUESTAS DE MEJORA DE SEÑAL DE VOZ EN PRESENCIA DE RUIDO Y REVERBERACIÓN	149
7.1 Procesador en array basado en filtrado de Wiener multicanal modificado por coherencia más derreverberador ciego basado en descomposición fase mínima-paso todo (MW-MPAP)	150
7.1.1 Descripción del sistema	150
7.1.2 Estimación recursiva de los espectros de señal y de ruido	155
7.1.3 Enventanado y reconstrucción de la señal temporal	156
7.1.4 Experimentos y resultados	160
7.2 Estimación objetiva de mejora de señal de habla en presencia de ruido y reverberación mediante el método E-RASTI (RASTI Emulado)	171
7.2.1 Descripción del método E-RASTI propuesto	171
7.2.2 Experimentos de validación del método E-RASTI	174
7.3 Procesador en array basado en supresión de ruido audible combinada con un filtro de Wiener modificado por coherencia (procesador ANS-MW)	180
7.3.1 Descripción del procesador ANS-MW	180
7.3.2 Experimentos y resultados	183
7.4 Recapitulación sobre los resultados obtenidos en las pruebas preliminares con las aportaciones propuestas	190
<i>PARTE 3: IMPLEMENTACIÓN SOBRE DSP DE UN PROTOTIPO EN TIEMPO REAL</i> ...	193
8 DESCRIPCIÓN DEL PROTOTIPO EN TIEMPO REAL IMPLEMENTADO	195
8.1 Elementos <i>hardware</i>	195
8.2 Medidas electroacústicas de los micrófonos del array	204

8.3 Elementos <i>software</i> del sistema basado en DSP PCI/C6600 + PMC/16I02	211
8.3.1 Programa residente en DSP	212
8.3.2 Software de comunicación y pruebas	217
9 PROTOTIPO DE ARRAY SUPERDIRECTIVO PERCEPTUAL (SD-ANS-MW) EN TIEMPO REAL	221
9.1 Descripción del conformador superdirectivo en la banda de baja frecuencia	221
9.2 Resultados electroacústicos del conformador de haz superdirectivo	230
9.2.1 Calibración	231
9.2.2 Directividad en cámara anechoica	233
9.3 Resultados de mejora de voz con el procesador SD-ANS-MW sobre una base de datos real	244
9.3.1 Base de datos multicanal UPM	244
9.3.2 Experimentos y resultados	247
EPÍLOGO	261
10 CONCLUSIONES Y LÍNEAS FUTURAS DE TRABAJO	263
10.1 Conclusiones	263
10.1.1 Propuestas de mejora de señal de voz mediante postfiltrado multicanal	263
10.1.2 Evaluación objetiva de los resultados a la salida del procesador	265
10.1.3 Localización de fuente	266
10.1.4 Directividad	266
10.1.5 Implementación en DSP de un prototipo de array microfónico	267
10.2 Líneas futuras de trabajo	268
BIBLIOGRAFÍA	271

PREÁMBULO

1 INTRODUCCIÓN

El término *Array* procede del inglés y se puede traducir como colección o conjunto ordenado de elementos. En el mundo de las comunicaciones, un array es una agrupación de emisores o receptores de señal, dispuestos en una determinada configuración geométrica, con el objetivo de mejorar, en el sentido más amplio, la comunicación entre dichos emisores y receptores. Es decir, un array ha de producir una mejora de la comunicación que se establecería en un sistema con un solo elemento emisor y/o receptor.

La propiedad básica que confiere a un array sus características beneficiosas es la directividad. Un receptor de señal es directivo cuando su capacidad de transducción es diferente según sea la posición espacial de la fuente de señal que está recibiendo. Normalmente, los receptores directivos producen una captación máxima en determinada dirección espacial (eje principal o de máxima captación), rechazando en mayor o menor medida la señal procedente de las demás direcciones espaciales. Por eso, un array de receptores, al ser directivo, mejorará la señal procedente del eje principal y atenuará las señales laterales, que normalmente se consideran fuentes de perturbación de la señal que se pretende captar. Entonces el término de directividad en un receptor es equivalente al concepto de “selectividad espacial”. Cuanto más directivo es un receptor de señal, más capacidad tendrá de discriminar la dirección principal de recepción con respecto a las demás no deseadas. La directividad está relacionada con la naturaleza ondulatoria de las señales que se transmiten, fundamentalmente de tipo radioeléctrico o acústico. De hecho, la directividad depende fuertemente de la relación entre la longitud de onda λ de la perturbación ondulatoria y el tamaño del receptor, de tal manera que cuanto mayor sea un receptor con relación a λ mayor será su directividad y por tanto su capacidad de rechazo de señales no deseadas. En este sentido, interesarán receptores de gran tamaño. Como esto puede ser complicado desde un punto de vista tecnológico, se diseñan y construyen arrays, ya que en la práctica un array constituye un receptor de gran tamaño compuesto por elementos más pequeños, y por tanto el conjunto es más directivo que cada uno de los elementos del array. Se desprende fácilmente que la directividad de un array depende fuertemente de dos características. Por una parte del tamaño del array y más concretamente de la distribución espacial de los integrantes del array y por otra parte de la frecuencia, de tal manera que un array será más directivo cuanto mayor sea su tamaño y menor sea la longitud de onda de la señal que recibe o lo que es lo mismo mayor sea la frecuencia.

Sin embargo, no es sólo la directividad lo que se aprovecha para sacar beneficio de la respuesta de un array. La captación de la fuente por múltiples receptores puede ser aprovechada para un mayor conocimiento del proceso de transmisión desde dicha fuente hasta el punto de recepción, y esto se puede usar para corregir el comportamiento no deseado del canal de transmisión y para atenuar con mayor perfección las perturbaciones ajenas a la señal original, que se pretende extraer más o menos libre de interferencias.

Son todas las características apuntadas las que hacen a la forma de captación en array especialmente adecuada para acometer la mejora de señal de voz en condiciones acústicas adversas, y por lo que en esta Tesis se proponen los arrays microfónicos para el fin señalado.

1.1 EL ARRAY DE MICRÓFONOS COMO DISPOSITIVO DE MEJORA DE VOZ

Los arrays tanto de emisores como de receptores se han empleado con gran profusión durante el siglo XX en comunicaciones radioeléctricas, especialmente en aplicaciones de radar. También en transmisiones acústicas formando parte de equipos sonar. En los últimos 20 años, el procesado en array está teniendo un gran auge en todos los campos de las comunicaciones, y especialmente en el tratamiento de señales acústicas. Este auge es consecuencia de la utilización masiva de ordenadores, cada vez más potentes y de más bajo precio, lo que hace posible alcanzar la gran capacidad de procesado necesaria para abordar el tratamiento digital de la señal multicanal asociada a un array. De esa manera, dentro del procesado de señal se puede considerar toda una rama de conocimiento basada en el empleo de arrays, pudiéndose hablar específicamente de procesado en array [Haykin 85] [Johnson 89] [Johnson 93] [Krim 96] [Chen 98] [Naidu 01]. Se acepta que en el momento actual el procesado en array es un campo de conocimiento maduro y completamente establecido.

En este trabajo de Tesis se estudian las bases teóricas, el uso y las aplicaciones de los arrays enfocados a la captación y mejora de señal acústica, especialmente de habla, por medio de transductores receptores de sonido, es decir micrófonos. Además, se ofrecen ideas originales de procesadores de voz basados en arrays de micrófonos.

En los últimos 20 años, se han venido proponiendo e implementando arrays tanto de emisores (altavoces) como de receptores (micrófonos) para tratar señales de habla. Durante los años 80 y 90 se desarrollaron de forma importante los principios básicos que sustentan los arrays microfónicos. Por ejemplo, en [Flanagan 85-a] [Flanagan 85-b] [Flanagan 85-c] y [Grenier 97-b] se ofrecen soluciones particulares sobre el apuntamiento de arrays de micrófonos en diferentes condiciones acústicas, y en [Hussain 97] y [Nordholm 99] sobre arrays microfónicos adaptativos. También ha tenido un gran desarrollo el empleo de arrays microfónicos como apoyo, mejora y solución particular a diferentes problemáticas relacionadas con la señal de habla, y en especial al reconocimiento de habla en [Van Compernolle 90-b] [Sullivan 93] [Giuliani 95] [Giuliani 96] [Sullivan 96] [Omologo 97-a] y [Fernández 99] y al reconocimiento de locutor [González-Rodríguez 97]. Si las bases del procesado en array están consolidadas, como se dijo anteriormente, la aplicación concreta a la captura de señales acústicas mediante arrays de micrófonos comienza a ser un capítulo de obligada presencia cuando se habla de aplicaciones acústicas [Silverman 87] [Gay 00], de tal manera que ya existen tratados específicos sobre arrays microfónicos [Brandstein 01]. Por tanto, se está en ese punto donde existen aplicaciones desarrolladas e implementadas por diferentes grupos de investigación en el mundo, pero que todavía no tienen reflejo amplio en el mercado, aunque ya se pueden encontrar productos comerciales de bajo coste que utilizan arrays microfónicos [Acousticmagic] [Andrea] [VocaLinks].

La señal de habla y el camino de transmisión acústica convencional tienen características propias que hacen que no sirvan plenamente los arrays propuestos para señales radioeléctricas. ¿Qué es lo que caracteriza a los arrays microfónicos que impide el aprovechamiento directo de los resultados obtenidos en las aplicaciones tradicionales de radar y sonar? En primer lugar la señal de habla tiene un ancho de banda relativo muy amplio. Se llama ancho de banda relativo a la relación entre la frecuencia central del espectro, en este caso de la voz humana, y el ancho de banda absoluto. Un ancho de banda relativo grande hace

que la directividad del array varíe mucho en su margen de frecuencias de utilización y esto incide negativamente en el provecho que se puede sacar de dicha directividad para mejorar la señal de voz. A esto se une el hecho de que las longitudes de onda típicas de las señales audibles son relativamente grandes (del orden de varios metros en baja frecuencia), lo que implicaría diseñar arrays de micrófonos de tamaño excesivo para conseguir una directividad suficiente, o al menos tan alta como se consigue en las aplicaciones radioeléctricas usuales. Otra característica que perjudica el funcionamiento de un array microfónico es la naturaleza del camino de transmisión acústica. En un escenario típico de captación de voz, prima la transmisión multirayecto. Es decir la señal acústica viaja desde la fuente sonora hasta el receptor por múltiples caminos. Estos caminos están constituidos por el llamado camino directo, que es el que une a la fuente con el micrófono, pero también por una multitud de caminos indirectos, que incluye las reflexiones de la señal acústica con las diferentes superficies del recinto en el que se desarrolla la comunicación, además de las posibles trayectorias no rectas debidas a la difracción con algún obstáculo pequeño. La naturaleza multirayecto de la transmisión acústica es conocida con el término genérico “reverberación” y constituye un serio inconveniente en el comportamiento de los arrays microfónicos. La reverberación disminuye drásticamente la calidad de la señal de voz, reduce la inteligibilidad del habla y es muy difícil luchar contra ella. Por todo ello, dificulta el funcionamiento adecuado de los esquemas de tratamiento en array tradicionales, sobre todo aquellos que utilizan un filtrado adaptativo para eliminar ruido, y que tan bien se comportan cuando la vía de comunicación entre la fuente y el array contiene sólo el camino directo y pocas reflexiones. Otra peculiaridad negativa relacionada con la vía de transmisión acústica, es que la distancia de la fuente al array puede variar mucho de forma relativa, y además suele ser comparable al tamaño del array, con lo que es necesario prever un amplio abanico de posibilidades a la hora de tratar la señal recibida. Esta dificultad geométrica no suele presentarse en las aplicaciones habituales de radar, usando la comunicación radioeléctrica, en las que la fuente se sitúa siempre a mucha distancia del receptor, de tal forma que el único parámetro de apuntamiento a tener en cuenta, a la hora de procesar la señal incidente, es el ángulo de llegada de la señal recibida, y la distancia no importa.

Las posibles aplicaciones de los arrays microfónicos son muchas. Quizás la más clara sea la mejora genérica de señal de habla, necesaria para múltiples propósitos. Usando las virtudes inherentes a la captación de sonido multicanal, en cuanto a mejora de señal y localización de fuente, surgen aplicaciones específicas que optimizan su funcionamiento con los arrays como herramienta de adquisición, por ejemplo los arrays de micrófonos para ayuda auditiva, para el reconocimiento de locutor, el reconocimiento de habla, la separación ciega de señales acústicas, etc.

Se trata de implementar dispositivos que perfeccionen la comunicación acústica en diversos ambientes, que sean capaces de proporcionar una señal con suficiente calidad y que funcionen con robustez frente al ruido de fondo, las señales interferentes y la reverberación. En este caso, la tecnología de arrays es apropiada en sistemas de conferencia, de telefonía manos libres, en sistemas con interacción hombre-máquina, o en general en cualquier entorno de captación de voz donde se prevean unas condiciones ambientales especialmente perjudiciales, como puede ser un automóvil o una fábrica. Otra aplicación, que constituye de por sí todo un campo de estudio, es la localización de forma automática de la fuente sonora. De hecho, muchas aplicaciones de mejora de habla con arrays, necesitan conocer de forma precisa la situación del hablante cuya señal se pretende mejorar. En este sentido, se pueden utilizar los arrays para averiguar el número y posición de posibles hablantes en una sala y así tener la posibilidad de seleccionar la fuente deseada atenuando las demás. Los arrays de micrófonos también son útiles en sistemas multimedia, donde la información espacial

proporcionada puede servir para apuntar de forma automática una cámara robotizada. Relacionada con esta última facultad, está la capacidad de utilizar un array para generar audio 3D [Bartlett 90] [Billingsley 90] [Klepko 97] [Williams 99], es decir aprovechar la información espacial y de localización de fuente proporcionada, para obtener una señal multicanal que suministre una sensación sonora tridimensional, de tal manera que pueda usarse en grabaciones sonoras de calidad, por ejemplo en producciones cinematográficas o en interpretaciones musicales.

En definitiva, existe un amplio abanico de posibilidades a la hora de dar uso a un array microfónico, aunque casi todas ellas se pueden englobar en cualquiera de los siguientes apartados, localización y seguimiento automáticos de fuente y mejora de señal de voz captada en condiciones acústicas adversas. El trabajo de Tesis que aquí se presenta se centra en esta última área temática.

1.2 OBJETIVOS DE LA TESIS

Según el diccionario de la Real Academia Española una tesis es una conclusión, una proposición que se mantiene con razonamientos. En este sentido, una Tesis Doctoral debe basarse en proposiciones razonadas de las que se desprendan una serie de conclusiones que aporten a la comunidad científica o a la tecnología de las aplicaciones una mejora del conocimiento sobre una determinada materia o, en el campo de la ingeniería, que permitan aplicar una determinada tecnología de manera más eficiente que como se estaba haciendo hasta el momento. De entre las múltiples posibilidades que ofrece el procesado de señal usando arrays de micrófonos, esta Tesis se centra en la aplicación de los mismos a la mejora de señal de habla (*Speech Enhancement*). Es decir, se tratará de mejorar el habla captado procedente de un locutor que se encuentre en condiciones acústicas adversas, especialmente de elevado ruido y reverberación.

Por otra parte se quiere implementar un prototipo que utilice una o varias de las propuestas y soluciones aportadas relativas a la mejora de habla. Eso dará pie a afrontar la problemática del procesado de una señal multicanal de forma real, y no sólo desde el punto de vista de las simulaciones por ordenador (o simulaciones *software*), que por otra parte es una práctica muy frecuente en el ámbito del procesado de señales acústicas. La implementación real de un prototipo hace que los algoritmos de procesado tengan que ser optimizados en cuanto a velocidad y a sencillez. A veces un incremento ligero de las prestaciones de un algoritmo supone un aumento exponencial de su complejidad y consecuentemente un funcionamiento más lento, características que lo pueden hacer inviable para su uso en el mundo real. Además, sería deseable que la implementación práctica de un prototipo se hiciese con tecnología convencional, que esté presente en el mercado actual y con un coste relativamente asequible.

Una característica primordial que debe tener un sistema multicanal de adquisición y procesado de señal de habla, es que funcione en tiempo real. La operación en tiempo real permite la escucha directa de la señal de audio procesada o el almacenamiento de la misma en formatos analógicos o digitales convencionales al mismo tiempo que la información analógica multicanal está entrando en el procesador. Si el funcionamiento no fuese en tiempo real, en cualquier aplicación se necesitaría una gran capacidad de almacenamiento de la señal multicanal preprocesada, normalmente en disco duro, en formato digital, para posteriormente ser procesada. Pero esta capacidad de almacenamiento, que aumenta linealmente con el número de canales del array, no está al alcance de la tecnología convencional con que se

pretende implementar un prototipo. Por tanto, el objetivo de funcionamiento en tiempo real es aquí primordial.

Agrupando el planteamiento general anterior en una serie de propósitos concretos, se proponen los siguientes objetivos para ser cubiertos con esta Tesis Doctoral:

1.- Revisión del estado del arte sobre el procesado en array usando micrófonos. En este punto se hará una especial incidencia en las técnicas de conformación de haz específicas empleando micrófonos. También se revisará su aplicación a la mejora de voz en condiciones de ruido y reverberación, y, aunque no será un objetivo primordial aquí, se examinarán las técnicas de localización de fuente mediante arrays de micrófonos más ampliamente propuestas en la literatura científica.

2.- Revisión de los métodos más utilizados para evaluar la calidad de habla de forma objetiva. Éste es un punto importante ya que la valoración del procesado de la señal de habla es difícil, sobre todo cuando la fuente de perturbación es la reverberación, y aunque de forma subjetiva es fácil decir si existe, por ejemplo, elevada reverberación en una señal, no es tan sencillo hacerlo mediante parámetros objetivos.

3.- Propuestas y aportaciones sobre mejora de voz utilizando arrays microfónicos. Se hará especial hincapié en la utilización de métodos de postfiltrado basados en el análisis y procesado de las características espacio temporales de la señal multicanal incidente al array. En concreto, en el filtrado de Wiener multicanal, en los métodos basados en coherencia intercanal y en el filtrado perceptual que utiliza las características del sistema auditivo humano. También se trabajará en la derreverberación ciega mediante descomposición cepstral.

4.- Se investigará la posibilidad de utilizar el índice RASTI de inteligibilidad como evaluador objetivo de calidad de la señal de habla, con el propósito de probar el funcionamiento de los algoritmos propuestos. Con este índice se intentará evaluar especialmente la capacidad de eliminar reverberación por parte de un procesador en array.

5.- Implementación práctica de un array microfónico en tiempo real, basado en DSP (*Digital Signal Processor*) y en ordenador personal de tipo PC. Este prototipo integrará una o varias de las soluciones propuestas, y deberá tener la posibilidad de trabajar en tiempo real y de grabar en formato digital la señal de audio multicanal implicada en el proceso, para el análisis posterior de los resultados. De deberá corroborar de la forma más amplia y precisa posible que el prototipo implementado funciona de acuerdo a la teoría.

6.- Difusión de los resultados. Uno de los objetivos primordiales de una Tesis Doctoral, que aquí también se persigue, es que sus conclusiones y resultados se difundan en los medios adecuados (congresos, revistas...) para su evaluación y refrendo por la comunidad científica internacional.

Con todo ello, los objetivos anteriores se resumen en uno principal: la propuesta de métodos de procesado en array usando micrófonos que en algún sentido sean novedosos y su implementación práctica en un prototipo que pueda funcionar de manera autónoma y que sirva para investigaciones futuras dentro de la misma línea de trabajo.

1.3 PRECEDENTES Y MARCO DE TRABAJO

El trabajo de investigación realizado en esta Tesis se ha desarrollado en el Área de Tratamiento de Voz y Señales (ATVS, www.atvs.diac.upm.es), dentro del Departamento de Ingeniería Audiovisual y Comunicaciones (DIAC), perteneciente a la Universidad Politécnica de Madrid (UPM). En el grupo de investigación ATVS se desarrolla toda una línea temática dedicada a la señal de voz a la que pertenece el procesado de señales de habla mediante arrays microfónicos y dentro de la cual se ha realizado el trabajo aquí presentado.

El ATVS surgió en 1994 como grupo de investigación y su campo de trabajo estaba especializado inicialmente en el tratamiento de señal de voz, y más específicamente en el área de identificación de locutores y acústica forense. Más adelante, el ATVS amplió su labor hacia la biometría, y en la actualidad trabaja además en el reconocimiento automático de firma escrita, huella dactilar y cara. El área de investigación de arrays microfónicos es parte de la línea de procesado de voz seguida por el grupo, y surgió como apoyo de ciertas aplicaciones forenses cuyo desarrollo ya se había iniciado dentro del ATVS.

El ATVS mantiene proyectos públicos de ámbito europeo (V y VI Programa Marco) y nacional (Ministerio de Ciencia y Tecnología) y diversos contratos con empresas líderes en el sector de la biometría, procesado de voz y seguridad. Además, cuenta con publicaciones en revistas y congresos de carácter internacional y participa activamente en foros nacionales e internacionales relativos a esos ámbitos.

Tiene especial interés para esta Tesis, el hecho de que el tema de procesado en array ha estado amparado por un proyecto de investigación del Ministerio de Ciencia y Tecnología que lleva por título “*Desarrollo e implementación de algoritmos de procesado de señal y computación de altas prestaciones para sistemas de emulación de entornos acústicos virtuales*” y que finalizó en Diciembre de 2003. Se trata de un proyecto coordinado liderado por la Universidad Politécnica de Valencia (UPV) en el que el grupo ATVS realiza el subproyecto denominado “*Sistemas de adquisición de señal y posición para audio 3D (AUD3D)*”. En este subproyecto se estudia la posibilidad de utilizar un array microfónico como parte integrante de un sistema multimedia de adquisición de señal de audio y vídeo, de tal manera que junto con otro tipo de sensores de posición, fundamentalmente de imagen, pueda hacerse un seguimiento automático de un locutor considerado como principal, a la vez que se mejore la señal de audio captada utilizando la interacción entre los parámetros de apuntamiento del array y del sensor de vídeo.

Adicionalmente el *Instituto de Estudios de Seguridad “Duque de Ahumada”*, adscrito a la Universidad Carlos III de Madrid, ha otorgado al grupo ATVS el Proyecto de Investigación en Equipo que lleva por título “*Sistema de captación y mejora de voz en condiciones acústicas adversas mediante arrays de micrófonos*”.

Recientemente, el Ministerio de Ciencia y Tecnología, dentro del programa MCYT’03, ha concedido al ATVS la dirección del proyecto titulado “Identificación de locutor combinando rasgos físico-acústicos, prosódicos y lingüísticos aplicada a la acústica forense” (TIC2003-09068-C02), en el que se incluye una línea de trabajo sobre “Arrays microfónicos aplicados a la acústica forense”. El desarrollo futuro de este proyecto ayudará a continuar la línea de investigación comenzada con esta Tesis.

1.4 ESTRUCTURA DE LA TESIS

Excluyendo el capítulo 1 de introducción, y la parte final a modo de epílogo, dedicada a las conclusiones finales, esta Tesis se divide en tres partes, Parte 1, Parte 2 y Parte 3. La estructura de contenidos expuestos está de acuerdo con la cadencia temporal del trabajo llevado a cabo en la misma a lo largo de aproximadamente cuatro años. Los trabajos se iniciaron a finales del año 1999 y surgieron como continuación de las investigaciones desarrolladas en la Tesis doctoral “Influencia y compensación del entorno acústico en sistemas de reconocimiento automático de locutores” [González-Rodríguez 99-a]. En esa Tesis se describe la mejora de voz utilizando arrays microfónicos y se hacen una serie de propuestas exitosas con aplicación específica al reconocimiento automático de locutores. A continuación se expone resumidamente la estructura de la Tesis que aquí se presenta.

En la “Parte 1: Mejora de habla con arrays de micrófonos” se describe el estado del arte sobre la utilización del procesado multicanal de señales acústicas aplicado a la mejora de señal de voz en condiciones de ruido y reverberación. Esta parte se subdivide en cuatro capítulos, del 2 al 5.

El capítulo 2 está dedicado a los tópicos más destacables sobre conformación de haz (*beamforming*) que puedan ser aplicados directamente a los arrays microfónicos. En él se desarrollan la nomenclatura y formulación básica que serán utilizadas de forma recurrente a lo largo de toda la Tesis. Se tratan los métodos más frecuentes de conformación de haz y su relación con la reducción de ruido y reverberación en el habla. En concreto, merecen especial atención en el punto 2.2 sobre arrays de banda ancha, los epígrafes dedicados al array anidado y al array superdirective, que tendrán especial relevancia para las ideas expuestas en la Tesis.

El capítulo 3 está dedicado a la localización de fuente mediante arrays de micrófonos. En él se describen los métodos normalmente utilizados para conocer dónde está la fuente de sonido a partir de la información multicanal recogida por un array de micrófonos. Son descritos y analizados desde los métodos más tradicionales de maximización de la potencia de salida y estimación de retardos por correlación cruzada, hasta los métodos que gozan de más atención en la actualidad, entre los que se encuentra el de localización basada en descomposición en subespacios vectoriales. Aunque a la postre el tema de localización de fuente no ha sido directamente tratado en los experimentos desarrollados por el autor con posterioridad, constituye una línea de trabajo futuro de gran interés y especialmente adaptada para ser aplicada a los arrays microfónicos.

El capítulo 4 está dedicado a las técnicas de mejora mediante postfiltrado. En él se describen todas las variantes de interés del filtrado de Wiener y la sustracción espectral, en especial la técnica de supresión auditiva de ruido (método ANS) que será utilizada profusamente en los trabajos y experimentos desarrollados en la Tesis. También se hace una especial incidencia, por la importancia práctica que tienen, en las técnicas de estimación en tiempo real de los espectros de potencia de voz y de ruido. Por otra parte, también en el capítulo 4, se describe la problemática de la derreverberación ciega, en especial la técnica de la derreverberación por descomposición cepstral, que también será utilizada en las propuestas del autor.

En el capítulo 5 se expone la formulación y soporte teórico de las técnicas más comúnmente utilizadas para la estimación objetiva de la calidad de la señal de habla. Téngase en cuenta, que uno de los principales escollos a los que se enfrentan las técnicas de mejora de habla es la validación de los resultados obtenidos. Aunque la apreciación subjetiva mediante índices formales de calidad o escuchas informales debe estar siempre presentes en la valoración de un procesador de habla, en esta Tesis se ha hecho una especial incidencia en los

evaluadores de tipo objetivo. En concreto se tratan dos grandes bloques de estimadores objetivos de mejora de habla, los basados en la mejora de la relación señal a ruido y los que utilizan el índice de inteligibilidad STI (*Speech Transmission Index*), que actualmente es un estándar en la valoración acústica de auditorios y sistemas electroacústicos.

A parte de lo que es el lógico trabajo de revisión y puesta en antecedentes sobre el estado del arte referente a los arrays microfónicos (Parte 1 de la Tesis), existen dos grandes períodos que pueden considerarse claves en el trabajo desarrollado a lo largo de esos cuatro años. En primer lugar el intervalo que abarca desde finales de 1999 hasta finales de 2001 en el que se plantean y difunden las propuestas preliminares sobre mejora de señal de voz usando un array microfónico, que conducirán finalmente a la implementación de un prototipo. Estos trabajos iniciales están descritos en la “Parte 2: Propuestas, experimentos y resultados preliminares”, y han sido difundidos en las publicaciones [González-Rodríguez 99-b] [González-Rodríguez 00] [Sánchez-Bote 00] [Sánchez-Bote 01-a] y [Sánchez-Bote 01-b]. El segundo de estos períodos está relacionado con la implementación de un prototipo de procesador digital multicanal en tiempo real, que pone en práctica las propuestas y experimentos llevados a cabo por el autor. Abarca desde finales de 2001 hasta mediados de 2003, cuando se dieron por finalizados los experimentos detallados en esta Tesis. La implementación del prototipo está descrita en la “Parte 3: implementación en DSP de un prototipo en tiempo real”. En esta parte se han vertido muchos esfuerzos, necesarios para superar los lógicos inconvenientes que supone la implementación de un prototipo de estas características, desde la selección y puesta en funcionamiento del material *hardware* y *software* preciso, hasta los detalles e inconvenientes relacionados con la implementación en tiempo real de dicho prototipo, bastante exigente desde el punto de vista computacional, como se verá. Finalmente, estos trabajos han concluido con éxito y se ha publicado un resumen de los mismos en [Sánchez-Bote 03-a].

La Parte 2 sobre experimentos preliminares contiene dos capítulos.

En el capítulo 6 se describen los elementos preliminares de trabajo, es decir el comportamiento teórico del array microfónico en la configuración que va a ser usada en las pruebas iniciales y también son descritas las bases de datos de señal de voz multicanal que van a ser evaluadas en dichos experimentos iniciales.

En el capítulo 7 se describen las propuestas y aportaciones originadas en este trabajo de Tesis, así como las pruebas y experimentos realizados sobre las bases de datos descritas en el capítulo 6. Básicamente, las propuestas realizadas en el capítulo 7 están constituidas por dos procesadores multicanal combinados (procesador MW-MPAP y procesador ANS-MW) y un método de evaluación objetiva de la calidad de habla (índice E-RASTI), basado en el cálculo de los índices de inteligibilidad que se usan actualmente en la evaluación de sistemas electroacústicos.

A continuación, la “Parte 3: Implementación en DSP de un prototipo en tiempo real” sobre el diseño del prototipo de array, contiene a su vez otros dos capítulos.

El capítulo 8 detalla las características de los elementos *hardware* utilizados y el *software* ideado por el autor, que componen el prototipo final de array microfónico implementado.

En el capítulo 9 se describe la propuesta final de mejora de señal de voz usando arrays microfónicos que ha sido especialmente diseñada para el prototipo en tiempo real. Aquí se consideran las limitaciones lógicas que impone la implementación en tiempo real en un procesador digital de señales (DSP), con capacidad limitada de computación. La propuesta

final de procesador en array contiene las ideas planteadas en las pruebas preliminares del capítulo 7, añadiendo como novedad la utilización de un conformador superdirective en la banda de baja frecuencia. Tienen gran interés en este capítulo las pruebas electroacústicas de directividad que, a la vista de los resultados obtenidos, corroboran la gran proximidad existente entre las predicciones teóricas destacadas por el autor y los resultados de dichas medidas. En el capítulo 9 además se describen las pruebas y experimentos con el prototipo de array sobre una base de datos en español, generada por el autor con el empleo del procesador implementado. Una versión resumida de los trabajos expuestos en la Parte 3 ha sido enviada en Julio de 2003 para su revisión a *IEEE Trans. on Acoust. Speech and Signal Processing* [Sánchez-Bote 03-b].

Por último el epílogo, con el capítulo 10 que contiene las aportaciones y conclusiones más destacables obtenidas del trabajo de investigación y las líneas de trabajo futuro propuestas. Dentro del epílogo, y para finalizar la Tesis, se relacionan las referencias bibliográficas usadas a lo largo del texto.

A la vista de los contenidos que son tratados en este trabajo de investigación queda patente el esfuerzo que se ha hecho en fijar las bases del procesado en array de señales acústicas, con especial énfasis en la implementación de un prototipo en tiempo real, de tal manera que la labor realizada pueda ser utilizada para mejoras y desarrollos futuros que complementen y continúen la línea de investigación seguida por esta Tesis.

PARTE 1

MEJORA DE HABLA CON ARRAYS DE MICRÓFONOS

2 CONFORMACIÓN DE HAZ (*BEAMFORMING*)

Beamforming es un término inglés que se puede traducir como conformación de haz. También es sinónimo de filtrado espacial o de directividad. 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 forma más sencilla de mejora de señal mediante un array [Van Been 88]. Si el eje de máxima captación de un array receptor (eje principal) se dirige hacia la fuente considerada como principal, se estará atenuando la señal procedente de otras fuentes no situadas en dicho eje principal, y que se consideran como fuentes de ruido. Si además el proceso se desarrolla en una sala, habrá reverberación, y el sonido procedente de las múltiples reflexiones con las paredes debido a la transmisión multirayecto procederá de todas direcciones. En ese sentido, el *beamforming* también reduce la reverberación, produciendo una señal de habla más seca y con más calidad. El principal objetivo del *beamforming* es por tanto producir una mejora inicial de ruido y reverberación en la señal de habla. El propósito ideal sería producir una señal de habla con calidad similar a la que se tendría si un único micrófono estuviese situado en el campo cercano del orador (de 2 a 10cm), pero cuando el array se sitúa en campo lejano (1 a 10m).

Se puede considerar que, para casi cualquier tipo de aplicación, una conformación de haz eficiente es el requerimiento mínimo exigible a un sistema de mejora de señal de habla mediante arrays microfónicos. Existe una bibliografía muy extensa sobre la conformación de haz, por ejemplo en [Griffiths 82] [Van Been 88] [Affes 96] se trata de forma genérica la conformación de haz, y en [Alvarado 90] [Flanagan 91] [Flanagan 93] [Fischer 96] [Flanagan 96] [Fischer 97] [Masgrau 99] [Brandstein 00] [Katkovnik 00] [Brandstein 01] y [Sachar 01] se afronta la conformación de haz con arrays microfónicos aplicada a señales acústicas.

Seguidamente se expone el planteamiento del problema de la conformación de haz con un array de micrófonos y se establece la nomenclatura con que serán tratadas más adelante las propuestas de la Tesis.

2.1 ARRAY DE MICRÓFONOS

2.1.1 Respuesta del array ante una fuente única

El modelo genérico que se va a considerar se muestra en la Figura 1. En éste, se representa una sola fuente emisora y un solo micrófono receptor. Este modelo se extenderá después a múltiples receptores. En lo sucesivo, los términos destacados en negrita se referirán a magnitudes vectoriales o matriciales (la diferencia entre una matriz y un vector deberá deducirse por el contexto). Cuando una magnitud definida en otro lugar como vectorial no vaya en negrita, se hará referencia a su módulo. En la Figura 1, el vector de posición de micrófono i es \mathbf{r}_i , \mathbf{r}_e es el vector de posición del emisor, relativo a dicho micrófono y \mathbf{r} es el vector de posición de la fuente. Se cumple además que $\mathbf{r}_i = \mathbf{r} - \mathbf{r}_e$. Se ha representado también

el sistema de referencia propio del micrófono (X' Y' Z'), necesario para poder definir adecuadamente las variables de posición de la fuente, relativas a dicho micrófono. Este sistema de referencia particular corresponde al sistema de referencia principal (X Y Z) desplazado en el vector de posición del micrófono, \mathbf{r}_i .

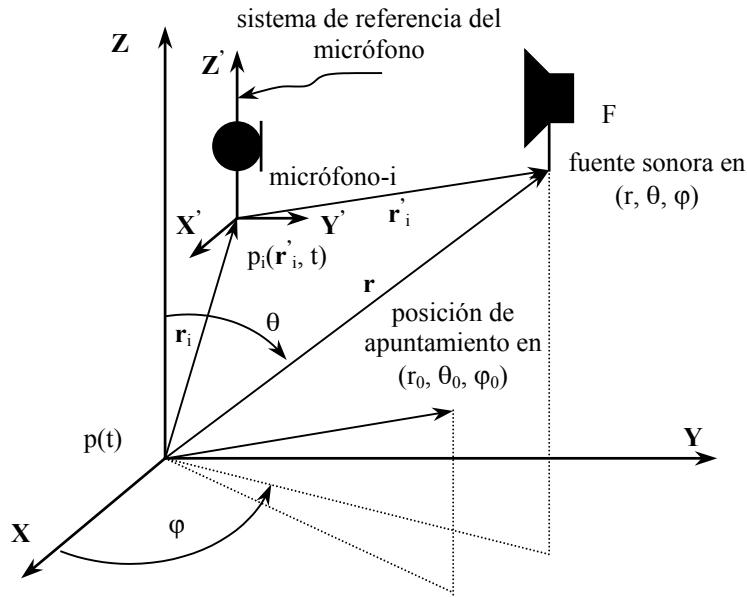


Figura 1. Sistema de referencia utilizado para definir un array de micrófonos.

Sea una fuente sonora omnidireccional F situada en el punto de coordenadas esféricas (r, θ, ϕ) . Dicha fuente está siendo captada por uno de los micrófonos integrantes del array, que se ha diseñado para apuntar a la posición (r_0, θ_0, ϕ_0) , y que se considera dentro del eje principal o de máxima captación del array. La posición de apuntamiento del array es conocida también como DOA principal (*Direction Of Arrival*) o simplemente DOA. La fuente produce una onda senoidal de presión acústica $p(t)$ en el centro de coordenadas, considerándose a esta señal libre de perturbaciones acústicas:

$$p(t) = p_0 \exp(j\omega t + \phi) \quad (1)$$

con p_0 la presión acústica (valor eficaz), ω la pulsación de la vibración acústica producida por la fuente y ϕ su fase inicial. Supóngase que sólo se considera el camino directo de transmisión acústica, que une a la fuente con el micrófono. La fuente, desde ese punto, creará una perturbación de naturaleza esférica en el micrófono i que, si la fuente es omnidireccional, estará relacionada con la presión en el centro coordinado por:

$$p_i(\mathbf{r}_i, t) = p(t) \frac{r}{r_i} \exp[-jk(r_i - r)] \quad (2)$$

con: $k = \omega/c$ (c es la velocidad del sonido) el número de onda de la transmisión acústica en el aire.

Cuando la onda de presión acústica llega al micrófono i , ésta es transducida, originando una señal de salida, $x_i(t)$ (que será normalmente una tensión eléctrica) mediante la siguiente ley:

$$x_i(t) = S_i D_i(\theta_i, \phi_i) p_i(r_i, t) = S_i D_i(\theta_i, \phi_i) p(t) \frac{r}{r_i} \exp[-jk(r_i - r)] \quad (3)$$

donde: S_i [V·Pa⁻¹] es la sensibilidad que tiene el micrófono i a la pulsación ω (se ha omitido expresar su dependencia con la frecuencia), $D_i(\theta_i, \phi_i)$ es la directividad del micrófono i , con θ_i, ϕ_i los ángulos coordenados esféricos (elevación y azimut) que marcan la posición de la fuente con respecto al micrófono i . Se llamará eje del micrófono a su dirección de máxima captación, normalmente $\theta_i = 0, \phi_i = 0$. Aquí, la característica de directividad del micrófono $D(\theta, \phi)$ expresa cómo varía su respuesta eléctrica al variar la posición (r, θ, ϕ) de la fuente, con relación a la respuesta que tendría si la fuente se situase en la posición de apuntamiento (r_0, θ_0, ϕ_0) :

$$D(\theta, \phi) = \frac{x(r, \theta, \phi)}{x(r_0, \theta_0, \phi_0)} \quad (4)$$

con $x(r, \theta, \phi)$ la respuesta del micrófono a una fuente en (r, θ, ϕ) y $x(r_0, \theta_0, \phi_0)$ la respuesta con la fuente situada en la posición de apuntamiento (Figura 1). En (4) se ha evitado considerar la dependencia de la directividad D con la distancia r , ya que en la práctica, esta dependencia sólo existe cuando dicha distancia es comparable con el tamaño del receptor.

Imagínese que uno de los micrófonos del array se considera de referencia. Este micrófono tiene índice cero (0) y está situado en el eje coordenado, es decir las coordenadas de la fuente F cumplen $(r, \theta, \phi) = (r_0, \theta_0, \phi_0)$. En (r_0, θ_0, ϕ_0) el subíndice 0 atiende por tanto al índice del micrófono de referencia, y no a la posición de apuntamiento (r_0, θ_0, ϕ_0) . El micrófono de referencia puede ser uno de los micrófonos del array, aunque no es necesario ya que se utiliza aquí como herramienta para las definiciones. Se le supone una sensibilidad S_0 y es omnidireccional $D_0(\theta, \phi) = 1$. Si ese micrófono recibiese la onda de presión $p(t)$ daría una respuesta eléctrica $x_0(t)$, es decir:

$$x_0(t) = S_0 p(t) \quad (5)$$

Entonces, despejando en (5), se puede relacionar la presión en el centro de coordenadas con la tensión entregada por el micrófono de referencia:

$$p(t) = \frac{x_0(t)}{S_0} \quad (6)$$

y consecuentemente la expresión (3) quedaría:

$$x_i(t) = \frac{S_i}{S_0} D_i(\theta_i, \phi_i) \frac{r}{r_i} \exp[-jk(r_i - r)] x_0(t) \quad (7)$$

que se puede escribir también como:

$$x_i(t) = a_i(r_i, \theta_i, \phi_i) x_0(t) \quad (8)$$

con,

$$a_i(r_i, \theta_i, \phi_i) = \frac{S_i}{S_0} D_i(\theta_i, \phi_i) \frac{r}{r_i} \exp[-jk(r_i - r)] \quad (9)$$

El factor $a_i(r_i, \theta_i, \phi_i)$ representa la respuesta particular del micrófono i ante una excitación de presión procedente de la posición r , y relativa a la respuesta del micrófono de referencia,

situado en el centro de coordenadas. Contiene información tanto del fenómeno de transducción electroacústica producido por el micrófono, a través de su sensibilidad y de su directividad, como del camino de transmisión acústica entre la fuente y el micrófono, aunque en este último caso sólo considera la respuesta relativa al centro del array. En su formulación se ha evitado expresar su dependencia con la frecuencia para destacar que depende fuertemente de la posición y orientación relativa entre el micrófono y la fuente, coordenadas $(\vec{r}_i, \theta_i, \phi_i)$.

Ahora se considera que existen I micrófonos, situados cada uno en una posición \vec{r}_i , y con sus características particulares de sensibilidad S_i y directividad $D_i(\theta_i, \phi_i)$. La expresión (9) se puede extender a todos los elementos del array microfónico, de tal manera que se hablará del vector de apuntamiento (*steering vector*):

$$\mathbf{a}(\vec{r}, \theta, \phi) = [a_1(\theta_1, \phi_1, \vec{r}_1), a_2(\theta_2, \phi_2, \vec{r}_2), \dots, a_I(\theta_I, \phi_I, \vec{r}_I)]^T \quad (10)$$

en el que se expresa la dependencia de la respuesta del array con la posición de la fuente (\vec{r}, θ, ϕ) referida al sistema de coordenadas **(XYZ)**. De la misma manera se puede obtener la salida eléctrica de cada micrófono en forma de vector de señales temporales:

$$\mathbf{x}(t) = [x_1(t), x_2(t), \dots, x_I(t)]^T \quad (11)$$

cumpliéndose,

$$\mathbf{x}(t) = \mathbf{a}(\vec{r}, \theta, \phi) x_0(t) \quad (12)$$

que es la ecuación de respuesta del array. Esta ecuación obtiene la salida eléctrica de cada micrófono a partir de la salida eléctrica del micrófono de referencia $x_0(t)$, y de la respuesta electroacústica del array, representada por el vector de apuntamiento $\mathbf{a}(\vec{r}, \theta, \phi)$, que incluye la respuesta de los micrófonos por medio de sus sensibilidades e incluye además el efecto del camino acústico entre fuente y array, que se traduce en atenuación y cambio de fase o retardo para cada uno de los micrófonos integrantes de dicho array.

2.1.2 Respuesta del array ante varias fuentes simultáneas

Ahora se extenderá el caso de una sola fuente incidente a varios emisores, como se representa en la Figura 2. Supóngase que existen M emisores diferentes, caracterizados por la presión temporal que cada uno de ellos proporciona al array de micrófonos. La presión temporal que ejerce cada fuente viene representada por la presión temporal de cada fuente en el centro de coordenadas:

$$p_m(t) = p_{0m} \exp(j\omega t + j\phi_m), \quad m = 1, 2, \dots, M \quad (13)$$

con ϕ_m la fase relativa de la fuente m . Nótese que ahora el índice m se refiere a cada una de las fuentes y no a cada uno de los micrófonos, como ocurría en (2), ya que ahora el array no está caracterizado por sus elementos sino por su posición con respecto a las fuentes y su posición de apuntamiento.

De igual manera que se hacía antes, se pueden definir las respuestas de referencia debidas a cada una de las M fuentes presentes:

$$x_{0m}(t) = S_0 p_m(t) \quad (14)$$

La respuesta conjunta del array sometido a las M fuentes será la suma del efecto de cada fuente, y estará representada por el vector $\mathbf{x}(t)$:

$$\mathbf{x}(t) = \sum_{m=1}^M \mathbf{a}_m(r_m, \theta_m, \varphi_m) x_{0m}(t) \quad (15)$$

donde el vector de apuntamiento \mathbf{a}_m está asociado a la posición $(r_m, \theta_m, \varphi_m)$ de cada una de las fuentes.

Se define ahora el vector respuesta de referencia del array como:

$$\mathbf{x}_0(t) = [x_{01}(t), \dots, x_{0m}(t), \dots, x_{0M}(t)]^T \quad (16)$$

y la matriz de apuntamiento [Haykin 85] [Naidu 01] (*steering matrix*) como:

$$\mathbf{am}(r, \theta, \varphi) = [\mathbf{a}_1(r_1, \theta_1, \varphi_1), \dots, \mathbf{a}_m(r_m, \theta_m, \varphi_m), \dots, \mathbf{a}_M(r_M, \theta_M, \varphi_M)] \quad (I \times M) \quad (17)$$

En esta matriz, cada columna representa la respuesta del array a una determinada posición espacial $(r_m, \theta_m, \varphi_m)$. Es decir, el elemento a_{ij} de esta matriz representa la respuesta electroacústica del micrófono i (fila) ante la fuente j (columna). Por tanto la expresión (15) se puede rescribir de forma matricial, como producto de matrices:

$$\mathbf{x}(t) = \mathbf{am}(r, \theta, \varphi) \mathbf{x}_0(t) \quad (18)$$

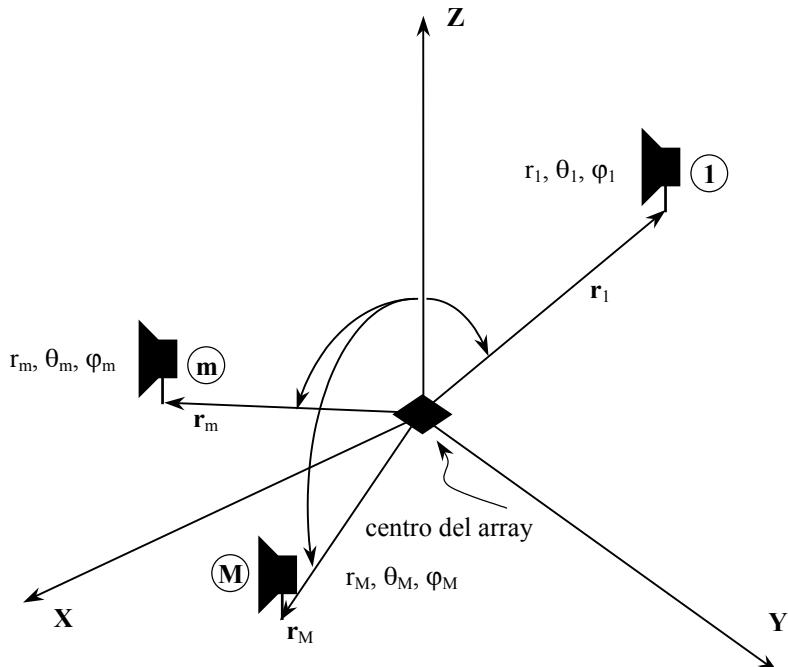


Figura 2. Array de micrófonos que capta M fuentes.

2.1.3 Efecto de la reverberación en la respuesta del array

La reverberación tiene el mismo efecto que un sistema multifuente. Imagínese una fuente origen situada en las coordenadas $(r_1, \theta_1, \varphi_1)$ según se muestra en la Figura 3. Esa fuente llegará al array de forma directa, por el camino \mathbf{r}_1 , pero también se reflejará en las

paredes del recinto para regresar al array por los caminos $\mathbf{r}_2, \dots, \mathbf{r}_q, \dots, \mathbf{r}_Q$. Estos Q caminos representan el sistema multirayecto que origina la reverberación. El número Q es muy grande en la práctica, tendente a infinito, ya que se incluyen reflexiones de orden cada vez mayor con las superficies límite del recinto en el que trabaja el array.

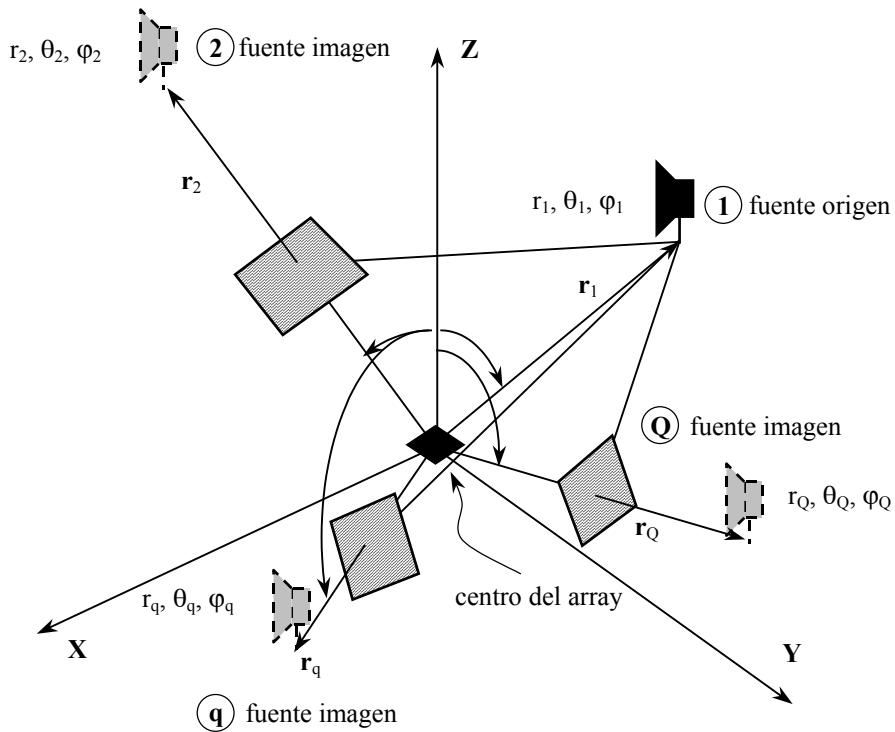


Figura 3. Array lineal de micrófonos que capta Q trayectos debidos a las reflexiones con las paredes originadas por la reverberación.

Es decir, si la fuente origen produce en el centro del array:

$$p_1(t) = p_{01} \exp(j\omega t) \quad (19)$$

las reflexiones con las paredes producirán

$$p_q(t) = p_{01} \frac{r_1}{r_q} \beta_q \exp[j\omega(t - \tau_q) + j\phi_q], \quad q = 2, \dots, Q \quad (20)$$

siendo β_q y ϕ_q respectivamente el coeficiente de reflexión (complementario a la absorción acústica) y el desfase acústico conjuntos que introduce la pared o paredes que aparecen en el camino q, y τ_q el retardo acústico del dicho camino q con respecto a la fuente origen. En (20) puede separarse el efecto del camino acústico y de la absorción de las paredes:

$$p_q(t) = p_1(t) \frac{r_1}{r_q} \exp(-j\omega\tau_q) \beta_q \exp(j\phi_q) \quad q = 2, \dots, Q \quad (21)$$

donde $p_1(t)$ es la presión (19) de la fuente origen en el centro de coordenadas. Por tanto, se podría hablar aquí también del vector de las respuestas eléctricas de referencia (en el centro del array) debidas a la reverberación:

$$\mathbf{x}_0(t) = [x_{01}(t), \dots, x_{0q}(t), \dots, x_{0Q}(t)]^T \quad (22)$$

con $x_{01}(t)$ la respuesta de referencia al camino directo (condiciones de campo libre) y $x_{02}(t), \dots, x_{0Q}(t)$ la respuesta de referencia a la reverberación. Si se mantiene el criterio de que $x_0(t)$ es la respuesta de referencia del array, equivalente a la señal limpia, sin reverberación en este caso, se cumplirá que $x_0(t) = x_{01}(t)$, la respuesta de referencia ante la señal directa. La expresión (14) que relaciona la respuesta eléctrica con la presión acústica sigue teniendo validez. Hay que destacar que cada elemento del vector $x_0(t)$ vendrá afectado por el camino acústico del recorrido que representa (atenuación por divergencia acústica y desfase por recorrido) y por el efecto de la pared (atenuación por el coeficiente de absorción de la pared y desfase por los efectos de la propia reflexión en la pared) de la misma forma que en (20). La respuesta del array ante la señal reverberante vendría dada ahora también por (18), como se proponía para el sistema multifuente. Ahora, cada elemento a_{ij} de la matriz de apuntamiento **am** representaría la respuesta electroacústica del micrófono i ante el camino j , teniendo en cuenta sólo la diferencia de respuesta con el centro del array. Por ejemplo, la salida eléctrica del micrófono i , cuando se consideren Q caminos se expresa por:

$$x_i(t) = \sum_{q=1}^Q a_{iq} x_{0q}(t) \quad (23)$$

con a_{iq} el factor de array del micrófono i ante el camino q , que como se sabe es en general un número complejo.

Es cierto que hasta ahora se ha considerado que el array recibe una sola frecuencia, representada por la excitación $p(t)$ de (1) o $p_i(t)$ de (19), sin embargo esta circunstancia puede generalizarse de forma inmediata. Si como norma general se asume que la excitación es de banda ancha, la ecuación (23) debe ponerse de la siguiente forma:

$$x_i(t) = \sum_{q=1}^Q x_{0q}(t) * a_{iq}(t) \quad (24)$$

en la que se ha aplicado el teorema de la convolución. El factor $a_{iq}(t)$ es la respuesta eléctrica al impulso del micrófono i ante el camino q , es decir la salida eléctrica del micrófono cuando la señal de referencia es una $\delta(t)$ (delta de Dirac). A pesar de esto, en lo sucesivo se preferirá expresar el comportamiento del array ante una sola frecuencia, o trabajar en el dominio de la frecuencia, ya que el tratamiento es más sencillo.

Hecha esta puntualización, tanto en el caso del tratamiento con una sola frecuencia como en el contrario, la reverberación puede considerarse como una perturbación aditiva. Es decir la señal captada por cada uno de los micrófonos del array será la suma de la señal directa (fuente origen) con todas reflexiones (fuentes imagen).

A partir de ahora se llamará $y(t)$ a la señal eléctrica “sucia” de salida del array o de cualquiera de sus micrófonos, según sea el caso, e incluye la suma aditiva de la referencia $x_0(t)$ con las posibles perturbaciones presentes junto a ella. Si se considera el micrófono de referencia en el centro del array, un hipotético $y_0(t)$ sería la suma de las respuestas temporales debidas a todos los caminos acústicos, incluido el de la fuente origen:

$$y_0(t) = \sum_{q=1}^Q x_{0q}(t) = x_{01}(t) + \sum_{q=2}^Q x_{0q}(t) \equiv \text{directa} + \text{reverberación} \quad (25)$$

y este mismo concepto se podrá aplicar a la salida de todos los micrófonos del array, $y_1(t), \dots, y_l(t)$, integrantes del vector de respuesta $y(t)$

$$\mathbf{y}(t) = \mathbf{a}_1 x_{01}(t) + [\mathbf{a}_2, \dots, \mathbf{a}_q, \dots, \mathbf{a}_Q] [x_{02}(t), \dots, x_{0q}(t), \dots, x_{0Q}(t)]^T \equiv \text{dir.} + \text{reverb.} \quad (26)$$

Se da la circunstancia además, de que el término debido a la reverberación tiene escasa coherencia con la señal directa, lo que puede ser aprovechado para su eliminación en diferentes esquemas de procesado.

2.1.4 Efecto del ruido aditivo

Supóngase ruido presente en el sistema, de carácter aleatorio. En principio, por simplificar, se admitirá que este ruido es captado por cada micrófono del array, de tal forma que existe total incoherencia intercanal en dicho ruido (se podría hablar de falta de coherencia espacial del ruido). La afirmación 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. En ese caso se dice que el ruido es altamente coherente. De hecho, más adelante en esta Tesis (capítulos 7 y 9), se plantearán el problema y las posibles soluciones a la presencia de ruido coherente. Sin embargo, a continuación se continuará con el tratamiento más clásico y por tanto básico que presupone una naturaleza incoherente del ruido aditivo.

Sea $\mathbf{y}(t)$ el vector columna de salida sin conformar del array, que contiene el ruido aditivo y que cuenta con tantos elementos como micrófonos I tenga el array. La expresión de $\mathbf{y}(t)$ (señal más ruido) viene representada por el vector columna:

$$\mathbf{y}(t) = [y_1(t), \dots, y_i(t), \dots, y_I(t)]^T \quad (27)$$

De forma general cuando existe una sola fuente sonora, la contaminación por ruido puede ser añadida como se muestra en la siguiente expresión:

$$\mathbf{y}(t) = \mathbf{a}(r, \theta, \phi) x_0(t) + \mathbf{n}(t) = \mathbf{x}(t) + \mathbf{n}(t) \quad (28)$$

con $\mathbf{a}(r, \theta, \phi)$ el vector de apuntamiento (*steering vector*) de (10) y $\mathbf{n}(t)$ el vector ruido incidente al array, con la misma estructura que (27):

$$\mathbf{n}(t) = [n_1(t), \dots, n_i(t), \dots, n_I(t)]^T \quad (29)$$

donde $n_i(t)$ es la salida eléctrica del micrófono i-ésimo debida sólo al ruido. La expresión (28) se generaliza para M fuentes incidentes como:

$$\mathbf{y}(t) = \mathbf{am}(r, \theta, \phi) x_0(t) + \mathbf{n}(t) = \mathbf{x}(t) + \mathbf{n}(t) \quad (30)$$

con $\mathbf{am}(r, \theta, \phi)$ la matriz de apuntamiento (*steering matrix*) de (17), que es una nueva versión de la ecuación del array (12) para incluir los efectos del ruido aditivo.

En definitiva, según lo obtenido en este apartado (28) y (30) y en los dos anteriores (18) y (26), tanto el ruido como la reverberación y en cierta manera la presencia de varias fuentes, pueden entenderse como perturbaciones aditivas que se añaden a la respuesta del array (12) cuando éste capta sólo una fuente. Esta fuente es la que genera la señal de voz que se pretende mejorar y es la que se llamará fuente principal. Por tanto la respuesta del array, considerando todos esos efectos descritos será como (28) pero ahora redefiniendo el ruido $\mathbf{n}(t)$ como:

$$\mathbf{n}(t) = \mathbf{n}_n(t) + \mathbf{n}_r(t) + \mathbf{n}_m(t) \quad (31)$$

para que incluya los efectos del ruido aditivo $\mathbf{n}_n(t)$, de la reverberación $\mathbf{n}_r(t)$ y de la presencia de fuentes ajenas a la principal $\mathbf{n}_m(t)$.

2.1.5 Momentos del array

Admitiendo que la señal captada por el array es de naturaleza aleatoria se pueden considerar algunas definiciones que serán útiles posteriormente.

La matriz de correlaciones cruzadas del vector $\mathbf{x}(t)$ es:

$$\mathbf{R}_{\mathbf{xx}}(\tau) = E\{\mathbf{x}(t) \mathbf{x}^H(t)\} \quad (32)$$

siendo $E\{\cdot\}$ el operador esperanza matemática y τ la variable retardo, implícita en una correlación entre señales estacionarias en sentido amplio. En (32) la multiplicación es matricial y el símbolo H expresa Hermitiano, equivalente al operador transpuesto conjugado. Esta matriz queda desarrollada en sus diferentes términos a continuación:

$$\mathbf{R}_{\mathbf{xx}}(\tau) = \begin{pmatrix} R_{x_1x_1}(\tau) & R_{x_1x_2}(\tau) & \cdots & R_{x_1x_I}(\tau) \\ R_{x_2x_1}(\tau) & R_{x_2x_2}(\tau) & \cdots & R_{x_2x_I}(\tau) \\ \vdots & \vdots & \ddots & \vdots \\ R_{x_Ix_1}(\tau) & R_{x_Ix_2}(\tau) & \cdots & R_{x_Ix_I}(\tau) \end{pmatrix} \quad (33)$$

con $R_{x_i x_j}(\tau)$ la correlación cruzada entre las señales temporales $x_i(t)$ y $x_j(t)$, salida de dos micrófonos del array, cuando no existe la presencia de una perturbación aditiva.

En caso de que se utilice el modelo más general de M fuentes y presencia de ruido o reverberación aditiva, $\mathbf{R}_{\mathbf{xx}}(\tau)$ debe reemplazarse por $\mathbf{R}_{\mathbf{yy}}(\tau)$ que es la matriz de correlaciones cruzadas del vector $\mathbf{y}(t)$ (28), que como se sabe contiene la salida múltiple del array considerando las perturbaciones acústicas. Si se considera al vector de ruido $\mathbf{n}(t)$ de carácter aleatorio su matriz de correlaciones cruzadas será:

$$\mathbf{R}_{\mathbf{nn}}(\tau) = \sigma^2 \mathbf{I}(\tau) \quad (34)$$

con σ^2 la potencia de ruido captado (igual en todos los canales del array) e $\mathbf{I}(\tau)$ la matriz identidad ($I \times I$), con I el tamaño del array. Se verifica que

$$\mathbf{R}_{\mathbf{yy}}(\tau) = \mathbf{R}_{\mathbf{xx}}(\tau) + \mathbf{R}_{\mathbf{nn}}(\tau) \quad (35)$$

En el dominio de la frecuencia, representada por la variable pulsación ω , se puede definir la matriz de espectros cruzados del vector de entrada:

$$\Phi_{\mathbf{XX}}(\omega) = \begin{pmatrix} \Phi_{X_1X_1}(\omega) & \Phi_{X_1X_2}(\omega) & \cdots & \Phi_{X_1X_I}(\omega) \\ \Phi_{X_2X_1}(\omega) & \Phi_{X_2X_2}(\omega) & \cdots & \Phi_{X_2X_I}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{X_Ix_1}(\omega) & \Phi_{X_Ix_2}(\omega) & \cdots & \Phi_{X_Ix_I}(\omega) \end{pmatrix} \quad (36)$$

siendo,

$$\Phi_{X_i X_j}(\omega) = E\{X_i(\omega) X_j^*(\omega)\} \quad (37)$$

el espectro cruzado entre las señales $X_i(\omega)$ y $X_j(\omega)$ proporcionadas por el array. Dicho espectro cruzado no es más que la transformada de Fourier de la correlación cruzada correspondiente:

$$\Phi_{X_i X_j}(\omega) = \mathcal{F}\left\{R_{x_i x_j}(\tau)\right\} \quad (38)$$

con $\mathcal{F}\{\cdot\}$ el operador transformada de Fourier.

Si existe una sola fuente, se puede demostrar que a partir de (12) que:

$$\Phi_{XX}(\omega) = \Phi_{X_0 X_0}(\omega) \mathbf{a}(\omega) \mathbf{a}^H(\omega) \quad (39)$$

donde $\Phi_{X_0 X_0}(\omega)$ representa el autoespectro de la señal eléctrica de referencia, captada en el centro del array y $\mathbf{a}(\omega)$ el vector de apuntamiento de (9) y (10), atendiendo a su variación con la pulsación ω .

Cuando se tienen M fuentes incorreladas, la ecuación (39) se puede generalizar como:

$$\Phi_{XX}(\omega) = \sum_{m=1}^M \Phi_{X_0 m X_0 m}(\omega) \mathbf{a}_m(\omega) \mathbf{a}_m^H(\omega) = \mathbf{am}(\omega) \Phi_{X_0 X_0}(\omega) \mathbf{am}^H(\omega) \quad (40)$$

con $\mathbf{am}(\omega)$ la matriz de apuntamiento de (17) atendiendo a la dependencia con la frecuencia y $\Phi_{X_0 X_0}(\omega)$ una matriz ($M \times M$), diagonal si las fuentes son incoherentes, que representa los autoespectros de las señales de referencia asociadas a cada una de las fuentes de señal que recibe el array. Si además existe ruido difuso incoherente:

$$\Phi_{YY}(\omega) = \mathbf{am}(\omega) \Phi_{X_0 X_0}(\omega) \mathbf{am}^H(\omega) + \sigma^2 \mathbf{I} \quad (41)$$

con

$$\Phi_{YY}(\omega) = \begin{pmatrix} \Phi_{Y_1 Y_1}(\omega) & \Phi_{Y_1 Y_2}(\omega) & \cdots & \Phi_{Y_1 Y_I}(\omega) \\ \Phi_{Y_2 Y_1}(\omega) & \Phi_{Y_2 Y_2}(\omega) & \cdots & \Phi_{Y_2 Y_I}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{Y_I Y_1}(\omega) & \Phi_{Y_I Y_2}(\omega) & \cdots & \Phi_{Y_I Y_I}(\omega) \end{pmatrix} \quad (42)$$

la matriz de espectros cruzados del vector de salida (señal más ruido).

2.1.6 Conformador de haz convencional. Estructura de filtrado y suma

La tarea básica del tratamiento en array de señales acústicas es obtener a la salida del procesador multicanal una señal $y(t)$, que represente fielmente la excitación acústica $p(t)$ que produce la fuente principal en el centro del array, eliminando o atenuando en mayor medida las fuentes de perturbación ajenas a la principal (ruido, reverberación o fuentes secundarias). 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 conformación de haz o *beamforming*, de tal manera que las perturbaciones indeseadas se atenúan por el fenómeno de la directividad o selectividad espacial del array.

La forma estándar más utilizada de conseguir esa conformación de haz es la de filtrado y suma, o de retardo y suma, en la versión más sencilla. Con la estructura de filtrado y suma, 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 .

El proceso de filtrado se puede describir tanto en el dominio del tiempo como en el dominio de la frecuencia y esto no cambia el resultado. Sin embargo, a la hora de la implementación práctica es muy diferente realizar un filtrado temporal mediante filtros FIR o IIR o un filtrado en frecuencia modificando directamente el espectro de la señal a filtrar. Para esto último hay que realizar la transformada de Fourier de la señal temporal, normalmente mediante el algoritmo FFT (*Fast Fourier Transform*), como fase previa al filtrado. Aquí el estudio se centrará en el filtrado en el dominio de la frecuencia, puesto que así se hace en las propuestas y experimentos de los capítulos 7 y 9, aunque se puede consultar el tratamiento temporal en [Flanagan 91].

Se supone inicialmente que se está procesando una señal senoidal (una sola frecuencia), en su versión de presión acústica $p(t)$ (1) o de tensión eléctrica $x_0(t)$ (5), que se puede expresar por tanto de forma fasorial. El filtro w_i que modifica al canal i del array, puede entenderse como un coeficiente complejo, constante o variable con la frecuencia. Si w_i no depende de la frecuencia, el proceso de filtrar equivale a multiplicar, tanto en el dominio del tiempo como en el dominio de la frecuencia. En caso contrario habría que considerar la operación convolución con la respuesta al impulso, cuando se trabaje en el dominio del tiempo. Pero, supóngase inicialmente que el módulo de los coeficientes w_i no depende de la frecuencia y su fase es proporcional a dicha frecuencia (equivalente a un retardo). Si se considera en un principio que no existe ruido ni ninguna perturbación acústica, el vector de entrada al array estará representado por $\mathbf{x}(t)$. Entonces w_i multiplica a la señal $x_i(t)$ saliente de cada micrófono. Se puede hablar entonces del vector de coeficientes de filtrado:

$$\mathbf{w} = [w_1, w_2, \dots, w_I]^T \quad (43)$$

Atendiendo al esquema de filtrado y suma, según se representa en la Figura 4, la salida conformada del array se expresará por:

$$y(t) = \sum_{i=1}^I w_i^* x_i(t) = \mathbf{w}^H \mathbf{x}(t) = \mathbf{w}^H \mathbf{a}(r, \theta, \varphi) x_0(t) \quad (44)$$

donde filtrar equivale a multiplicar por el conjugado de cada coeficiente w_i . Nótese que $y(t)$ no se escribe en negrita porque ya no es un vector sino una señal monocanal.

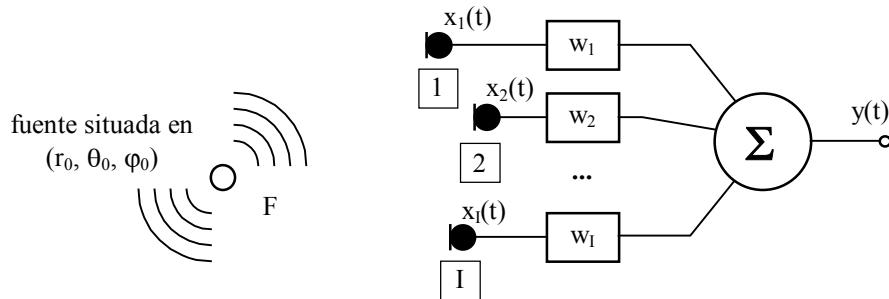


Figura 4. Conformación de haz mediante filtrado y suma.

La potencia de salida del array viene dada por:

$$P = E\{y(t)^2\} = E\{\mathbf{w}^H \mathbf{x}(t) \mathbf{w} \mathbf{x}^H(t)\} = \mathbf{w}^H \mathbf{R}_{xx}(\tau) \mathbf{w} \quad (45)$$

De forma más general, en el caso de que se considere como entrada al vector $\mathbf{y}(t)$ (señal más ruido y/o reverberación), la potencia a la salida del conformador será igualmente:

$$P = E\{\mathbf{w}^H \mathbf{y}(t) \mathbf{w} \mathbf{y}^H(t)\} = \mathbf{w}^H E\{\mathbf{y}(t) \mathbf{y}^H(t)\} \mathbf{w} = \mathbf{w}^H \mathbf{R}_{yy}(\tau) \mathbf{w} \quad (46)$$

Si la fuente sonora principal está situada en las coordenadas espaciales (r_0, θ_0, ϕ_0) correspondientes a la DOA y es única (no existen fuentes de perturbación ni reverberación), la técnica de conformación de haz más sencilla se basa en maximizar, mediante la variación de los coeficientes w , la potencia de salida del array. Supóngase un array de micrófonos omnidireccionales, es decir $D_i(\theta_i, \phi_i) = 1$, y una fuente sonora situada en las coordenadas (r_0, θ_0, ϕ_0). De momento no se considera ruido ni reverberación. El vector complejo de ponderaciones que maximiza la potencia de salida para la fuente principal en (r_0, θ_0, ϕ_0), viene dado [Krim 96] [Naidu 01] por:

$$\mathbf{w}_{RS}(r_0, \theta_0, \phi_0) = \frac{\mathbf{a}(r_0, \theta_0, \phi_0)}{\mathbf{a}^H(r_0, \theta_0, \phi_0) \mathbf{a}(r_0, \theta_0, \phi_0)} \quad (47)$$

siendo \mathbf{w}_{RS} el vector de ponderaciones para la configuración de conformación de retardo (R) y suma (S). En esta expresión el denominador es un factor de normalización. Al resultado de [Naidu 01] se le ha modificado dicho factor de normalización para que los coeficientes \mathbf{w}_{RS} no amplifiquen ni atenúen la salida conformada $y(t)$ con respecto a la señal entregada por cada micrófono, $x_i(t)$. En ese caso se dice que el conformador produce una respuesta sin distorsión (*distortionless response*), concepto que será desarrollado con más detalle en el punto 2.2.3 sobre superdirecividad. Para comprobar este hecho, si se aplica el filtro multicanal (47) a la respuesta del array (12) según la expresión (44) particularizada para la DOA principal, la salida del conformador apuntado a la posición (r_0, θ_0, ϕ_0) es:

$$\begin{aligned} y(t) &= \mathbf{w}_{RS}^H \mathbf{a}(r_0, \theta_0, \phi_0) x_0(t) = \\ &= \left[\frac{\mathbf{a}(r_0, \theta_0, \phi_0)}{\mathbf{a}^H(r_0, \theta_0, \phi_0) \mathbf{a}(r_0, \theta_0, \phi_0)} \right]^H \mathbf{a}(r_0, \theta_0, \phi_0) x_0(t) = x_0(t) \end{aligned} \quad (48)$$

es decir, la señal de referencia $x_0(t)$, considerada precisamente como la salida sin distorsión del array.

Los elementos del vector de coeficientes de filtrado $\mathbf{w}_{RS}(r_0, \theta_0, \phi_0)$ tienen la misma fase que sus homólogos del vector de apuntamiento $\mathbf{a}(r_0, \theta_0, \phi_0)$ particularizados para la DOA principal. Esta característica hará que, en la operación de filtrado y suma expresada en (44), la fase de cada elemento de $\mathbf{w}_{RS}(r_0, \theta_0, \phi_0)$ se invierta (por la operación conjugación) con respecto a cada elemento correspondiente de $\mathbf{a}(r_0, \theta_0, \phi_0)$, con lo que finalmente en (44) se tiene una suma de productos de módulos, cuando las coordenadas genéricas de apuntamiento (r, θ, ϕ) coincidan con la posición de la fuente en (r_0, θ_0, ϕ_0), considerada como principal. La fase de cada elemento a_i del vector de apuntamiento representa el retardo acústico que tiene el micrófono correspondiente, con respecto a un punto de referencia, el origen de coordenadas en este caso. Es decir, cada elemento del filtro multicanal w_{RSi} asociado a un micrófono del array consiste en un retardo que compensa el retardo acústico desde la fuente hasta el micrófono sobre el que se sitúa dicho filtro. Por tanto, se produce una alineación temporal de todos los micrófonos, por medio del desfase eléctrico proporcionado por los filtros \mathbf{w}_{RS} . Por eso, la estructura representada en (44) se llama también de retardo y suma, RS (o DS *Delay and Sum*), aunque también es conocida como conformador convencional.

La fase de los elementos \mathbf{w}_{RS} de (47) hace que el eje principal del array o eje de máxima captación, se oriente hacia la dirección desde la que llega la fuente principal, por eso se

maximiza la potencia de salida, ya que el diagrama polar de directividad del array $D(\theta, \phi)$ apunta a la fuente. El módulo de los elementos w_{RS} dado en (47) es responsable de la forma que tiene el patrón de directividad del array, sobre todo de dos parámetros importantes como son la anchura del lóbulo de captación principal y la atenuación de los lóbulos secundarios, pero no influye en el apuntamiento del array. La elección de w_{RS} según (47) junto con la operación de filtrado y suma de (44) asegura un máximo de captación del array (directividad máxima) para la dirección de apuntamiento (r_0, θ_0, ϕ_0) , cualquiera que sean las sensibilidades S_i de los micrófonos, contenidas en el módulo del vector de apuntamiento $a(r, \theta, \phi)$, incluso si éstas no son iguales o de diferente signo, o también si las distancias de los micrófonos a la fuente son muy diferentes.

En la Figura 5 se representa el patrón polar de directividad $D(\theta, \phi)$ de dos arrays lineales uniformes con la misma geometría y apuntamiento pero distinta ponderación, en módulo, aplicada a los micrófonos. Puede observarse cómo en ambos casos es diferente la amplitud angular del lóbulo principal y la atenuación de los lóbulos secundarios. El módulo de los elementos w_i puede servir además para corregir defectos de los micrófonos en cuanto a su sensibilidad o directividad individual, consistentes especialmente en las diferencias de valor de estos parámetros entre los micrófonos del array.

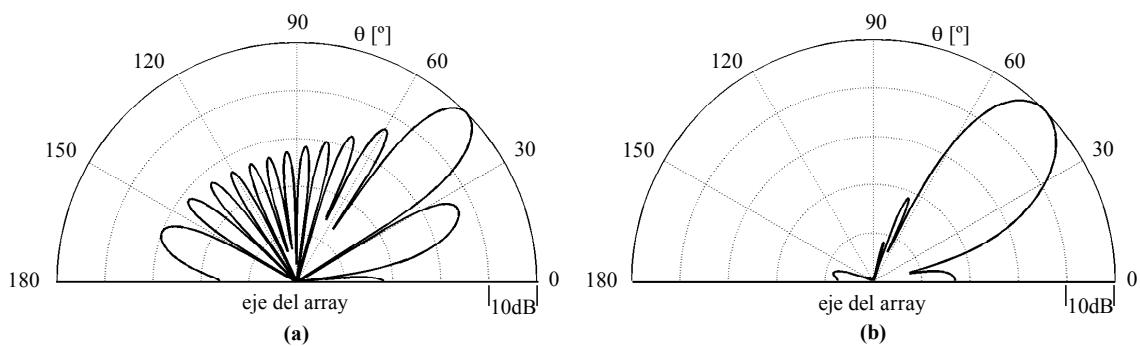


Figura 5. Patrón de directividad $D(\theta, \phi)$ de un array lineal uniforme de 15 micrófonos, con separación intermicrófónica $\Delta z=0.04\text{m}$, $f_0=4\text{kHz}$, $r_0=10\text{m}$. **(a)** Con pesos (módulos) como los dados en (47). **(b)** Con pesos siguiendo una ponderación tipo Hanning.

Nótese que si en (47) se considera que todos los módulos de los vectores de apuntamiento cumplen $|a(r_0, \theta_0, \phi_0)| = 1$, entonces el conformador convencional se transforma en:

$$w_{RS}(r_0, \theta_0, \phi_0) = \frac{1}{I} a(r_0, \theta_0, \phi_0) \quad (49)$$

con I el número de elementos del array como factor de normalización. La condición de módulo unidad $|a(r_0, \theta_0, \phi_0)| = 1$, no es difícil de cumplir si en el array se trabaja con micrófonos de características similares (pareados) y además se realiza, previamente a la conformación de haz, una igualación de niveles entre canales que compense las diferentes distancias de los micrófonos a la fuente.

2.1.7 Algunos tópicos sobre directividad

A continuación se desarrollan algunos de los conceptos más comúnmente utilizados para caracterizar la respuesta angular de un array y que son utilizados frecuentemente a lo largo de esta Tesis. Un mayor detalle de esto puede encontrarse en [Sánchez-Bote 02-a].

Se define “Directividad” $D(r_0, \theta, \varphi)$ de un array de micrófonos según una determinada posición espacial (r_0, θ, φ) como la respuesta eléctrica del micrófono respecto a la respuesta máxima. Se considera que, si el array está correctamente apuntado y está constituido por micrófonos omnidireccionales, dará respuesta máxima en las coordenadas $(r_0, \theta_0, \varphi_0)$, es decir en la DOA principal. Para un esquema de filtrado y suma, como el definido en (44) la directividad quedará:

$$D(r_0, \theta, \varphi) = \frac{\mathbf{w}^H(r_0, \theta_0, \varphi_0) \mathbf{a}(r_0, \theta, \varphi)}{\mathbf{w}^H(r_0, \theta_0, \varphi_0) \mathbf{a}(r_0, \theta_0, \varphi_0)} \quad (50)$$

En (50) la directividad depende del ángulo (θ, φ) y se calcula para una distancia r_0 de apuntamiento del array, a la que se sitúa la fuente principal. La dependencia con el ángulo es muy fuerte, pero la dependencia con la distancia es muy débil. Quiere decir que cuando r_0 sea suficientemente grande con respecto al tamaño del array, la directividad dependerá exclusivamente del ángulo, debido a que la influencia de la distancia en el numerador de (50) compensa su influencia en el denominador y $D(r_0, \theta, \varphi) = D(\theta, \varphi) \forall r_0$. El módulo máximo de la directividad de un array es unidad, por definición.

El “Factor de Directividad” $Q(r_0, \theta, \varphi)$, es la relación entre la energía captada por el array cuando apunta a la posición (r_0, θ, φ) y el promedio espacial de energía captada, promedio que se refiere a todas las direcciones espaciales a la distancia r_0 . De forma analítica:

$$\begin{aligned} Q(r_0, \theta, \varphi) &= \frac{\left[|\mathbf{w}^H(r_0, \theta_0, \varphi_0) \mathbf{a}(r_0, \theta, \varphi)| \right]^2}{\left\langle \left[|\mathbf{w}^H(r_0, \theta_0, \varphi_0) \mathbf{a}(r_0, \theta, \varphi)| \right]^2 \right\rangle_{\text{promedio espacial}}} = \\ &= \frac{\left[|\mathbf{w}^H(r_0, \theta_0, \varphi_0) \mathbf{a}(r_0, \theta, \varphi)| \right]^2}{\frac{1}{4\pi} \int_{\varphi=0}^{2\pi} \int_{\theta=0}^{\pi} \left[|\mathbf{w}^H(r_0, \theta_0, \varphi_0) \mathbf{a}(r_0, \theta, \varphi)| \right]^2 \sin\theta d\theta d\varphi} \end{aligned} \quad (51)$$

Se puede demostrar fácilmente [Sánchez-Bote 02-a] que el factor de directividad está relacionado con la directividad por:

$$Q(r_0, \theta, \varphi) = \frac{1}{\frac{1}{4\pi} \int_{\varphi=0}^{2\pi} \int_{\theta=0}^{\pi} |D(r_0, \theta, \varphi)|^2 \sin\theta d\theta d\varphi} \quad (52)$$

El factor de directividad es un parámetro que puede ser mayor o menor que la unidad. Se suele considerar el factor de directividad en la posición de apuntamiento, que si coincide con la DOA, cuando el array esté bien apuntado, será el factor de directividad máximo:

$$Q_{MAX}(r_0) = Q(r_0, \theta_0, \varphi_0) \quad (53)$$

siendo:

$$Q_{\text{MAX}}(r_0) \geq Q(r_0, \theta, \varphi) \quad \forall (\theta, \varphi) \quad (54)$$

Aquí la dependencia con la posición de apuntamiento r_0 tiene la misma consideración que se hacía con $D(\theta, \varphi)$, es decir si la fuente principal se sitúa en un punto suficientemente alejado con respecto al tamaño del array $Q(r_0, \theta, \varphi) = Q(\theta, \varphi) \quad \forall r_0$ y $Q_{\text{MAX}}(r_0) = Q_{\text{MAX}} \quad \forall r_0$.

Normalmente el factor de directividad se expresa en unidades logarítmicas, mediante el índice de directividad máximo DI_{MAX} (*Directivity Index*):

$$\text{DI}_{\text{MAX}} = 10 \log Q_{\text{MAX}} \quad (55)$$

Cuanto mayor sea Q_{MAX} o DI_{MAX} , más directivo será el array, y mayor rechazo tendrá ante señales acústicas que incidan lateralmente, ya procedan de ruido, de reverberación o de fuentes ajena a la señal de voz principal. Por tanto, desde el punto de vista de la mejora de señal de voz interesan arrays muy directivos, con elevado factor de directividad máximo.

2.1.8 Array lineal

Aunque en la captación de señal de voz mediante arrays existe la posibilidad de utilizar diferentes configuraciones (arrays lineales, circulares, planos, uniformes, no uniformes...) –consúltese el capítulo 2 de [Naidu 01]–, en esta Tesis se tratarán los arrays lineales de micrófonos. El término lineal atiende al hecho de que los micrófonos se disponen espacialmente sobre una línea, manteniendo diferentes distancias entre ellos. En el caso de que la distancia intermicrofónica sea constante se hablará de array lineal uniforme de micrófonos.

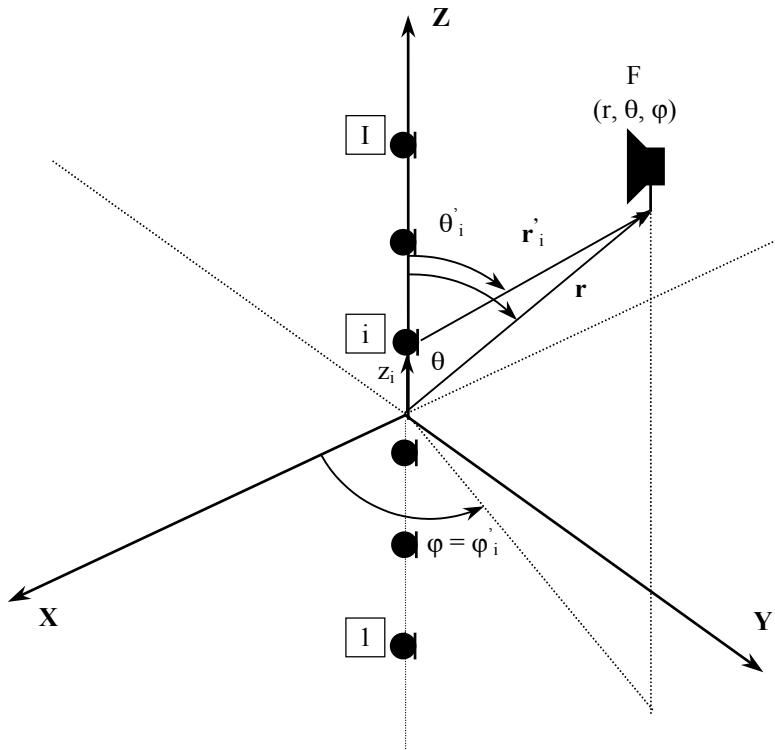


Figura 6. Array lineal de micrófonos con una fuente situada en las coordenadas (r, θ, φ) .

Supóngase que se dispone una array microfónico compuesto por I elementos iguales, de tal manera que se sitúen en una línea sobre el eje **Z** (Figura 6), y con sus ejes paralelos. En este caso y con la geometría de la Figura 6, el vector de posición \mathbf{r}_i del micrófono i-ésimo es:

$$\mathbf{r}_i = z_i \hat{\mathbf{z}} \quad (56)$$

con z_i la posición del micrófono sobre el eje **Z** y $\hat{\mathbf{z}}$ el vector unitario dicho eje. El módulo de la distancia particular de cada micrófono al punto genérico del campo acústico de coordenadas (r, θ, ϕ) se expresa por:

$$r'_i = \sqrt{(r \sin \theta)^2 + (r \cos \theta - z_i)^2} \quad (57)$$

Aproximación de campo lejano

Si se utiliza la aproximación de Fraunhofer o de campo lejano, es decir, la fuente está muy alejada del array, o lo que es igual $r \gg |z_l - z_i|$ ($|z_l - z_i|$ es el tamaño del array), se puede desarrollar (57) alrededor de $z_i = 0$ con lo que:

$$\frac{r'_i}{r_i} \approx r - z_i \cos \theta \quad , \quad \frac{1}{r'_i} \approx \frac{1}{r} \quad y \quad \theta'_i \approx \theta \quad (58)$$

y además $\phi'_i = \phi$ por la simetría de revolución alrededor del eje **Z** que impone la geometría array. Es decir, con la aproximación de campo lejano, las coordenadas angulares genéricas de la fuente coinciden con las coordenadas angulares particulares de cada uno de los micrófonos del array y entonces $(\theta'_i, \phi'_i) = (\theta, \phi)$. Con estas consideraciones, el elemento a_i del vector de apuntamiento (9), quedaría:

$$a_i(\theta, \phi) = \frac{\mathbf{S}_i}{\mathbf{S}_0} D_i(\theta, \phi) \exp(jk z_i \cos \theta) \quad (59)$$

En realidad la aproximación de Fraunhofer no es necesaria para el tratamiento de los arrays microfónicos, pero simplifica mucho su estudio. De hecho, en la práctica la fuente estará cerca del array, aunque eso no cambia en nada lo fundamental.

Si además, los micrófonos del array son omnidireccionales, es decir $D_i(\theta, \phi) = 1 \quad \forall i$ y de sensibilidades \mathbf{S}_i también iguales (si no es así se puede corregir), las expresiones (47) o (49) que dan los coeficientes w_i para la estructura de retardo y suma de un array apuntado a la dirección θ_0 , quedarían:

$$\mathbf{w}_{RS}(\theta_0) = \frac{1}{I} \exp(jk z_i \cos \theta_0) \quad (60)$$

La directividad, según fue expresada en (50), de un array lineal de micrófonos omnidireccionales apuntando a la dirección (θ_0) y en campo lejano será de forma general:

$$D(\theta) = \frac{\mathbf{w}^H(\theta_0) \mathbf{a}(\theta)}{\mathbf{w}^H(\theta_0) \mathbf{a}(\theta_0)} = \frac{\sum_{i=1}^I w_i^*(\theta_0) \frac{\mathbf{S}_i}{\mathbf{S}_0} \exp(jk z_i \cos \theta)}{\sum_{i=1}^I w_i^*(\theta_0) \frac{\mathbf{S}_i}{\mathbf{S}_0} \exp(jk z_i \cos \theta_0)} = \frac{\sum_{i=1}^I C_i \exp(jk z_i \cos \theta)}{\sum_{i=1}^I C_i \exp(jk z_i \cos \theta_0)} \quad (61)$$

Esta expresión muestra cómo la directividad de un array lineal es equivalente a la suma de I fasores con diferentes pesos C_i , que pueden ser reales o complejos. El denominador es un factor de normalización. La expresión (61) tiene mucho parecido con la transformada de Fourier, y esto será aprovechado más adelante para explicar la relación existente entre las ponderaciones C_i junto a la distribución geométrica de los elementos del array, con la forma del patrón de directividad del array, $D(\theta)$.

En la exposición hecha hasta ahora se ha supuesto que los micrófonos del array son omnidireccionales $D_i(\theta, \varphi) = 1 \forall i$, pero esta hipótesis no es estrictamente necesaria. Para poder sacar alguna relación entre la distribución geométrica y los pesos de los micrófonos del array, con la forma de su patrón de directividad, es suficiente con que los micrófonos sean iguales y con sus ejes colineares. Si los micrófonos son directivos, con el mismo patrón polar, $D_i(\theta, \varphi) = D_\mu(\theta, \varphi) \forall i$, y se disponen con los ejes paralelos, su directividad influye igualmente en cada elemento a_i del vector de apuntamiento, y se puede sacar como factor común. Entonces, la directividad del array apuntado a la dirección (θ_0, φ_0) quedará:

$$D(\theta, \varphi) = \frac{D_\mu(\theta, \varphi)}{D_\mu(\theta_0, \varphi_0)} \frac{\sum_{i=1}^I C_i \exp(jk z_i \cos \theta)}{\sum_{i=1}^I C_i \exp(jk z_i \cos \theta_0)} = \frac{1}{D_\mu(\theta_0, \varphi_0)} D_\mu(\theta, \varphi) D_{FA}(\theta, \varphi) \quad (62)$$

donde $D_{FA}(\theta, \varphi)$ es la directividad del factor de array o simplemente la directividad del array suponiendo que los micrófonos son omnidireccionales, y equivale a (61). La expresión (62) es conocida frecuentemente como teorema del producto y su significado es que la directividad de un array de receptores directivos iguales es el producto de la directividad de cada micrófono por el factor de array. El término $1/D_\mu(\theta_0, \varphi_0)$ es una constante de normalización que mantiene el valor máximo de $D(\theta, \varphi)$ en uno.

Directividad y transformada de Fourier

El análisis de la directividad de un receptor puede hacerse mediante la transformada de Fourier. En óptica se conoce desde hace tiempo [Goodman 68] [Haykin 75] la relación existente entre la imagen difractada por una abertura y la transformada de Fourier bidimensional de la abertura. Del mismo modo, se puede establecer una relación entre la directividad asociada a un array de micrófonos y la transformada de Fourier del array, entendiéndose que un array es una señal en el dominio espacial (coordenada z para un array lineal). Es decir, la variable tiempo t , que se utiliza para el análisis de Fourier de señales temporales, pasa a ser variable espacial z , (que no hay que confundir con la variable z , asociada a la transformada Z).

La transformada de Fourier de una secuencia discreta $f(z)$ en la variable espacial z :

$$f(z) = \sum_{i=1}^I C_i \delta(z - z_i) \quad (63)$$

viene dada por,

$$F(\Omega) = \sum_{i=1}^I C_i \exp(-j\Omega z_i) \quad (64)$$

Si se hace

$$\Omega = -k \cos\theta \quad (65)$$

la expresión (64) se identifica fácilmente con la (61) que es la directividad de un array de micrófonos omnidireccionales. Es decir, la directividad de un array lineal de micrófonos omnidireccionales se puede entender como la transformada de Fourier de una serie discreta en la coordenada espacial z . La fase de los elementos de la serie discreta viene determinada por las posiciones z_i de los micrófonos del array. Los coeficientes complejos C_i están asociados a la transducción electroacústica que produce cada micrófono y a la ponderación eléctrica proporcionada por los coeficientes w_i . Por tanto, las posiciones de los micrófonos en el eje Z están representadas por las posiciones en el eje espacial z de las $\delta(z)$ de (63), asociadas a la serie discreta que representa al array. La variable transformada Ω , que aquí se llamará pulsación o frecuencia angular, equivale por (65) al ángulo θ que es el parámetro con el que varía la directividad. Por tanto, la directividad de un array lineal de I micrófonos omnidireccionales, situados en las coordenadas z_i del eje z , y con ponderaciones electroacústicas C_i viene dada por:

$$D(\Omega) = \frac{\mathcal{F} \left[\sum_{i=1}^I C_i \delta(z - z_i) \right]}{\mathcal{F} \left[\sum_{i=1}^I C_i \delta(z - z_i) \right]_{\Omega=\Omega_0 = -k \cos\theta_0}} \quad \text{con } \Omega = -k \cos\theta \quad (66)$$

donde $\mathcal{F}\{\cdot\}$ representa, como se sabe, la transformada de Fourier. El denominador de (66) normaliza la directividad y se calcula mediante la transformada de Fourier particularizada $F(\Omega_0)$ de la secuencia $f(z)$ del array, con Ω_0 el valor de la pulsación angular para el eje principal del array, que resulta de sustituir $\theta = \theta_0$ en (64).

La nueva variable pulsación angular Ω , asociada al ángulo θ , tiene un dominio de existencia limitado. Por la propia definición (65), $\Omega \in [-k, k]$ ya que $\cos\theta \in [-1, 1]$. A la pulsación $\Omega = -k$ le corresponde un ángulo $\theta = 0^\circ$ y a $\Omega = k$ un ángulo $\theta = 180^\circ$. Hay que considerar que la directividad de un array lineal de micrófonos omnidireccionales tiene simetría de revolución alrededor del eje $\theta = 0^\circ$, por lo que será suficiente con considerar $\theta \in [0^\circ, 180^\circ]$, y el dominio de existencia de Ω es compatible con la variación posible del ángulo θ .

La asimilación de la directividad de un array a la transformada de Fourier de su geometría no añade nada nuevo al cálculo de dicha directividad, sin embargo ofrece una perspectiva muy útil para el enfoque del problema de optimización en el diseño del array. La teoría sobre transformadas de señales discretas está muy asentada y puede emplearse en el diseño de un array de características directivas óptimas.

Array lineal uniforme

Un array lineal es uniforme cuando las distancias intermicrofónicas Δz son iguales, es decir

$$z_i = n_i \Delta z \quad (67)$$

con n_i un número entero e $i = 1, 2, \dots, I$, el índice microfónico.

Según lo expuesto en el punto anterior, un array lineal uniforme puede asociarse a una serie discreta $f(z)$ en el dominio de la coordenada espacial z , con periodo de muestreo espacial Δz . Esta serie discreta tiene longitud finita, que equivale a la longitud total en metros del array, en el eje **Z**. Se puede también decir que la serie discreta $f(z)$ está enventanada con una ventana espacial del tamaño de la longitud del array. Por tanto $f(z)$ viene dada por:

$$f(z) = \sum_{i=1}^I C_i \delta(z - n_i \Delta z) \quad (68)$$

que se convierte en $F(\Omega)$ en el dominio transformado, función que representa la directividad como ya se ha explicado. La pulsación equivalente de muestreo será:

$$\Omega_s = \frac{2\pi}{\Delta z} \quad (69)$$

y los coeficientes C_i , asociados a la ponderación electroacústica de cada micrófono son los pesos que representan la envolvente del array en el dominio de la coordenada espacial z . Por la teoría del muestreo, la directividad $F(\Omega)$ es periódica con periodo Ω_s .

Esta analogía entre directividad y transformada de Fourier, va a servir para predecir de forma sencilla las características principales en cuanto a directividad de un array lineal uniforme. A continuación se relacionan las características más destacables del patrón de directividad de un array lineal uniforme y su relación con la geometría del mismo.

1.- Aliasing espacial

Un array lineal uniforme manifiesta *aliasing* espacial (también *aliasing* angular) cuando además del lóbulo principal de captación existen uno o más lóbulos laterales de igual amplitud que el principal, en el intervalo $\theta \in [0^\circ, 180^\circ]$. Existirá *aliasing* espacial cuando

$$k + |\Omega_0| \geq \Omega_s \quad (70)$$

Esta condición asegura que el periodo de repetición Ω_s del espectro de $F(\Omega)$ es menor que la distancia entre el lóbulo principal representado por Ω_0 y los extremos del intervalo de definición de la variable Ω , con lo que el lóbulo principal se repetirá en dicho intervalo de definición originándose *aliasing* espacial. En la Figura 7 puede observarse cómo la repetición del espectro $D(\Omega)$ en el intervalo $\Omega \in [-k, k]$ produce *aliasing* espacial. La condición expresada en (70) indica que los arrays que sufrirán un menor *aliasing* espacial serán aquellos con $\Omega_0 = 0$, equivalente a $\theta_0 = 90^\circ$. Es decir la configuración *broadside* ($\theta_0 = 90^\circ$) tiene menor posibilidad de *aliasing* espacial que la *endfire* ($\theta_0 = 0^\circ$ equivalente a $\Omega_0 = -k$) a la hora de configurar un array lineal uniforme de micrófonos omnidireccionales. La Figura 10 ilustra la producción de *aliasing* espacial a medida que varía la dirección de apuntamiento en el array. Comparando la Figura 10(b) con la Figura 10(a) se verifica cómo un mismo array manifiesta *aliasing* espacial cuando su eje principal pasa de ser $\theta_0 = 90^\circ$ a $\theta_0 = 45^\circ$ debido a que $1 + |\Omega_0|/k > \Omega_s/k$ por el aumento de $|\Omega_0|$.

La condición (70) puede desarrollarse con más detalle para los dos casos considerados de apuntamiento *broadside* y *endfire*. Es decir se tiene *aliasing* espacial si:

$$k \geq \Omega_s \text{ ó } f \geq \frac{c}{\Delta z} \quad \text{para la configuración } \textit{broadside} \quad (71)$$

$$k \geq \frac{\Omega_s}{2} \text{ ó } f \geq \frac{c}{2\Delta z} \quad \text{para la configuración } endfire$$

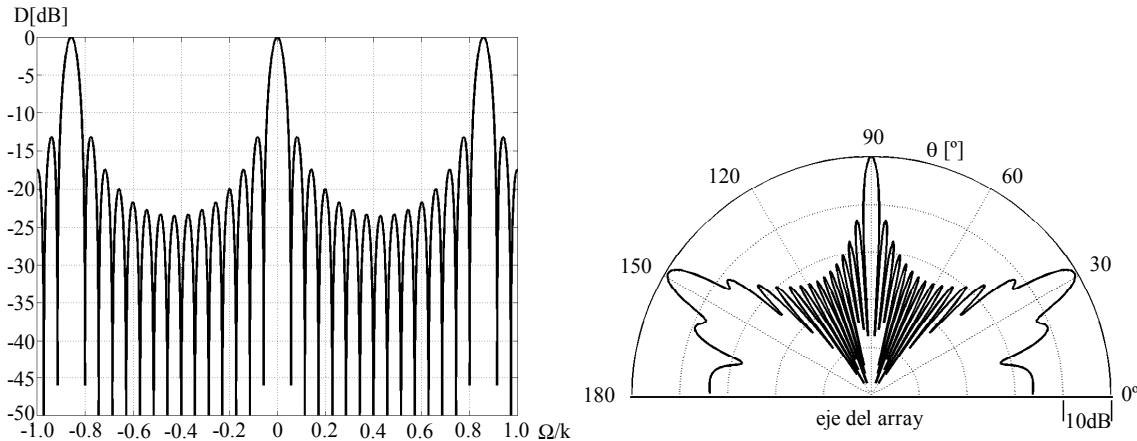


Figura 7. *Aliasing* espacial. Se muestra la directividad $D(\Omega)$ (izquierda) y $D(\theta)$ (derecha) para un array lineal uniforme equiponderado de 15 micrófonos, $\Delta z=0.08\text{m}$, $\theta_0=90^\circ$, $f_0=5\text{kHz}$ ($\Omega_s/k=0.9$).

2.- Amplitud angular del lóbulo principal y número de lóbulos secundarios

Estos dos parámetros se controlan conjuntamente con la longitud del array en el eje espacial **Z** (o tamaño del array), que como se expresó antes equivale a la longitud de una supuesta ventana espacial, para seguir utilizando la analogía con el procesado de señales discretas. La longitud del array es $(I - 1) \Delta z$. Cuanto mayor longitud tenga el array $f(z)$, más limitada en frecuencia será su transformada de Fourier $F(\Omega)$ y por tanto más estrechos y numerosos serán los lóbulos de su diagrama polar de directividad. Esto se muestra en la Figura 8.

El criterio de resolución de Rayleigh establece el ancho del lóbulo principal de un array lineal uniforme de I fuentes separadas por Δz , para una ventana rectangular de ponderación (array equiponderado, véase la Figura 8). En ese caso, el primer nulo de captación Ω_{nulo} que delimita el lóbulo principal, tiene una separación angular con respecto al eje principal Ω_0 del array, dada por la siguiente expresión:

$$\Omega_{\text{nulo}} - \Omega_0 = \pm \frac{2\pi}{I \Delta z} \quad (72)$$

Esta última relación se ha calculado aprovechando la similitud entre directividad y transformada de Fourier de la ventana rectangular. La posición de primer nulo de la función sinc $(\Omega I \Delta z / 2\pi)$ que es la transformada de Fourier de la ventana rectangular, apropiada para representar al array equiponderado, verifica (72). Sustituyendo el valor Ω de (65), esta última expresión puede rescribirse como:

$$\cos \theta_{\text{nulo}} = \cos \theta_0 \mp \frac{\lambda}{I \Delta z} \text{ para } \left| \cos \theta_0 \mp \frac{\lambda}{I \Delta z} \right| \leq 1 \quad (73)$$

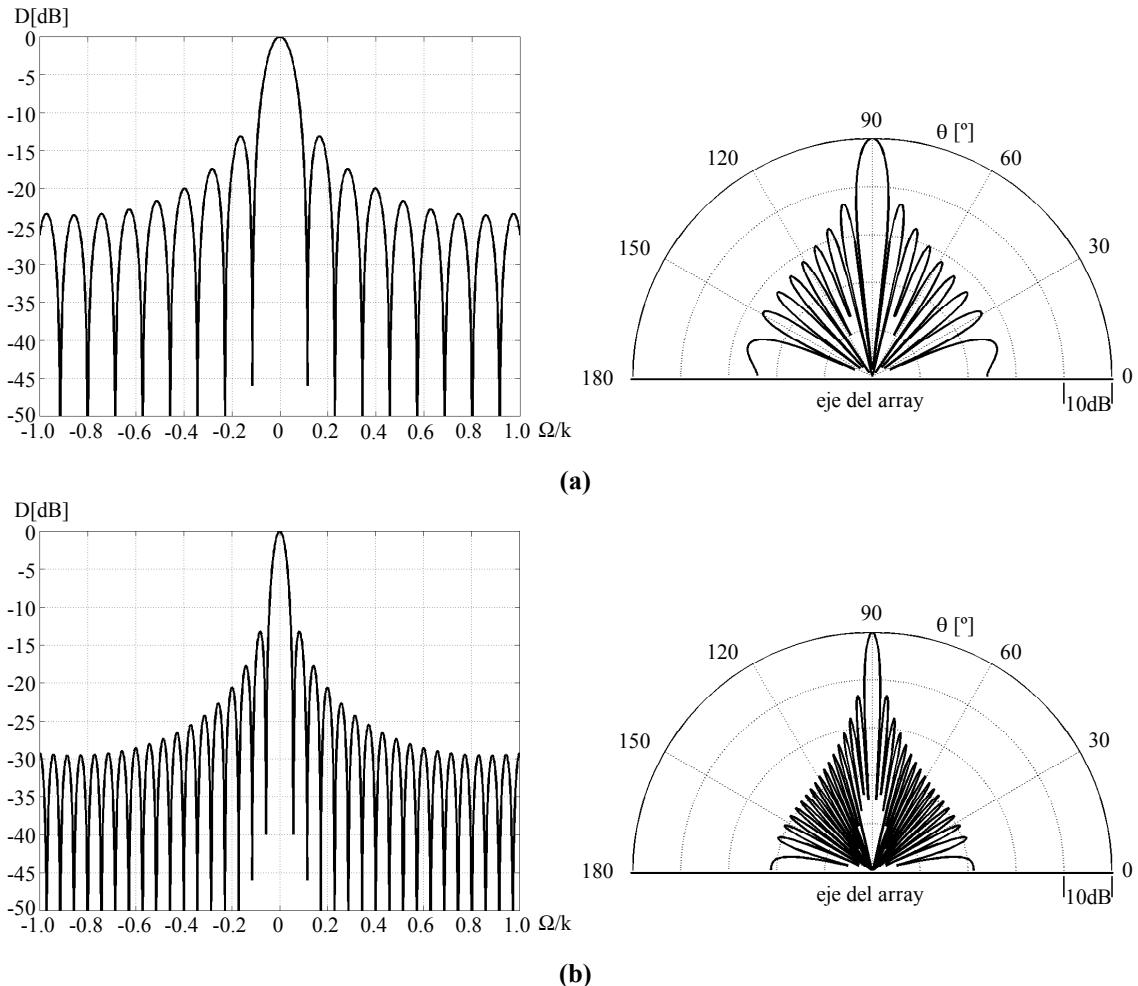


Figura 8. Efecto del tamaño del array en la amplitud angular del lóbulo principal y número de lóbulos secundarios. Se muestra la directividad $D(\Omega)$ y $D(\theta)$ para un array lineal uniforme equiponderado con $\Delta z=0.08\text{m}$, $\theta_0=90^\circ$ y $f_0=2.5\text{kHz}$ ($\Omega_s/k=1.7$). **(a)** $I=15$ micrófonos, $(I-1)\Delta z=1.12\text{m}$. **(b)** $I=30$ micrófonos, $(I-1)\Delta z=2.32\text{m}$.

Cuando la configuración es *broadside* ($\theta_0 = 90^\circ$)

$$\cos \theta_{\text{nulo-bs}} = \mp \frac{\lambda}{I \Delta z} \text{ para } \frac{\lambda}{I \Delta z} \leq 1 \quad (74)$$

y cuando es *endfire* ($\theta_0 = 0^\circ$)

$$\cos \theta_{\text{nulo-ef}} = 1 - \frac{\lambda}{I \Delta z} \text{ para } \frac{\lambda}{I \Delta z} \leq 1 \quad (75)$$

Las expresiones (74) y (75) permiten calcular el valor de $\lambda/(I \Delta z)$ que establece una selectividad espacial suficiente determinada por la posición más o menos próxima de los dos nulos adyacentes a la dirección de apuntamiento. Analíticamente, si se quiere que la respuesta de un array lineal equiponderado con apuntamiento arbitrario tenga al menos un cero en el intervalo $\theta \in [0, \pi]$, se debe cumplir que:

$$\left| \cos \theta_{\text{nulo}} \right| \leq 1 \Rightarrow \left| \cos \theta_0 \mp \frac{\lambda}{I \Delta z} \right| \leq 1 \quad (76)$$

Así para el array equiponderado en configuración *broadside*, se debe cumplir:

$$\frac{\lambda}{I \Delta z} \leq 1 \text{ ó } f \geq \frac{c}{I \Delta z} \quad (77)$$

Para el apuntamiento *endfire* la expresión (76) obliga a usar el signo menos, entonces:

$$\left| 1 - \frac{\lambda}{I \Delta z} \right| \leq 1 \Rightarrow \frac{\lambda}{I \Delta z} \leq 2 \text{ ó } f \geq \frac{c}{2 I \Delta z} \quad (78)$$

Siguiendo los mismos razonamientos originados en (72), puede establecerse la anchura del lóbulo principal o lo que es lo mismo la distancia angular $\Delta\theta$ entre los dos nulos más próximos a θ_0 . Las frecuencias Ω correspondientes a los dos nulos más próximos al eje principal se determinan a partir de (72) según:

$$\begin{aligned} \Omega_{\text{nulo-1}} &= \Omega_0 - \frac{2\pi}{I \Delta z} \\ \Omega_{\text{nulo-2}} &= \Omega_0 + \frac{2\pi}{I \Delta z} \end{aligned} \quad (79)$$

estando limitados esos dos valores al intervalo de convergencia de Ω .

A la hora de determinar $\Delta\theta$ existen tres casos:

- 1.- Si $-k \leq \Omega_{\text{nulo-1}} \leq k$ y $-k \leq \Omega_{\text{nulo-2}} \leq k$ –Figura 9(a)– entonces

$$\Delta\theta = \theta_{\text{nulo-2}} - \theta_{\text{nulo-1}} = a \cos\left(\cos\theta_0 - \frac{\lambda}{I \Delta z}\right) - a \cos\left(\cos\theta_0 + \frac{\lambda}{I \Delta z}\right) \quad (80)$$

- 2.- Si $\Omega_{\text{nulo-1}} \leq -k$ –Figura 9(b)– entonces:

$$\Delta\theta = 2\theta_{\text{nulo-2}} = 2 a \cos\left(\cos\theta_0 - \frac{\lambda}{I \Delta z}\right) \quad (81)$$

- 3.- Si $\Omega_{\text{nulo-2}} \geq k$ entonces

$$\Delta\theta = 2(\pi - \theta_{\text{nulo-1}}) = 2 \left[\pi - a \cos\left(\cos\theta_0 + \frac{\lambda}{I \Delta z}\right) \right] \quad (82)$$

Por tanto las expresiones (80), (81) y (82) permiten calcular la anchura $\Delta\theta$ del lóbulo principal de captación del array.

Como casos particulares de especial interés, la anchura del lóbulo principal se establece según (80) en

$$\Delta\theta_{\text{bs}} = a \cos\left(-\frac{\lambda}{I \Delta z}\right) - a \cos\left(\frac{\lambda}{I \Delta z}\right) = \pi - 2 a \cos\left(\frac{\lambda}{I \Delta z}\right) \quad \text{para } \textit{broadside} (\theta_0 = 90^\circ) \quad (83)$$

y según (81) en

$$\Delta\theta_{\text{ef}} = 2\theta_{\text{nulo-2}} = 2 a \cos\left(1 - \frac{\lambda}{I \Delta z}\right) \quad \text{para } \textit{endfire} (\theta_0 = 0^\circ) \quad (84)$$

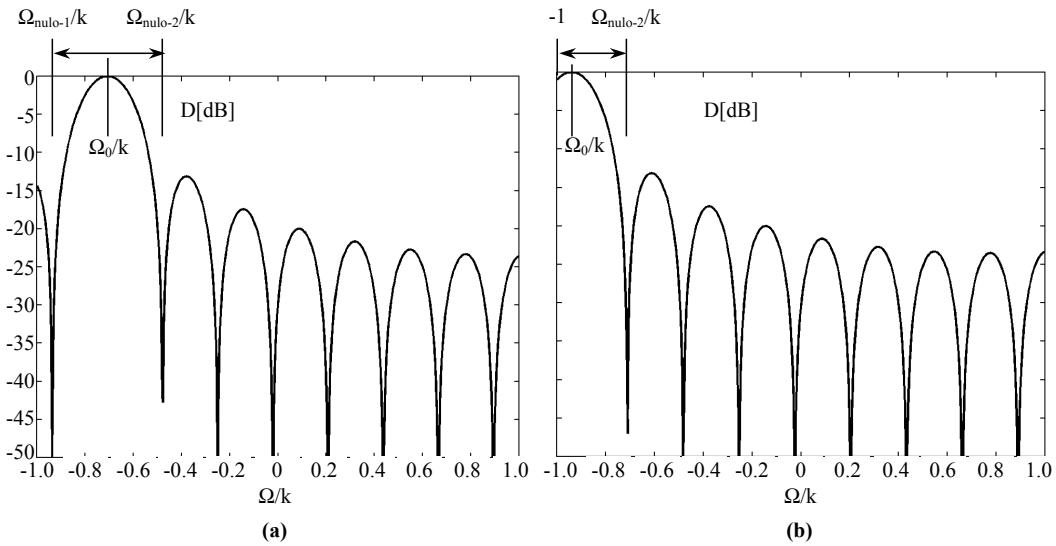


Figura 9. Determinación de la anchura del lóbulo principal para un array lineal uniforme equiponderado. **(a)** Los dos primeros nulos θ_{nulo-1} y θ_{nulo-2} pertenecen al intervalo $[0, \pi]$. **(b)** Sólo uno de los nulos pertenece al intervalo $[0, \pi]$.

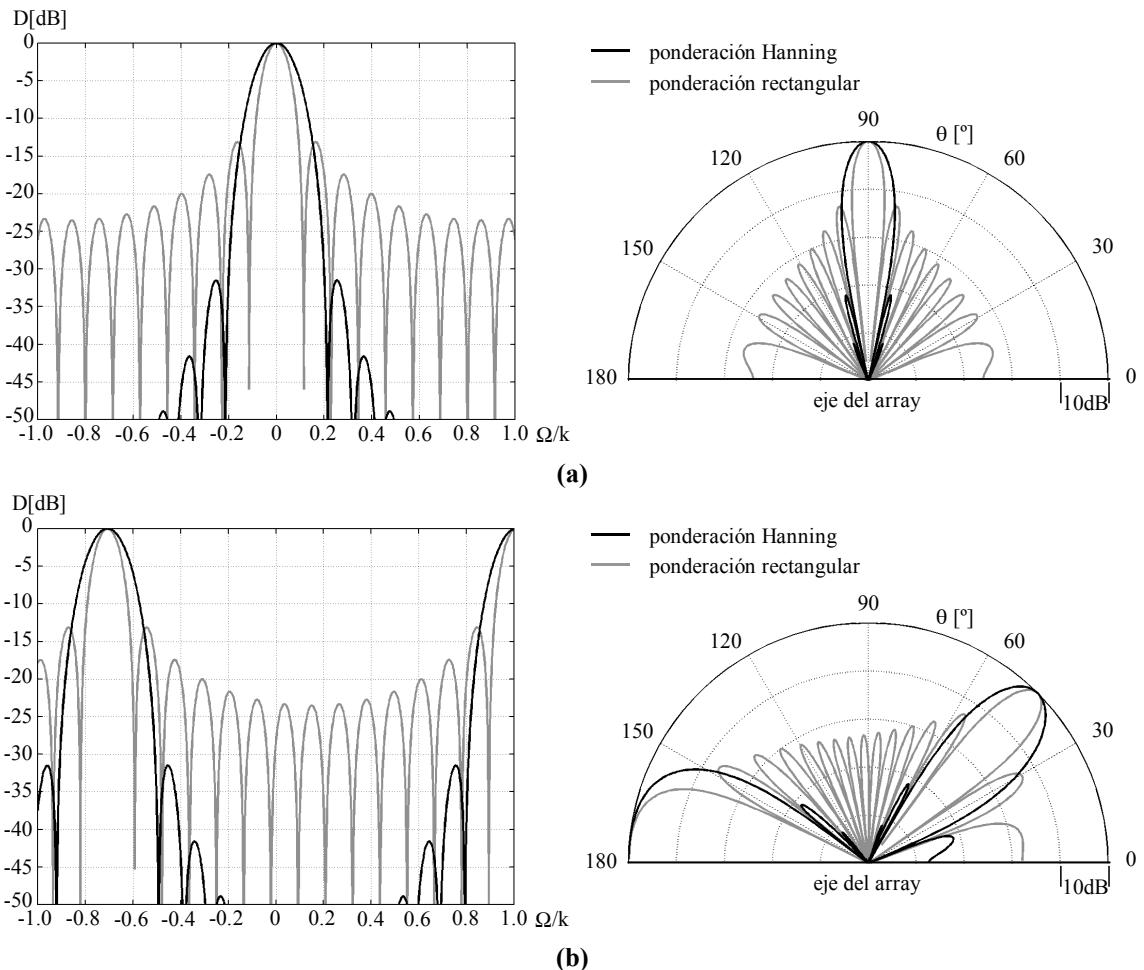


Figura 10. Efecto de la ponderación de los micrófonos del array proporcionada por los coeficientes C_i . Se muestra el efecto de las ponderaciones Hanning y rectangular en la directividad $D(\Omega)$ (izquierda) y $D(\theta)$ (derecha) para un array lineal uniforme de 15 micrófonos, con $\Delta z=0.08\text{m}$ y $f_0=2.5\text{kHz}$ ($\Omega_s/k=1.7$). **(a)** Apuntamiento *broadside* ($\Omega_0/k=0$ ó $\theta_0=90^\circ$). **(b)** Apuntamiento lateral ($\Omega_0/k=-0.71$ ó $\theta_0=45^\circ$).

El exhaustivo tratamiento anterior sólo es válido para una ponderación rectangular de los elementos microfónicos del array. Efectivamente, si la ponderación no es rectangular la transformada de Fourier de $f(z)$ no será la función sinc, y las posiciones de los lóbulos y número de éstos en el diagrama polar serán diferentes. En la Figura 10 se muestra cómo esto es así. La ponderación Hanning aplicada a los micrófonos del array induce menor número de lóbulos secundarios, pero al mismo tiempo, todos estos lóbulos, incluido el principal son angularmente más amplios. Un ejemplo práctico puede encontrarse en [Abe 84].

3.- Diferencia [dB] entre el lóbulo principal y el primer secundario

Esta característica viene determinada por la forma de la envolvente, o ventana de ponderación, de la secuencia $f(z)$ que representa al array, es decir por el módulo de los coeficientes C_i .

Interesa que la diferencia de amplitud entre lóbulos sea grande para que el array sea selectivo espacialmente. En un mismo array lineal uniforme hay cierto tipo de ventanas espaciales de ponderación que producen, como es sabido por teoría de la señal [Harris 78], una buena atenuación del primer lóbulo secundario con respecto al principal. Sin embargo, a medida que la diferencia de amplitud entre lóbulos aumenta, también lo hace la amplitud angular $\Delta\theta$ de los mismos. Hay que encontrar por tanto, una solución de compromiso entre diferencia de nivel y amplitud angular de los lóbulos de captación del array. En concreto, la ponderación tipo Hanning (Figura 10) mantiene un buen compromiso entre estas dos características que se relacionan de forma antagónica con la selectividad espacial del array.

2.2 ARRAY DE BANDA ANCHA

Un array microfónico está diseñado habitualmente para captar el espectro de audio, especialmente la voz humana. El ancho de banda relativo de la señal de habla es muy amplio, abarcando varias octavas de frecuencia, lo que constituye un inconveniente para el *beamforming*. Como se ha visto, las características directivas de un array son muy dependientes de la frecuencia, expresada por el número de onda k . Al aumentar k , el dominio de variación de la frecuencia angular Ω aumenta –compárese por ejemplo, la Figura 7 con la Figura 8(a), en las que se representa el mismo array a dos frecuencias diferentes– con lo que la directividad expresada por $D(\Omega)$ contiene mayor número de lóbulos y éstos son más estrechos. Además, al aumentar la frecuencia, puede aparecer también el fenómeno indeseable del *aliasing* espacial. Si se alcanza una frecuencia en la que se produce *aliasing* espacial, una buena manera de evitarlo, es aumentar la frecuencia de muestreo angular (Ω_s) mediante la disminución de la separación intermicrofónica Δz (o periodo espacial). Esto hace que se deban emplear arrays cada vez más pequeños (al disminuir Δz) a medida que la frecuencia aumenta. Por otra parte un array muy pequeño funciona mal en baja frecuencia. En este caso, la pulsación de muestreo angular Ω_s puede ser excesivamente grande, de modo que en el dominio de existencia de $D(\Omega)$, desde $\Omega = -k$ hasta $\Omega = +k$ apenas pueda contenerse una fracción del lóbulo principal de la transformada de Fourier del array $f(z)$. Como consecuencia el array será muy poco directivo. La escasa directividad en baja frecuencia puede corregirse con un aumento del número de micrófonos –compárese la Figura 8(a) con la Figura 8(b)– para elevar la cantidad de lóbulos y la selectividad espacial de los mismos. En cualquier caso queda claro, que a la hora de diseñar un array que funcione de forma uniforme en un espectro amplio, debe adoptarse alguna estrategia que trate de forma distinta las bajas frecuencias y las

altas frecuencias, con el objetivo de configurar lo que es conocido como array de banda ancha. Un array de banda ancha será por tanto aquél que ofrezca una directividad igual, o aproximadamente igual, a todas las frecuencias. Existen muchas soluciones [Brandstein 01] para conseguir este propósito, aunque aquí se desarrollan sólo las más utilizadas en la práctica.

2.2.1 Array de directividad constante (CDB o *Constant Directivity Beamformer*)

Un array de directividad constante –CDB ó *Constant Directivity Beamformer* [Brandstein 01]– es aquél que presenta un patrón polar de directividad $D(\theta, \phi)$ con características similares en un ancho de banda extenso. La teoría de arrays de banda ancha fue introducida en [Doles 88] y consolidada finalmente en [Ward 95].

Según Ward [Ward 95], la directividad de un array discreto de receptores, cualquiera que sea su disposición geométrica, se puede hacer independiente de la frecuencia, expresada por el número de onda k . Para ello basta con elegir adecuadamente los pesos $C_i(k)$ asociados a cada micrófono del array. Como se ve, a partir de este momento se admite que los coeficientes puedan variar con la frecuencia, ya que en el punto anterior se habían considerado constantes para cada uno de los micrófonos del array.

Si se considera el valor de los coeficientes $C_i(k)$:

$$C_i(k) = k \Delta z_i H_i(k) \exp(j\Omega_0 z_i) \quad (85)$$

con Δz_i la longitud del subintervalo i -ésimo de separación intermicrofónica,

$$\Delta z_i = \frac{z_{i+1} - z_{i-1}}{2} \quad (86)$$

y $H_i(k)$ la función de transferencia en frecuencia, asociada a un filtrado del micrófono i -ésimo, se puede demostrar que $F(\Omega)$ ó $F(\theta)$ –recuérdese que $F(\Omega)$ representa la transformada de Fourier de $f(z)$ y por tanto la directividad del array– dada por

$$F(\Omega) = \sum_{i=1}^I k \Delta z_i H_i(k) \exp(j\Omega_0 z_i) \exp(-j\Omega z_i) = F(\theta) \quad (87)$$

depende sólo del ángulo θ , si las funciones de transferencia $H_i(k)$ y $H_j(k)$ asociadas a dos micrófonos cualesquiera i y j se calculan de la siguiente manera,

$$H_i(k) = H_j \left(\frac{z_i}{z_j} k \right) \quad (88)$$

con z_i/z_j un factor de compresión o expansión del eje de frecuencia, asociado a cada micrófono. La ecuación (88) equivale a un escalado en frecuencia, de tal manera que, para mantener la directividad constante con k , a medida que la posición z_i del micrófono i se va haciendo mayor, la función de transferencia H se va comprimiendo hacia las bajas frecuencias. En (88) si $z_i > z_j$ la función de transferencia $H_j(k)$ equivale a una compresión en frecuencias de la función $H_i(k)$.

Implementación de un array lineal de directividad constante (CDB)

El procedimiento para construir este tipo de arrays está descrito en [Brandstein 01] y consiste en, partiendo de un array lineal uniforme que cubra sin *aliasing* espacial toda la banda de frecuencias prevista (puede ser logarítmico aunque aquí se propone uno uniforme), ir añadiendo a ambos lados micrófonos de forma logarítmica, con un $H(k)$ asociado cada vez más comprimido utilizando (88). Se supone que los micrófonos del array de partida no se filtran en frecuencia, aunque bajo cierto punto de vista puede considerarse que se les aplica un filtro paso bajo de frecuencia de corte igual al ancho de banda de trabajo, el de la señal de voz en este caso $-B = [0\text{Hz}, 8\text{kHz}]$ es el valor del ancho de banda de la voz utilizado en esta Tesis-. Entonces, la compresión en frecuencia de los nuevos micrófonos que se añaden a ambos lados, equivale a un filtrado paso bajo de frecuencia de corte cada vez menor. El procedimiento se resume en los siguientes pasos.

1.- Seleccionar un array lineal (puede ser uniforme o no) de partida, de longitud o apertura $Q/2$ veces la longitud de onda menor, asociada a la frecuencia mayor (Q es el factor de apertura). Si se elige un array uniforme:

$$z_n = n \Delta z \text{ con } n = \dots -2, -1, 0, 1, 2, \dots \text{ y } |n| \leq \frac{Q}{2} \quad (89)$$

$$\Delta z = \frac{1}{2} \lambda_{\min} = \frac{1}{2} \frac{c}{f_{\max}} = \frac{1}{2} \frac{2\pi}{k_{\max}}$$

Nótese que se ha introducido el índice n que puede adquirir valores negativos, cuando se refiere a los micrófonos situados en la parte negativa del eje Z . Por contra, recuérdese que el índice i utilizado anteriormente indicaba el orden del micrófono dentro del array, y sólo toma valores naturales. La separación intermicrofónica Δz del array lineal uniforme de partida evita la existencia de *aliasing* espacial para la más alta frecuencia de trabajo, ya que sustituyendo el valor Δz de (89) en la definición (69), la frecuencia de muestreo angular vale $\Omega_s = 2\pi/\Delta z = 2k_{\max}$, y según (70) se impide esta posibilidad para cualquier ángulo θ_0 de apuntamiento del array (apartado 2.1.8. de esta Tesis). El factor de apertura Q está relacionado con la selectividad espacial del array de partida y por tanto del array CDB. Si Q es pequeña la selectividad espacial del array será pequeña y viceversa.

2.- Añadir logarítmicamente micrófonos a ambos lados del array lineal uniforme de partida, en las posiciones:

$$z_{n+1} = \frac{Q}{Q-1} z_h \text{ con } n = \dots -2, -1, 0, 1, 2, \dots \text{ y } |n| > \frac{Q}{2} \quad (90)$$

$$\text{hasta } z_n < (Q-1) \frac{\lambda_{\max}}{2} = (Q-1) \frac{c}{2 f_{\min}}$$

siendo f_{\min} la frecuencia menor desde la que se quiere directividad constante. Evidentemente, si f_{\min} es muy pequeña será necesario añadir muchos micrófonos, y el array CDB resultante será muy grande

3.- Cada micrófono añadido en los extremos se pondrá con un coeficiente C_n de módulo:

$$|C_n| = k \Delta z_n H_n(f) \quad (91)$$

siendo Δz_i la separación correspondiente al micrófono n -ésimo, definida en (86) (cambiando el índice, de i á n) y $H_n(f)$ un filtro paso bajo asociado a cada uno de ellos, de frecuencia de corte

$$f_n = \frac{Q c}{2 |z_n|} \quad (92)$$

La fase de los coeficientes C_n determina el apuntamiento, para todas las frecuencias, del lóbulo principal del array, según fue expresado en (85), donde $\Omega_0 = -k \cdot \cos\theta_0$. Es decir, se puede configurar a voluntad la dirección de apuntamiento.

Ejemplo: diseño de un array CDB de 15 micrófonos

Se elige la frecuencia máxima de trabajo $f_{max} = 8\text{kHz}$, y la apertura del array lineal de base $Q = 2$. Por (89) $\Delta z = 0.021\text{m}$. El array lineal de partida (Figura 11) tendrá tres elementos situados en $z_{-1} = -1 \cdot \Delta z$ ($n = -1$), $z_0 = 0 \cdot \Delta z$ ($n = 0$) y $z_1 = 1 \cdot \Delta z$ ($n = 1$). Aplicando (90) se siguen colocando micrófonos hasta completar el último elemento del array que vendrá determinado por la frecuencia menor de trabajo. En concreto, si se quiere $f_{min} = 125\text{Hz}$, el z_n máximo obtenido según (90) será de $64 \cdot \Delta z$, y el array quedará finalmente configurado como en la Figura 11.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
n	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
z _n [m]	-64Δz	-32Δz	-16Δz	-8Δz	-4Δz	-2Δz	-1Δz	0Δz	1Δz	2Δz	4Δz	8Δz	16Δz	32Δz	64Δz
f _n [kHz]	0.25	0.5	1	2	4	8	16	∞	16	8	4	2	1	0.5	0.25

Figura 11. Array de directividad constante (CDB) de 15 micrófonos.

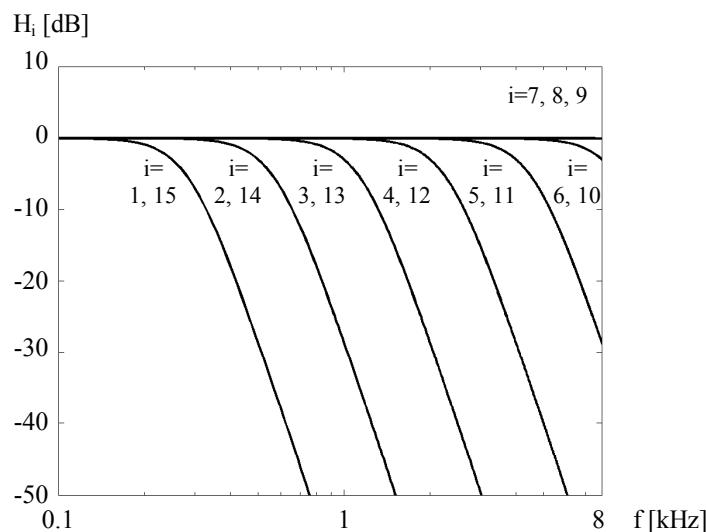


Figura 12. Filtros paso bajo $H_i(f)$ asociados a los micrófonos del array CDB diseñado en el ejemplo propuesto.

En la Figura 11 se dan también las frecuencias de corte de los filtros paso bajo $H_n(f)$ utilizados para cada micrófono. Si se quiere obtener una respuesta en frecuencia plana en el eje $\theta = \theta_0$ de apuntamiento del array [Brandstein 01], estos filtros no deben ser excesivamente abruptos a la hora de alcanzar la banda atenuada. En la Figura 12 se muestran las funciones de transferencia de los filtros seleccionados para este ejemplo (en este caso se utiliza el índice i). Los filtros son de orden 6 y tienen una inflexión suave desde la banda de paso a la banda atenuada.

En la Figura 13 se representa la directividad $D(\theta, f)$ en forma de mapa de grises del array diseñado en el ejemplo. Puede observarse cómo se mantiene la directividad aproximadamente constante en la banda de 125Hz a 8kHz y cómo la respuesta en frecuencia del array es también constante para la dirección *broadside* de apuntamiento, donde éste presenta una captación máxima.

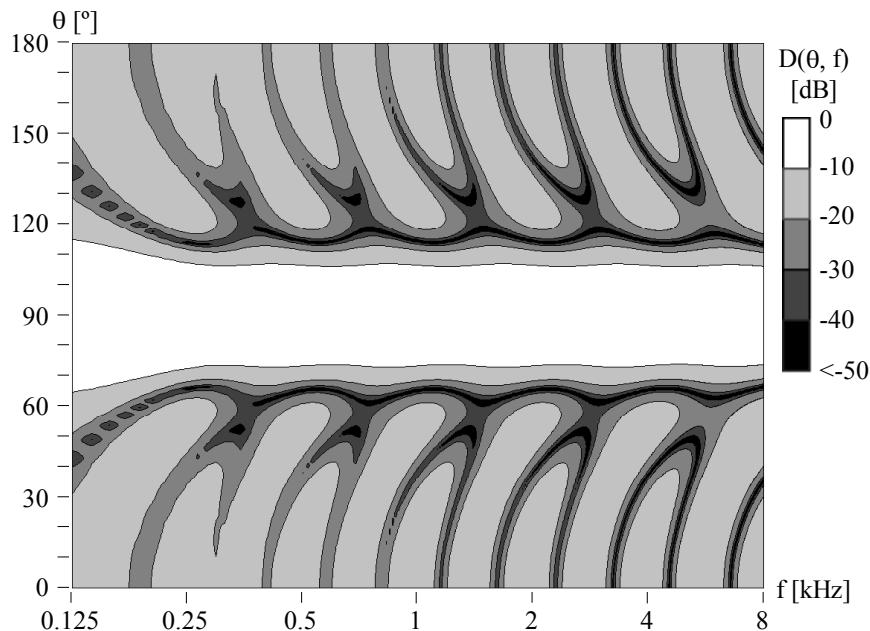


Figura 13. $D(\theta, f)$ del array de directividad constante (CDB) de 15 micrófonos diseñado en el ejemplo propuesto, siguiendo la Figura 11 y con los filtros de la Figura 12. El array está apuntado hacia $\theta_0=90^\circ$ (*broadside*).

2.2.2 Array anidado

En el apartado anterior se ha descrito la construcción de un array lineal mediante la adición, a ambos lados de un array lineal uniforme, de micrófonos distribuidos logarítmicamente. Se puede construir también una respuesta directiva constante a partir del anidamiento de varios arrays lineales uniformes de longitud cada vez mayor. En [Mahieux 96] se aplican estas ideas a una implementación práctica en un procesador digital.

En un array lineal uniforme, si a medida que aumenta la frecuencia f (o el número de onda k) se va aumentando la pulsación de muestreo angular Ω_s , mediante una disminución de la distancia intermicrófónica Δz , se estará manteniendo la directividad constante. En concreto, para un mismo número de micrófonos I , dos arrays A y B tendrán la misma directividad a dos

frecuencias diferentes ω_1 y ω_2 si sus distancias intermicrofónicas Δz_A y Δz_B se relacionan de forma inversa que las frecuencias, es decir

$$D_A(\omega_1) = D_B(\omega_2) \quad \text{si} \quad \frac{\Delta z_B}{\Delta z_A} = \frac{\omega_1}{\omega_2} \quad (93)$$

Entonces, si se quiere que un array lineal tenga directividad constante se tiene que hacer una división en subbandas de frecuencia. A medida que la banda sea de mayor frecuencia, menor será la separación intermicrofónica Δz requerida.

Sea NB el número de subbandas de frecuencia en las que se va a dividir el espectro de audio, entonces será necesario tener NB arrays lineales uniformes. Si cada array tiene I_0 micrófonos, el número total de micrófonos será de $I = I_0 \times NB$. Si la división del espectro se hace en forma de bandas de $1/n$ octava ($1/1$ octava, $1/3$ octava...) es posible que algunos micrófonos puedan ser comunes para varias bandas con lo que $I < I_0 \times NB$. Es lo que se llama array anidado [Flanagan 91] [Kellerman 91] [Khalil 94]. En realidad un array anidado es un caso particular del array logarítmico de directividad constante (CDB) descrito en el apartado 2.2.1, puesto que al decrecer la frecuencia de cada subbanda, los micrófonos implicados van separándose cada vez más, y por tanto se distribuyen logarítmicamente alrededor del subarray asociado a la subbanda de mayor frecuencia. En concreto un array CDB para una apertura $Q = 2$ (el ejemplo de la Figura 11) tiene una distribución geométrica de los micrófonos igual a un array anidado de 7 subarrays y tres micrófonos cada uno.

Sin embargo, el tratamiento en frecuencia (filtros H_i) del array CDB es diferente al considerado para un array anidado. La diferencia estriba en que los filtros asociados al array anidado son de tipo paso banda para seleccionar adecuadamente cada una de las subbandas implicadas. Por el contrario en un array CDB como el de la Figura 11, estos filtros son paso bajo. En la práctica, esto comporta diferencias poco significativas en las curvas de directividad obtenidas con ambos métodos, para distribuciones geométricas iguales de los micrófonos en ambos casos.

Ejemplo: diseño de un array anidado de 5 bandas de octava

Se quiere construir un array anidado en bandas de tal manera que se cubran las cinco octavas de frecuencia central desde 250Hz hasta 4kHz (desde $250 / \sqrt{2} = 177$ Hz hasta $250 \cdot \sqrt{2} = 5657$ Hz), por lo que se necesitarán 5 octavas para cubrir el espectro requerido. Si se elige la separación intermicrofónica de 4cm para la banda de más alta frecuencia se tendrá:

$$\Delta z_{4k} = 0.04m, \Delta z_{2k} = 0.08m, \Delta z_{1k} = 0.16m, \Delta z_{500} = 0.32m, \Delta z_{250} = 0.64m$$

Si se desea construir un array de dimensiones manejables, la separación máxima (baja frecuencia) limita mucho el número de micrófonos I_0 que se pueden utilizar por banda. En este caso como $\Delta z_{250} = 0.64m$, se podría construir cada subarray con 7 elementos ($I_0 = 7$), con lo que el tamaño ($I_0 - 1$) del array anidado sería $\Delta z_{250} = 3.84m$, correspondiente a la subbanda de baja frecuencia. En la Figura 14 se representa la configuración geométrica del array anidado diseñado en este ejemplo.

Atendiendo a la Figura 14, el número necesario de micrófonos para este ejemplo es de 23, y no de 35 (7×5), que serían los requeridos si no se anidasesen las subbandas, es decir si no existiesen micrófonos comunes en bandas de frecuencia diferentes. En efecto, el micrófono central del array (que ocupa el lugar duodécimo), es común para las cinco bandas de octava.

Además existen otros micrófonos comunes a dos bandas diferentes, con el consiguiente ahorro de receptores.

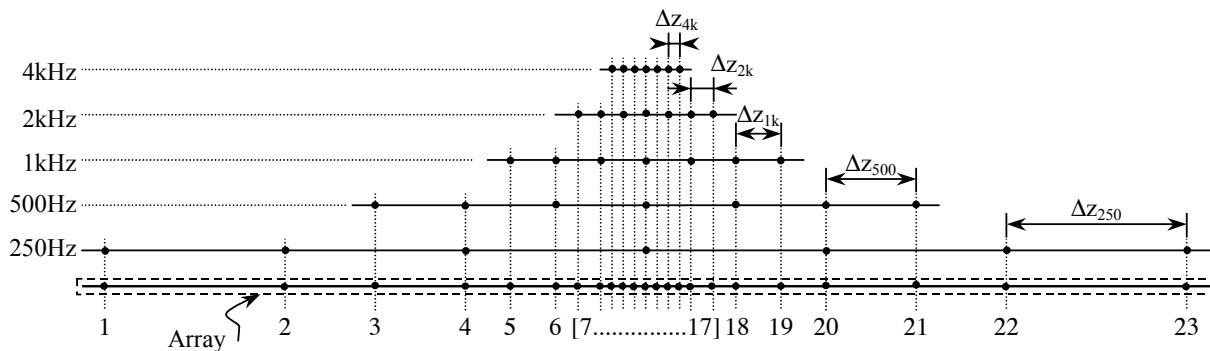


Figura 14. Array anidado de 5 subbandas de 1/1 octava.

En la Figura 15 se representa, en forma de mapa de niveles, la directividad $D(\theta, f)$ del array propuesto en el ejemplo. Cada uno de los cinco subarrays constituye un array lineal uniforme en una banda de octava. Dentro de cada banda de octava la directividad no es uniforme, pero la directividad del array total es la misma cuando se comparan dos frecuencias cualesquiera separadas por una octava. Su comportamiento directivo es por tanto peor que el array logarítmico puro presentado anteriormente (array CDB de la Figura 11) ya que en ese caso la directividad no tenía ninguna variación local, como le pasa al array anidado dentro de cada subbanda de frecuencia.

Se puede mejorar la uniformidad, dentro de cada subbanda, de la directividad del array anidado –Figura 15 (b)– aplicando unos filtros de subbandas no excesivamente abruptos con las mismas pendientes de atenuación que las mostradas anteriormente en la Figura 12. Este tipo de filtrado hace que se mezcle la parte más directiva de un determinada subbanda con la parte menos directiva de la siguiente, suavizando el mapa de directividad resultante.

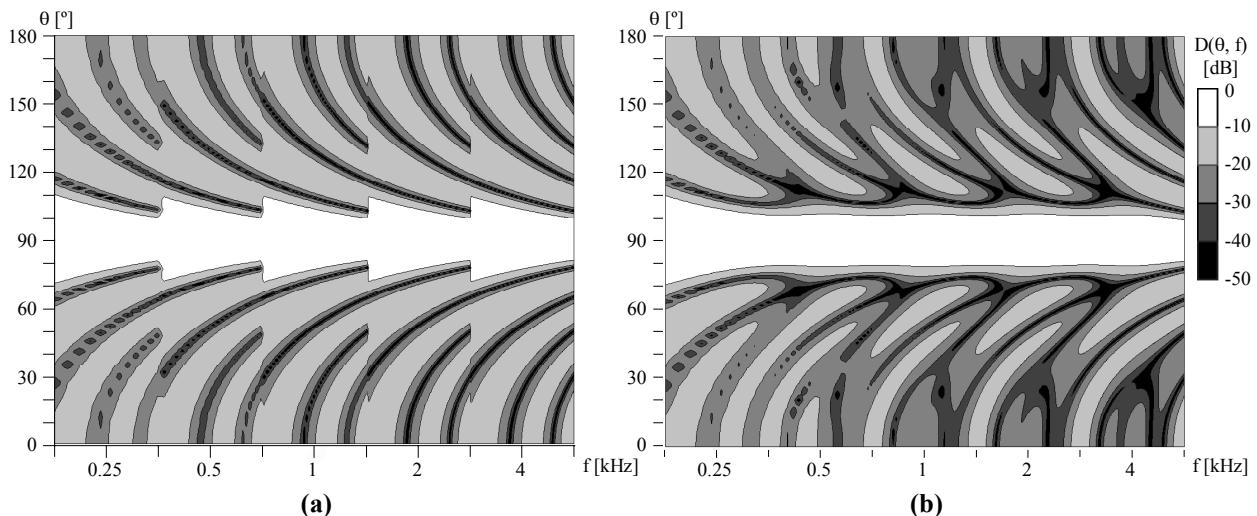


Figura 15. Directividad $D(\theta, f)$ de un array anidado de 5 octavas, apuntado hacia $\theta_0=90^\circ$ con Δz mínima ($f_0=250\text{Hz}$) de 0.04m. **(a)** Aplicando un filtrado infinitamente abrupto en cada subbanda. **(b)** Aplicando filtros de 6º orden en cada subbanda.

A pesar de que la directividad que produce un array anidado es muy uniforme, está muy limitada en frecuencia, sobre todo en la parte baja del espectro, que es precisamente donde resulta más interesante una elevada selectividad espacial para tratar la señal de voz. Si se desea aumentar la directividad en baja frecuencia se requiere diseñar un array de dimensiones muy elevadas. Este inconveniente también lo tiene el array CDB visto en el punto 2.2.1. Por tanto, el principal problema que presentan las configuraciones de las Figuras 11 y 14 es su gran tamaño, necesario para tratar las bajas frecuencias con suficiente selectividad espacial. Se necesita otro esquema de diseño de un array microfónico, que aumenten la directividad en baja frecuencia sin que para ello sea preciso un array de tamaño mayor a 1 metro, por ejemplo. A continuación se describe la solución más comúnmente utilizada para solventar este problema.

2.2.3 Array superdirective

Los esquemas básicos de conformación de haz tratados hasta ahora se basan en la maximización, utilizando (47), de la potencia de salida del array en un escenario en el que una fuente única está situada en una determinada dirección espacial (r_0, θ_0, φ_0), pero esta condición no exige nada especial al resto de las direcciones del espacio. La maximización de la respuesta en la dirección principal del array conduce a las estructuras convencionales de retardo y suma desarrolladas en los apartados anteriores, en las que las fases de los coeficientes w_i asociados a cada micrófono tan sólo compensan el retardo acústico entre la fuente sonora y el array, pero no se ocupan de configurar una determinada directividad, que sea suficientemente elevada en baja frecuencia. En efecto, el principal inconveniente de la estructura de retardo y suma es la baja selectividad espacial en la parte baja del espectro de audio, lo que obliga a diseñar arrays excesivamente grandes.

Se puede establecer una configuración de coeficientes diferente a (47) mediante una nueva condición que mejore la respuesta directiva del array. La nueva condición consiste precisamente en minimizar la potencia de salida del array para aquellas direcciones de llegada diferentes a la principal. Es decir, ahora no sólo se necesita que la respuesta en la DOA principal sea máxima, sino además será necesario que la respuesta en las demás direcciones espaciales sea mínima, o cuando menos suficientemente pequeña para atenuar el ruido y la reverberación de procedencia no axial.

El principio de funcionamiento de un array superdirective se entiende muy bien cuando se aplica a un dipolo acústico. Imagíñese un par de micrófonos –Figura 16(a)– separados una distancia Δz , apuntando en configuración *end-fire* ($\theta = 0^\circ$) y operando en baja frecuencia. La estructura de retardo y suma tan sólo asegura que la potencia en la DOA principal sea máxima, mediante la compensación del retardo acústico. Si $\Delta z \ll \lambda$, la selectividad espacial será prácticamente nula, con lo que el array-dipolo será omnidireccional. La cuestión ahora es la siguiente. ¿Existe alguna combinación de los dos micrófonos que sea más directiva que la configuración anterior, pero que además siga teniendo un máximo de captación en $\theta = 0^\circ$? Efectivamente así es, de hecho existen muchas de estas combinaciones. Por ejemplo, si se restan las respuestas de los dos micrófonos sin retardar, resulta un patrón polar en forma de 8 o bidireccional –Figura 16(b)–, muy conocido en microfonía convencional, apuntado a la dirección $\theta = 0^\circ$ deseada. Existen otras combinaciones de los dos micrófonos que consiguen también respuestas más directivas que la configuración de retardo y suma, por ejemplo se puede obtener [Sánchez-Bote 02-a] un diagrama polar cardioide, supercardioide, etc., siempre mediante combinaciones de sumas, restas y retardos de los micrófonos. De forma más general, en un array con más de dos elementos, es posible buscar combinaciones de

micrófonos de tal manera que, manteniendo una respuesta máxima en el eje principal de captación, el array tenga una respuesta global mínima en las otras direcciones espaciales, que es por donde llegará preferentemente el ruido, la reverberación u otras fuentes presentes, ajenas a la señal de voz que se quiere mejorar y considerados genéricamente como perturbación acústica.

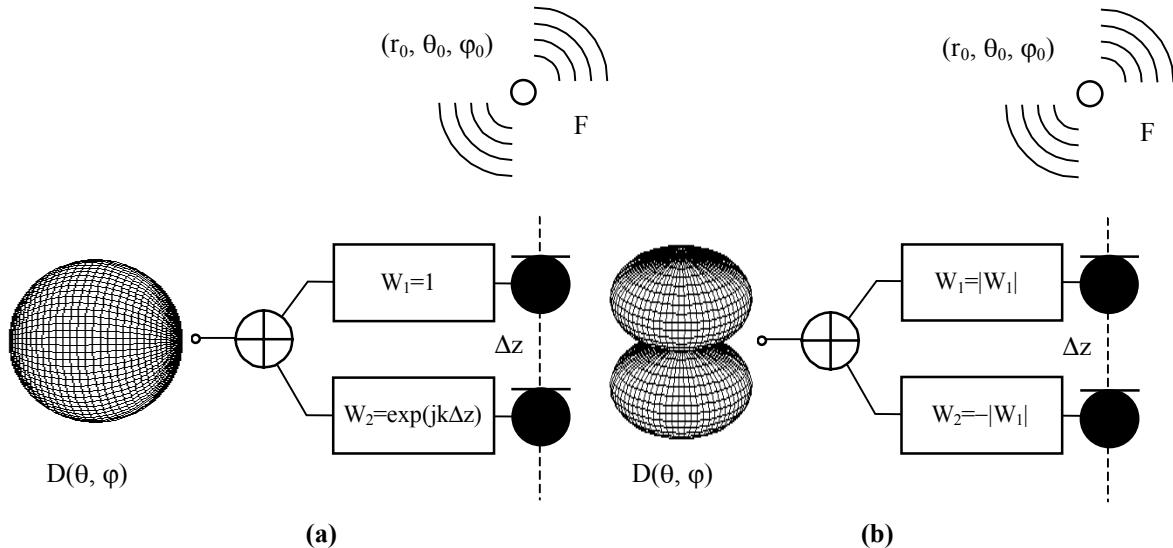


Figura 16. Comportamiento de un array de dos micrófonos (dipolo acústico) en baja frecuencia. **(a)** Estructura de retardo y suma. **(b)** Dipolo bidireccional.

El principal problema asociado a la configuración de la Figura 16(b) es que, al haber una resta de por medio, la amplitud de la señal proporcionada por el array para incidencia desde θ_0 es muy baja. Si se quiere mantener la respuesta en el eje principal del array en un valor suficiente, se deberá amplificar bastante cada uno de los micrófonos mediante los filtros $|W_i(\omega)|$. Esta amplificación excesiva se debe producir en baja frecuencia, como se muestra a continuación, aumentando el ruido externo de procedencia acústica o el ruido interno de procedencia eléctrica (ruido propio de los micrófonos) que por su baja coherencia intercanal no sea eliminado por la combinación de los elementos del array.

Puesto que el objetivo de una elevada selectividad espacial es eliminar ruido o reverberación de procedencia no axial, y puesto que un array solo puede ser selectivo cuando capta señales con elevada coherencia intercanal, el objetivo de un array superdirective será el de ofrecer elevada directividad sólo cuando las señales acústicas no deseadas tengan elevada coherencia intercanal. Por eso se suele supeditar la condición de mínima captación no axial a la existencia de un campo acústico de ruido con determinadas características de coherencia espacial.

La teoría de superdirectividad se conoce desde mediados del siglo XX y no ha sido aplicada a arrays de micrófonos hasta la década de los 90 [Elko 00].

Diseño de un array superdirective

Se supone ahora que cada uno de los micrófonos del array está captando cierta cantidad de ruido $n_i(t)$ –que constituye el vector $\mathbf{n}(t)$ –, como se anticipó en el apartado 2.1.4 de esta Tesis. El vector de ruido $\mathbf{n}(t)$ (31) representa a todas las perturbaciones acústicas distintas de

la señal original, simbolizadas por el vector $\mathbf{x}(t)$. Incluye el ruido aditivo, la reverberación y las fuentes ajenas a la original, cuya participación en la respuesta del array es aditiva. Entonces se debe considerar la ecuación (28) (caso de una sola fuente incidente con presencia de ruido), y la salida del conformador de haz (44) se transforma en:

$$\mathbf{y}(t) = \sum_{i=1}^I w_i^* y_i(t) = \mathbf{w}^H \mathbf{y}(t) = \mathbf{w}^H \mathbf{a}(r, \theta, \phi) \mathbf{x}_0(t) + \mathbf{w}^H \mathbf{n}(t) \quad (94)$$

que expresada en el dominio de la frecuencia queda,

$$\mathbf{Y}(\omega) = \mathbf{W}^H(\omega) \mathbf{A}(\omega) \mathbf{X}_0(\omega) + \mathbf{W}^H(\omega) \mathbf{N}(\omega) \quad (95)$$

donde se han eliminado por simplicidad las dependencias angulares.

El objetivo principal en este momento es la minimización de la potencia de salida, por lo que hay que establecer las siguientes definiciones. La densidad espectral de potencia (autoespectro) de la salida conformada del array viene dada por:

$$\Phi_{YY}(\omega) = \mathbf{W}^H(\omega) \Phi_{YY}(\omega) \mathbf{W}(\omega) \quad (96)$$

con $\Phi_{YY}(\omega)$ la matriz de espectros cruzados intermicrofónicos captados por el array.

El diseño de un array superdirective consiste en la búsqueda del juego de coeficientes microfónicos $\mathbf{W}(\omega)$ que minimice la potencia de salida del array $-\Phi_{YY}(\omega)$ de (96) – con la condición de que dicha salida –representada aquí por la referencia $x_0(t)$ – salga sin distorsión, es decir,

$$\min_{\mathbf{W}} \mathbf{W}^H(\omega) \Phi_{YY}(\omega) \mathbf{W}(\omega) \text{ con la condición de que } \mathbf{W}^H(\omega) \mathbf{A}(\omega) = 1 \quad (97)$$

La solución de este problema de minimización se llama Respuesta de Mínima Varianza sin Distorsión (MVDR ó *Minimum Variance Distortionless Response*) o Mínima Varianza Linealmente Restringida (LCMV ó *Linearly Constrained Minimum Variance*) y el resultado es bien conocido [Cox 87] [Van Been 88]. En lo sucesivo será referenciado indistintamente como conformador superdirective o MVDR. El conjunto de coeficientes solución de (97) es:

$$\mathbf{W}_{SD}(\omega) = \frac{\Phi_{NN}^{-1}(\omega) \mathbf{A}(\omega)}{\mathbf{A}^H(\omega) \Phi_{NN}^{-1}(\omega) \mathbf{A}(\omega)} \quad (98)$$

con $\Phi_{NN}(\omega)$ la matriz de espectros cruzados de ruido,

$$\Phi_{NN}(\omega) = \begin{pmatrix} \Phi_{N_1 N_1}(\omega) & \Phi_{N_1 N_2}(\omega) & \cdots & \Phi_{N_1 N_I}(\omega) \\ \Phi_{N_2 N_1}(\omega) & \Phi_{N_2 N_2}(\omega) & \cdots & \Phi_{N_2 N_I}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{N_I N_1}(\omega) & \Phi_{N_I N_2}(\omega) & \cdots & \Phi_{N_I N_I}(\omega) \end{pmatrix} \quad (99)$$

Cuando el ruido $\mathbf{N}(\omega)$ es homogéneo –todos los micrófonos del array reciben la misma potencia de ruido es decir $\Phi_{N_i N_j}(\omega) = \Phi_{N_j N_i}(\omega) \forall i, j$ –, la solución anterior se simplifica como:

$$\mathbf{W}_{SD}(\omega) = \frac{\Gamma_{NN}^{-1}(\omega) \mathbf{A}(\omega)}{\mathbf{A}^H(\omega) \Gamma_{NN}^{-1}(\omega) \mathbf{A}(\omega)} \quad (100)$$

con $\Gamma_{NN}(\omega)$ la matriz coherencia de ruido,

$$\boldsymbol{\Gamma}_{\text{NN}}(\omega) = \begin{pmatrix} 1 & \Gamma_{N_1 N_2}(\omega) & \cdots & \Gamma_{N_1 N_1}(\omega) \\ \Gamma_{N_2 N_1}(\omega) & 1 & \cdots & \Gamma_{N_2 N_1}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{N_1 N_1}(\omega) & \Gamma_{N_1 N_2}(\omega) & \cdots & 1 \end{pmatrix} \quad (101)$$

siendo $\Gamma_{N_i N_j}(\omega)$ la función coherencia [Carter 87] del ruido captado por los micrófonos i y j.

La expresión (100) resuelve el problema de la minimización sin distorsión del ruido n_i que llega a cada uno de los micrófonos. Ese ruido, entendido como señal no deseada puede asociarse al llamado ruido difuso (usando la terminología acústica), a la reverberación o incluso a la señal de voz producida por una fuente que quiera ser eliminada. En el último sentido, puede usarse el conformador MVDR como un filtro que elimine fuentes de voz situadas en posiciones relativamente cercanas a la DOA principal.

Otra ventaja del conformador superdirective sobre el convencional es su superior selectividad angular. El conformador convencional (47) con estructura de retardo y suma carece de buena resolución angular cuando se pretende separar dos fuentes con DOA muy próximas. Si la conformación de haz necesita discernir entre dos fuentes muy cercanas angularmente se prefiere el conformador MVDR. La potencia de salida del conformador en función del ángulo de apuntamiento θ_0 es conocida a veces como “espectro de salida” [Naidu 01]. Dicho espectro de salida es la directividad del array en unidades de potencia. El espectro de salida del conformador MVDR es conocido también como espectro de Capon [Capon 69] [Lacoss 71] ya que fue este autor el que primero propuso esa forma de obtener la respuesta de un array.

Supóngase que están presentes junto al array dos fuentes, con DOA's θ_1 y θ_2 . Si se quiere separar (o eliminar) la fuente procedente de la DOA θ_2 , se tendrá que crear un mínimo de captación en esta dirección y un máximo relativo en la primera dirección. Por tanto, la configuración óptima de los coeficientes $\mathbf{W}_{SD}(\omega)$ será aquélla que obtenga la mínima potencia global a la salida del array con la condición de que se mantenga fija, en un valor alto, la componente de la señal de salida correspondiente a la DOA principal. Ésta es la condición del conformador MVDR, que en esta situación configurará un patrón de directividad que ofrezca un mínimo abrupto en la posición θ_2 y un máximo relativo en la DOA θ_1 , para conseguir que la salida sea sin distorsión. Para que en este caso el conformador MVDR funcione correctamente, el ruido considerado en el vector $\mathbf{n}(t)$ debe incluir la información de la señal interferente en θ_2 .

La visión dada anteriormente a la conformación MVDR corresponde al objetivo de la eliminación de una fuente próxima a la fuente principal, considerada como ruido. Sin embargo, el enfoque más clásico se asocia a la eliminación de ruido acústico de tipo difuso. En efecto, a la vista de (100), la solución del array superdirective depende fuertemente de la coherencia espacial del ruido captado por los micrófonos. Si $\boldsymbol{\Gamma}_{\text{NN}}(\omega) = \mathbf{I}(\omega)$ se obtiene la solución de retardo y suma de (47). Es decir, si no existe coherencia espacial en el ruido (o el ruido captado por dos micrófonos diferentes es totalmente incoherente), prima la condición de no distorsión expresada por $\mathbf{W}_{SD}^H(\omega) \mathbf{A}(\omega) = 1$ y no importa que el array sea muy selectivo espacialmente en direcciones no axiales a la hora de conseguir una respuesta óptima. Cuando existe cierto grado de coherencia espacial en el ruido y $\boldsymbol{\Gamma}_{\text{NN}}(\omega) \neq \mathbf{I}(\omega)$, el numerador de (100) occasionará que el array fuerce la mayor directividad posible, incluso en baja frecuencia, usando fundamentalmente inversiones de signo del tipo a la de la Figura 16(b) para el par bidireccional, atenuándose el ruido con elevada coherencia intercanal mediante una mayor

selectividad espacial. El denominador de (100) es el responsable de que el módulo de los coeficientes $\mathbf{W}_{SD}(\omega)$ aumente cuando la selectividad espacial del array sea elevada, con el objetivo de seguir manteniendo la condición de no distorsión $\{\mathbf{W}_{SD}^H(\omega) \mathbf{A}(\omega) = 1\}$. Normalmente, la coherencia de ruido es alta en baja frecuencia lo que obliga a que el denominador de (100) sea muy pequeño, haciendo crecer el valor en módulo de los coeficientes $\mathbf{W}_{SD}(\omega)$. Por ello el conformador superdirectivo suele presentar problemas de funcionamiento en baja frecuencia, ya que la excesiva amplificación proporcionada aumenta las componentes con baja coherencia espacial de ruido presente no previsto, por ejemplo el ruido acústico incoherente o el ruido eléctrico interno de los micrófonos del array. Esta señal incoherente no puede ser eliminada por directividad y es muy amplificada por el conformador MVDR. Para aminorar esto existen las soluciones restringidas de conformación MVDR, como se explicará más adelante en este mismo apartado.

Para arrays de configuración fija (no adaptativos), es muy importante tener una buena estimación a priori de cuál va a ser la coherencia espacial del ruido o de la reverberación presentes en la sala de operación, expresada por la matriz $\Gamma_{NN}(\omega)$. Históricamente se ha considerado en este sentido el campo acústico de ruido isotrópico, por ser este tipo de ruido el que se puede equiparar con mayor aproximación a las condiciones de ruido y reverberación existentes con más frecuencia en una sala. Se considera que en una sala existe un ruido isotrópico si en cada punto del espacio se recibe una señal temporalmente incorrelada (aleatoria) procedente simultáneamente de todas las direcciones espaciales. En muchas ocasiones la naturaleza del ruido acústico y la reverberación existentes en una sala equivalen en primera aproximación a esa condición de ruido isotrópico. Los cálculos iniciales de correlación de ruidos acústicos captados por micrófonos omnidireccionales en situaciones reales fueron publicados por R.K. Cook *et al* en [Cook 55]. Como conclusión principal de esos trabajos se puede establecer la coherencia $\Gamma_{N_i N_j}(\omega)$ de la señal recibida por dos micrófonos omnidireccionales situados en las posiciones z_i y z_j cuando están captando un campo de ruido isotrópico como:

$$\Gamma_{N_i N_j}(\omega) = \frac{\frac{\sin\left(\frac{\omega}{c}|z_i - z_j|\right)}{\frac{\omega}{c}|z_i - z_j|}}{\text{sinc}\left(\frac{\omega}{\pi c}|z_i - z_j|\right)} = \text{sinc}\left(2 \frac{|z_i - z_j|}{\lambda}\right) \quad (102)$$

Según (102) la coherencia intercanal crece cuando la separación intermicrofónica se hace pequeña con relación a la longitud de onda (baja frecuencia), puesto que si esto es así disminuye el factor $|z_i - z_j|/\lambda$ y la función sinc aumenta. En la situación de alta coherencia espacial (baja frecuencia y Δz pequeño), el ruido coherente de direcciones no axiales se podrá cancelar haciendo inversiones de fase alternativas en los micrófonos del array, para que el numerador de (100) corresponda a una elevada selectividad espacial. Al mismo tiempo habrá que elevar el módulo de los coeficientes microfónicos $|\mathbf{W}_{SD}(\omega)|$ para compensar la pérdida de nivel que producen esas inversiones de fase. En alta frecuencia desaparece la coherencia intercanal del ruido y el array superdirectivo se convierte en el convencional que usa la configuración de retardo y suma.

Existen otros tipos de ruido susceptibles de consideración en el tratamiento en array de señales acústicas. Por ejemplo el ruido isotrópico cilíndrico [Cron 62], o incluso se puede sopesar la influencia de la directividad de los micrófonos, cuando éstos no sean omnidireccionales –capítulo 4 de [Brandstein 01]–.

Solución restringida

El elevado valor que adquiere el módulo de los coeficientes superdirectivos $\mathbf{W}_{SD}(\omega)$ en baja frecuencia hace imposible la implementación de un array superdirectivo adoptando la solución (100). En la práctica se utilizan soluciones restringidas (*constrained*), de tal manera que se limite la amplificación del ruido incoherente a un valor máximo. Esto se consigue con soluciones alternativas a (100) como la que se propone [Gilbert 55] a continuación:

$$\mathbf{W}_{SD}(\omega) = \frac{[\Gamma_{NN}(\omega) + \mu(\omega) \mathbf{I}(\omega)]^{-1} \mathbf{A}(\omega)}{\mathbf{A}^H(\omega) [\Gamma_{NN}(\omega) + \mu(\omega) \mathbf{I}(\omega)]^{-1} \mathbf{A}(\omega)} \quad (103)$$

El parámetro de restricción $\mu(\omega)$ impide que el denominador de (103) responsable del elevado valor de la respuesta $|\mathbf{W}_{SD}(\omega)|$ se haga demasiado pequeño, limitando la amplificación de baja frecuencia del conformador. Por tanto $\mu(\omega)$ permite optimizar la respuesta del array superdirectivo en baja frecuencia. Normalmente es constante con ω , aunque puede no serlo. En la práctica $\mu[\text{dB}]$ tiene un valor comprendido entre -10dB y -30dB siendo:

$$\mu[\text{dB}] = 10 \log \mu \quad (104)$$

Ejemplo de diseño de un array superdirectivo de 5 micrófonos

A continuación se propone el diseño de un array lineal uniforme de 5 micrófonos basado en superdirecividad, apuntado hacia $\theta_0 = 90^\circ$ en la aproximación de campo lejano $r_0 = \infty$ y con una separación intermicrofónica de $\Delta z = 0.04\text{m}$. Para ello se calculan los coeficientes $\mathbf{W}_{SD}(\omega)$ utilizando la solución no restringida (100) y considerando un ruido difuso con función de coherencia espacial según (102).

En la Figura 17(a) se representa el módulo de la respuesta de los filtros del conformador superdirectivo $W_{SDi}(\omega)$ asociados a cada micrófono y en la Figura 17(b) la fase de los mismos. Puede verse que la ganancia que hay que proporcionar a los micrófonos en baja frecuencia es muy elevada ($\approx 100\text{dB}$ aproximadamente!), lo que en la práctica es irrealizable ya que daría muchos problemas por la amplificación de ruido eléctrico de baja frecuencia producido por los micrófonos (ruido eléctrico interno), con poca coherencia intercanal. En baja frecuencia la diferencia de fase entre micrófonos consecutivos es siempre de 180° –Figura 17(b)–, de ahí que la directividad en baja frecuencia del array sea una combinación óptima de pares bidireccionales, lo que se puede constatar en la Figura 18(b), con lóbulos de captación máxima en $\theta = 90^\circ$ y $\theta = 180^\circ$. A partir de unos 4kHz la coherencia intercanal es muy pequeña y el array superdirectivo deriva a uno convencional de retardo y suma, con $|W_{SDi}(\omega)| = 1/I = 1/5$ (equivalente a -14dB).

En la Figura 18 se representa la directividad $D(\theta, f)$ en forma de mapa de grises del array de cinco micrófonos propuesto, en sus versiones de retardo y suma –Figura 18(a)– y superdirectiva –Figura 18(b)–. A partir de unos 4kHz las dos soluciones convergen ya que la coherencia espacial del ruido difuso para la distancia intermicrofónica considerada se hace prácticamente nula.

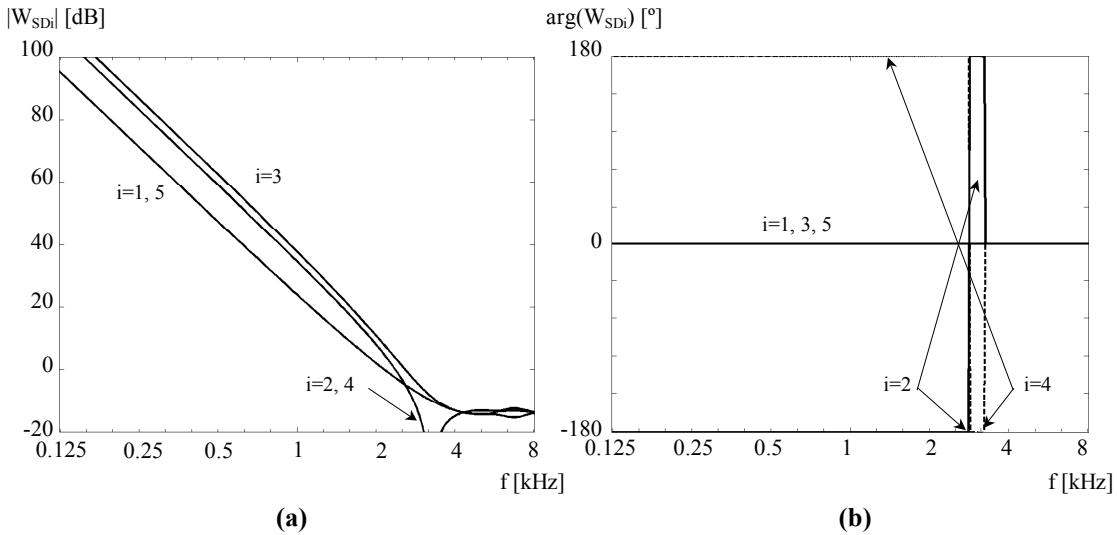


Figura 17. Módulo y fase de los filtros $W_{SDi}(\omega)$ asociados a cada micrófono del array superdirective de cinco micrófonos propuesto, y cuyo mapa de directividad se muestra en la Figura 18. Los coeficientes se han calculado utilizando la solución no restringida (100) considerando un ruido difuso con función de coherencia espacial según (102). **(a)** Módulo. **(b)** Fase.

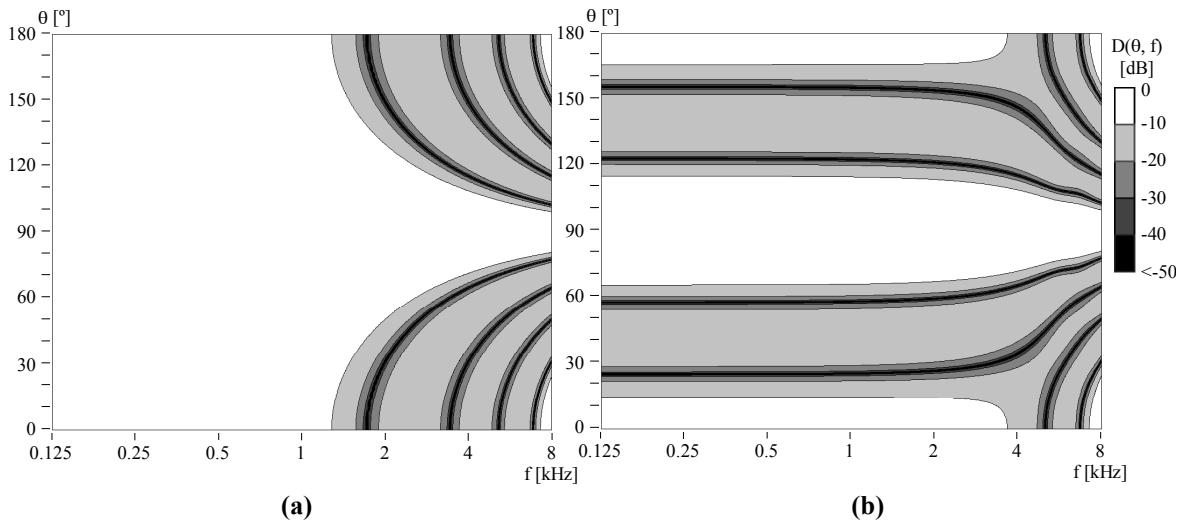


Figura 18. Mapa de directividad $D(\theta, f)$ de un array lineal uniforme de cinco micrófonos, apuntado hacia $\theta_0=90^\circ$ con $\Delta z = 0.04m$. **(a)** Estructura de retardo y suma con coeficientes $W_{RS}(\omega)$ según (47). **(b)** Array superdirective con coeficientes $W_{SD}(\omega)$ según (100) diseñado para un ruido difuso con función de coherencia espacial según (102).

Array superdirective mediante un conformador fijo más matriz de bloqueo

Una implementación alternativa [Buckley 86] del array superdirective representado por la expresión (100) y que se puede aplicar también a la solución restringida de (103) se ilustra en la Figura 19. Si se usa la estructura allí representada, la salida conformada del array $Y_{SD}(\omega)$, en el dominio de la frecuencia, se puede expresar como:

$$Y_{SD}(\omega) = \underbrace{W_S^H(\omega) Y_R(\omega)}_{Y_{RS}(\omega)} - \underbrace{H^H(\omega) [B Y_R(\omega)]}_{Y_B(\omega)} \quad (105)$$

En esta expresión, $\mathbf{Y}_R(\omega)$ es el vector de entrada al array (señal original más perturbación) en el dominio de la frecuencia, una vez aplicados los retardos de compensación para alinear en tiempo todas las señales captadas por los micrófonos. Considerando que los micrófonos están en campo lejano (aproximación de Fraunhofer), son omnidireccionales y con la misma sensibilidad \mathbf{S}_i :

$$\mathbf{Y}_R(\omega) = \left[Y_1(\omega) \cdot \exp\left(-j\frac{\omega}{c} z_1 \cos \theta_0\right), \dots, Y_I(\omega) \cdot \exp\left(-j\frac{\omega}{c} z_I \cos \theta_0\right) \right]^T \quad (106)$$

para la dirección de apuntamiento $\theta = \theta_0$.

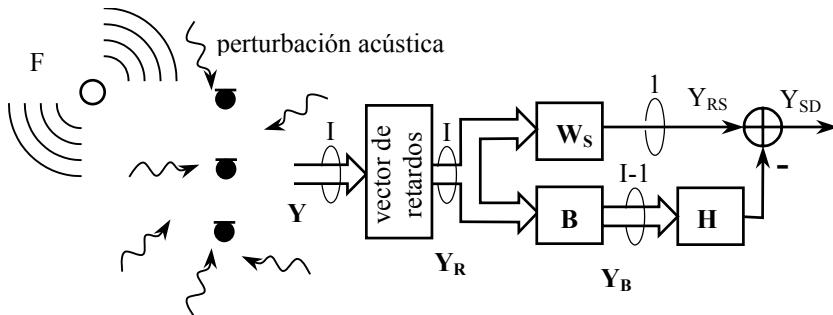


Figura 19. Diagrama de bloques para la implementación de un array superdirective mediante una matriz de bloqueo.

$\mathbf{W}_S(\omega)$ es el vector de coeficientes necesario para implementar la suma de todos los canales alineados en tiempo $\mathbf{Y}_R(\omega)$ y tiene por expresión:

$$\mathbf{W}_S = \frac{1}{I} \underbrace{\begin{bmatrix} 1, 1, \dots, 1 \end{bmatrix}}_I^T = \frac{1}{I} \mathbf{1} \quad (107)$$

($\mathbf{1}$ es una columna de unos) donde se ha considerado la división por el factor de normalización I (número de micrófonos del array) para que el conjunto de las etapas de retardo (106) y suma (107) proporcionen la salida de un conformador convencional como en (60). Por tanto $Y_{RS}(\omega)$ equivale a la salida del conformador convencional de retardo y suma, con los posibles problemas de poca directividad en baja frecuencia si el array no es lo suficientemente grande.

En (105), \mathbf{B} es la matriz de bloqueo. Se encarga de realizar las combinaciones entre canales para conseguir pares bidireccionales necesarios para configurar el array superdirective. La matriz de bloqueo \mathbf{B} tiene las siguientes características:

- Es de dimensión $(I - 1) \times I$, con I el número de micrófonos del array.
- La suma de todos los valores por cada fila es cero.
- El rango de \mathbf{B} ha de ser $I-1$.

La salida de \mathbf{B} es por tanto el vector de bloqueo $\mathbf{Y}_B(\omega)$ de longitud $I - 1$. Cada elemento de $\mathbf{Y}_B(\omega)$ es la salida de un array que ofrece un cero en la dirección de apuntamiento $\theta = \theta_0$, como puede comprobarse fácilmente y debido a la segunda condición de \mathbf{B} .

En (105) $\mathbf{H}(\omega)$ es el vector de filtrado –de dimensión $(I - 1) \times 1$ –, encargado de acomodar los niveles de cada elemento de $\mathbf{Y}_B(\omega)$ para cumplir la condición de mínima potencia de salida sin distorsión, expresada en (100).

La solución (105) del conformador superdirectivo mediante matriz de bloqueo tiene dos importantes ventajas prácticas respecto a la implementación más convencional de (100). Por una parte, el número de filtros $\mathbf{H}(\omega)$ que es necesario calcular y aplicar es de $I - 1$, uno menos que el número de micrófonos del array, a diferencia de los I coeficientes $\mathbf{W}_{SD}(\omega)$ de (100). Por otra parte siempre se dispone de forma separada de los canales alineados en tiempo $\mathbf{Y}_R(\omega)$ que pueden utilizarse en un procesador para otras tareas relacionadas con la mejora de habla.

En condiciones de ruido uniforme se puede obtener $\mathbf{H}(\omega)$ [Nordholm 92] mediante la siguiente expresión:

$$\mathbf{H}(\omega) = \left[\mathbf{B}(\omega) \Gamma'_{NN}(\omega) \mathbf{B}^H(\omega) \right]^{-1} \mathbf{B}(\omega) \Gamma'_{NN}(\omega) \mathbf{W}_S \quad (108)$$

para que la estructura mediante matriz de bloqueo mostrada en la Figura 19 sea equivalente al array superdirectivo genérico de la expresión (100). En (108) $\Gamma'_{NN}(\omega)$ es la matriz de coherencia de ruido una vez que éste ha pasado por la alineación temporal de (106). También se puede considerar una solución restringida de (108) sin más que añadir el parámetro de restricción $\mu(\omega)$ a la diagonal principal de $\Gamma'_{NN}(\omega)$,

$$\mathbf{H}(\omega) = \left[\mathbf{B}(\omega) \left[\Gamma'_{NN}(\omega) + \mu(\omega) \mathbf{I}(\omega) \right] \mathbf{B}^H(\omega) \right]^{-1} \mathbf{B}(\omega) \left[\Gamma'_{NN}(\omega) + \mu(\omega) \mathbf{I}(\omega) \right] \mathbf{W}_S \quad (109)$$

para configurar una conformación superdirectiva mediante matriz de bloqueo similar a (103).

Desde un punto de vista más formal, la matriz de bloqueo proyecta el vector de entrada $\mathbf{Y}_R(\omega)$ en el subespacio vectorial generador de la señal de ruido. Esta proyección contiene ruido y no señal, y por lo tanto se puede restar de la salida del conformador convencional, como se ha representado en la Figura 19. La estructura de la Figura 19 se llama también “cancelador generalizado del lóbulo lateral” (GSC o *Generalized Sidelobe Canceller*) puesto que elimina el ruido que no pertenece al subespacio de la señal, es decir el ruido que procede de direcciones laterales.

Merece especial atención por su importancia histórica el conformador de Griffith-Jim (GJBF, *Griffiths-Jim BeamFormer*) [Griffiths 82] en el que la matriz de bloqueo es del tipo:

$$\mathbf{B} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \quad (110)$$

En este caso, la matriz \mathbf{B} configura $I-1$ dipolos acústicos (pares bidireccionales), eligiendo las $I-1$ parejas de micrófonos consecutivos existentes en el array. El conformador de Griffith-Jim puede usarse como conformador fijo o también como conformador adaptativo, que es para lo que inicialmente fue concebido.

Para unificar más los resultados del array superdirectivo se puede poner la ecuación (105) de la siguiente forma:

$$\mathbf{Y}_{SD}(\omega) = \left[\mathbf{W}_S^H(\omega) - \mathbf{H}^H(\omega) \mathbf{B} \right] \mathbf{Y}_R(\omega) = \mathbf{W}'_{SD}(\omega) \mathbf{Y}_R(\omega) \quad (111)$$

con

$$\mathbf{W}'_{SD}(\omega) = \mathbf{W}_S^H(\omega) - \mathbf{H}^H(\omega) \mathbf{B} \quad (112)$$

el vector de ponderaciones que es necesario aplicar en cada canal después de realizada la alineación temporal. Es decir, la solución superdirective o MVDR también se puede factorizar en dos partes, por una parte la alineación temporal y por otra el filtrado superdirective con los coeficientes $\mathbf{W}'_{SD}(\omega)$. Esta nueva versión de la conformación superdirective se representa en la Figura 20. Los elementos del vector $\mathbf{W}'_{SD}(\omega)$ de (112) son diferentes a los ya conocidos de (100) y (103) porque en estas dos últimas expresiones la alineación temporal está contenida en el propio cálculo de los coeficientes superdirectivos.

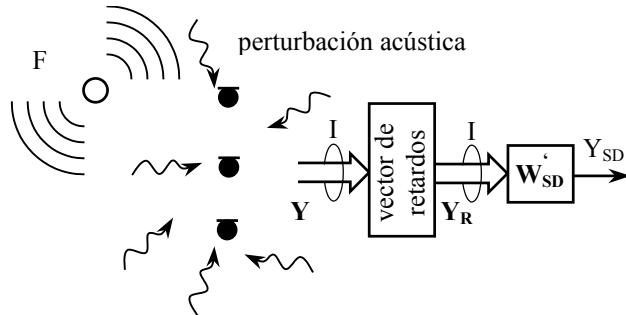


Figura 20. Diagrama de bloques de la implementación de un array superdirective usando los coeficientes $\mathbf{W}'_{SD}(\omega)$ para ser aplicados después de la alineación temporal.

2.3 CONFORMACIÓN ADAPTATIVA

Se llama conformador adaptativo a aquél que modifica el patrón de directividad en función de las condiciones de ruido y reverberación presentes en la sala donde se sitúa el array de micrófonos. El esquema de conformación propuesto en el apartado anterior, es decir, el array superdirective, es en cierta medida de tipo adaptativo, pero el concepto de adaptación no es dinámico. Un array superdirective funciona óptimamente para una señal interferente de ruido y/o reverberación con determinadas características de coherencia espacial, previamente aceptadas y establecidas. Lo que ahora se propone es una adaptación dinámica a las condiciones de ruido y reverberación existentes, con el objeto de minimizar la potencia de salida de las señales interferentes. Las bases de la conformación adaptativa fueron establecidas con el conformador GJBF [Griffiths 82] también conocido a veces como “cancelador generalizado de lóbulo lateral” (GSC).

La estructura básica de un GJBF contiene un array lineal uniforme de I micrófonos, en la disposición ya conocida. Dada una dirección de llegada de la fuente principal (DOA principal), se ha de configurar el vector de retardos que realice la alineación temporal. En el dominio de la frecuencia la alineación temporal se efectúa mediante (106). A la señal multicanal alineada temporalmente se le aplica la matriz de bloqueo propuesta en (110), a la salida de la cual se tiene un array lineal uniforme de $I-1$ parejas bidireccionales, con un cero teórico de captación dirigido a la DOA.

En la Figura 21 se representa esquemáticamente un GJBF en el dominio de la frecuencia. Los filtros fijos $H_{Fi}(\omega)$ sirven para implementar un conformador fijo. Por ejemplo, si se quiere una estructura de retraso y suma, $H_{Fi}(\omega) = 1/I \forall i$, como en (107). No obstante se pueden adoptar otras estructuras más complejas de filtrado y suma [Fischer 96]. Con el conformador en funcionamiento, los filtros adaptativos $H_{Ai}(\omega)$ varían dinámicamente hasta obtener un mínimo de potencia en la salida del conformador $Y(\omega)$. Eso es equivalente a que la

salida $Y_B(\omega)$ es máxima, siendo $Y_B(\omega)$ la salida bloqueada del cancelador múltiple. Si el array está bien apuntado hacia θ_0 , $Y_B(\omega)$ será una versión conformada de toda la señal que llega al array y no procede de la DOA principal θ_0 , ya que la salida del cancelador múltiple tiene un cero teórico hacia θ_0 . Es decir, $Y_B(\omega)$ contiene sólo ruido y reverberación. Por tanto el restador implementado en el cancelador múltiple tan sólo resta ruido y reverberación de la salida del conformador fijo.

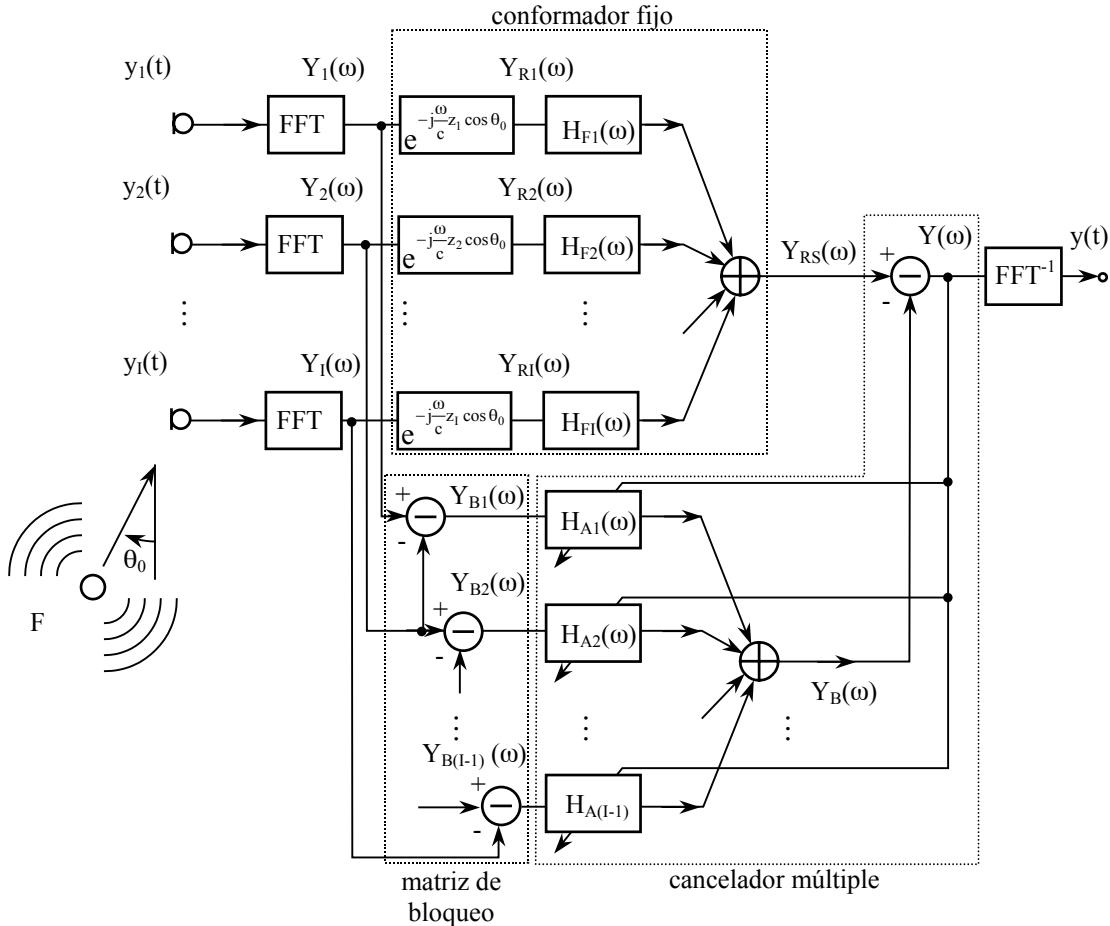


Figura 21. Esquema de funcionamiento de un conformador adaptativo de Griffiths-Jim, implementado en el dominio de la frecuencia. $H_{Fi}(\omega)$: filtros fijos para el conformador fijo de retardo y suma. $H_{Ai}(\omega)$: filtros adaptativos.

En la Figura 22 se representa la directividad a tres frecuencias de un GJBF adaptado a una señal interferente que procede de una fuente de ruido situada a la misma distancia que la fuente principal pero procedente de la dirección θ_n . Por simplicidad se han considerado la fuente principal y la de ruido en condiciones de campo libre. Puede observarse cómo la condición de adaptación supone que se produce un cero en la dirección de la interferencia θ_n . Los filtros adaptativos se implementan de tal manera que se cumpla la condición de no distorsión de la fuente principal, que equivale a que la directividad para el ángulo de apuntamiento θ_0 vale 0dB a todas las frecuencias. En este caso tan ideal, con sólo dos fuentes en campo libre, puede ser que el diagrama de directividad valga más de 0dB para otras direcciones distintas de θ_0 y θ_n , como se manifiesta en la Figura 22. Esto es debido a que la condición de adaptación tan sólo fuerza $D(\theta_0) = 1$ y $D(\theta_n) = 0$, sin importar lo que ocurre en otros ángulos, ya que se presupone que el ruido procede sólo de una dirección del espacio. En

una situación real, con ruido y reverberación procedente de todas las direcciones espaciales la condición de adaptación suele ser equivalente a mínimos no absolutos en el patrón polar, en las direcciones en las que se sitúan las fuentes de ruido y reverberación y a máximos que normalmente no superan los 0dB, puesto que se podría amplificar el ruido o reverberación lateral, presentes en $Y_B(\omega)$ (Figura 21), que en una situación real procederá de todas las direcciones espaciales, en mayor o menor cuantía.

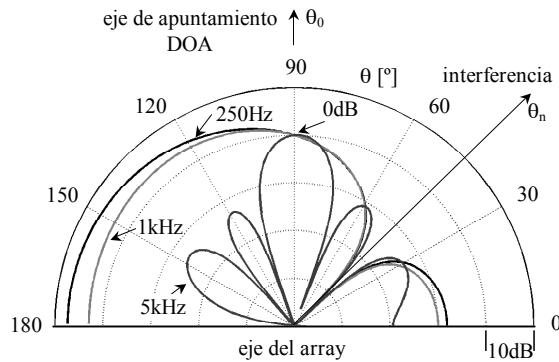


Figura 22. Directividad $D(\theta)$ a tres frecuencias de un conformador adaptativo de Griffiths-Jim constituido por un array lineal uniforme de cinco micrófonos (con $\Delta z=0.04m$), apuntado a $\theta_0=90^\circ$ y adaptado para suprimir interferencias en $\theta_n=45^\circ$.

En la Figura 23 se representa la directividad del mismo array del ejemplo para todas las frecuencias de interés en forma de mapa, $D(\theta, f)$. Puede observarse cómo se mantiene, a todas las frecuencias, una captación unidad en la dirección θ_0 y una captación nula en la dirección de interferencia θ_n .

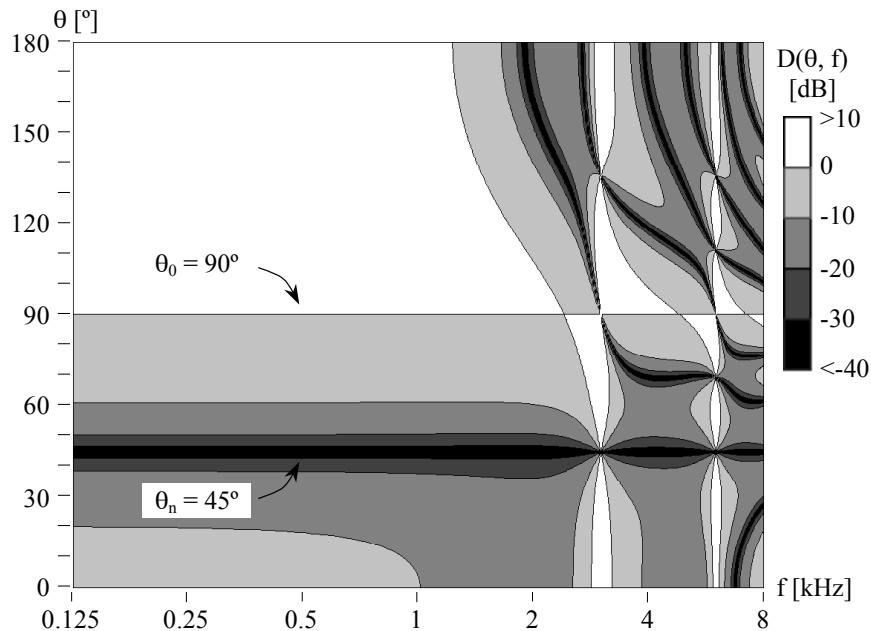


Figura 23. Mapa de directividad $D(\theta, f)$ de un lineal uniforme de cinco micrófonos con $\Delta z=0.04m$, apuntado hacia $\theta_0=90^\circ$ y adaptado para cancelar interferencias en $\theta_n=45^\circ$ (como en la Figura 22).

La estructura GJBF se comporta de manera muy buena en teoría, pero tiene algunos problemas de robustez. El principal deriva de que la salida $Y_B(\omega)$ pueda contener algo de la señal principal, procedente de la DOA de apuntamiento. Esto en teoría no debería ocurrir, ya que la matriz de bloqueo anula toda la señal de la dirección θ_0 . En la práctica esta anulación nunca es completa. Los errores en el pareamiento de módulo y de fase de cada par bidireccional de micrófonos que implementa la matriz de bloqueo, hacen que no se anule completamente la señal principal, y se reste como si fuese ruido en la salida del conformador fijo.

Otro error que produce las mismas consecuencias es el fallo en la estimación de la DOA, sobre todo cuando la fuente de señal principal se mueve con respecto al array. Existe una gran variedad de técnicas para conseguir una conformación adaptativa suficientemente robusta. Las técnicas más utilizadas consisten en la reducción de la señal principal (pérdidas de señal principal) que pueda contener la salida de la matriz de bloqueo [Claeson 92] y la implementación óptima de los filtros adaptativos $H_{Ai}(\omega)$ para que eliminen las posibles pérdidas de señal principal de la salida de la matriz de bloqueo [Hoshuyama 99]. En [Angus 93] se plantea de forma incipiente una aplicación práctica usando un DSP y la conformación adaptativa con un array de 7 micrófonos omnidireccionales. En [Brandstein 01] se hace una revisión bastante completa de las técnicas de conformación robusta aplicadas a arrays microfónicos. No obstante, el campo de los arrays adaptativos se aleja un poco de los objetivos de esta Tesis, por lo que en ésta no se hace un estudio más exhaustivo de esas técnicas.

3 LOCALIZACIÓN DE FUENTE

Hasta el momento, cuando se ha hablado de conformación de haz, se ha presupuesto que se conoce la dirección de apuntamiento del array, determinado por las coordenadas esféricas r_0, θ_0, ϕ_0 de la DOA principal, que corresponden al locutor cuya voz se quiere mejorar. Sin embargo, esta información no es necesariamente conocida, y la localización de la fuente principal constituye por sí misma una tarea con elevada dificultad, que es necesario abordar si se quiere un apuntamiento del array con cierto grado de automatismo. Además, téngase en cuenta que en ambientes de captación acústicamente adversos, que son los tratados en esta Tesis, la dificultad aumenta, ya que aparte de la posible presencia de ruido o reverberación difusos que enmascaren la presencia de la fuente principal, pueden existir otras fuentes secundarias de voz, que en el mejor de los casos estarán incorreladas con la principal, pero que normalmente tendrán alto grado de correlación con la señal de voz a mejorar, por ejemplo si proceden de reflexiones con las superficies del recinto escenario de la captación con arrays. Un análisis detallado sobre la problemática de la robustez en la localización de fuente usando arrays de micrófonos se trata en [Strobel 00].

Aparte del número y naturaleza de las fuentes a situar por el array, la configuración geométrica de los elementos del array determina fuertemente la estrategia a seguir en la tarea de localización. Si los micrófonos del array son omnidireccionales, para determinar simultáneamente los ángulos de elevación θ_0 y azimut ϕ_0 se necesita normalmente un array bidimensional, ya que con un array lineal no se puede distinguir, por razones de simetría, la DOA de fuentes situadas a diferentes azimuts ϕ_0 alrededor de la línea del array.

El buen funcionamiento de la conformación de haz tratada en el capítulo 2 es bastante dependiente de la localización de fuente. Algunos esquemas de conformación son muy sensibles a la falta de precisión en el apuntamiento del array. Por ejemplo los conformadores adaptativos de tipo GJBF son mucho más sensibles a las imprecisiones en la DOA que el conformador convencional de tipo filtrado y suma. Además, téngase en cuenta que los esquemas de conformación adaptativa no sólo necesitan una correcta ubicación de la DOA, sino que también necesitan conocer la existencia de posibles fuentes de interferencia, y por supuesto discriminarlas respecto al locutor principal.

Dependiendo de los usos que vaya a tener el array de micrófonos, la determinación de la DOA puede valerse de muchas estrategias. En un caso sencillo, en el que el array vaya a situarse en posición fija y a gran distancia de las posibles fuentes de voz, la DOA de apuntamiento puede hacerse fija, con $(\theta_0, \phi_0) = (0^\circ, 0^\circ)$ y $r_0 = \infty$. Se entiende que la distancia es grande, cuando supera suficientemente al tamaño del array, por ejemplo en una relación de 5 a 10 veces. Este esquema de DOA fija puede ser válido para estructuras de conformación poco exigentes en cuanto al apuntamiento, por ejemplo el conformador convencional. También puede usarse un apuntamiento manual en el que un operador del array introduce directamente en el procesador las coordenadas de la fuente, de forma numérica o ayudado por algún método externo, como la imagen proporcionada por una cámara de video. Cuando la fuente de señal se acerca al array, las condiciones de apuntamiento se hacen más difíciles

porque un pequeño movimiento lateral de dicha fuente cambia de forma relativamente muy elevada las coordenadas de apuntamiento, haciendo que el conformador de haz funcione incorrectamente.

Los esquemas de localización de fuente más sofisticados son los de apuntamiento automático, bien utilizando exclusivamente la señal acústica del array o combinando otras informaciones, como la procedente de otros micrófonos más alejados o, como en el localizador manual, la imagen proporcionada por una cámara de vídeo. En el último caso la información de localización se puede usar en los dos sentidos [Strobel 01], es decir, se puede apuntar una cámara de vídeo utilizando la información sonora del array o al revés, el apuntamiento del array se realiza mediante la imagen de vídeo. Existen esquemas de seguimiento automático de fuente que son capaces de discriminar a uno entre varios locutores interiores [Di Claudio 00] [Kamiyanagida 01] [Tanaka 01] e incluso realizar un seguimiento del locutor principal [Cao 95] [Vermaak 01].

Otro argumento adicional es la complejidad computacional. Interesa que los esquemas de seguimiento automático de fuente funcionen en tiempo real. Téngase en cuenta que la determinación de la DOA suele ser complementaria en un procesador global en array y por tanto debe trabajar simultáneamente junto con el resto de procesado de habla necesario, bien sea el correspondiente a la conformación de haz solamente o también a otros esquemas adicionales de postfiltrado y mejora. Esto hace que algunos esquemas de determinación automática de DOA, muy complicados computacionalmente, sean difícilmente realizables para el trabajo en tiempo real.

En esta Tesis, el presente capítulo sobre localización de fuente es accesorio en cierta medida, ya que el grueso del trabajo desarrollado se centra en los aspectos de conformación de haz y mejora de voz mediante postfiltrado utilizando arrays microfónicos. No obstante, a continuación se hace una revisión del estado del arte sobre los métodos más usados tanto en localización monofuente, como en localización multifuente, haciendo mención de los métodos basados en subespacios que están siendo utilizados con gran profusión en la actualidad.

Los esquemas más usados de localización de fuente se agrupan en tres categorías [Brandstein 97-b]:

- Maximización de la potencia de salida del array (SRP ó *Steered Response Power*). Estos métodos analizan la potencia de salida del array para diferentes apuntamientos, entendiéndose que la DOA se sitúa donde dicha potencia de salida sea máxima.
- Determinación de la diferencia de tiempo de llegada (TDOA ó *Time Difference Of Arrival*). Estos métodos estudian la correlación cruzada entre parejas de micrófonos para determinar la DOA.
- Estimación espectral de alta resolución. Aquí se analiza la matriz de correlaciones cruzadas de la señal multicanal del array y como resultado se obtiene la DOA y la situación de posibles fuentes interiores, por lo que esta estrategia puede ser usada para localización multifuente. Los métodos basados en subespacios se pueden englobar dentro de este tipo de estimación.

3.1 MAXIMIZACIÓN DE LA POTENCIA DE SALIDA

Se trata de buscar alguna función de verosimilitud obtenida de la señal de salida del array y que esté relacionada con el apuntamiento correcto del mismo, de tal manera que analizando las posibles localizaciones de la fuente se alcance el valor máximo de la misma. Normalmente la función de verosimilitud es la potencia de salida de un conformador convencional de retardo y suma, que será considerada como un “localizador de máxima verosimilitud” (ML o *Maximum Likelihood*). Por lo tanto, cuando se busca el apuntamiento correcto del conformador convencional se exploran los retardos asociados a los pesos w de cada micrófono de tal manera que se maximice la potencia de la salida conformada del array, $y(t)$.

La maximización de la potencia de salida tiene dos importantes frentes de estudio. Por una parte la articulación de estrategias de conformación, de tal manera que con un array limitado en número de micrófonos y en separación intermicrofónica, se pueda conseguir suficiente resolución espacial para extraer sin muchas dudas la DOA principal del array y resolver con suficiente precisión la presencia de otras fuentes de perturbación existentes junto a la fuente principal. Téngase en cuenta que el ruido y la reverberación reducirán la capacidad de éxito de la localización de fuente. Por otra parte se necesita resolver el problema de maximización con suficiente rapidez y ahorro computacional para que pueda ser implementado en tiempo real.

En cuanto a la implementación del problema de maximización, en [Wax 83] se revisan los métodos iterativos tipo Newton más tradicionales, encontrándose en la bibliografía otros métodos más sofisticados pero de difícil aplicación en arrays que funcionen en tiempo real.

Otra disyuntiva es el método de conformación que se debe utilizar para conseguir que la potencia de salida del array sea suficientemente selectiva espacialmente. Como ya se ha visto, la condición de maximización de la potencia de salida que viene expresada en (47), da los pesos w_{RS} necesarios para configurar un conformador convencional. Cuando se busca a ciegas la posición de una determinada fuente mediante la maximización de la potencia de salida, en realidad se estará buscando un máximo de la directividad cuadrática $|D(\theta, \phi)|^2$ del array (50). El máximo de potencia está condicionado a la posición de apuntamiento, determinada por la distancia r_0 y el ángulo θ_0 , que para un array lineal son suficientes. Quiere decir que la forma que presente la función directividad $D(\theta, \phi)$ del array, determinará lo rápido que se llegue al apuntamiento correcto y la existencia de alguna ambigüedad en el problema de maximización. Puede desprenderse que para arrays poco directivos, la maximización de la potencia de salida presenta la dificultad de que su diagrama polar no ofrece un máximo local abrupto, por lo que en condiciones acústicas adversas se hace de difícil implementación. También constituye un problema el *aliasing* espacial según se expuso en el punto 2.1.8 o la existencia de lóbulos laterales de amplitud elevada, que podrían complicar la solución del problema de maximización. En el caso de que la localización de fuente se haga mediante un array lineal de micrófonos omnidireccionales, hay que considerar que la simetría de revolución de la directividad alrededor del eje del array limita el ángulo θ_0 de búsqueda en el intervalo $[0, \pi]$, puesto que la directividad del mismo se repite en el intervalo $[\pi, 2\pi]$. Las mismas razones de simetría hacen imposible la determinación del azimut ϕ_0 .

La localización/discriminación de la dirección de apuntamiento se complica cuando existen varias fuentes de señal de voz. En este caso habrá que considerar el problema de maximización de la directividad conjunta del array. Si existen dos fuentes que atacan al array y éstas son incoherentes, la directividad conjunta a una determinada frecuencia (o en una

determinada banda) vendrá determinada por la suma de potencia (suma no coherente) de las directividades individuales obtenidas al apuntar al array a cada una de las dos fuentes:

$$D(\theta, \varphi) = \sqrt{|D_1(\theta, \varphi)|^2 + |\alpha D_2(\theta, \varphi)|^2} \quad (113)$$

entendiendo que $D_1(\theta, \varphi)$ es la directividad que tendría el array cuando se apunta a la fuente 1 y $D_2(\theta, \varphi)$ si se apunta a 2. El factor α determina la atenuación de la fuente 2 con respecto a la fuente 1. Por ejemplo, supóngase que se emplea al array anidado de 5 octavas visto en el punto 2.2.2 –Figura 15(a)–. Si $\alpha = 1$ (las dos fuentes a localizar son igual de potentes) la directividad conjunta del array se representa en la Figura 24. Según esta figura, y entendiendo que sólo se busca la maximización de la potencia para el ángulo de apuntamiento θ (no se considera la distancia r), la presencia de la fuente 2 originaría dos ángulos diferentes de apuntamiento, θ_1 y θ_2 . Hay que considerar que la presencia de la fuente 2 puede deberse a una fuerte reflexión de la fuente 1 en alguna de las paredes del recinto donde se encuentra el array (aunque en ese caso no sería cierta la incoherencia interfuente apuntada anteriormente).

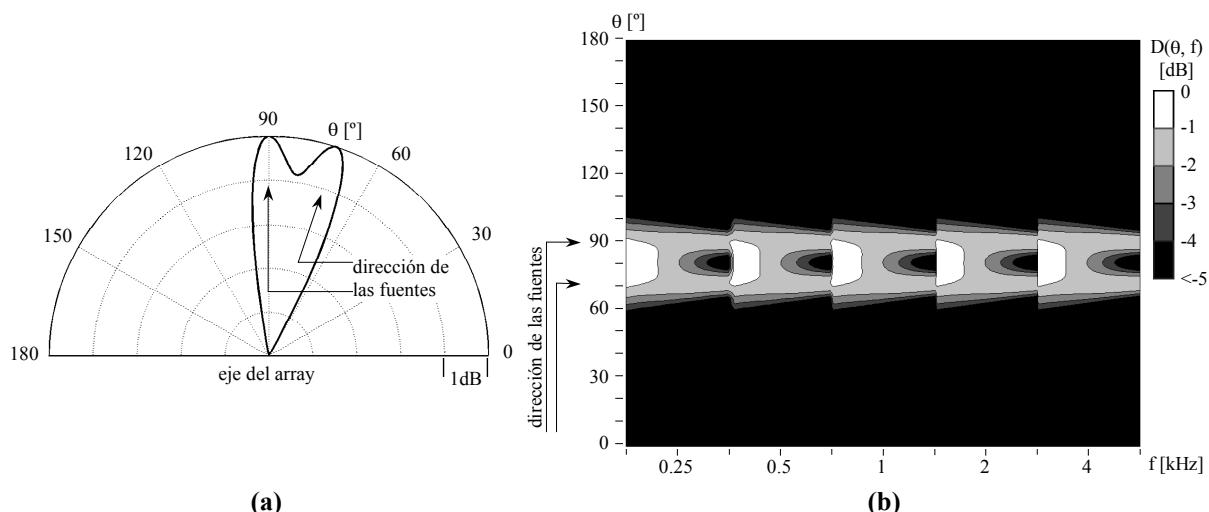


Figura 24. Directividad conjunta de un array anidado *broadside* de cinco octavas en presencia de dos fuentes incoherentes de igual amplitud separadas $\theta_2 - \theta_1 = 17.83^\circ$. **(a)** Curva polar $D(\theta)$ a $f=1\text{kHz}$. **(b)** Mapa de niveles de $D(\theta, f)$ mostrando todas las frecuencias.

Si la fuente secundaria es de menor amplitud se hace más fácil la localización de la fuente principal por simple discriminación de la amplitud del máximo de la potencia captada, a la hora de resolver el problema de la maximización, como se muestra en la Figura 25 (nótese la diferencia de escalas respecto a la figura anterior).

Tanto en la Figura 24 como en la Figura 25 puede apreciarse cómo dentro de cada subbanda, al reducirse la frecuencia disminuye la capacidad de discriminación (el array se hace menos selectivo) y por lo tanto empeoran las condiciones del problema de maximización.

El criterio de resolución de Rayleigh establece el ancho del lóbulo principal de un array lineal uniforme equiponderado (ponderación de tipo rectangular) de I receptores separados por Δz (véase la Figura 8). Si el array es *broadside* la ecuación (74) proporciona el primer nulo de captación adyacente al principal y permite establecer la proximidad angular máxima de la fuente secundaria respecto a la principal para que ésta pueda resolverse. La capacidad de resolución según la condición de Rayleigh –(74) para el un apuntamiento *broadside*– equivale

a que la fuente secundaria se sitúa en el primer mínimo de la directividad $D_1(\theta, \varphi)$ que tiene el array apuntando a la fuente 1.

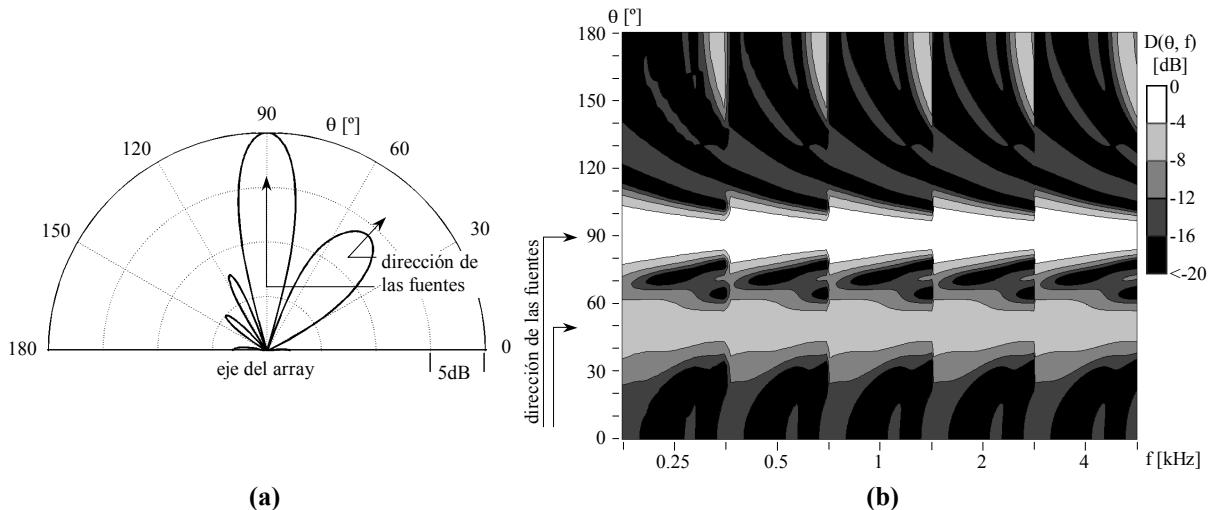


Figura 25. Directividad conjunta de un array anidado *broadside* de cinco octavas en presencia de dos fuentes incoherentes separadas $\theta_2 - \theta_1 = 40^\circ$ y con una diferencia de nivel de 6dB ($\alpha=0.5$). **(a)** Curva polar $D(\theta)$ a $f=1\text{kHz}$. **(b)** Mapa de niveles de $D(\theta, f)$ mostrando todas las frecuencias.

Se pueden utilizar otras estructuras de conformación diferentes, teniendo en cuenta en cualquier caso, que aunque no se utilice un conformador convencional, el método de localización de fuente por maximización de la potencia de salida exige respuesta máxima para la dirección principal. Podría emplearse el array superdirective fijo visto en el punto 2.2.3, diseñado para minimizar ruido isotrópico o difuso, que también presenta un máximo local en la dirección principal (Figura 18). El conformador MVDR que da lugar al array superdirective confiere una mayor resolución angular a la hora de localizar la fuente principal, y de discriminarla de otras fuentes secundarias. Esto es así porque en definitiva es más directivo. La utilización de un array receptor superdirective para discriminar dos fuentes angularmente próximas a veces se conoce como conformador de Capon [Capon 69] y existen técnicas para conseguir aun mayor resolución [Munier 87] basadas en la modificación de la matriz de correlación de la señal o del ruido estimado.

Existen también otros métodos diversos de localización de fuente entre los que se puede destacar el fundamentado en la maximización de la entropía de la salida del array [Naidu 01].

En resumen, las técnicas de localización tipo SRP son las más simples y fáciles de implementar, bastante adaptadas al funcionamiento en tiempo real con un array microfónico basado en DSP, aunque en el caso más simple del conformador convencional, la maximización de la potencia puede carecer de suficiente resolución espacial y capacidad de discriminación de fuentes secundarias, además de ser poco robusta frente al ruido y la reverberación.

3.2 LOCALIZACIÓN BASADA EN ESTIMACIÓN DE RETARDOS (TDOA)

La localización de fuente basada en la estimación de retardos consiste en buscar, los retardos intermicrófonicos τ con los que llega la señal de habla captada por el array. Para calcular el TDOA se utilizan medidas de correlación entre las señales entregadas por las

diferentes parejas de micrófonos del array. A partir de esa información puede obtenerse la localización de la fuente (coordenadas r_0, θ_0, ϕ_0). En [Knapp 76] se describen los principios fundamentales del método y en [Brandstein 95] una aplicación basada en TDOA.

Cuando se ha estimado una serie de retardos intermicrofónicos la siguiente tarea será obtener la posición estimada de la fuente. En [Gay 00] (pp.247-257) se exponen algunos métodos para este propósito. En lo que respecta a esta Tesis, la tarea principal será la estimación de retardos, puesto que éstos serán los parámetros de entrada utilizados para, mediante conformación de haz, apuntar a la fuente principal, siendo irrelevantes las coordenadas espaciales de posición de dicha fuente.

3.2.1 TDOA (*Time Difference Of Arrival*)

En la Figura 26 se representa el modelo de array lineal de micrófonos para estimar el retardo en una pareja de micrófonos.

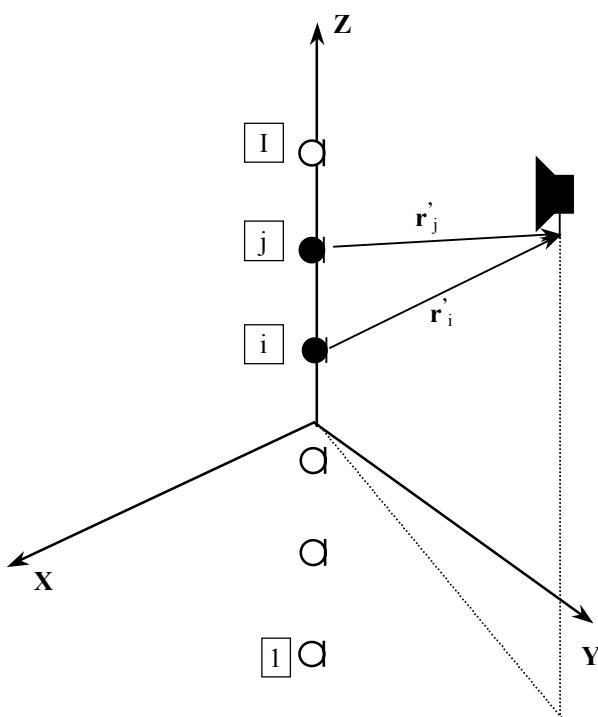


Figura 26. Estimación de retardos en un array lineal de micrófonos.

En condiciones de campo libre, cuando incide una onda senoidal de presión acústica $p(t)$, la tensión entregada por cada el micrófono i viene dada por la ecuación (3). Se reproduce aquí dicha ecuación:

$$x_i(t) = S_i D_i(\theta_i, \phi_i) p(t) \frac{r}{r_i} \exp[-jk(r_i - r)] \quad (3)$$

Si además se supone que los micrófonos son omnidireccionales:

$$x_i(t) = S_i p(t) \frac{r}{r_i} \exp[-jk(r_i - r)] = S_i p(t) \frac{r}{r_i} \exp(-jk\tau_i) \quad (114)$$

con

$$\tau_i = \frac{r'_i - r}{c} \quad (115)$$

el tiempo que tarda en llegar la señal acústica desde la fuente hasta el micrófono i , referido a la posición central del array en el eje de coordenadas, considerado como referencia en toda esta Tesis. Las propiedades de la transformada de Fourier hacen a la expresión (114), válida para señales senoidales, equivalente a

$$x_i(t) = S_i \frac{r}{r'_i} p(t - \tau_i) \quad (116)$$

para una presión acústica $p(t)$ de banda ancha. El objetivo de la estimación de retardo será la obtención de τ_{ij} , dado por:

$$\tau_{ij} = \tau_i - \tau_j \quad (117)$$

al que también se conoce como TDOA, o tiempo de retardo entre los micrófonos “ i ” y “ j ”. Aunque la ecuación (116) representa un modelo muy simplificado de la captación multimicrófonica del array, puede servir como punto de partida para el cálculo del TDOA. En la práctica se deberá considerar necesariamente el efecto de la reverberación y el ruido aditivo, que casi siempre introducen perturbación en el cálculo correcto de los retardos del array. En [Champagne 96] se discute la falta de robustez frente a la reverberación del cálculo de TDOA.

3.2.2 Método de la Correlación Cruzada Generalizada (GCC o *Generalized Cross-Correlation*)

Para la obtención del retardo intermicrofónico τ_{ij} , el método GCC [Knapp 76] maximiza la correlación cruzada entre dos versiones filtradas de las señales $x_i(t)$ y $x_j(t)$, captadas por los dos micrófonos objeto de estudio. De forma general, la función a maximizar o función de correlación cruzada generalizada viene dada por:

$$c_{x_i x_j}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \psi(\omega) \Phi_{X_i X_j}(\omega) \exp(i\omega\tau) d\omega \quad (118)$$

donde $\Phi_{X_i X_j}(\omega)$ es el espectro cruzado intermicrófono y $\Psi(\omega)$ es la función de filtrado o de peso. Si se elige la función de peso $\Psi(\omega) = 1$ se estará ante el método clásico de la correlación cruzada ya que en ese caso $c_{x_i x_j}(\tau) = R_{x_i x_j}(\tau)$ o correlación cruzada simplemente, entre $x_i(t)$ y $x_j(t)$.

La idea básica que manifiesta el método GCC expresado en (118) es que dadas dos señales $x_i(t)$ y $x_j(t)$ y con $x_j(t)$ una versión retardada y amplificada (o atenuada) de $x_i(t)$, la correlación entre las mismas tiene un máximo para $\tau = \tau_{ij}$. La condición de máximo se alcanza en (118) cuando el término $\exp(i\omega\tau)$ compensa a la fase de $\Phi_{X_i X_j}(\omega)$. En ese caso la correlación cruzada generalizada quedaría para el retardo τ_{ij} o TDOA:

$$\max_{\tau} c_{x_i x_j}(\tau) = c_{x_i x_j}(\tau_{ij}) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \psi(\omega) |\Phi_{X_i X_j}(\omega)| d\omega \quad (119)$$

Para evitar que la variación con ω de los espectros cruzados $|\Phi_{X_i X_j}(\omega)|$ introduzca una elevada fluctuación en $c_{x_i x_j}(\tau)$, que produzca máximos locales de la correlación cruzada, normalmente se elige la función de peso:

$$\Psi(\omega) = \frac{1}{|\Phi_{X_i X_j}(\omega)|} \quad (120)$$

En ese caso el método GCC pasa a llamarse GCC-PHAT o simplemente PHAT (*PHAt Transform*) [Knapp 76] [Omologo 96].

En la Figura 27(a) se representa la correlación cruzada generalizada GCC –con $\Psi(\omega) = 1$ – entre dos micrófonos de un array anidado de 15 micrófonos, en condiciones acústicas muy adversas, con mucho ruido de banda ancha (SNR = 6dB) y operando en una sala muy reverberante ($T_{60} \approx 1s$). Puede observarse cómo aparece un máximo para el retardo de $\tau = \tau_{ij} \approx 1.2ms$, que corresponde al TDOA estimado entre ambos micrófonos, con la configuración geométrica dada. Además aparecen otros máximos locales que pueden deberse a reflexiones de la señal de voz con las paredes del recinto, a errores propios debido a la contaminación por ruido o a la variación de $|\Phi_{X_i X_j}(\omega)|$ con ω . En la Figura 27(b) se representa la GCC por el método PHAT con la función de peso $\Psi(\omega)$ (120), para la misma muestra de señal de voz. Puede comprobarse cómo el método PHAT resuelve mucho mejor los candidatos a τ_{ij} .

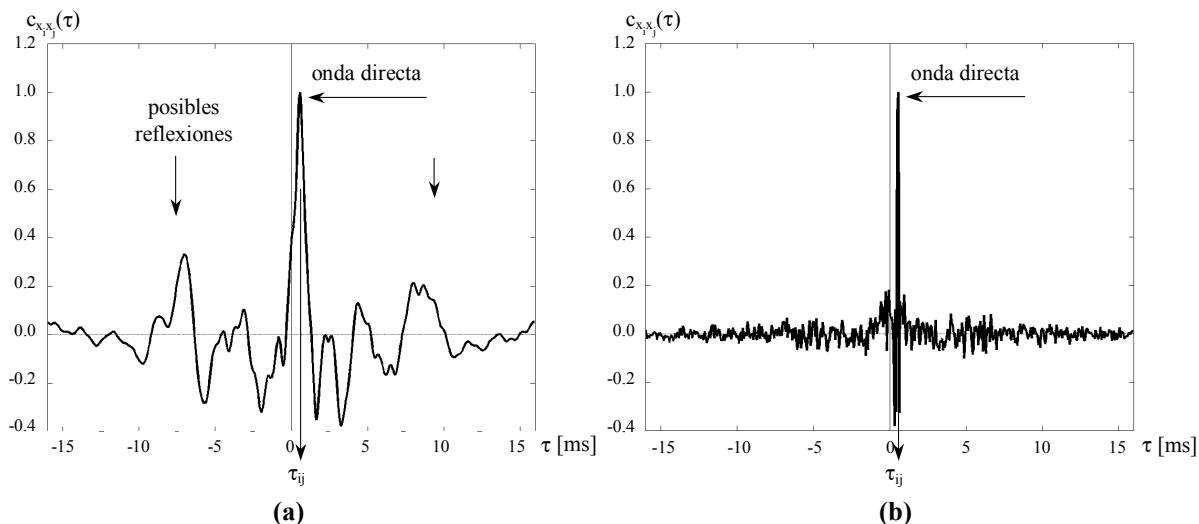


Figura 27. Correlación cruzada generalizada (GCC) entre los micrófonos $i=1$ e $i=8$ de un array anidado de 15 micrófonos (separación de $z_8-z_1=48cm$). La fuente está situada a 3m del centro del array con $\theta_0=45^\circ$. Se está captando una señal de voz mezclada con ruido difuso (SNR=6dB). Se han promediado varias ventanas de 32ms hasta completar una trama de voz de 1s. (a) Método GCC clásico, con $\Psi(\omega)=1$. (b) Método PHAT con $\Psi(\omega)=1/|\Phi_{X_i X_j}(\omega)|$.

El método GCC-PHAT se ha venido usando con profusión en los últimos años [Brandstein 97-b] [Omologo 97-b] [González-Rodríguez 99-a], siendo su implementación relativamente sencilla y con poco gasto computacional para una aplicación en tiempo real.

Aunque en condiciones de poco ruido y reverberación el método GCC-PHAT funciona muy bien, tiene sus limitaciones en condiciones de baja SNR y sobre todo cuando la reverberación es alta en el recinto [Bédard 94] [Brandstein 97-a]. Existen varias aproximaciones para obtener una estimación de la TDOA en condiciones adversas sobre todo

de alta reverberación. Algunas explotan las características particulares de la señal de voz para diseñar una función de peso $\Psi(\omega)$ adecuada, y otras realizan una deconvolución previa de la función de transferencia del recinto, para después hacer la estimación TDOA por métodos convencionales.

En definitiva, el método GCC-PHAT se ha utilizado en el pasado y se utiliza ampliamente en la actualidad, en la forma descrita en este punto o con diversas variantes, diferenciadas normalmente por la función de peso empleada. Es computacionalmente sencillo para ser usado en aplicaciones multimicrófono con un funcionamiento en tiempo real, sobre un soporte DSP. Se puede utilizar en arrays de micrófonos con un número relativamente grande de captadores (mayor de 10), constituyendo la sección de localización y apuntamiento automático para muchas aplicaciones de mejora de señal de habla multicanal. Sus principales inconvenientes son la falta de robustez ante la reverberación en el recinto y la presencia de más de una fuente sonora, que tiene un efecto parecido a la reverberación. La utilización conjunta del método PHAT con el método SRP descrito en el apartado 3.1 mejora mucho los dos inconvenientes apuntados anteriormente.

3.3 LOCALIZACIÓN DE ALTA RESOLUCIÓN BASADA EN ESTIMACIÓN ESPECTRAL

Bajo este epígrafe se engloban aquellos métodos de localización de fuente basados en el análisis de la matriz de covarianzas o correlaciones cruzadas de la señal de habla captada por un array de micrófonos, según quedó ésta definida en (32) o (35). A este tipo de localizadores espectrales de alta resolución pertenecen varias estrategias como el modelado autorregresivo (AR *Auto Regressive*), varianza mínima (MV *Minimum Variance*), máxima verosimilitud clásica (ML *Maximum Likelihood*) y las técnicas basadas en subespacios (métodos MUSIC, SPRIT, MIN-NORM). Aunque estos métodos se han venido usando con éxito en el ámbito de las comunicaciones radioeléctricas (antenas), tienen algunas limitaciones inherentes a la voz para ser usados con arrays microfónicos. La no estacionariedad y elevado ancho de banda de la señal de habla, la localización cercana del emisor y la presencia de reverberación, juegan un papel negativo en la localización de fuente mediante estimación espectral. De entre estos métodos, los basados en subespacios son los que mejor se adaptan a las condiciones de recepción típicas de señal de habla.

3.3.1 Métodos de banda estrecha

Desde un punto de vista estadístico un modelo de captación en array está bien definido –[Brandstein 01], pág. 182– sólo para señales de banda estrecha, es decir cuando el ancho de banda de la señal captada es como mucho un 1% de la frecuencia central. Esta condición no la cumple evidentemente la señal de habla; sin embargo, a continuación se reseñan los métodos de banda estrecha más utilizados para la localización de fuente, ya que como se verá posteriormente, pueden ser extrapolados bajo ciertas condiciones al caso de banda ancha.

Métodos basados en subespacios. Método MUSIC

Si se observan las características peculiares de la matriz de correlaciones cruzadas $\mathbf{R}_{yy}(\tau)$ (covarianzas cruzadas si se supone que la salida del array es de media nula) expresada

en (35), asociada al vector de salida $\mathbf{y}(t)$ de un array receptor que capta a una serie de fuentes sonoras, es posible obtener la posición de dichas fuentes. Basándose en esto, se vienen desarrollando recientemente los llamados métodos de procesado en array basados en subespacios [Cazdow 90] [Stoica 90-a] [Stoica 90-b] [Asano 00] [Tanaka 01] [Jabloun 01], y muy especialmente el algoritmo conocido como MUSIC (*MUltiple SIgnal Classification*) [Bienvenu 80] [Schmidt 86]. Aunque el algoritmo MUSIC fue introducido por primera vez para la estimación de DOA principal cuando existen varias fuentes sonoras [Stoica 91] [Viberg 91-a], ha sido muy utilizado posteriormente para otras aplicaciones como la separación de fuentes e incluso la mejora de señal de voz [Ephraim 95]. El método MUSIC en su formulación más clásica necesita hacer un barrido del vector de posición (r, θ, ϕ) de las posibles fuentes para encontrar las coordenadas (r_0, θ_0, ϕ_0) correspondientes a la/s fuente/s, sin embargo existen métodos derivados que evitan este paso, obteniéndose de forma directa la posición de la fuente, mediante la exploración de las características de la matriz de covarianzas $\mathbf{R}_{yy}(\tau)$ (35) del array. A pesar de la dependencia genérica de $\mathbf{R}_{yy}(\tau)$ con la variable temporal de retardo τ , aquí se considerará la particularización \mathbf{R}_{yy} para $\tau = 0$.

Sea la señal multicanal $\mathbf{y}(t)$ (28) a la salida de un array de micrófonos que capta M fuentes, representada por su matriz de covarianzas \mathbf{R}_{yy} (35). Esta pertenece a un subespacio vectorial de dimensión M (el número de fuentes), que en principio debe ser menor que I (el número de micrófonos) para que el método funcione. Dicho de otra forma, la matriz de covarianzas \mathbf{R}_{yy} genera el espacio vectorial (de dimensión I) al que pertenece la señal $\mathbf{y}(t)$ captada por el array (señal + ruido), que a su vez se puede descomponer en dos subespacios vectoriales ortogonales, el de la señal (de dimensión M) al que pertenece $\mathbf{x}(t)$ (15), que contiene la suma de las M fuentes libres de ruido captadas por el array, y el del ruido (de dimensión $I-M$) al que pertenece el ruido aditivo $\mathbf{n}(t)$ captado también por el array. Según este planteamiento, aquí todas las fuentes se consideran como principales, porque se va a intentar localizar la posición de todas ellas, a diferencia del razonamiento expresado en el punto 2.1.4 de esta Tesis, donde las fuentes secundarias se consideraban también como ruido, y podían ser incluidas en el vector $\mathbf{n}(t)$. Analíticamente:

$$\mathbf{R}_{yy} = \mathbf{U}_y \Lambda_y \mathbf{U}_y^H = \mathbf{U}_x \Lambda_x \mathbf{U}_x^H + \mathbf{U}_n \Lambda_n \mathbf{U}_n^H = \mathbf{U}_x \Lambda_x \mathbf{U}_x^H + \sigma^2 \mathbf{I} \quad (121)$$

En esta expresión, las columnas de las matrices \mathbf{U} son los autovectores que generan los subespacios vectoriales correspondientes, bien sea el de la señal \mathbf{U}_x (matriz $I \times M$), o el del ruido \mathbf{U}_n (matriz $I \times I - M$). La matriz Λ_y es una versión irreducible de \mathbf{R}_{yy} , en cuya diagonal se encuentran los autovalores de \mathbf{R}_{yy} : $\lambda_{y1}, \lambda_{y2}, \dots, \lambda_{yl}$. En (121), los autovectores de \mathbf{R}_{xx} son los autovectores del subespacio vectorial correspondiente a la señal $\mathbf{x}(t)$, con autovalores $\lambda_{x1}, \lambda_{x2}, \dots, \lambda_{xM}$. Los autovalores σ^2 que generan el subespacio de ruido $\mathbf{n}(t)$ (que aquí se supone espacialmente incoherente), representado por su matriz \mathbf{R}_{nn} , son todos iguales, en una primera suposición (para ruido uniforme). Los autovalores de \mathbf{R}_{yy} son a su vez $\lambda_{yi} = \lambda_{xi} + \sigma^2$ para $i = 1..M$ y $\lambda_{yi} = \sigma^2$ para $i = M+1..I$.

El concepto fundamental subyacente en (121), es que la señal limpia representada por el vector $\mathbf{x}(t)$, pertenece al subespacio vectorial de la señal incluido en el espacio vectorial que genera al vector $\mathbf{y}(t)$ (señal + ruido). Este subespacio tiene dimensión M (número de fuentes) y está generado por los autovectores de \mathbf{U}_x –columnas en (121)–. Dicho subespacio es además ortogonal a aquél generado por los autovectores presentes en \mathbf{U}_n –columnas en (121)– que se llama subespacio vectorial del ruido, el cual constituye el complemento ortogonal al subespacio vectorial de la señal. De acuerdo con esto, el vector de apuntamiento $\mathbf{a}_m(r_m, \theta_m, \phi_m)$ que localiza a la posición (r_m, θ_m, ϕ_m) donde se sitúa la fuente de señal m , será ortogonal a cualquiera de los autovectores \mathbf{U}_n que generan el subespacio correspondiente al

ruido. Por tanto, la proyección de $\mathbf{a}_m(r_m, \theta_m, \varphi_m)$ sobre cualquiera de los autovectores del subespacio ruido verifica:

$$\mathbf{u}_n^H \mathbf{a}_m(r_m, \theta_m, \varphi_m) = 0 \quad (122)$$

donde \mathbf{u}_n es un autovector ($I \times 1$) generador del subespacio ruido, o igualmente:

$$\mathbf{U}_n^H \mathbf{a}_m(r_m, \theta_m, \varphi_m) = (\underbrace{0, 0, \dots, 0}_{I-M})^T \quad (123)$$

con \mathbf{U}_n la matriz ($I \times I - M$) compuesta por los autovectores del subespacio generador del ruido.

La forma de detectar cuál de los autovalores corresponde al subespacio asociado al ruido y cuál pertenece al subespacio generador de la señal consiste precisamente en explorar la cantidad y multiplicidad de los autovalores de valor mínimo, que son los que pertenecerán al subespacio vectorial del ruido. Es decir como $\lambda_{\min} = \sigma^2 < \lambda_{x_i} + \sigma^2 = \lambda_{y_i}$, queda claro que los autovalores de valor mínimo serán los correspondientes al ruido, y que su multiplicidad será la dimensión del subespacio de ruido, pudiéndose determinar en ese momento el número M de fuentes incidentes. Por tanto, el procedimiento de descomposición de \mathbf{R}_{yy} en los dos términos de (121) se inicia mediante la búsqueda de sus autovectores, para posteriormente determinar cuáles son los que generan el subespacio del ruido y cuáles el de la señal.

Una vez conocidos los autovectores \mathbf{U}_n del subespacio vectorial del ruido, el método MUSIC busca las posiciones de las M fuentes presentes, representadas por las coordenadas de posición $(r_m, \theta_m, \varphi_m)$, mediante la maximización de una función que incluya operaciones de proyección como la (122). Consecuentemente, el espectro MUSIC se define como sigue:

$$P_{MU}(r, \theta, \varphi) = \frac{1}{\mathbf{a}^H(r, \theta, \varphi) \mathbf{U}_n \mathbf{U}_n^H \mathbf{a}(r, \theta, \varphi)} \quad (124)$$

donde el denominador expresa la potencia de la proyección del vector de apuntamiento genérico $\mathbf{a}(\theta, \varphi, r)$ sobre el subespacio de ruido. La expresión (124) se puede escribir también [Naidu 01] en función de los autovectores \mathbf{U}_x de la señal:

$$P_{MU}(r, \theta, \varphi) = \frac{1}{\mathbf{a}^H(r, \theta, \varphi) (\mathbf{I} - \mathbf{U}_x \mathbf{U}_x^H) \mathbf{a}(r, \theta, \varphi)} \quad (125)$$

donde \mathbf{U}_x es la matriz ($I \times M$) cuyas columnas son los autovectores del subespacio señal.

El denominador de (124) o de (125) es idealmente cero para las direcciones de la/s fuente/s, situadas en $(r, \theta, \varphi) = (r_m, \theta_m, \varphi_m)$. Por tanto, el método MUSIC consiste en buscar el conjunto de M posiciones (r, θ, φ) que minimice el espectro MUSIC, $P_{MU}(r, \theta, \varphi)$. Hay que advertir que el espectro MUSIC sólo tendrá máximos relativos muy nítidos para todas las DOA's de las M fuentes presentes, cuando dichas fuentes sean totalmente incoherentes entre sí y con el ruido de fondo.

En la Figura 28 se esquematiza el proceso a seguir cuando se usa el método MUSIC. En primer lugar, se estima en cada trama de voz captada con el array la matriz de covarianzas \mathbf{R}_{yy} . El proceso se puede hacer en el dominio del tiempo, considerando un determinado ancho de banda, o se puede hacer un análisis en frecuencia, extrayendo en ese caso la matriz de espectros cruzados $\Phi_{YY}(\omega)$ definida en (42), previa realización de la FFT, y particularizada para un intervalo estrecho de frecuencias. A continuación se analiza la estructura de la matriz \mathbf{R}_{yy} . Se calculan los autovalores y se decide cuáles son de ruido y cuáles son de señal,

mediante la exploración de los autovalores mínimos y su multiplicidad. Despues, por métodos de álgebra numérica se extraen los autovectores del ruido, representados por la matriz \mathbf{U}_n , si se va a utilizar (124); o los autovectores de la señal \mathbf{U}_x , si se va a usar (125). A continuación se explora el espectro MUSIC –(124) o (125)– para encontrar el juego de posiciones espaciales $(r, \theta, \varphi) = (r_m, \theta_m, \varphi_m)$ que maximice P_{MU} .

La gran ventaja del método MUSIC frente a otros es su buena resolución espacial (máximos de P_{MU} muy nítidos) y su falta de ambigüedad en la decisión (*unbiased*). El principal inconveniente es que el espectro MUSIC según (124) (125) carece de buena resolución espacial cuando existe mucha coherencia entre fuentes próximas. Este es el caso que se manifiesta en un ambiente de mucha reverberación donde las reflexiones de mayor energía son tratadas como fuentes coherentes. Sin embargo, este problema ha sido tratado y solucionado en cierta medida en trabajos relativamente recientes [Buckley 90] [Farrier 90] [Kaveh 90] [Xu 93]. Otro inconveniente que afecta negativamente al método MUSIC es la existencia de *aliasing* espacial, inherente especialmente en un array lineal uniforme, sobre todo cuando está presente más de una fuente de señal de voz. Uno de los métodos que existe para superar este problema es distribuir aleatoriamente los micrófonos en el array lineal. También es un inconveniente el ancho de banda tan extenso de la señal de habla, aspecto que ha sido abordado en la literatura científica [Su 83] [Wang 85].

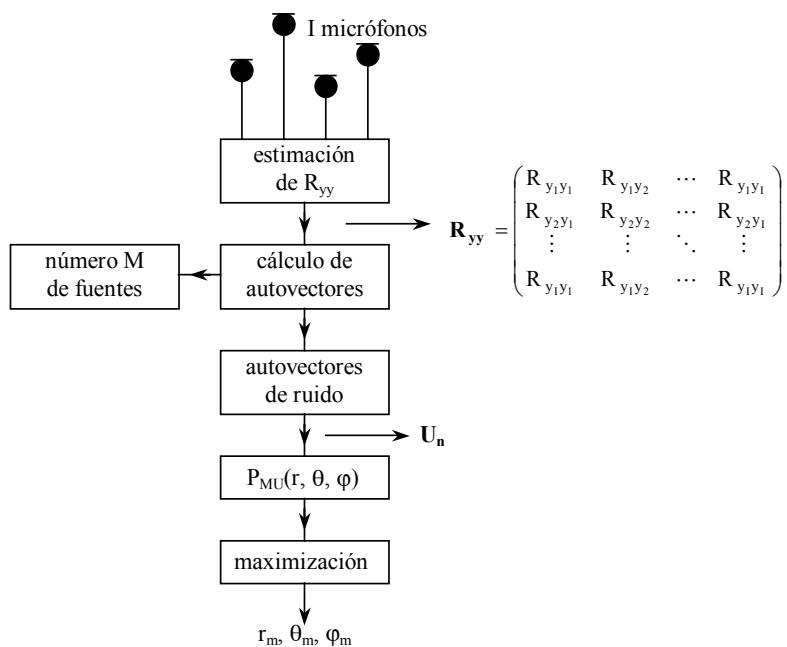


Figura 28. Método MUSIC.

Otros métodos basados en subespacios

Cuando se utiliza el método MUSIC clásico se ha de testar la función $P_{MU}(r, \theta, \varphi)$ para encontrar la posición de las fuentes. Existen métodos que utilizan la misma base teórica de la descomposición en subespacios para localizar a las fuentes de forma directa. Entre estos métodos pueden destacarse el ROOT MUSIC y el MIN-NORM.

El método ROOT MUSIC [Barabell 83] es inicialmente igual que el método MUSIC. Se debe encontrar un desarrollo de la matriz de covarianzas \mathbf{R}_{yy} en subespacios ortogonales, y

por lo tanto se parte del conocimiento de los autovectores \mathbf{u}_n que generan el subespacio del ruido. Para un array lineal uniforme, en condiciones de campo lejano y con micrófonos omnidireccionales equisensibles ($\mathbf{S}_i = \mathbf{S}_0 \forall i$), el vector de apuntamiento hacia una dirección genérica θ , se puede expresar (59) como:

$$\mathbf{a}(\theta) = \exp(jk z_1 \cos \theta) [1, s, s^2, \dots, s^{I-1}]^T \quad (126)$$

con

$$s = \exp(jk \Delta z \cos \theta) \quad (127)$$

y z_1 la coordenada z del primer micrófono del array.

Sean los $I - M$ autovectores \mathbf{u}_n del subespacio ruido, con cada uno de ellos se cumplirá lo mismo que (122),

$$\mathbf{u}_n^H \mathbf{a}(\theta) = 0 \Rightarrow \mathbf{u}_n^H [1, s, s^2, \dots, s^{I-1}]^T = 0 \quad (128)$$

y por tanto también se cumplirá con la suma de todos ellos, como se escribe a continuación:

$$\sum_{i=1}^{I-M} \mathbf{u}_{ni}^H [1, s, s^2, \dots, s^{I-1}]^T = 0 \quad (129)$$

con \mathbf{u}_{ni} el autovector i -ésimo de ruido. Las raíces s de la ecuación (129) que estén sobre la circunferencia unidad corresponderán a las fuentes de señal, cuya posición $\theta = \theta_m$ se podrá saber despejando θ de (127).

El método ROOT MUSIC simplifica el problema de maximización de (124) o de (125) y es adecuado para arrays lineales uniformes que captan fuentes situadas en campo lejano, ya que la ecuación (126) asume esta suposición. En la práctica esta situación no es la normal con arrays microfónicos de tamaño mayor a un metro, puesto que los locutores se situarán en posiciones relativamente cercanas al array. Además, en su formulación original es un método de banda estrecha, por lo que necesita ser replanteado para su aplicación a banda ancha (caso de la voz), como se explica más delante en el punto 3.3.3.

En el método MIN-NORM (norma mínima) se busca una solución a la ecuación

$$\min_{\theta} \mathbf{a}^H(\theta) \mathbf{R}_{yy} \mathbf{a}(\theta), \text{ con la condición } \mathbf{a}^H(\theta) \mathbf{a}(\theta) = 1 \quad (130)$$

El vector $\mathbf{a}(\theta)$ solución será un autovector de \mathbf{R}_{yy} correspondiente a su más pequeño autovalor [Naidu 01]. Por tanto, el método MIN-NORM ofrece una alternativa al método MUSIC más directa a la hora de encontrar las fuentes de señal presentes junto con el array, aunque la ecuación (130) presentará generalmente una mayor complejidad computacional en su solución.

Los dos métodos anteriores sólo pueden usarse a partir de la señal $y(t)$ entregada por un array lineal uniforme. El método ESPRIT [Roy 86] (*Estimation of Signal Parameters via Rotation Invariance Technique*) se utiliza con arrays de dipolos receptores, cuyos integrantes deben tener ejes paralelos. Cada dipolo estará formado por dos receptores, uno se considera el elemento par y otro el impar. El método consiste en utilizar las propiedades de los subespacios generados por la señal de salida de los arrays formados por los elementos de los dipolos del mismo tipo (según sea un elemento par o impar). Estos dos subespacios vectoriales están relacionados por rotación.

El método WSF (*Weighted Subspace Fitting*) [Viberg 91-a] [Viberg 91-b] utiliza un procedimiento de mínimos cuadrados (*Least Squares*) para intentar encajar una combinación lineal de vectores de apuntamiento $\mathbf{a}(\theta, \varphi)$ sobre los autovectores \mathbf{U}_x del subespacio vectorial de la señal.

Los métodos basados en subespacios más aceptados para la localización de la/s fuente/s incidentes utilizando la señal de salida de un array, han sido descritos de forma concisa. Éstos, se han venido proponiendo y desarrollando desde mediados de la década de los 80 hasta la actualidad, para cubrir objetivos más amplios que la localización de fuente, entre los que se encuentra la mejora de la señal captada por el array. No obstante, por su complejidad computacional, su uso no se ha extendido suficientemente en prototipos de arrays de micrófonos que trabajen en tiempo real.

3.3.2 Métodos paramétricos

La efectividad de los métodos basados en subespacios anteriormente descritos se sustenta en una supuesta incorrelación de las M fuentes de señal representadas por el vector de referencias $\mathbf{x}_0(t)$ de (16). Esta incorrelación es necesaria en especial cuando estas fuentes proceden de direcciones muy próximas. En la práctica, esto no es así frecuentemente y para superar el efecto pernicioso de este fenómeno se han propuesto recientemente algoritmos alternativos basados en técnicas de máxima verosimilitud (*Maximum Likelihood*, ML), que aunque computacionalmente son más complejos, pueden ser en determinadas circunstancias más eficientes y robustos que los anteriores, y constituyen los llamados métodos paramétricos. Básicamente se encuentran en la literatura dos clases de algoritmos, los basados en ML determinista [Ziskind 88] [Cadzow 90] [Viberg 94] y los basados en ML estocástica [Stoica 90-c] [Ottersten 02]. Ambos métodos tratan de encontrar la máxima verosimilitud ML en el vector de observaciones $\mathbf{y}(t)$, suponiendo que los elementos de éste son variables aleatorias independientes.

En el caso de ML determinista la función densidad de probabilidad asociada a cada canal del array es de tipo Gaussiana y determinista. El método se implementa minimizando el menor logaritmo de la verosimilitud intercanal, I_{DML} o ML determinista logarítmica:

$$I_{DML} [(r, \theta, \varphi), \mathbf{x}_0(t), \sigma^2] = I \log \sigma^2 + \frac{1}{\sigma^2} E \left\{ \| \mathbf{y}(t) - \mathbf{a}m(r, \theta, \varphi) \mathbf{x}_0(t) \|^2 \right\} \quad (131)$$

donde $\|\cdot\|$ representa la norma euclídea. La minimización de (131) estima [Wax 92] las fuentes de señal representadas por el vector $\mathbf{x}_0(t)$, la varianza del ruido σ^2 y posición de las fuentes (r, θ, φ) , es decir, las DOA's.

El método ML estocástico considera que la señal observada $\mathbf{y}(t)$ es un proceso gaussiano aleatorio. El procedimiento es paralelo al caso determinista, con la salvedad de que ahora la señal está caracterizada por sus parámetros estadísticos, en concreto por la matriz de covarianza de la fuente, $\mathbf{R}_{x_0 x_0}$ (o una estimación, $\hat{\mathbf{R}}_{x_0 x_0}$ de la misma):

$$\mathbf{R}_{x_0 x_0} = E \left\{ \mathbf{x}_0(t) \mathbf{x}_0^H(t) \right\} \quad (M \times M) \quad (132)$$

Se ha de minimizar una función de máxima verosimilitud estocástica SML dependiente de la matriz covarianza $\mathbf{R}_{x_0 x_0}$, de la varianza del ruido σ^2 y de la dirección de llegada de las fuentes (θ, φ) . Los detalles de esta minimización están referidos en [Bohme 86] y [Jaffer 88].

Los métodos basados en ML requieren un conocimiento de la estadística de la señal incidente al array, que es la señal de habla contaminada por ruido y reverberación. Por supuesto, la estadística de la señal de habla se aparta bastante de la suposición Gaussiana, por lo que la aplicación de ML clásica no es óptima. Por eso los métodos paramétricos aplicados al procesado en array de la señal de habla, y a la detección de la DOA se han desarrollado de forma combinada con los métodos basados en subespacios. En [Stoica 90-a] y [Stoica 90-b] se desarrollan métodos tipo MUSIC que aprovechan las ventajas de los métodos paramétricos para poder tratar eficientemente fuentes de señal con alto grado de correlación. Como se ha explicado, esta propiedad de las fuentes puede ser tratada de forma apropiada con la minimización de una función de máxima verosimilitud. Estos métodos paramétricos tipo MUSIC tienen sus versiones específicas para el caso de arrays lineales y uniformes, por ejemplo los ya comentados ROOT-MUSIC [Barabell 83], SPRIT [Paulraj 86], WSF-raíz (ROOT-WSF) [Stoica 90-a], y además el IQML (*Iterative Quadratic Maximum Likelihood*) [Bresler 86].

3.3.3 Métodos de banda ancha

Los métodos explicados anteriormente no se pueden aplicar de forma directa a la señal de habla, ya que ésta es de banda ancha. Si se parte de un análisis temporal, habrá que filtrar la señal en N bandas suficientemente estrechas, cada una con frecuencia central o pulsación ω_n , con $n = 1, 2,..N$, hablándose entonces de la matriz de covarianzas para la pulsación ω_n , $\mathbf{R}_{yy}(\omega_n)$. Entonces se tendrá una descomposición en subespacios, y un juego de autovalores y autovectores dependiente de la frecuencia: $\lambda_{yi}(\omega_n)$, $\mathbf{U}_x(\omega_n)$, $\mathbf{U}_n(\omega_n)$, etc.

Si se trabaja por el contrario en el dominio de la frecuencia, se deberá realizar previamente una FFT de cada trama de voz, y a partir de ahí trabajar con la matriz de espectros cruzados $\Phi_{YY}(\omega)$ de (42). La descomposición en subespacios de $\Phi_{YY}(\omega)$, tal y como se ha expuesto en el apartado 3.3.1 anterior, es válida pero ahora se tendrá un juego de autovalores y autovectores para cada frecuencia. Se tendrá genéricamente (obviando la discretización en frecuencia de la FFT): $\lambda_{yi}(\omega)$, $\mathbf{U}_x(\omega)$, $\mathbf{U}_n(\omega)$, etc.

El método de banda ancha más sencillo para la localización de fuente corresponde al conocido como “aproximación incoherente”. Básicamente consiste en hacer una búsqueda de la DOA y del número de fuentes en cada banda de frecuencias, para después agrupar los resultados del apuntamiento en las diferentes bandas (mediante la búsqueda de la mediana o la media de los mismos), porque evidentemente la posición de la fuente o fuentes no puede depender de la frecuencia. Debido a las características de no estacionariedad y gran anchura de banda de la señal de voz, cuando la SNR es media o baja, la DOA para diferentes frecuencias suele dispersarse mucho, siendo necesaria una estimación a muy largo plazo, lo cual no es siempre posible. La aproximación incoherente puede ser adecuada para situaciones donde se conozca que determinadas frecuencias de la señal captada tienen una SNR alta (por ejemplo en los formantes de la señal de voz). En ese caso, el apuntamiento del array se realiza considerando sólo esas bandas.

Sin embargo los métodos de banda ancha que se han demostrado más efectivos son los llamados “métodos coherentes”, como se tratará a continuación. La idea básica consiste en realizar un análisis de banda estrecha en diferentes frecuencias, para después aplicar determinada estadística que permita encontrar la DOA con una determinada verosimilitud, a partir de la información dispersa de banda estrecha, previamente disponible.

Método de subespacio de señal coherente (CSSM, *Coherent Signal Subspace Method*)

El método CSSM [Wang 85] consiste en hallar la posición de las M fuentes presentes a partir de N soluciones, una para cada frecuencia, obtenidas mediante cualquiera de los métodos de banda estrecha ya vistos, por ejemplo el método MUSIC. Por tanto se parte de N matrices (17) de apuntamiento $\mathbf{am}_n(r, \theta, \varphi)$, una para cada pulsación ω_n . El coste computacional será muy alto si N es un número grande. El estudio a diferentes frecuencias tiene que proporcionar un número M único de fuentes presentes, aunque dará una cierta dispersión en las DOA's encontradas, representada por una dispersión en las matrices $\mathbf{am}_n(r, \theta, \varphi)$. A continuación se define el array de referencia, representado por su matriz de apuntamiento $\mathbf{am}_0(r, \theta, \varphi)$ como aquél asociado a una frecuencia central ω_0 :

$$\mathbf{am}_0(r, \theta, \varphi) = [\mathbf{a}_{10}(r_1, \theta_1, \varphi_1), \dots, \mathbf{a}_{m0}(r_m, \theta_m, \varphi_m), \dots, \mathbf{a}_{M0}(r_M, \theta_M, \varphi_M)] (I \times M) \quad (133)$$

Se han de encontrar N transformaciones lineales (una para cada frecuencia) representadas por N matrices de transformación \mathbf{T}_n ($I \times I$), de tal manera que cada una ha de cumplir:

$$\mathbf{T}_n \mathbf{am}_n = \mathbf{am}_0 \text{ con } n = 1, 2, \dots, N \quad (134)$$

Se llama matriz universal de covarianza espacial \mathbf{R}_{yy_0} (*USCM Universal Spatial Covariance Matrix*) al promedio ponderado de las matrices de covarianza \mathbf{R}_{yy_n} a cada frecuencia:

$$\mathbf{R}_{yy_0} = \sum_{n=1}^N \alpha_n \mathbf{T}_n \mathbf{R}_{yy_n} \mathbf{T}_n^H \quad (135)$$

A continuación se calcula la matriz universal de covarianza espacial de ruido \mathbf{R}_{nn_0} mediante la expresión:

$$\mathbf{R}_{nn_0} = \sum_{n=1}^N \alpha_n \mathbf{T}_n \mathbf{T}_n^H \quad (136)$$

y después su matriz raíz cuadrada \mathbf{N}_0 despejándola de la siguiente igualdad,

$$\mathbf{R}_{nn_0} = \mathbf{N}_0 \mathbf{N}_0^H \quad (137)$$

Por último se realiza una descomposición en subespacios por el método de banda estrecha, es decir, se buscan de nuevo los autovectores y autovalores, en esta ocasión tomando como punto de partida la matriz transformada:

$$\mathbf{N}_0^{-1} \mathbf{R}_{yy_0} (\mathbf{N}_0^{-1})^H \quad (138)$$

para minimizar finalmente el espectro MUSIC de banda estrecha $P_{MU}(r, \theta, \varphi)$, según (124) o (125), teniendo en cuenta que ahora los vectores de apuntamiento son las columnas del array de referencia $\mathbf{am}_0(r, \theta, \varphi)$.

Existen otros métodos que combinan de forma parecida al método CCSM la información obtenida de un análisis espectral de banda estrecha; de hecho casi todos los métodos de banda estrecha comentados en el apartado 3.3.1 tienen su versión de banda ancha, por ejemplo el WSF de banda ancha [Cazdow 90] ó WB-WSF, o el ROOT-MUSIC expuesto más arriba. Sin embargo no se entra en más detalle, ya que como se ha explicado anteriormente, la localización de fuente no va a ser tratada en los experimentos de esta Tesis.

4 MEJORA DE HABLA MEDIANTE POSTFILTRADO

La mejora de señal de habla consiste en la atenuación de ruido y reverberación presentes en la voz captada por uno o varios micrófonos. La mejora de habla monocanal, en su versión de minimización de ruido es un tema clásico que ha sido acometido con diferentes técnicas [Faucon 89] como la minimización del error cuadrático medio [Ephraim 85], los estimadores de tipo espectral [Le Bouquin 97], los filtros de Kalman [Gabrea 00] [Yong 00], la transformada KLT [Mittal 00] [Reyayee 01], o diversos tipos de postfiltrado [Kim 00] [Tilp 00]. Una primera aproximación a la mejora de voz multicanal es el tratamiento binaural [González-Rodríguez 99-a] [Gómez 00]. No obstante la captación de señal de voz con más de dos canales es el método que proporciona mejores resultados, al precio de aumentar la complejidad y el coste del sistema.

Como se ha visto, cuando se utilizan arrays microfónicos para captar señal de habla, se debe implementar una primera etapa de conformación de haz. La conformación de haz o *beamforming*, que también se llama filtrado espacial, puede ser considerada como un primer paso de mejora de la señal de habla. Es decir, cuando el haz de captación principal del array microfónico es dirigido hacia la fuente principal, se está atenuando la captación de fuentes que no proceden de dicha dirección principal. Esto significa una atenuación de las posibles fuentes de ruido y reverberación. Lo último es cierto porque la reverberación se origina por reflexiones de la señal principal con las superficies límite de una sala, reflexiones que en general no llegan al array microfónico desde la dirección principal. Las señales procedentes de otros locutores que no sean el principal, pueden ser consideradas también como fuentes de ruido y generalmente son de difícil atenuación. Sin embargo, la mejora de habla asociada sólo al filtrado espacial o *beamforming* es en general insuficiente, y se hace necesario implementar otras formas de procesado para la mejora de la señal de habla, basadas en la captación mediante arrays de micrófonos, y que permitan atenuar en mayor medida el ruido y la reverberación presentes en la señal vocal.

La mejora de señal de habla empleando arrays microfónicos tiene dos aspectos diferentes. Por una parte, lo que sería la eliminación de ruido [Kaneda 86] [Omologo 93] [Ihle 00] [Saruwatari 00] en ambientes reverberantes [Zelinski 88] [Van Compernolle 90-a] y por otra la eliminación de reverberación [Marro 98]. El trabajo de esta Tesis se centra en la eliminación de ruido mediante postfiltrado y la mayoría de los experimentos y propuestas de las partes 2 y 3 del texto así lo corroboran. La idea básica consiste en filtrar o atenuar las zonas del espectro donde esté presente el ruido y dejar pasar aquellas frecuencias donde predomine la señal. Desgraciadamente, en la mayoría de las ocasiones los espectros de señal y de ruido se solapan por lo que es necesario aprovechar las ventajas especiales que tiene la captación mediante un array para optimizar los métodos de mejora monocanal. Por ejemplo, la coherencia intercanal puede servir para dilucidar qué es ruido y qué es señal, aunque el empleo de arrays tiene otras muchas ventajas adicionales.

La reducción de reverberación es un problema bastante específico en la señal de habla que no es asimilable a otro tipo de aplicaciones no acústicas. Tanto es así que en la mayoría

de las situaciones típicas de captación sonora, la reverberación tiene más energía que la señal directa. Efectivamente, en una sala de conferencias normal, con dimensiones típicas y con tiempos de reverberación del orden de $T_{60} = 0.6 - 0.8$ s, el campo sonoro reverberante puede superar en más de 5dB SPL al campo sonoro directo, en las bandas de frecuencias medias, de máxima importancia en la inteligibilidad de la voz. Esto produce una gran degradación de la señal de habla, que en el ámbito acústico y electroacústico siempre se ha relacionado con una pérdida de inteligibilidad.

La conformación de haz produce, como ya se ha mencionado, una primera etapa de derreverberación, pero será necesario hacer algo más, aprovechando la información espacial disponible en un array de micrófonos. Por tanto, en este capítulo también se expondrán algunas de las técnicas más usadas para derreverberar una señal acústica en su versión monocanal y haciendo una extensión a los arrays microfónicos.

4.1 REDUCCIÓN DE RUIDO MEDIANTE POSTFILTRADO

Los métodos monocanal para la reducción de ruido han sido y siguen siendo muy populares y utilizados en el mundo del procesado digital de señales acústicas. De entre éstos los más usados son el filtrado de Wiener y la sustracción espectral y muchos otros que de forma directa o indirecta se relacionan con los dos anteriores. Éstos métodos se fundamentan en el filtrado de la señal acústica contaminada en función de la cantidad y la composición espectral del ruido que la perturba, teniendo que ser estimados estos parámetros. Existe una generalización inmediata de los métodos de reducción de ruido monocanal cuando se aplican a un array de captadores. La extensión multicanal de las técnicas de postfiltrado monocanal tiene un gran paralelismo con la teoría de superdirecividad, tratada en el punto 2.2.3 de esta Tesis.

La principal dificultad que poseen las técnicas de postfiltrado para la reducción de ruido en que deben ser ciegas. Es decir, el análisis se hace a posteriori ya que se dispone sólo de la señal acústica captada y contaminada por ruido sin tener otra referencia de qué es señal (voz) y qué es ruido. Por tanto una de las principales tareas de cualquier postfiltro es la determinación de en qué instantes de tiempo predomina la señal y en cuáles lo hace el ruido. Esta tarea es conocida como segmentación o detección de actividad de voz (*VAD, Voice Activity Detection*).

Este apartado se organiza de la siguiente forma. En primer lugar se trata el filtrado de Wiener y su extensión multicanal a un array de micrófonos, de gran interés aquí puesto que en las propuestas y experimentos desarrollados por el autor se parte de técnicas de este tipo. En segundo lugar se estudia la sustracción espectral, que puede considerarse como una generalización del filtrado de Wiener. A continuación se estudian los métodos de postfiltrado perceptual-auditivo, es decir aquéllos que intentan producir la menor distorsión audible en la señal de salida utilizando las propiedades del sistema auditivo humano. También se hace una pequeña introducción a la segmentación de habla y a la estimación de señal y de ruido, y finalmente se da una visión generalista de otros métodos de reducción de ruido que están siendo propuestos por diversos autores para el tratamiento de una señal multicanal.

4.1.1 Filtrado de Wiener multicanal

El concepto de filtrado de Wiener se propuso por vez primera en los trabajos de Norbert Wiener [Wiener 49] sobre predicción de trayectorias de objetos móviles. Son famosas las

contribuciones a los trabajos de Levinson [Levinson 47] y Durbin [Durbin 60] con sus algoritmos optimizados de cálculo. Además, tienen gran interés para esta Tesis las soluciones multicanal de Burg [Burg 64], Wiggins y Robinson [Wiggins 65].

Si una señal de voz contaminada por ruido entra en un filtro de Wiener, la salida de dicho filtro debe parecerse, con el menor error cuadrático posible (MMSE *Minimum Mean Squared Error*), a la señal de entrada limpia de ruido, que normalmente no se conoce y debe ser estimada. En pocas palabras, un filtro de Wiener realiza un filtrado dependiente de la relación SNR estimada y presente en la señal de entrada. Volviendo a la teoría de superdirectividad, tratada en el punto 2.2.3, la ecuación (100) correspondiente a los coeficientes $\mathbf{W}_{SD}(\omega)$ óptimos de un array superdirectivo, proporciona la respuesta de mínima varianza (MVDR) del array ante un ruido cuyas características de coherencia espacial se conocen. Por lo tanto, esta solución produce la salida de mayor SNR con máxima verosimilitud (ML) para una señal de entrada de banda estrecha [Brandstein 01]. Sin embargo, la solución del array superdirectivo no maximiza la SNR en banda ancha, como exige el filtrado de Wiener. El filtrado de Wiener permite cierto grado de distorsión lineal, que está restringida en el conformador superdirectivo (respuesta MVDR), pero sin embargo mejora la SNR y la reverberación si se aplica a la salida de un array superdirectivo o MVDR.

La extensión del filtrado Wiener monocanal a la versión multicanal es conocida [Burg 64], aunque la aplicación a arrays de micrófonos es un poco más reciente [Kaneda 86] [Zelinski 88]. En la década de los 90 han aparecido numerosas propuestas y aplicaciones basadas en el postfiltrado de una señal acústica, bien sea monocanal o multicanal, combinadas normalmente con otras técnicas que aprovechan la información espacial extra proporcionada por la captación multicanal, sobre todo la coherencia intermicrofónica, además de otros tópicos usados en el procesado de señales acústicas [Fischer 96] [Hussain 97] [Le Bouquin 97] [Mahmoudi 98] [González-Rodríguez 00].

Puesto que el trabajo desarrollado en esta Tesis en cuanto a propuestas y experimentos utiliza el filtrado de Wiener, se ha creído conveniente describir los fundamentos de este método de mejora aplicado a una señal multicanal [Naidu 01] procedente de un array lineal y uniforme de micrófonos.

El planteamiento del problema es el siguiente. Sea un array de I micrófonos, que capta señal de voz procedente de M fuentes en presencia de ruido y reverberación. Sólo se considera a una de esas fuentes como la principal, por eso se utilizará (28) como ecuación de partida, según quedaba definido en el capítulo 2 sobre conformación de haz. Se reproduce aquí dicha ecuación:

$$\mathbf{y}(t) = \mathbf{a}(r, \theta, \phi) \mathbf{x}_0(t) + \mathbf{n}(t) = \mathbf{x}(t) + \mathbf{n}(t) \quad (28)$$

Se recuerda que $\mathbf{y}(t)$, $\mathbf{x}(t)$ y $\mathbf{n}(t)$ son respectivamente los vectores ($I \times 1$) de señal más ruido, señal limpia y ruido (incluye reverberación y las fuentes ajenas a la principal) captados, que tienen su versión en frecuencia en $\mathbf{Y}(\omega)$, $\mathbf{X}(\omega)$ y $\mathbf{N}(\omega)$. El tratamiento espectral puede hacerse mediante descomposición en subbandas por medio de un banco de filtro de análisis óptimo o mediante la transformada STFT (*Short Time Fourier Transform*) usando la FFT previo enventanado de la señal en tramas de corta duración. En cualquier caso ese aspecto no preocupa en este momento.

La misión del filtrado de Wiener multicanal será obtener una versión mejorada $y_w(t)$ (un solo canal) de la señal multicanal $\mathbf{y}(t)$ que produce el array, con el menor error cuadrático posible con respecto a una señal de referencia. Por supuesto en un escenario con múltiples fuentes habrá que elegir cuál es la fuente principal (señal de referencia) y por tanto la señal

que se quiere mejorar. Las fuentes secundarias de voz aquí serán tratadas como ruido y por lo tanto quedarán atenuadas por el postfiltrado. Se considerará señal de referencia del postfiltro de Wiener a la señal eléctrica limpia $x_0(t)$ (5) que entregaría el canal de referencia (o canal 0) si no existiese ruido ni reverberación.

Para obtener la señal mejorada $y_w(t)$ se ha de filtrar cada uno de los canales del array, de índice i , con el filtro óptimo $h_i(\tau)$ o $H_i(\omega)$. En la versión temporal, el vector respuesta al impulso del filtro de Wiener multicanal será:

$$\mathbf{h}(\tau) = [h_1(\tau), h_2(\tau), \dots, h_I(\tau)]^T \quad (139)$$

y el vector salida $\mathbf{y}_w(t)$ antes de la conformación de haz,

$$\mathbf{y}_w(t) = [y_{w1}(t), y_{w2}(t), \dots, y_{wI}(t)]^T \quad (140)$$

con

$$y_{wi}(t) = \int_0^\infty h_i^*(t-\tau) y_i(\tau) d\tau \quad i = 1, 2, \dots, I \quad (141)$$

la salida filtrada de cada canal, que es la convolución del micrófono i con la respuesta al impulso del filtro correspondiente. La salida total filtrada $y_w(t)$ será la suma de cada uno de los canales del array:

$$y_w(t) = \sum_{i=1}^I \int_0^\infty h_i^*(t-\tau) y_i(\tau) d\tau = \int_0^\infty \mathbf{h}^H(t-\tau) \mathbf{y}(\tau) d\tau \quad (142)$$

Esta fórmula es similar a (44), que expresaba la conformación de haz convencional (estructura de filtrado y suma). Los coeficientes \mathbf{h} son totalmente equiparables a los coeficientes \mathbf{w} de (44). Es decir, el filtrado Wiener multicanal es equivalente a un conformador tipo filtrado y suma cuya salida requiere la condición MMSE con respecto a una señal de referencia. De hecho, si se rescribe la ecuación (142) en el dominio de la frecuencia:

$$\mathbf{Y}_w(\omega) = \mathbf{H}^H(\omega) \mathbf{Y}(\omega) \quad (143)$$

es mucho más equiparable a (44), siendo

$$\mathbf{H}(\omega) = [H_1(\omega), H_2(\omega), \dots, H_I(\omega)]^T \quad (144)$$

el filtro de Wiener multicanal en frecuencia.

El objetivo que persigue el filtrado de Wiener para el caso multicanal también es, como en el caso monocanal, minimizar la esperanza matemática del error cuadrático medio (MSE), que es la diferencia entre la salida filtrada y conformada $\mathbf{Y}_w(\omega)$ con la estimación del canal de referencia $X_0(\omega)$ [Naidu 01].

Ahora se define

$$\Phi_{YX_0}(\omega) = [\Phi_{Y_1X_0}(\omega), \Phi_{Y_2X_0}(\omega), \dots, \Phi_{Y_I X_0}(\omega)]^T \quad (145)$$

como el vector de espectros cruzados (37) de todos los canales con el de referencia.

También se define el error cometido $E(\omega)$ de la salida $\mathbf{Y}_w(\omega)$ con respecto a la señal de referencia $X_0(\omega)$:

$$E(\omega) = X_0(\omega) - Y_w(\omega) = X_0(\omega) - \mathbf{H}^H(\omega) \mathbf{Y}(\omega) \quad (146)$$

con potencia:

$$\Phi_{EE}(\omega) = E\left\{ \left[X_0(\omega) - \mathbf{H}^H(\omega) \mathbf{Y}(\omega) \right] \left[X_0^*(\omega) - \mathbf{Y}^H(\omega) \mathbf{H}(\omega) \right] \right\} \quad (147)$$

Esta ecuación se puede desarrollar de la siguiente manera:

$$\Phi_{EE}(\omega) = \Phi_{X_0 X_0}(\omega) - \mathbf{H}^H(\omega) \Phi_{YX_0}(\omega) - \Phi_{YX_0}^H(\omega) \mathbf{H}(\omega) + \mathbf{H}^H(\omega) \Phi_{YY}(\omega) \mathbf{H}(\omega) \quad (148)$$

siendo $\Phi_{YY}(\omega)$ la matriz de espectros cruzados del vector de salida del array, como en (42). La solución buscada del filtro \mathbf{H} de Wiener es aquélla que minimiza la potencia del error cometido (MMSE), es decir se cumple que el gradiente de la potencia de ruido respecto a las coordenadas \mathbf{H} del filtro de Wiener, tiene un valor nulo (condición de mínimo local):

$$\nabla_{\mathbf{H}} \Phi_{EE}(\omega) = \mathbf{0} \quad (149)$$

El autoespectro $\Phi_{EE}(\omega)$ es una función cuadrática de $\mathbf{H}(\omega)$ y por tanto tiene un mínimo global respecto de las coordenadas \mathbf{H} . La solución de (149) verifica la ecuación multicanal de Wiener-Hopf:

$$\Phi_{YY}(\omega) \mathbf{H}_W(\omega) = \Phi_{YX_0}(\omega) \quad (150)$$

donde $\mathbf{H}_W(\omega)$ es el filtro óptimo o filtro de Wiener buscado. Entonces el filtro de Wiener multicanal viene dado por:

$$\mathbf{H}_W(\omega) = \Phi_{YY}^{-1}(\omega) \Phi_{YX_0}(\omega) \quad (151)$$

El filtro de Wiener multicanal $\mathbf{H}_W(\omega)$ minimiza el ruido en la salida del array, o lo que es lo mismo maximiza la SNR. Para calcularlo hay que estimar la señal de referencia $x_0(t)$ ó $X_0(\omega)$. La referencia se obtiene frecuentemente a partir de una estimación del ruido $n(t)$ ó $N(\omega)$ presente en la salida conformada del array. Normalmente se supondrá ruido difuso homogéneo, con igual potencia en todos los micrófonos del array. Ahora bien, la solución (151) no puede estar en contradicción con la solución MVDR de (100) que daba los coeficientes $\mathbf{W}(\omega)$ de un conformador óptimo en el sentido de poca captación de ruido. De hecho, como se va a demostrar a continuación, la solución de Wiener es una generalización de la solución MVDR. Para ello se ha de factorizar el resultado (151), como se hace seguidamente.

Se debe ahora rescribir la ecuación fundamental del array (28) (dominio temporal) en el dominio de la frecuencia:

$$\mathbf{Y}(\omega) = \mathbf{A}(\omega) X_0(\omega) + \mathbf{N}(\omega) \quad (152)$$

donde, recuérdese que $X_0(\omega)$ es el espectro de la señal eléctrica de referencia captada por un micrófono omnidireccional, de sensibilidad $\mathbf{S}_0(\omega)$ y situado en el origen de coordenadas. El vector de espectros cruzados (145) de todos los canales del array con la referencia puede expresarse como:

$$\Phi_{YX_0}(\omega) = \Phi_{X_0 X_0}(\omega) \mathbf{A}(\omega) \quad (153)$$

donde se ha considerado que el ruido $\mathbf{N}(\omega)$ y la señal de referencia $X_0(\omega)$ están incorrelados. La matriz de espectros cruzados (42) se puede expresar como,

$$\Phi_{YY}(\omega) = \Phi_{X_0 X_0}(\omega) \mathbf{A}(\omega) \mathbf{A}^H(\omega) + \Phi_{NN}(\omega) \quad (154)$$

y el filtro de Wiener óptimo de (151) quedaría,

$$\mathbf{H}_W(\omega) = \left[\Phi_{X_0 X_0}(\omega) \mathbf{A}(\omega) \mathbf{A}^H(\omega) + \Phi_{NN}(\omega) \right]^{-1} \Phi_{X_0 X_0}(\omega) \mathbf{A}(\omega) \quad (155)$$

que se puede descomponer aplicando un poco de álgebra de matrices (fórmula de Sherman-Morrison-Woodbury), en :

$$\begin{aligned} \mathbf{H}_W(\omega) &= \left[\frac{\Phi_{X_0 X_0}(\omega)}{\Phi_{X_0 X_0}(\omega) + [\mathbf{A}^H(\omega) \Phi_{NN}^{-1}(\omega) \mathbf{A}(\omega)]^{-1}} \right] \frac{\Phi_{NN}^{-1}(\omega) \mathbf{A}(\omega)}{\mathbf{A}(\omega)^H \Phi_{NN}^{-1}(\omega) \mathbf{A}(\omega)} = \\ &= H_W(\omega) \mathbf{W}_{SD}(\omega) \end{aligned} \quad (156)$$

que constituye la solución de Wiener multicanal factorizada. En esta última expresión $\mathbf{W}_{SD}(\omega)$ corresponde a la solución superdirectiva o MVDR dada en (98) o (100), para ruido homogéneo en todos los micrófonos del array.

A simple vista, la ecuación (156) consta de dos factores. El primer factor es un filtro monocanal que se puede interpretar a su vez como el filtro de Wiener óptimo $H_W(\omega)$ aplicado a la salida conformada del array. El segundo factor constituye la solución obtenida para el conformador superdirectivo con $\mathbf{W}_{SD}(\omega)$ el vector de I coeficientes equivalente a la solución MVDR.

A continuación se demuestra que $H_W(\omega)$ de (156) es el filtro de Wiener monocanal óptimo aplicado a la salida conformada $Y_{SD}(\omega)$ del array superdirectivo. Efectivamente, si se aplican los coeficientes $\mathbf{W}_{SD}(\omega)$ del conformador MVDR a (152) se obtiene la salida conformada $Y_{SD}(\omega)$:

$$\begin{aligned} Y_{SD}(\omega) &= \mathbf{W}_{SD}^H(\omega) \mathbf{Y}(\omega) = \mathbf{W}_{SD}^H(\omega) \mathbf{A}(\omega) X_0(\omega) + \mathbf{W}_{SD}^H(\omega) \mathbf{N}(\omega) = \\ &= X_0(\omega) + \mathbf{W}_{SD}^H(\omega) \mathbf{N}(\omega) = X_0(\omega) + N(\omega) \end{aligned} \quad (157)$$

donde se ha aplicado la condición de no distorsión expresada en (97) según la cual:

$$\mathbf{W}_{SD}^H(\omega) \mathbf{A}(\omega) X_0(\omega) = X_0(\omega) \quad (158)$$

Se ha llamado:

$$N(\omega) = \mathbf{W}_{SD}^H(\omega) \mathbf{N}(\omega) \quad (159)$$

al ruido conformado con el array superdirectivo. $N(\omega)$ corresponde al ruido minimizado espacialmente por directividad según la condición MVDR. Si ahora se calcula mediante (157) la potencia de salida $\Phi_{Y_{SD} Y_{SD}}(\omega)$ sustituyendo $\mathbf{W}_{SD}(\omega)$ por su valor de (98), y si se admite que el ruido conformado $N(\omega)$ es incoherente con respecto a la señal $X_0(\omega)$ de referencia, se obtiene:

$$\begin{aligned} \Phi_{Y_{SD} Y_{SD}}(\omega) &= \Phi_{X_0 X_0}(\omega) + \mathbf{W}_{SD}^H(\omega) \Phi_{NN}(\omega) \mathbf{W}_{SD}(\omega) = \\ &= \Phi_{X_0 X_0}(\omega) + [\mathbf{A}^H(\omega) \Phi_{NN}^{-1}(\omega) \mathbf{A}(\omega)]^{-1} = \Phi_{X_0 X_0}(\omega) + \Phi_{NN}(\omega) \end{aligned} \quad (160)$$

que es el denominador del primer factor $H_W(\omega)$ de (156) con $\Phi_{NN}(\omega)$ la potencia de ruido conformado. Efectivamente, si se sustituye (160) en el denominador de $H_W(\omega)$ en (156) resulta que $H_W(\omega)$ es el filtro de Wiener monocanal óptimo aplicado a la salida conformada del array superdirectivo, tomando como señal de referencia la salida eléctrica $X_0(\omega)$:

$$\begin{aligned}
 H_W(\omega) &= \frac{\Phi_{X_0 X_0}(\omega)}{\Phi_{X_0 X_0}(\omega) + [\mathbf{A}^H(\omega) \Phi_{NN}^{-1}(\omega) \mathbf{A}(\omega)]^{-1}} = \frac{\Phi_{X_0 X_0}(\omega)}{\Phi_{Y_{SD} Y_{SD}}(\omega)} = \\
 &= \frac{\Phi_{X_0 X_0}(\omega)}{\Phi_{X_0 X_0}(\omega) + \Phi_{NN}(\omega)}
 \end{aligned} \tag{161}$$

Esta expresión es la de un filtro de Wiener monocanal y se puede poner también en términos de la relación señal a ruido a priori,

$$H_W(\omega) = \frac{\text{SNR}(\omega)}{1 + \text{SNR}(\omega)} \tag{162}$$

con

$$\text{SNR}(\omega) = \frac{\Phi_{X_0 X_0}(\omega)}{\Phi_{NN}(\omega)} \tag{163}$$

la relación señal a ruido a la salida del conformador.

En la Figura 29 se esquematiza el diagrama de bloques funcional de un filtro de Wiener multicanal. Una vez captada la señal sucia multicanal $\mathbf{Y}(\omega)$ se conforma por el método MVDR aplicando los coeficientes superdirectivos $\mathbf{W}'_{SD}(\omega)$ de (112), después de la alineación temporal de los I canales del array. Téngase en cuenta que para esta versión del conformador superdirectivo se necesitan estimar la matriz de espectros cruzados de ruido $\Phi_{NN}(\omega)$ –o la matriz de coherencia $\Gamma_{NN}(\omega)$ – y los retardos asociados τ_i de cada micrófono con respecto al centro del array. Asimismo, para calcular el filtro Wiener monocanal $H_W(\omega)$ se necesita estimar la potencia de la señal de referencia $\Phi_{X_0 X_0}(\omega)$ y la potencia de ruido remanente después del conformador $\Phi_{NN}(\omega)$.

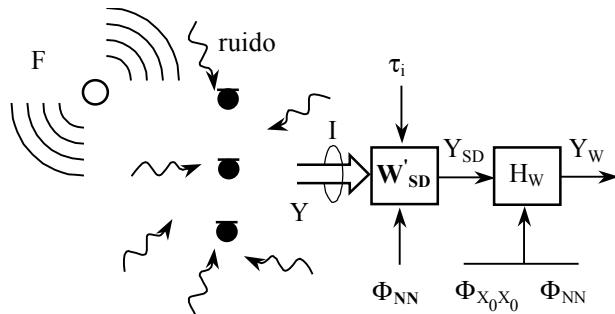


Figura 29. Factorización del filtro de Wiener multicanal en un conformador superdirectivo y un filtro de Wiener monocanal.

En la Figura 30 se representa en detalle el diagrama de bloques operativo de un filtro Wiener multicanal. La conformación superdirectiva se ha hecho según el esquema de la expresión (111) y la Figura 20. Como se ve se aplica el filtro de coeficientes $\mathbf{W}'_{SD}(\omega)$ de (112) a la señal alineada en tiempo $\mathbf{Y}_R(\omega)$ para obtener la solución superdirectiva $\mathbf{Y}_{SD}(\omega)$ y sobre ésta se aplica el filtro de Wiener monocanal $H_W(\omega)$. En la Figura 30 se propone que las estimaciones de los espectros de señal $\Phi_{X_0 X_0}(\omega)$ y de ruido $\Phi_{NN}(\omega)$ se pueden hacer a partir de la información multicanal que proporciona el array. Los coeficientes $\mathbf{W}'_{SD}(\omega)$ del conformador superdirectivo se pueden obtener a partir de la señal multicanal captada

–haciendo una estimación de $\Phi_{NN}(\omega)$ o $\Gamma_{NN}(\omega)$ – con lo que la conformación sería adaptativa (ver el apartado 2.3 de esta Tesis) o se pueden preestablecer para unas condiciones de ruido dadas. En este último caso lo normal es presuponer una matriz de coherencia espacial de ruido $\Gamma_{NN}(\omega)$ en condiciones de ruido difuso.

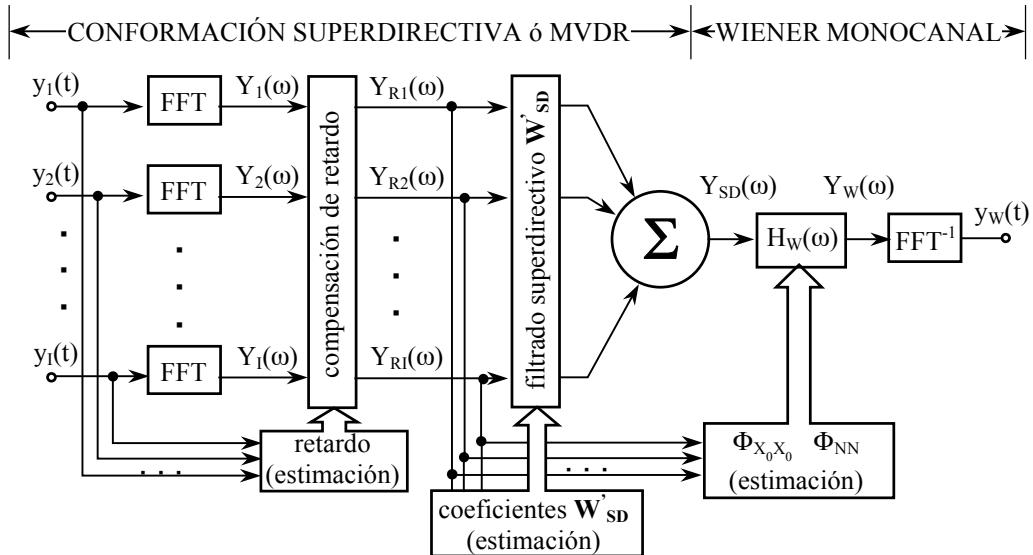


Figura 30. Esquema de implementación del filtrado Wiener multicanal en el dominio de la frecuencia, a partir de un conformador superdirective y un filtrado Wiener aplicado a la salida del conformador.

Para finalizar este apartado se hace una breve recapitulación sobre las propiedades y efectos del filtrado de Wiener multicanal.

Un filtro de Wiener multicanal está compuesto por dos etapas. Una primera etapa de conformación, por el método de filtrado y suma, según los coeficientes de un array superdirective o MVDR. Este conformador obtiene mediante selectividad espacial, la salida con mínimo ruido para una determinada frecuencia, pero está sujeto a la condición de que la salida sea sin distorsión, es decir, que a esa frecuencia la señal –representada por $X_0(\omega)$ – no se verá amplificada ni atenuada por el array. Es una condición de normalización. Si lo que se desea es maximizar la relación SNR a la salida, la condición de normalización debe romperse. Es decir, ahora sí que se debe atenuar a la señal cuando el ruido es muy alto, para mantener alta la relación SNR. Para ello se aplica, a la salida conformada $Y_{SD}(\omega)$ del array, el filtro de Wiener monocanal $H_w(\omega)$ de (161) o (162). Este filtro vale la unidad cuando la SNR es muy alta, con lo que para esa frecuencia concreta el filtrado de Wiener multicanal constituye aproximadamente un array MVDR. Cuando la SNR sea muy baja, el filtro $H_w(\omega)$ atenúa fuertemente la señal y el ruido. Por ello se dice que la solución MVDR minimiza el ruido en banda estrecha, y el filtro de Wiener multicanal minimiza el ruido (maximiza la SNR) en banda ancha, aun a pesar de distorsionar linealmente la salida, atenuando unas frecuencias con respecto a otras.

Si para conocer el valor de los pesos $W_{SD}(\omega)$ – $W'_{SD}(\omega)$ – del conformador MVDR es necesario conocer, estimar o prever la matriz de covarianzas o de espectros cruzados del ruido presente en los micrófonos del array, con el postfiltro de Wiener además hay que conocer o estimar el valor de la potencia de la señal, representada por $\Phi_{X_0X_0}(\omega)$. En el caso de que la señal de voz fuere estacionaria, no será muy difícil hacer esas estimaciones. Sin embargo ese no es el caso y habrá que realizar las estimaciones a corto plazo, durante instantes de tiempo

en los que la voz se pueda considerar como estacionaria. Como en esta Tesis se trabaja en el dominio de la frecuencia, usando la FFT de tramas cortas de la señal multicanal, habrá que elegir convenientemente el tamaño de la ventana temporal de estimación, de tal manera que sea lo suficientemente corta para que manifieste las características espetrales instantáneas de las estimaciones y lo suficientemente larga para que no se refleje excesiva variabilidad intertrama en el postfiltro $H_w(\omega)$ calculado, lo que daría lugar a un sonido artificial o distorsionado.

4.1.2 Sustracción espectral

La teoría del filtrado de Wiener tiene un ámbito teórico de aplicación en señales estacionarias. Pero la voz no es una señal estacionaria. Tampoco el ruido es estacionario en muchos casos, y aunque así lo sea puede ser difícil de estimar. Cuando se aplican métodos de postfiltrado analizando el espectro de forma casi instantánea mediante la transformada STFT, el filtrado de Wiener se transforma en sustracción espectral, que puede considerarse como una generalización del primero. La extensión de las técnicas monocanal de sustracción espectral a la captación multicanal es equivalente a la que se hacía en el apartado anterior para el filtrado de Wiener. Un filtrado óptimo multicanal, en este caso mediante la sustracción espectral, es equivalente a un filtrado monocanal de la señal $y_{SD}(t)$ previamente conformada de forma óptima mediante un array superdirectivo.

La sustracción espectral [Boll 79] [Vaseghi 96] es un método para recuperar el módulo del espectro de una señal en presencia de ruido aditivo. Como su nombre indica se resta el ruido, que es estimado durante los intervalos de no-actividad de la señal de habla. Se supondrá que el ruido es quasi estacionario, en el sentido de que su espectro no cambia sustancialmente en los intervalos en que éste no puede ser estimado, y por eso se puede hacer una estimación del mismo con un gran periodo de latencia ante las nuevas actualizaciones que van surgiendo en el transcurso del tiempo.

La sustracción espectral procesa espectros de potencia y por lo tanto no restaura la fase, que es la misma de la señal sucia, antes del filtro. Esto no debe considerarse importante ya que el oído humano es poco sensible al ruido de fase. En este sentido la sustracción espectral es equivalente al filtrado de Wiener ya que $H_w(\omega)$ en (161) es real y siempre positivo, luego sólo procesa el módulo de la señal a la salida del conformador $Y_{SD}(\omega)$.

A partir de ahora se considerará el procesado monocanal sobre la salida conformada de un array de micrófonos, dando como buena la factorización del filtro de Wiener multicanal hecha en el apartado anterior (4.1.1) de esta Tesis. Efectivamente, si se considera una ecuación paralela a la (156) se podría establecer, en el dominio de la frecuencia, un filtro $H_S(\omega)$ de sustracción espectral multicanal como sigue:

$$H_S(\omega) = H_S(\omega) W(\omega) \quad (164)$$

con $W(\omega)$ el conformador y $H_S(\omega)$ el filtro monocanal que se aplica a la salida conformada del array, de tal manera que,

$$Y_S(\omega) = [H_S(\omega) W(\omega)]^H Y(\omega) = H_S(\omega) Y(\omega) \quad (165)$$

con $Y_S(\omega)$ la salida filtrada por sustracción espectral e $Y(\omega)$ la salida del conformador. Nótese que ahora $W(\omega)$ se refiere a una conformación genérica y no particularmente a una conformación superdirectiva, ya que la sustracción espectral no constituye, en general, un

filtrado óptimo en el sentido MMSE, que era necesario para la estimación del filtro de Wiener multicanal.

Se consideran ahora por separado el módulo y la fase tanto de $Y(\omega)$ como de $Y_S(\omega)$, por medio de las siguientes expresiones:

$$Y(\omega) = |Y(\omega)| \exp[j\phi_Y(\omega)] \quad (166)$$

$$Y_S(\omega) = |Y_S(\omega)| \exp[j\phi_{Y_S}(\omega)] \quad (167)$$

con

$$\phi_Y(\omega) = \arg[Y(\omega)] \quad (168)$$

$$\phi_{Y_S}(\omega) = \arg[Y_S(\omega)] \quad (169)$$

las fases respectivas de la señal conformada por el array antes y después del filtro de sustracción espectral.

A continuación se exponen la formulación y planteamientos básicos en los que se sustenta la sustracción espectral. La sustracción espectral persigue obtener a la salida del filtro $H_S(\omega)$ una estimación del espectro de la señal de referencia $\hat{X}_0(\omega)$, usando la siguiente ecuación genérica:

$$|Y_S(\omega)|^\gamma = |\hat{X}_0(\omega)|^\gamma = |Y(\omega)|^\gamma - \alpha |N(\omega)|^\gamma \quad (170)$$

donde $|N(\omega)|$ es el espectro de ruido en módulo, después de la conformación de haz. El parámetro $\gamma > 1$ controla la figura de ganancia asociada al postfiltro y determina los diferentes tipos de sustracción espectral –sustracción espectral en potencia ($\gamma = 2$), en magnitud ($\gamma = 1$) o en posiciones intermedias– y el parámetro α controla la cantidad de ruido que va a ser sustraído. Si se sustrae más ruido del presente (o el estimado), se hablará de sobresustracción ($\alpha > 1$) o en caso contrario será una infrasustracción. En (170) se ha considerado que el resultado $Y_S(\omega)$ del filtro de sustracción es una estimación de la señal de referencia $\hat{X}_0(\omega)$, puesto que esta ecuación, a diferencia del filtro de Wiener, no producirá en general la referencia $X_0(\omega)$, aun estimando perfectamente el valor del ruido $N(\omega)$ y suponiendo su incorrelación total con $X_0(\omega)$. En la práctica eso no importa mucho, ya que para un postfiltrado ciego, siempre habrá que utilizar estimaciones de la señal o del ruido, y (170) permite una mayor versatilidad que el filtrado de Wiener.

Está claro que en (170) existe una incongruencia, puesto que si $|Y(\omega)|^\gamma < \alpha |N(\omega)|^\gamma$ se obtendría un módulo negativo a la salida del sustractor, lo cual no es posible. Este efecto se produce normalmente cuando se elige un factor α muy alto que casi siempre es consecuencia de una SNR muy mala, o cuando se estima incorrectamente el ruido. Si es así, se tiene que limitar el segundo miembro de (170) a valores positivos, mediante la truncación o la extracción del módulo del mismo. En cualquier caso esto siempre ocasiona distorsión audible.

Cuando se trabaja en el dominio de la frecuencia, el proceso de la sustracción espectral se ilustra en la Figura 31. La salida $Y(\omega)$ del conformador se descompone en módulo y fase. El módulo se trata según (170) para obtener una estimación del módulo de la señal de referencia $X_0(\omega)$. Éste sirve para restaurar, junto con la fase $\phi_Y(\omega)$ de la señal a la salida del conformador, la estimación de la señal de referencia, que aquí se ha llamado $y_S(t)$. Se considera por tanto que la fase de la señal sucia es un buen estimador de la fase de la señal recuperada. Aparte de que la carencia de sensibilidad auditiva a la fase de la señal acústica

justifica lo anterior, se ha demostrado [Ephraim 84] que esta suposición es conveniente para la sustracción espectral cuando es usada la transformada STFT que es el método que se empleará aquí para el de cálculo del espectro en frecuencia.

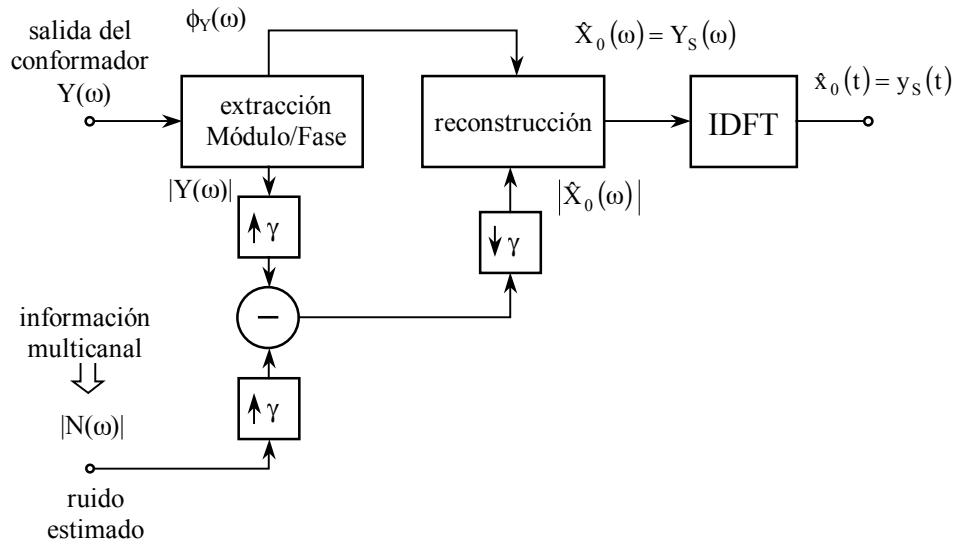


Figura 31. Esquema funcional que representa el postfiltrado mediante sustracción espectral de la salida conformada de un array.

Si se organiza (170) de la siguiente manera:

$$|\hat{X}_0(\omega)|^\gamma = |Y(\omega)|^\gamma - \alpha |N(\omega)|^\gamma = |H_S(\omega)|^\gamma |Y(\omega)|^\gamma \quad (171)$$

se puede obtener una expresión para el módulo del filtro restaurador $H_S(\omega)$ que se aplica a la salida $Y(\omega)$ del conformador:

$$|H_S(\omega)| = \left[\frac{|Y(\omega)|^\gamma - \alpha |N(\omega)|^\gamma}{|Y(\omega)|^\gamma} \right]^{\frac{1}{\gamma}} = \left[1 - \alpha \left[\frac{|N(\omega)|}{|Y(\omega)|} \right]^\gamma \right]^{\frac{1}{\gamma}} = \frac{|\hat{X}_0(\omega)|}{\left[|\hat{X}_0(\omega)|^\gamma + \alpha |N(\omega)|^\gamma \right]^{\frac{1}{\gamma}}} \quad (172)$$

La expresión (172) se suele dar en función de la relación señal a ruido. En la práctica, puesto que antes de filtrar no se conoce la estimación de la señal de referencia $\hat{X}_0(\omega)$, frecuentemente se considera la relación señal a ruido a posteriori $\text{SNR}_{\text{post}}(\omega)$:

$$\text{SNR}_{\text{post}}(\omega) = \frac{\Phi_{YY}(\omega)}{\Phi_{NN}(\omega)} = \frac{|Y(\omega)|^2}{|N(\omega)|^2} \quad (173)$$

Dividiendo numerador y denominador de (172) por el módulo del ruido $|N(\omega)|$ se puede expresar también $|H_S(\omega)|$ mediante la relación señal a ruido $\text{SNR}(\omega)$ a priori (163):

$$|H_S(\omega)| = \left[1 - \frac{\alpha}{\text{SNR}_{\text{post}}^{\frac{1}{2}}(\omega)} \right]^{\frac{1}{\gamma}} = \frac{\text{SNR}^{\frac{1}{2}}(\omega)}{\left[\text{SNR}^{\frac{1}{2}}(\omega) + \alpha \right]^{\frac{1}{\gamma}}} \quad (174)$$

A partir de (174) se puede establecer, para la sustracción espectral genérica de (170) o (171), la relación entre $\text{SNR}(\omega)$ y $\text{SNR}_{\text{post}}(\omega)$ siendo:

$$\text{SNR}_{\text{post}}^{\frac{\gamma}{2}}(\omega) = \text{SNR}^{\frac{\gamma}{2}}(\omega) + \alpha \quad (175)$$

Un caso particular muy empleado en la práctica es la sustracción espectral de potencia [Ephraim 84] [Etter 94], así llamada cuando los parámetros de forma valen, $\gamma = 2$ y $\alpha = 1$:

$$|H_S(\omega)| = \left[1 - \frac{1}{\text{SNR}_{\text{post}}(\omega)} \right]^{\frac{1}{2}} = \left[\frac{\text{SNR}(\omega)}{1 + \text{SNR}(\omega)} \right]^{\frac{1}{2}} \quad (176)$$

La ecuación (176) debe compararse con (162) que representa el filtrado de Wiener tradicional. El filtro para la sustracción espectral de potencia equivale (considerando en ambos casos estimaciones a corto plazo de SNR) a la raíz cuadrada del filtro óptimo de Wiener, con lo cual la sustracción espectral de potencia no maximiza la SNR en banda ancha, como lo hacía el filtrado de Wiener. Eso no importa ya que en cualquier caso se trabajará con estimaciones, y la solución óptima no se obtendrá nunca. De todas formas, en [Gay 00] se demuestra cómo la sustracción espectral de potencia es un estimador óptimo en cuanto a máxima verosimilitud de la varianza de la señal, mientras que el filtrado de Wiener es una estimación óptima en cuanto a mínimo error cuadrático medio (MMSE) del espectro de la señal de referencia.

A parte del exponente, la diferencia entre el filtrado de Wiener (162) y la sustracción espectral (176) es bastante teórica y poco práctica [Vaseghi 96]. El filtrado de Wiener utiliza la autocorrelación (o densidad espectral de potencia) para estimar la señal y el ruido, es decir promedios muestrales suponiendo diferentes realizaciones del proceso de generación de señal y ruido. Por contra, un filtro de sustracción espectral, en su versión más clásica, utiliza el espectro instantáneo de la señal ruidosa y el promedio temporal del ruido a más largo plazo. Para procesos ergódicos, los promedios temporales coinciden con los promedios muestrales, con lo que el filtro de Wiener coincidirá con la raíz cuadrada del de sustracción espectral de potencia. La señal de habla es claramente no estacionaria y por eso ambos procesos no son coincidentes. En la práctica tanto el filtrado de Wiener como la sustracción espectral y los estimadores asociados suelen utilizar la transformada STFT cuando se aplican a voz, debido a la no estacionariedad intrínseca de la misma, ya que a partir de ésta se puede reconstruir en tiempo la señal filtrada –véase [Cochiere 83] para más detalles–. Si se elige adecuadamente la duración temporal de los promedios espectrales de la señal y el ruido, la sustracción espectral se aproximará bastante al filtrado óptimo de Wiener, y por tanto las realizaciones prácticas del filtrado de Wiener y del sustractor espectral serán básicamente iguales.

Otra versión muy utilizada de la sustracción espectral es el sustractor de magnitud [Boll 79] que reconstruye la señal estimada mediante la simple diferencia de módulos, según expresa la siguiente ecuación, para $\gamma = 1$ y $\alpha = 1$:

$$|H_S(\omega)| = 1 - \frac{1}{\frac{1}{\text{SNR}_{\text{post}}^{\frac{1}{2}}(\omega)}} = \frac{\text{SNR}^{\frac{1}{2}}(\omega)}{\text{SNR}^{\frac{1}{2}}(\omega) + 1} \quad (177)$$

En la Figura 32 se representa la ganancia del filtro de sustracción espectral H_S en función de la relación señal a ruido a priori de (163), y en función de los parámetros α y γ . Se compara también con el filtro de Wiener tradicional. Se puede observar que los parámetros α

y γ controlan la severidad de la atenuación del postfiltro $H_S(\omega)$. Si α es muy grande, el sustractor espectral atenúa mucho la señal $Y(\omega)$ a la salida del conformador, incluso para relaciones SNR(ω) buenas. A α se le llama normalmente factor de sobresustracción. Parece que un α alto sólo convendrá cuando la SNR(ω) sea muy mala, ya que si no la señal de salida va a distorsionarse por una sustracción excesiva. Los parámetros γ y α controlan el valor de SNR(ω) a partir del cual la sustracción espectral comienza a ser pequeña, es decir, el punto de inflexión de las curvas de la Figura 32. Por ello parece conveniente, si SNR(ω) es muy mala, modificar conjuntamente α y γ , para así sobresustraer sólo las tramas de voz con SNR(ω) bajas.

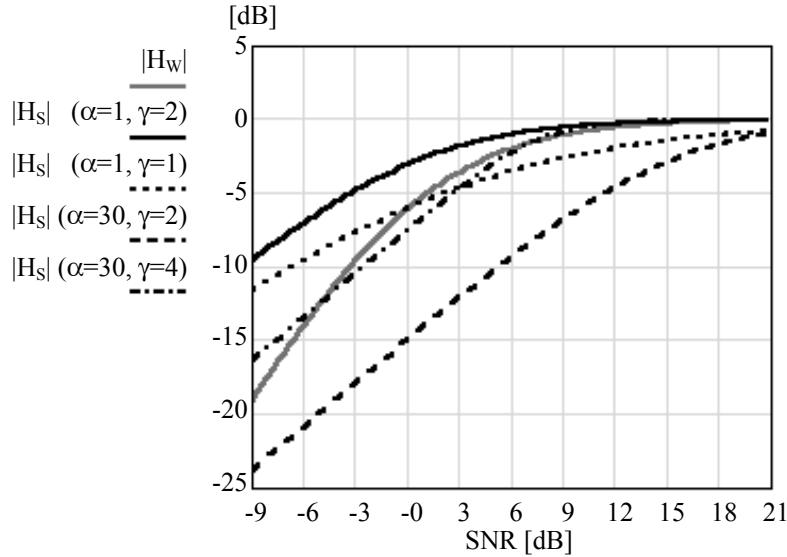


Figura 32. Ganancia del filtro de sustracción espectral de (172) o (177) para distintos valores de los factores α y γ en función de SNR a priori. Se compara con el filtro de Wiener de (162).

Existen versiones más sofisticadas del esquema básico de sustracción espectral mostrado en la ecuación (172), como la sustracción espectral no lineal (NSS, *Non-linear Spectral Subtraction*). Ésta establece una dependencia no lineal de la ganancia del filtro de sustracción $H_S(\omega)$ con la relación a ruido a priori SNR(ω) [Lockwood 92]. La versión clásica del sustractor no lineal consiste en sobresustraer aquellas componentes espectrales que estén muy contaminadas por el ruido, es decir, con bajas SNR(ω). La razón es que en estas condiciones el sustractor espectral convencional sólo va a introducir distorsión y por lo tanto es mejor sustraer en exceso. En (178) se presenta la expresión correspondiente a un sustractor no lineal, controlado por la relación señal a ruido a posteriori estimada.

$$|H_S(\omega)| = \begin{cases} \left[1 - \alpha \frac{1}{\text{SNR}_{\text{post}}^{\frac{1}{2}}(\omega)} \right]^{\frac{1}{\gamma}} & \text{si } \text{SNR}_{\text{post}}^{\frac{1}{2}}(\omega) > \alpha + \beta \\ \left[\frac{\beta}{\text{SNR}_{\text{post}}^{\frac{1}{2}}(\omega)} \right]^{\frac{1}{\gamma}} & \text{en otro caso} \end{cases} \quad (178)$$

o bien en función de la relación señal a ruido a priori:

$$|H_S(\omega)| = \begin{cases} \frac{\text{SNR}^{\frac{1}{2}}(\omega)}{\left[\frac{\gamma}{\text{SNR}^{\frac{1}{2}}(\omega) + \alpha} \right]^{\frac{1}{\gamma}}} & \text{si } \text{SNR}^{\frac{1}{2}}(\omega) > \beta \\ \left[\frac{\beta}{\frac{\gamma}{\text{SNR}^{\frac{1}{2}}(\omega) + \alpha}} \right]^{\frac{1}{\gamma}} & \text{en otro caso} \end{cases} \quad (179)$$

donde se ha considerado (175) para establecer la última relación. En estas dos últimas expresiones, el parámetro β se llama factor de suelo espectral (*spectral flooring*) y sirve para controlar la ganancia del filtro H_S cuando la relación señal a ruido es muy baja.

En la Figura 33 se muestra la ganancia del filtro de sustracción (178) ó (179) en función de diversos valores de los parámetros α , β y manteniendo $\gamma=2$.

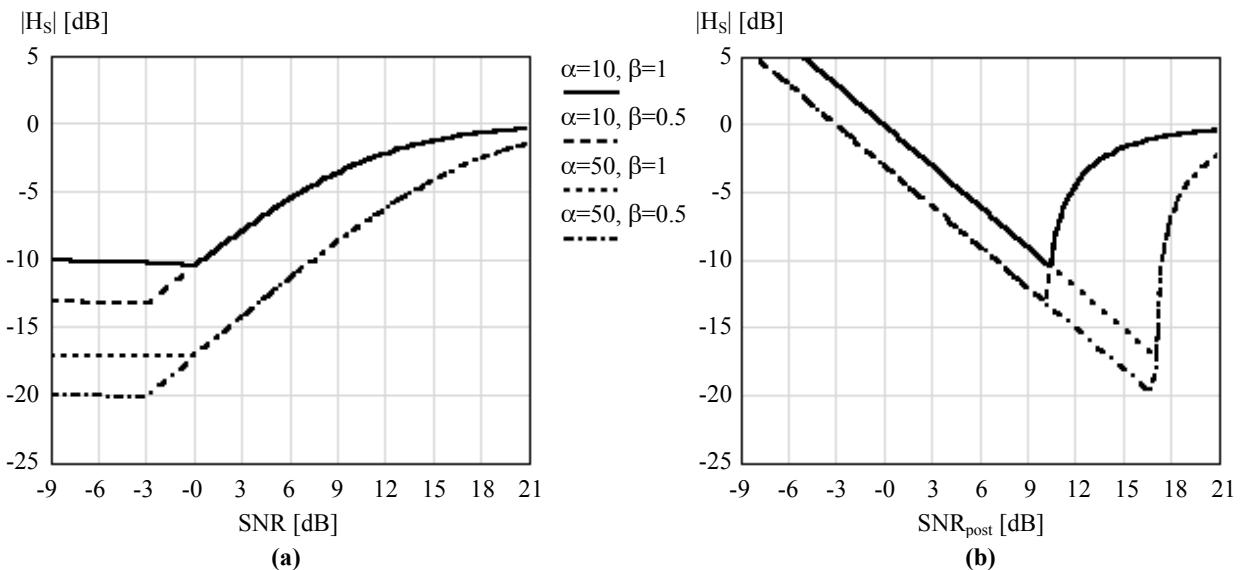


Figura 33. Ganancia del filtro de sustracción espectral de (178) y (179) para distintos valores del factor de sobresustracción α y del factor de suelo espectral β , manteniendo $\gamma=2$. **(a)** En función de la relación señal a ruido a priori, SNR . **(b)** En función de la relación señal a ruido a posteriori, SNR_{post} .

A continuación se explica cómo funciona un sustractor no lineal. Para SNR alta –véanse las expresiones (178)-a ó (179)-a– se aplica el sustractor espectral convencional de (174). En este caso es conveniente sobresustraer (factor de sobresustracción $\alpha \gg 1$) y así se evita reflejar en la señal de salida la varianza del ruido aditivo. Pero cuando la SNR sea baja –véanse (178)-b ó (179)-b– no se aplicará la sustracción espectral, dejando pasar el ruido multiplicado por el factor β de suelo espectral (*spectral flooring*) [Berouti 79]. Es decir, en buenas condiciones (alta SNR) se sobresustrae, y en malas condiciones (baja SNR) se infrasustrae, dejando pasar atenuado el ruido estimado. En (179) queda patente que el factor β es el umbral de $\text{SNR}(\omega)$ a partir del cual se aplica la sustracción espectral convencional. Esto también se manifiesta en la Figura 33(a), donde se aprecia que la inflexión de las curvas se

sitúa en $\text{SNR}(\omega) = -3\text{dB}$ ($\beta = 0.5$) y en $\text{SNR}(\omega) = 0\text{dB}$ ($\beta = 1$). Existen en la literatura científica múltiples alternativas propuestas para el esquema general de sustracción espectral expuesto en (178) ó (179) [Lockwood 92], con diferentes soluciones para la obtención de la estimación de ruido y con diferentes funciones de transferencia [Tsoukalas 97] a las expresadas en (178) ó (179).

La sustracción espectral produce distorsión audible en la señal filtrada de salida. Esta distorsión, que puede ser muy elevada cuando se sobresustrae ($\alpha \gg 1$), se produce cuando el filtro de sustracción posee componentes espectrales muy pequeños o incluso negativos numéricamente. Esto se da en malas condiciones de ruido (baja SNR), fundamentalmente porque los promedios temporales del ruido no coinciden con sus valores instantáneos. El efecto audible de esta distorsión se llama ruido musical y se reconoce por un sonido metálico de fondo que acompaña a la señal de voz. El origen de este efecto básicamente está en que el espectro de la señal procesada posee líneas espectrales que varían rápidamente en el tiempo, siguiendo los cambios aleatorios del espectro del ruido. Existen varias estrategias para disminuir este ruido musical. Por ejemplo, reduciendo la varianza del ruido instantáneo mediante un filtrado recursivo sobre la señal contaminada $Y(\omega)$ [Vaseghi 96]. Para señal de habla esto no tiene gran utilidad porque las variaciones de espectro de la voz se producen con gran velocidad, y no es posible realizar un suavizado temporal del mismo. Existen otros métodos en los que se analiza instantáneamente el espectro y se eliminan aquellas barras espectrales candidatas a ser ruido musical [Vaseghi 96]. Sin embargo, el método más profusamente adoptado hasta ahora para reducir el ruido musical y otras distorsiones inherentes al sustractor espectral es la sustracción espectral no lineal expuesta anteriormente. No obstante, existen en este sentido otras soluciones, que se desvían en cierta medida de la sustracción espectral, y que se exponen a continuación.

4.1.3 Filtrado perceptual. Método ANS

Para mejorar señal de habla con ruido aditivo, los métodos de limpieza más usados son los ya mencionados de sustracción espectral y el filtrado de Wiener. Básicamente estos sistemas consiguen la mejora estimando el ruido $N(\omega)$ en las tramas temporales de no-actividad de voz. Como se ha visto, el sustractor espectral no lineal modifica su función de transferencia dependiendo de la relación señal a ruido para no introducir excesiva distorsión en la señal filtrada, inherente al procesado basado en la STFT. Parece evidente, que si lo que interesa es disminuir la sensación subjetiva de distorsión habrá que considerar las características del sistema auditivo humano. En los últimos años ha cobrado gran pujanza el filtrado basado en las propiedades de enmascaramiento auditivo. Esta técnica, conocida como supresión de ruido audible (*Audible Noise Suppression* ó ANS) se ha aplicado con éxito en señales de voz monocanal [Akbari 95] [Gustafsson 98] [Tsoukalas 97] [Virag 95] [Virag 99] y también a arrays microfónicos [Sánchez-Bote 01-a] [Sánchez-Bote 03-a].

El método ANS para la supresión de ruido se basa en el procesado por bandas críticas, de tal manera que se produzca una reducción del mismo tan sólo hasta el umbral en el que dicho ruido se manifiesta como audible. Dicho umbral es el llamado umbral de enmascaramiento. En [Johnston 88] [Hwang 96] [Zwicker 99] [Pohlmann 00] e [ISO/IEC 13818-3] se exponen los principios básicos del enmascaramiento auditivo humano. El umbral de enmascaramiento depende de la frecuencia y de la cantidad de señal libre de ruido que haya en una determinada banda crítica. Se llamará aquí $T(\omega)$ (AMT *Auditory Masking Threshold*) y tiene dimensiones de potencia como los autoespectros de señal $\Phi_{YY}(\omega) = |Y(\omega)|^2$

y de ruido $\Phi_{NN}(\omega) = |N(\omega)|^2$. El objetivo fundamental que persigue el filtrado perceptual o filtrado auditivo es que el ruido residual inherente a todo procesado de limpieza de ruido sea “perceptualmente blanco”, es decir percibido como espectralmente blanco desde el punto de vista subjetivo. De esta forma se consigue reducir el ruido musical inherente a los sustractores espectrales clásicos desarrollados en el apartado 4.1.2. Si se compara subjetivamente una señal de voz procesada mediante sustracción espectral no lineal basada en SNR con otra procesada mediante el método auditivo o ANS la última será apreciada como de mejor calidad, aunque ambas tengan cuantitativamente la misma tasa de mejora en la SNR resultante.

El umbral de enmascaramiento $T(\omega)$ es un nivel definido para cada una de las bandas críticas auditivas, e indica una cota superior por debajo de la cual el ruido queda enmascarado por la señal. Tan sólo es necesario encontrar un filtro que realice una supresión de ruido teniendo en cuenta el nivel del ruido en cada banda crítica, relativo al umbral de enmascaramiento $T(\omega)$. De este modo, si el umbral de enmascaramiento es alto, no será necesaria mucha supresión de ruido y viceversa. Por tanto la respuesta subjetiva del procesador será mucho mejor que el caso del sustractor espectral no lineal clásico (178) ó (179) o el filtro de Wiener (162).

Existen múltiples posibilidades a la hora de aplicar esquemas de postfiltrado que utilicen los métodos auditivos. Así se podría aplicar [Virag 99] el sustractor no lineal de (179) con los parámetros α y β dependientes de los umbrales de enmascaramiento $T(\omega)$. Sin embargo, por la importancia que tiene para el desarrollo y resultados expresados en los capítulos posteriores de esta Tesis, se presta especial atención al postfiltro $H_{ANS}(\omega)$ de supresión perceptual monocanal adaptado de [Tsoukalas 97] y dado por:

$$\begin{aligned} H_{ANS}(\omega) &= \frac{1}{\left[\left[\delta(\omega) \frac{|N(\omega)|^2}{|Y(\omega)|^2} \right]^\varepsilon + 1 \right]^{\frac{1}{2}}} = \\ &= \frac{1}{\left[\frac{\delta^\varepsilon(\omega)}{\text{SNR}_{\text{post}}^\varepsilon(\omega)} + 1 \right]^{\frac{1}{2}}} = \frac{1}{\left[\frac{\delta^\varepsilon(\omega)}{[\text{SNR}(\omega) + 1]^\varepsilon} + 1 \right]^{\frac{1}{2}}} \end{aligned} \quad (180)$$

En esta expresión se ha considerado la relación de (175) entre $\text{SNR}(\omega)$ y $\text{SNR}_{\text{post}}(\omega)$ con $\alpha = 1$ y $\gamma = 2$, equivalente a sustracción espectral de potencia. El parámetro $\delta(\omega)$ controla la ganancia del filtro $H_{ANS}(\omega)$ por lo que debe depender del valor del ruido relativo al umbral de enmascaramiento $T(\omega)$, como se verá más adelante. El parámetro ε es un factor de forma que suele valer la unidad. Aunque la función de transferencia (180) no se ajusta exactamente al esquema del sustractor espectral convencional (176) ó (177), de forma general puede considerarse equivalente. En la Figura 34 se representa la función de transferencia del filtro $H_{ANS}(\omega)$ según la relación señal a ruido a priori SNR y a posteriori SNR_{post} . El parámetro δ controla el umbral de SNR por debajo de la cual la sustracción es elevada y el parámetro ε controla la velocidad de supresión con respecto a la SNR. Además, a partir de la Figura 34(a), se observa cierto grado de parecido entre las curvas de supresión auditiva y de sustracción no lineal con suelo espectral de la Figura 33(a). Es decir, el filtro H_{ANS} , cumple la condición de

no sustraer excesivamente cuando la relación señal a ruido a priori es muy mala, ya que el grado de supresión permanece constante para SNR baja.

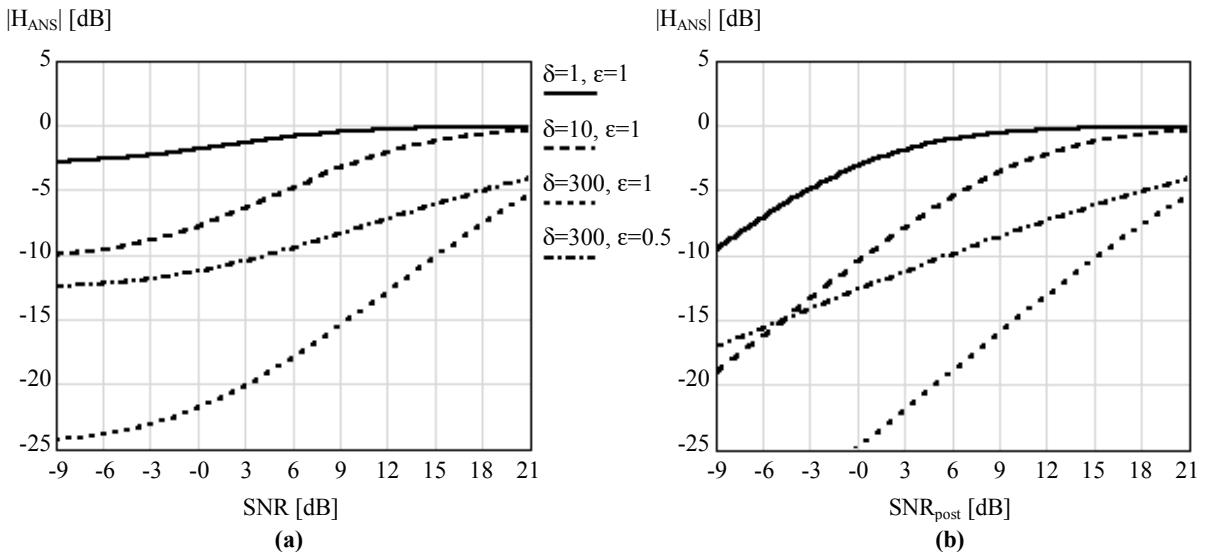


Figura 34. Ganancia del filtro $H_{ANS}(\omega)$ (180) para distintos valores de los factores δ y ϵ . **(a)** En función de la relación señal a ruido a priori, $SNR(\omega)$. **(b)** En función de la relación señal a ruido a posteriori, $SNR_{post}(\omega)$.

Se ha dicho, que el parámetro principal que controla la cantidad de supresión perceptual según el método ANS es el parámetro δ . Ese parámetro debe depender del umbral de enmascaramiento $T(\omega)$. A continuación se explica cómo debe ser esa dependencia.

Supresión de ruido audible usando los valores del umbral de enmascaramiento $T(\omega)$

En [Tsoukalas 97] se hace un análisis exhaustivo de las posibles dependencias que puede tener $\delta(\omega)$ con $T(\omega)$ para hacer una supresión audible eficiente. De entre todas las alternativas propuestas, aquí se estudia con más detalle la siguiente expresión, por la importancia que tiene en el desarrollo posterior de esta Tesis:

$$\delta(\omega) = \left[1 + \frac{T(\omega)}{|N(\omega)|^2} \right] \left[\frac{|N(\omega)|^2}{T(\omega)} \right]^{\frac{1}{\epsilon}} = \left[1 + \frac{1}{NMR(\omega)} \right] [NMR(\omega)]^{\frac{1}{\epsilon}} \quad (181)$$

En esta expresión $NMR(\omega)$ es la relación de ruido a umbral de enmascaramiento (*Noise to Masking Ratio*), definida como sigue:

$$NMR(\omega) = \frac{|N(\omega)|^2}{T(\omega)} \quad (182)$$

En la Figura 35 se representa el factor $\delta(\omega)$ en función de $NMR(\omega)$ según la expresión (181). El efecto de la fórmula (181) es aumentar el parámetro $\delta(\omega)$ de sobresustracción cuando el ruido supera ampliamente al umbral de enmascaramiento –ó $NMR(\omega) \gg 1$ –. Cuando el ruido es igual o está por debajo de $T(\omega)$ –ó $NMR(\omega) \ll 1$ –, entonces $\delta(\omega)$ vale cero y por tanto $H_{ANS}(\omega) = 1$, que no aplicará ninguna atenuación a la señal de entrada del supresor auditivo

—ver (180)—. El parámetro ϵ controla de nuevo la velocidad de cambio de $\delta(\omega)$ respecto al valor de NMR(ω).

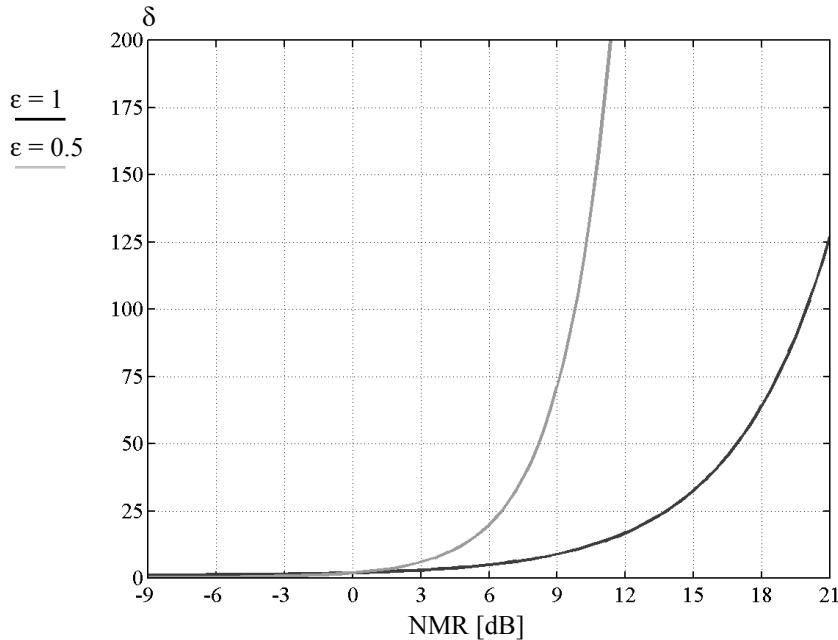


Figura 35. Parámetro δ en función de la relación NMR según (181).

Cálculo de los umbrales de enmascaramiento $T(\omega)$ o $T(b)$

La obtención de $T(\omega)$ es el paso previo necesario para que el método ANS pueda ser aplicado. Aunque hasta ahora se ha definido el parámetro $T(\omega)$ como función densidad espectral de potencia, en lo sucesivo se trabajará con $T(b)$, que representa los umbrales de enmascaramiento (con unidades de potencia) para cada banda crítica auditiva, b . El cálculo de $T(b)$ debe hacerse partiendo de una muestra de voz limpia que tiene que ser estimada previamente. Aquí se volverá a considerar, como en la sustracción espectral, la estimación del canal de referencia $\hat{X}_0(\omega)$ en el dominio de la frecuencia. El método de cálculo de $T(\omega)$ desarrollado a continuación procede de J. D. Johnston [Johnston 88] aunque también aparece en [Tsoukalas 97] y [Virág 99] y en la norma [ISO/IEC 13818-3]. A continuación se resume brevemente la forma de obtención del umbral de enmascaramiento $T(\omega)$.

1.- En primer lugar debe hacerse un análisis en bandas críticas auditivas. Es decir se debe calcular el espectro por bandas críticas de la señal de referencia $\hat{X}_0(\omega)$. Aquí ese espectro se llamará $|\hat{X}_0(b)|^2$, donde b es el índice de cada banda crítica. De forma más rigurosa b es el índice de frecuencia de la escala bark [Zwicker 99], aunque como la descomposición en frecuencias bark coincide con la descomposición en bandas críticas auditivas, se admite que b es un entero desde 1 hasta B , siendo B el número de bandas críticas consideradas. Así:

$$|\hat{X}_0(b)|^2 = \int_{\omega_{ib}}^{\omega_{fb}} |\hat{X}_0(\omega)|^2 d\omega \quad (183)$$

con ω_{ib} y ω_{fb} son las pulsaciones inicial y final de la banda crítica b . Por supuesto, si el espectro $|\hat{X}_0(b)|^2$ se obtiene mediante la transformada FFT, la expresión (183) se convertirá

en una suma de elementos discretos. En la Tabla 1 se especifican las frecuencias correspondientes a las bandas críticas auditivas. En esta tabla puede comprobarse que para una frecuencia máxima de análisis de 8kHz correspondiente a una frecuencia de muestreo $f_s = 16\text{kHz}$, se necesitan $B = 22$ bandas críticas.

b	1	2	3	4	5	6	7	8	9	10	11	12	13
f_{ib} [kHz]	0.00	0.10	0.20	0.30	0.40	0.51	0.63	0.77	0.92	1.08	1.27	1.48	1.72
f_{fb} [kHz]	0.10	0.20	0.30	0.40	0.51	0.63	0.77	0.92	1.08	1.27	1.48	1.72	2.00

b	14	15	16	17	18	19	20	21	22	23	24	25
f_{ib} [kHz]	2.00	2.32	2.70	3.15	3.70	4.40	5.30	6.40	7.70	9.50	12.0	15.5
f_{fb} [kHz]	2.32	2.70	3.15	3.70	4.40	5.30	6.40	7.70	9.50	12.0	15.5	20.0

Tabla 1. Frecuencias inferior y superior de las bandas críticas hasta 20kHz.

2.- Despues debe convolucionarse el espectro por bandas críticas $|\hat{X}_0(b)|^2$ con una función de ensanchamiento (*Spreading Function*) SF(b). Con esto se consideran los fenómenos de enmascaramiento, es decir cómo la percepción subjetiva de nivel de señal captado en una determinada banda de frecuencia se ve influida por las bandas críticas adyacentes [Zwicker 99]. De esa forma se obtiene el espectro C(b) (que tiene dimensiones de potencia):

$$C(b) = |\hat{X}_0(b)|^2 * SF(b) \quad (184)$$

En la Tabla 2 y la Figura 36 se tienen los valores más significativos de la función de ensanchamiento, obtenidos de [Virág 99].

b'	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
SF [dB]	-74.6	-51.3	-28.0	-8.0	0.0	-4.0	-12.0	-21.0	-30.9	-40.8	-50.7	-60.6	-70.5

Tabla 2. Valores más significativos de la función de ensanchamiento SF(b'). El parámetro b' aquí es un número de banda crítica relativo.

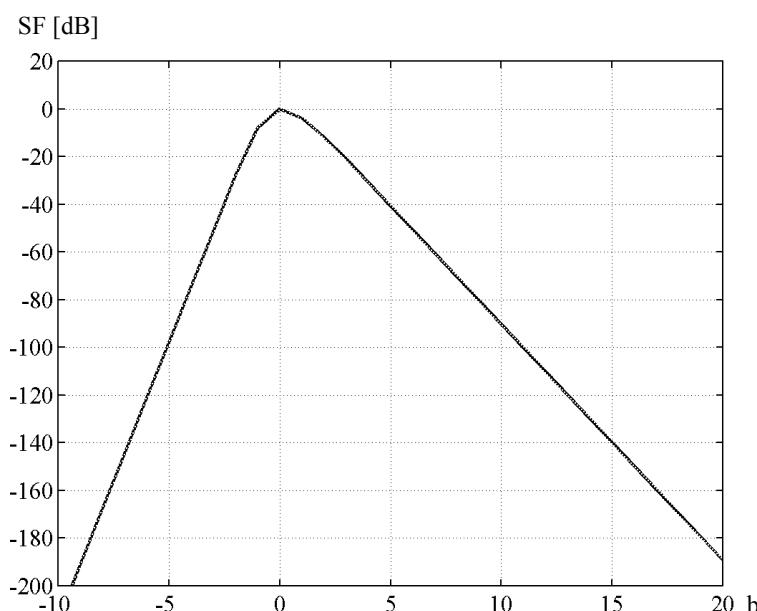


Figura 36. Función de ensanchamiento SF(b) en versus el índice de banda crítica relativo b' .

El objetivo de la convolución (184) es modificar el espectro para que en cada banda crítica influyan adecuadamente las adyacentes. Sin embargo no se debe incrementar el nivel global de $C(b)$ en cada banda crítica con respecto a $|\hat{X}_0(b)|^2$. Por eso hay que atenuar $C(b)$ según un coeficiente de renormalización $D(b)$, que se obtiene hallando el incremento de nivel que proporciona en cada banda crítica la función de ensanchamiento sobre una entrada de referencia, con potencia unidad en todas las bandas críticas:

$$D(b) = 1 * SF(b) \quad (185)$$

Según la función de ensanchamiento de la Tabla 2 y la Figura 36 se obtiene un factor de renormalización que se representa en la Tabla 3 para las 22 primeras bandas críticas.

b	1	2	3	4	5	6	7	8	9	10	11
D(b)	1.16	1.56	1.62	1.63	1.63	1.63	1.63	1.63	1.63	1.63	1.63
b	12	13	14	15	16	17	18	19	20	21	22
D(b)	1.63	1.63	1.63	1.63	1.63	1.63	1.63	1.63	1.63	1.63	1.47

Tabla 3. Factor de renormalización $D(b)$ para las 22 primeras bandas críticas, hasta 8kHz. Se ha considerado la función de ensanchamiento de la Tabla 2 y la Figura 36.

3.- A continuación se calcula el umbral de enmascaramiento mediante la siguiente expresión que incluye la renormalización a través de $D(b)$:

$$T(b) = \frac{C(b)}{D(b)} 10^{\frac{O(b)[\text{dB}]}{10}} \quad (186)$$

donde $O(b)$ [dB] es un parámetro de corrección (*offset*), con unidades en decibelios y que representa una atenuación sobre $C(b)$ para tener en cuenta las características espectrales de la señal de voz. Así el tratamiento es diferente si la trama de voz en análisis es de tipo sonoro (tonal o *tonelike*) o sorda (*noiselike*). El valor de *offset* $O(b)$ [dB], viene dado por la siguiente expresión:

$$O(b) [\text{dB}] = -\text{ton}(14.5 + b) - (1 - \text{ton}) 5.5 \quad (187)$$

con “ton” la tonalidad de la señal de habla que entra al procesador. El parámetro “ton” mide si el espectro de la señal es muy armónico, para tramas de voz sonoras, o es muy plano, para tramas de voz sordas. La tonalidad de un determinado espectro se calcula según:

$$\text{ton} = \min \left\{ \frac{\text{SFM} [\text{dB}]}{-60}, 1 \right\} \quad (188)$$

donde SFM [dB] (*Spectral Flatness Measure*) es una medida de la característica tonal del espectro de potencia de la señal de referencia $|\hat{X}_0(\omega)|^2$. El valor de -60dB se asigna a la SFM [dB] máxima permitida, correspondiente a un tono puro. El parámetro SFM [dB] se calcula mediante la comparación del promedio aritmético con el promedio geométrico del espectro de potencia $|\hat{X}_0(\omega)|^2$:

$$\text{SFM} [\text{dB}] = 10 \log \frac{\left\langle |\hat{X}_0(\omega)|^2 \right\rangle_{\text{pa}}}{\left\langle |\hat{X}_0(\omega)|^2 \right\rangle_{\text{pg}}} \quad (189)$$

donde el subíndice “pa” se refiere a promedio aritmético y el subíndice “pg” a promedio geométrico. De esa forma la tonalidad varía entre $\text{ton} = 1$ para $\text{SFM} [\text{dB}] \geq -60\text{dB}$ (correspondiente a un tono puro) y $\text{ton} = 0$ para $\text{SFM} [\text{dB}] = 0\text{dB}$ (correspondiente a un ruido blanco). En la práctica suele escogerse un valor de “ton” que represente en promedio a la señal que se está procesando, aquí la señal de habla. Por ejemplo $\text{ton} = 0.8$ puede considerarse un buen valor para la señal se habla promedio. Otros autores [Sinha 93] prefieren ser más específicos ya que el habla es muy tonal en baja frecuencia y muy poco tonal en alta frecuencia. En la Figura 37 se representa la función de offset $O(b)$ [dB] para varios valores de tonalidad y para el habla promedio, según [Sinha 93]. En la Figura 38 se representa un ejemplo de cálculo de los umbrales de enmascaramiento $T(\omega)$ según tres formas de calcular el offset $O(b)$ [dB]: considerando el parámetro “ton” medido aplicando (187), el parámetro “ton” fijo ($\text{ton} = 0.8$), y el correspondiente al habla promedio según [Sinha 93]. En dicha figura se han considerado dos casos de voz limpia: una trama de voz sonora y una trama de voz sorda.

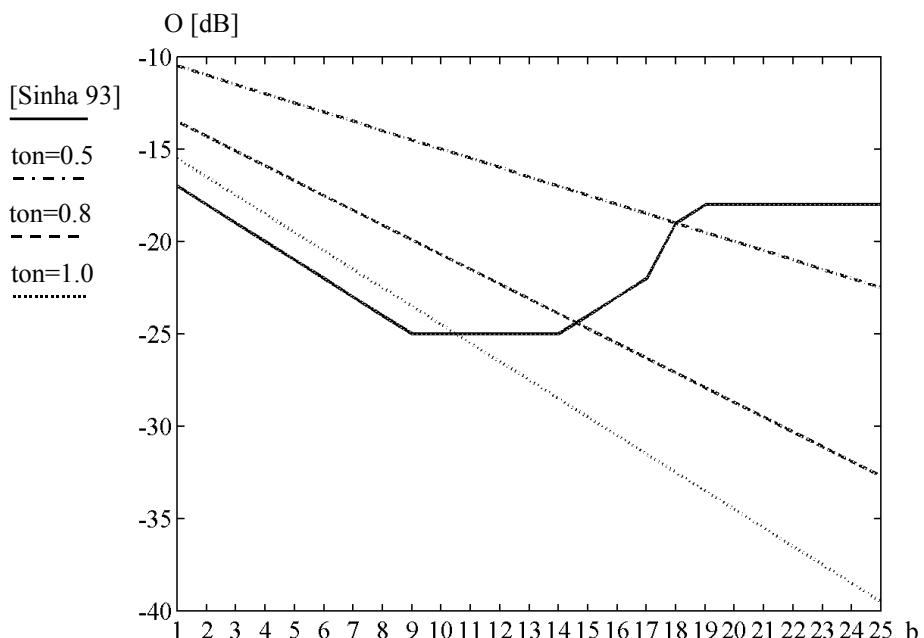


Figura 37. Offset $O(b)$ [dB] para varios valores de tonalidad –“ton” de (187)– y para habla promedio según [Sinha 93].

Hay que decir finalmente, que los umbrales $T(b)$ calculados mediante el proceso anterior serían totalmente válidos si la señal de partida para su cálculo, es decir la señal de referencia $X_0(\omega)$ fuese una señal de presión acústica, ya que en rigor los umbrales $T(b)$ son auditivos y por tanto se deben calcular sobre señales de presión acústica. En lo que concierne al trabajo que se desarrolla en esta Tesis eso no es importante, ya que aunque $X_0(\omega)$ no sea una señal acústica, la estimación de umbrales sólo se hará para compararlos con el ruido $N(\omega)$, y lo que importará aquí será la ganancia relativa del ruido frente al umbral, es decir la relación $NMR(\omega)$.

Por otra parte, también hay que considerar aquí, cómo se hace el paso de $T(b)$ calculado por banda crítica a $T(\omega)$ para cada frecuencia individual, ya que ésta es la forma de T necesaria para ser introducida en las expresiones de supresión audible (181) del método ANS. Este paso se hace de la siguiente forma:

$$T(\omega) = T(b) \text{ para } \omega \in [\omega_{ib}, \omega_{fb}] \text{ y } b = 1..B \quad (190)$$

con ω_{ib} y ω_{fb} las pulsaciones inicial y final de la banda crítica b. La expresión (190) simplemente considera constante el umbral $T(\omega)$ dentro de cada banda crítica, y le asigna el valor $T(b)$.

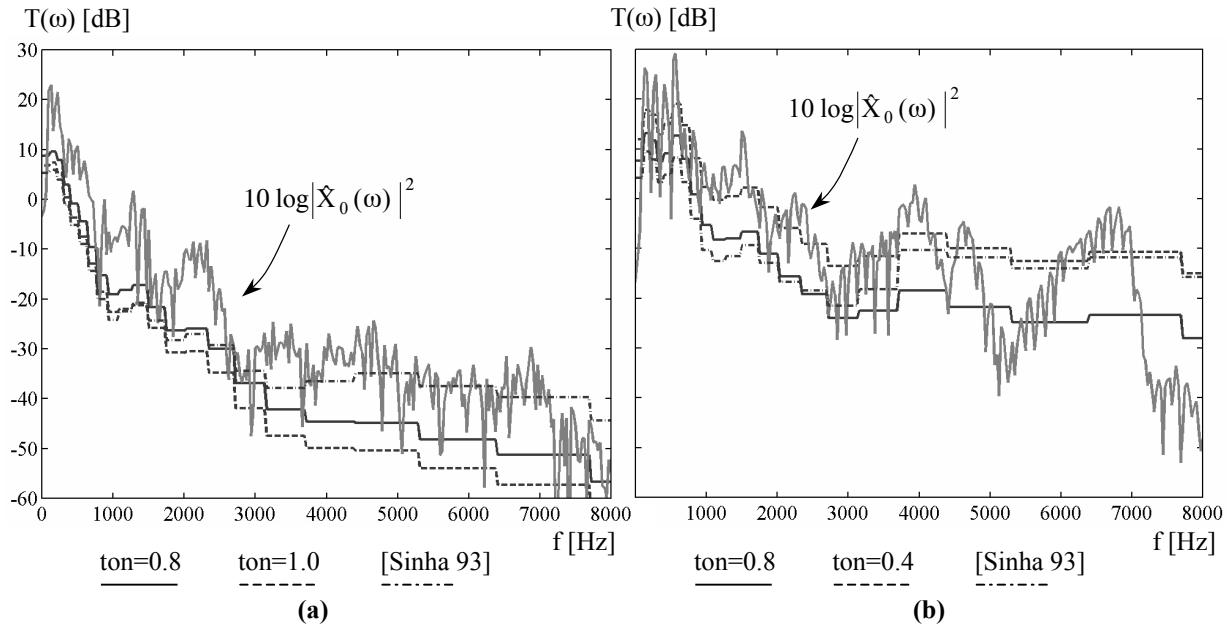


Figura 38. Cálculo de los umbrales de enmascaramiento $T(\omega)$ aplicado sobre dos tramas diferentes de voz limpia. **(a)** Para una trama de voz sonora $-\text{ton}=1.0$, calculado a partir de SFM [dB]-. **(b)** Para una trama de voz sorda $-\text{ton}=0.4$, calculado a partir de SFM [dB]-. En ambos casos se comparan con los umbrales calculados con $\text{ton}=0.8$ y según el *offset* de [Sinha 93].

En la Figura 39 se representa esquemáticamente la implementación del filtrado ANS, en la que se incluyen todos los pasos necesarios desarrollados hasta aquí.

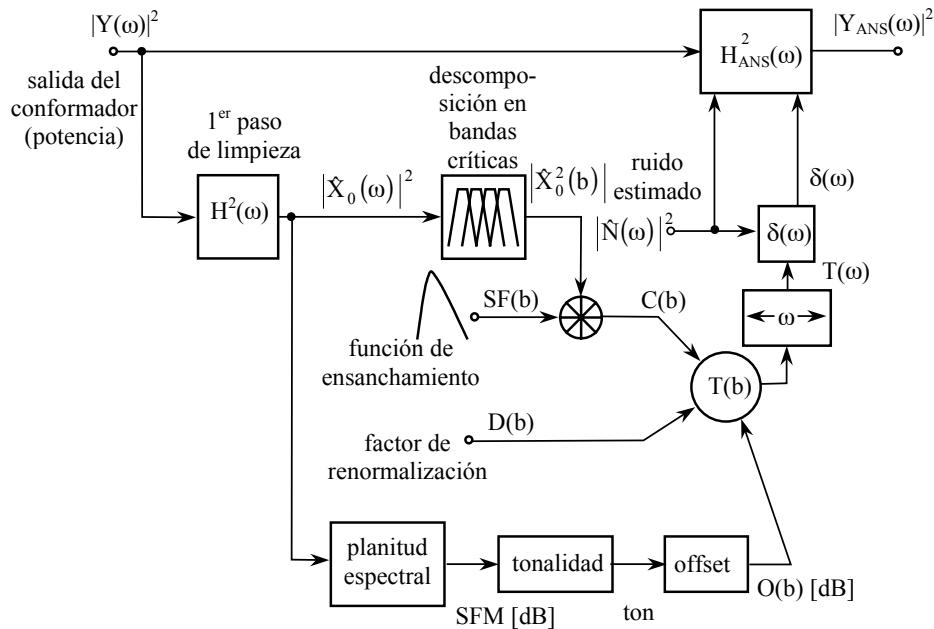


Figura 39. Filtrado perceptual. Método ANS.

Para finalizar, una consideración al margen es cuál es la señal de entrada que se utiliza para el cálculo de los umbrales de enmascaramiento. La obtención de una buena estimación de la señal limpia $|\hat{X}_0(\omega)|^2$ es fundamental ya que los umbrales de enmascaramiento $T(b)$ se refieren a voz limpia. Por lo tanto un paso previo al filtrado ANS debe incluir una primera etapa de limpieza. Existen varios métodos para realizar esta primera limpieza. Se pueden calcular los umbrales a partir de la voz sucia de entrada al procesador (que será la salida conformada del array) utilizando el método ANS de forma iterativa [Tsoukalas 97] hasta obtener la voz finalmente limpia. Para afrontar esta primera etapa de limpieza, lógicamente son también válidos los métodos más tradicionales ya comentados como el filtrado de Wiener o la sustracción espectral [Sánchez-Bote 01-a] [Sánchez Bote 02-b] [Sánchez-Bote 03-a] [Sánchez-Bote 03-b]. En cualquier caso, la bondad del método utilizado depende mucho de la cantidad de ruido y/o reverberación de la señal de voz a mejorar, teniéndose que adoptar soluciones diferentes según la calidad de partida. Una vez obtenida una primera estimación de la señal de voz limpia se puede proceder con el filtrado final según $H_{ANS}(\omega)$ de (180).

4.1.4 Técnicas de estimación de potencia de ruido y de señal de habla

Los esquemas de postfiltrado básicos propuestos hasta el momento, por ejemplo el filtrado de Wiener (161), la sustracción espectral (172) o el filtrado perceptual (180), utilizan estimaciones de la potencia de ruido $|\hat{N}(\omega)|^2$ o de la potencia de la señal limpia $|\hat{X}_0(\omega)|^2$, que deben haberse obtenido previamente. Suponiendo que el postfiltrado es en el dominio de la frecuencia mediante la STFT, al procesador llegarán los espectros de tramas de voz de corta duración (típicamente decenas de milisegundos). Es evidente que las estimaciones anteriores no deben hacerse a partir del espectro instantáneo, puesto que las variaciones momentáneas de la señal $Y(\omega)$ de entrada al procesador inducirán distorsión y ruido musical. Para evitar esto, en todas las estimaciones hay que usar suavizados del valor instantáneo estimado, que tengan en cuenta los valores en tramas anteriores de voz.

Una tarea adicional muy importante a la hora de realizar las estimaciones de señal y de ruido es la detección de actividad de habla (VAD), es decir, la estimación de forma automática de los períodos de tiempo en los que no está presente la señal de habla y por tanto se puede pasar a estimar el ruido y viceversa. Existen múltiples esquemas para realizar esta tarea, desde los más sencillos fundamentados en la detección de cruces por cero o en estimaciones de la energía de la señal de habla [Junqua 91], pasando por técnicas basadas en el reconocimiento de habla, o que utilizan otros fundamentos diversos como medidas de distancias en parámetros LPC [Rabiner 77], medidas de periodicidad en la señal de voz [Tucker 92] o de carácter estadístico [Martin 00], o mediante la evaluación de la coherencia intercanal entre varios micrófonos captadores [Guérin 00]. Si se considera que el proceso VAD está superado una vez la señal sucia $Y(\omega)$ ha llegado al filtro de mejora, cuando entra una trama de voz k al procesador, se debe conocer de antemano si ésta es de voz o de ruido. Más adelante se volverá sobre el asunto de la detección VAD.

La forma más sencilla de hacer el suavizado en las estimaciones de señal o de ruido es la llamada media móvil (MA, *Moving Average*) [Boll 79] [McAulay 80]. En el caso de que se esté estimando la potencia de ruido $|\hat{N}(\omega)|^2$ la media móvil se expresa por:

$$|\hat{N}(\omega, k)|^2 = \frac{1}{U} \sum_{u=1}^U |\hat{N}(\omega, k-u)|^2 \quad (191)$$

donde aquí k expresa el número de trama temporal actual correspondiente al ruido (es decir sólo se incluyen las tramas de ruido detectadas por el VAD) y U es el número de tramas anteriores que se consideran para la calcular la media móvil. Aunque en (191) el ruido es siempre una estimación se ha llamado $|\hat{N}(\omega, k)|^2$ a la estimación tras la media móvil y a $|N(\omega, k)|^2$ al ruido estimado a la salida del conformador, sin suavizado. Cuanto mayor sea el valor de U mayor será el suavizado del espectro estimado. También se pueden utilizar estructuras del tipo anterior para estimar la señal limpia con solo sustituir en (191) a $|\hat{N}(\omega, k)|^2$ por $|\hat{X}_0(\omega, k)|^2$, aunque en este caso se deben adoptar valores de U más pequeños que para el ruido. La razón es que el ruido suele considerarse estacionario y la voz no. Por lo tanto a la hora de estimarlo es válido hacerlo con tiempo de relajación alto, mientras que esto no es válido para la voz, cuyas características espectrales varían muy rápidamente, aunque algún suavizado sí es conveniente. En la Figura 40 se representa una figura de estimación de la potencia de ruido $|\hat{N}(\omega, k)|^2$ para dos valores de U , cuando $|N(\omega, k)|^2$ tiene forma de impulso cuadrado con respecto al índice de trama k . Cuando U es grande, la variación del ruido estimado es más lenta con respecto a la variación real del ruido.

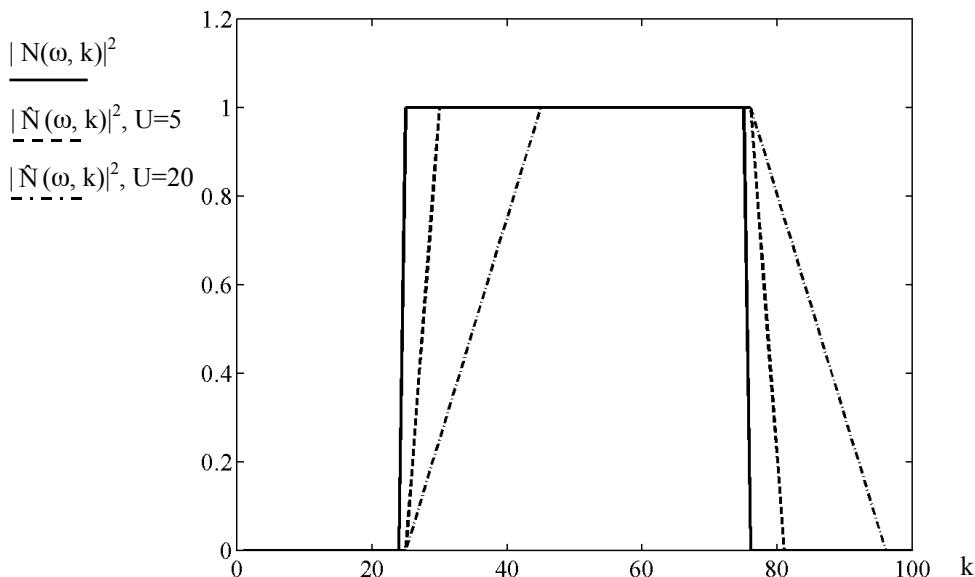


Figura 40. Estimación de la potencia de ruido mediante la media móvil. Se representa la potencia de ruido estimado (normalizada a la unidad) para dos valores de U , comparada con la potencia de ruido instantáneo por trama.

La media móvil requiere promediar en cada instante temporal los valores correspondientes a U tramas, sin ninguna ponderación. Un método alternativo de estimación es el conocido como recursión de polo único [McAulay 80] [Ephraim 84] que consiste en ponderar las diferentes tramas según la proximidad temporal con la actual. Se expresa de forma general con:

$$|\hat{N}(\omega, k)|^2 = \lambda |\hat{N}(\omega, k-1)|^2 + (1-\lambda) |N(\omega, k)|^2 \quad (192)$$

con λ un parámetro que gobierna la velocidad de actualización ($0 < \lambda < 1$). Se considera que la actualización es muy rápida si λ es pequeña, y viceversa. La expresión (192) representa un filtro paso bajo en el dominio de la variable de trama k . Cuanto mayor sea el valor de λ más lentamente se actualizará la estimación del ruido y más baja será la frecuencia de corte del filtro equivalente. En la Figura 41 se representa la respuesta temporal de un estimador de polo único para diferentes valores de la constante λ .

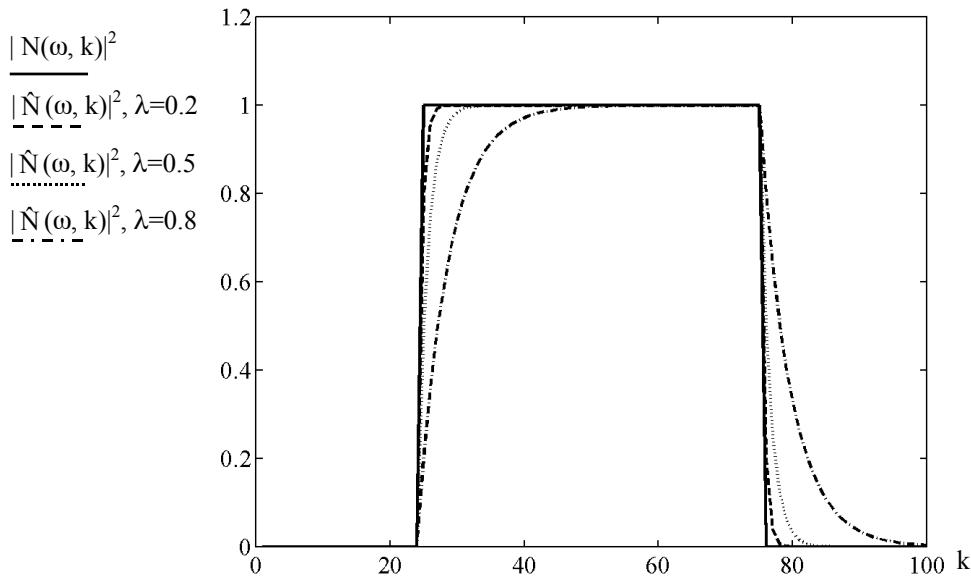


Figura 41. Estimación de la potencia de ruido (normalizada a la unidad) mediante recursión de polo único. Se representa la potencia de ruido estimado para tres valores de λ , comparada con la potencia de ruido instantáneo por trama.

Hablando en términos de energía, la respuesta al impulso de (192) es:

$$|\hat{N}(\omega, k)|^2 = (1 - \lambda)\lambda^k \quad (193)$$

que corresponde a la respuesta temporal de un filtro paso bajo. Es decir, si el estimador de ruido recibe a su entrada un impulso de potencia de ruido $|N(\omega, k)|^2 = \delta(\omega, k)$ (con δ la delta de Dirac), la salida del estimador será $|\hat{N}(\omega, k)|^2$ de (193). Normalmente la constante λ se relaciona con el tiempo de caída t_Δ del filtro paso bajo equivalente, como se muestra a continuación. Ante un impulso $\delta(\omega, k)$ en $k = 0$, cuando hayan transcurrido k tramas de la estimación, el incremento de nivel [dB] en dicha estimación de ruido será:

$$\Delta|\hat{N}(\omega, k)|^2 [\text{dB}] = 10 \log \frac{|\hat{N}(\omega, k)|^2}{|\hat{N}(\omega, 0)|^2} = k 10 \log \lambda \quad (194)$$

Ahora bien, el índice de trama se puede expresar también como:

$$k = \frac{t}{\Delta t} \quad (195)$$

donde t es la variable tiempo y Δt es el intervalo de actualización de la trama, que no tiene por qué coincidir con la longitud temporal de una trama, ya que en el análisis mediante STFT se suele considerar cierta cantidad de solapamiento entre tramas. Entonces se puede sustituir (195) en (194) para despejar el λ necesario para que en el tiempo $t = t_\Delta$ se produzca una caída de $\Delta|\hat{N}(\omega, k)|^2$ [dB] resultando:

$$\lambda = 10^{\frac{\Delta t}{10 t_\Delta} \Delta|\hat{N}(\omega, k)|^2 [\text{dB}]} \quad (196)$$

siendo t_Δ es el tiempo de caída necesario para la reducción de $\Delta|\hat{N}(\omega, k)|^2$ [dB] en la estimación del ruido. Normalmente se considera como valor representativo de atenuación en

el tiempo t_Δ , el expresado por la igualdad: $\Delta|\hat{N}(\omega, k)|^2 [\text{dB}] = 10 \log e^{-1} [\text{dB}] = -4.3 \text{dB}$ (equivalente a 1 neperio). Sustituyendo este valor en (196) se obtendría:

$$\lambda = e^{-\frac{\Delta t}{t_\Delta}} \quad (197)$$

Por ejemplo, se considera una velocidad rápida de actualización de la estimación cuando t_Δ es del orden de algunos milisegundos, que puede aceptarse como un tiempo de caída apropiado para estimaciones de señal, cuyas características espectrales varían de forma rápida. Así para un periodo de actualización de trama de $\Delta t = 10 \text{ms}$ si $t_\Delta = 10 \text{ms}$ entonces $\lambda = 0.37$. Por contra, un periodo muy lento de actualización puede ser de algunos segundos. Así, con el mismo valor de $\Delta t = 10 \text{ms}$ si se considera un tiempo de caída $t_\Delta = 1 \text{s}$ entonces $\lambda = 0.99$, valor que puede ser válido para la estimación del ruido, que suele calificarse como casi estacionario.

Otra posibilidad muy usada para las estimaciones de señal y de ruido, utilizando la misma filosofía, consiste en el método conocido como recursión de polo único y dos lados [Etter 94]. Esta forma de estimación responde a la misma expresión (192) pero ahora con un doble valor para la constante λ , es decir:

$$\lambda = \begin{cases} \lambda_a & \text{si } |\hat{N}(\omega, k)|^2 \geq |\hat{N}(\omega, k-1)|^2 \\ \lambda_c & \text{si } |\hat{N}(\omega, k)|^2 < |\hat{N}(\omega, k-1)|^2 \end{cases} \quad (198)$$

donde λ_a y λ_c son respectivamente las constantes de ataque y caída del estimador de potencia de ruido para este caso. Es decir, si el nivel de potencia estimada está creciendo, se considera una constante de ataque λ_a diferente de la constante de caída λ_c , aplicada cuando el nivel decrece. Si el estimador de doble lado representado por (192) y (198) se usa para estimar ruido $\hat{N}(\omega, k)$, se considerará normalmente que $\lambda_a > \lambda_c$, de tal manera que se sea muy lento al ataque. Es decir, cuando la estimación de ruido esté contaminada por algún proceso transitorio, como captación de habla o algún ruido espurio, este estimador no se actualizará. Por contra cuando el estimador de doble lado se use para estimar habla representada por $\hat{X}_0(\omega, k)$, se considera normalmente $\lambda_a < \lambda_c$, puesto que se necesita atender a los ataques rápidos de la señal de voz.

Existen otros esquemas de estimación [Etter 94] frecuentemente derivados de la recursión de polo único vista en (192) con diferentes consideraciones sobre la actualización de la estimación de ruido y señal y las constantes de ataque y caída definidas anteriormente.

Detección de actividad de habla basada en estimación recursiva

Los esquemas de recursión vistos en este apartado pueden utilizarse, además de como estimadores de señal y de ruido, también como detectores VAD de actividad de habla. Esta posibilidad se explica a continuación.

Considérese una recursión de polo único y dos lados a partir de la siguiente expresión:

$$\begin{aligned} |\hat{X}_0(\omega, k)|^2 &= \lambda_S |\hat{X}_0(\omega, k-1)|^2 + (1-\lambda_S) |Y(\omega, k)|^2 \\ |\hat{N}(\omega, k)|^2 &= \lambda_N |\hat{N}(\omega, k-1)|^2 + (1-\lambda_N) |Y(\omega, k)|^2 \end{aligned} \quad (199)$$

con λ_S una constante de actualización para la señal y λ_N una constante de actualización para el ruido. Esta expresión constituye en sí misma un estimador recursivo de señal y de ruido más un detector de actividad de habla simultáneo, ya que a partir de la señal de habla sucia $|Y(\omega, k)|^2$ es capaz de estimar de forma independiente la señal y el ruido. La distinción de si la salida del estimador es señal $|\hat{X}_0(\omega, k)|^2$ o ruido $|\hat{N}(\omega, k)|^2$ se hace por la elección de los valores de las constantes de ataque y caída, diferentes para la señal λ_{Sa} y λ_{Sc} y para el ruido λ_{Na} y λ_{Nc} . Si se toma λ_{Na} muy lento y λ_{Nc} , λ_{Sa} y λ_{Sc} muy rápidos se tendrá un VAD basado en los cambios de potencia de la señal entrante $|Y(\omega, k)|^2$. Es decir si la potencia aumenta, es signo de que llega una trama de voz, la señal se actualiza (λ_{Sa} pequeño) pero el ruido no (λ_{Na} grande). Por contra, cuando la potencia decrece, tanto el ruido como la señal se actualizan rápidamente (λ_{Sc} y λ_{Nc} pequeños). En la Figura 42 se representa un estimador-detector VAD de este tipo.

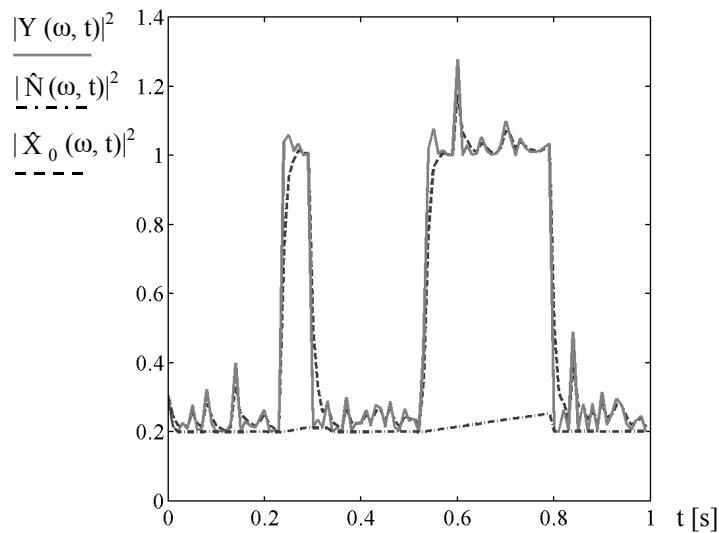


Figura 42. Detección de actividad de voz y estimador simultáneo de las potencias de ruido y señal mediante recursión de polo único y doble lado, para una frecuencia ω cualquiera. Se ha considerado un tiempo de trama de $\Delta t=10\text{ms}$ y unas constantes de tiempo características $t_{\Delta Sa}=10\text{ms}$, $t_{\Delta Sc}=10\text{ms}$, $t_{\Delta Na}=4\text{s}$ y $t_{\Delta Nc}=1\text{ms}$. Valores de potencia normalizados a la unidad.

Se ha considerado un periodo de trama de $\Delta t = 10\text{ms}$ con tiempos característicos $t_{\Delta Sa} = 10\text{ms}$, $t_{\Delta Sc} = 10\text{ms}$, $t_{\Delta Na} = 4\text{s}$ y $t_{\Delta Nc} = 1\text{ms}$. Los valores de las constantes λ se obtienen mediante (197). A la vista de la Figura 42, se verifica que la estimación de ruido no sigue los incrementos rápidos en la potencia de la señal entrante $|Y(\omega, k)|^2$, pero sí se actualiza rápidamente ante descensos bruscos de señal. La estimación de la voz se actualiza rápidamente en ambos casos, tanto si se produce un incremento o un decremento brusco de la potencia entrante.

La gran ventaja de un detector VAD diseñado de esta forma es su eficiencia y sencillez desde el punto de vista de su implementación práctica, ya que con pocos requerimientos de cálculo consigue realizar una buena estimación de señal y ruido, a la vez que realiza la segmentación voz-ruido.

4.1.5 Otros esquemas de reducción de ruido

Como se ha visto, las técnicas más utilizadas en la reducción de ruido se basan en el postfiltrado de la señal conformada una vez que se ha estimado la cantidad de ruido remanente en dicha señal monocanal. Existen otros métodos de postfiltrado con una filosofía de funcionamiento diferente. Entre ellos se pueden destacar los basados en el filtrado de la señal $y(t)$ a la salida del conformador utilizando los métodos de descomposición en subespacios, cuyos principios teóricos ya se han tratado en el capítulo de localización de fuente (apartado 3.3.1) sobre la implementación del método MUSIC para la determinación de las DOA's de las fuentes presentes. Este método, cuando se aplica a la mejora de habla también es conocido como descomposición generalizada en valores singulares GSVD (*Generalized Singular Value Decomposition*). La implementación del método GSVD es relativamente reciente, tanto en aplicaciones monocanal [Ephraim 95] como en arrays de micrófonos [Asano 00] [Jabloun 01]. En principio, se ha demostrado superior a los métodos tradicionales como el filtrado de Wiener multicanal o la sustracción espectral, aunque como contrapartida su mayor complejidad lo hace menos adecuado para una implementación en sistemas en tiempo real. El método GSVD se basa en la descomposición del filtro de Wiener multicanal óptimo dado en (151) en valores singulares (por medio de los autovectores y autovalores) a partir de promedios de observaciones de la matriz de covarianzas de ruido $\mathbf{R}_{nn}(\tau)$ en los instantes de no-actividad de la voz y de la matriz de covarianzas de la señal $\mathbf{R}_{yy}(\tau)$ en los instantes de presencia de voz. La descomposición en valores singulares permite una elección adecuada del filtro óptimo de Wiener para una mejor cancelación del ruido. El método GSVD puede encontrarse descrito en detalle en [Brandstein 01].

Por último existen métodos de mejora de voz que se alejan de las técnicas de postfiltrado óptimo presentadas anteriormente. Entre ellos se encuentran los que utilizan el modelado de habla. Básicamente estos métodos se basan en la extracción de las características de la señal de habla captada por el array de micrófonos y en el cálculo de un filtro adecuado en función del modelo de habla calculado. Por ejemplo, el más sencillo de estos métodos utiliza el modelo de habla de excitación dual (DE, *Dual Excitation*) que implementa una descomposición multicanal [Brandstein 98] de la señal incidente en habla sorda y sonora, aplicando el filtrado apropiado para cada tipo de señal. También existen esquemas más complejos que utilizan la reconstrucción de los parámetros LPC residuales (*Linear Predictive Coding*) [Brandstein 99] según los cuales se procesan los residuos remanentes en la codificación LPC con el objetivo de atenuar el efecto de las perturbaciones acústicas para posteriormente recodificar una versión mejorada de la señal de habla sucia. Este esquema más convencional de reconstrucción LPC se puede hacer también usando un análisis multirresolución (mediante descomposición en ondículas o *wavelets*). Todos estos métodos que utilizan el modelado de habla pertenecen al campo de la mejora de habla no lineal “basada en modelo” (*Nonlinear Model-Based Techniques*) [Gay 00].

4.2 DERREVERBERACIÓN

La reverberación es un fenómeno inherente a la transmisión acústica en espacios cerrados, donde la señal acústica viaja desde el emisor hasta el receptor por un canal multitrayecto, es decir, recorriendo múltiples caminos [Allen 77-b] [Tohyama 95] [Kahrs 98] [Beltrán 99]. Este escenario de transmisión acústica se conoce como sistema acústico multitrayecto. En cada uno de estos caminos el nivel de la señal acústica transmitida normalmente diminuye, bien por divergencia esférica bien por absorción en su reflexión con las paredes del recinto. De igual manera, cada trayecto cambia la fase original de la señal

acústica, bien por retardo acústico, bien por efecto de la impedancia acústica de las paredes en las que se producen las reflexiones. En el apartado 2.1.3 de esta Tesis se expusieron los aspectos fundamentales de la señal reverberante y de su captación por un array microfónico (véase la Figura 3). Aquí se incidirá en las características particulares de la reverberación que puedan servir para atenuarla de tal manera que se mejore la calidad del habla a la salida de un procesador.

La reverberación es uno de los fenómenos más difíciles de tratar. Quizás, si ésta se mantiene en unos niveles no excesivos puede sonar de forma natural y ser mucho más tolerable que el ruido aditivo. Sin embargo, cuando la reverberación sube tanto que impide la correcta inteligibilidad de la señal de habla, se hace muy molesta y es muy difícil de eliminar puesto que su contenido espectral es casi coincidente con la señal de habla directa y es simultánea a la misma, de tal manera que no se puede reducir mediante un simple filtrado.

Supóngase que se desea calcular la salida eléctrica $x_i(t)$ de uno cualquiera de micrófonos del array cuando está recibiendo una onda de presión directa $p_1(t)$ como en (19) medida en el centro de coordenadas. Como existe reverberación, además de la presión $p_1(t)$ debida a la onda directa, el array estará sometido a las $Q-1$ reflexiones $p_2(t), \dots, p_q(t), \dots, p_Q(t)$. La expresión (5) proporciona la señal de referencia a la salida del array debida a cada una de las Q presiones incidentes, pudiéndose hablar de $x_{01}(t), \dots, x_{0q}(t), \dots, x_{0Q}(t)$. Teniendo en cuenta (21), la salida de referencia del camino q viene dada por:

$$x_{0q}(t) = \sum_0 p_q(t) = \sum_0 p_1(t) * \delta(t - \tau_q) * \frac{r_1}{r_q} \beta_q(t) \quad q = 2, \dots, Q \quad (200)$$

donde, para mayor generalidad, se ha considerado que la señal directa $p_1(t)$ es de banda ancha y por tanto los productos de (21) se han transformado en convoluciones temporales. En (200) $\beta_q(t)$ es la respuesta al impulso del fenómeno de reflexión con las paredes del recinto, cuando la onda acústica recorre el camino q . Incluye respectivamente la atenuación por absorción y el desfase, β_q y ϕ_q de (21). El cociente r_1/r_q representa la atenuación por distancia del camino q respecto al camino directo, representado por r_1 y $\delta(t - \tau_q)$ simboliza el retardo del camino acústico q , también respecto al directo. La salida eléctrica del micrófono i será la suma de las respuestas de los Q caminos implicados en el sistema multirayecto, como ya expresaba (24):

$$y_i(t) = \sum_{q=1}^Q x_{0q}(t) * a_{iq}(t) = \sum_{q=1}^Q \sum_0 p_1(t) * \delta(t - \tau_q) * \frac{r_1}{r_q} \beta_q(t) * a_{iq}(t) = \sum_{q=1}^Q p_1(t) * h_{iq}(t) \quad (201)$$

donde recuérdese que $a_{iq}(t)$ era la respuesta al impulso del micrófono i excitado por una señal de referencia procedente del camino q . Se sobreentiende que para el camino directo (fuente origen) $\beta_1(t) = \delta(t)$ y $\tau_1 = 0$. Como en (25) y (26) la salida del array se expresa por $y(t)$ –no $x(t)$ –, para denotar que contiene la perturbación debida a la reverberación. En (201) $h_{iq}(t)$ es la respuesta al impulso del micrófono i ante el camino q y tiene dimensiones de $V \cdot Pa^{-1}$. Analíticamente:

$$h_{iq}(t) = \sum_0 * \delta(t - \tau_q) * \frac{r_1}{r_q} \beta_q(t) * a_{iq}(t) = b_{iq}(t) * \delta(t - \tau_q) \quad (202)$$

La expresión (201) representa por tanto la respuesta eléctrica del sistema multirayecto ante una excitación de presión acústica. En (202) $b_{iq}(t)$ simboliza la respuesta al impulso del micrófono i ante el camino de transmisión q excluyendo los efectos del retardo acústico y τ_q el retardo relativo de dicho camino acústico q con respecto a la fuente origen, para $q = 1$. La

respuesta al impulso $b_{iq}(t)$ incluye los efectos de atenuación y desfase debidos tanto al camino acústico q como a la reflexión con las paredes y los efectos de la respuesta acústico eléctrica del micrófono del array que está captando la fuente imagen q . Nótese que, a diferencia del vector de apuntamiento $a_{iq}(t)$ de (24) –versión temporal del expresado en (9)–, el vector $b_{iq}(t)$ no sólo tiene en cuenta la respuesta electroacústica de cada micrófono con respecto a la referencia x_{0q} , en el centro de coordenadas sino que también considera los posibles efectos de atenuación y desfase del sistema multirayecto. Esta diferencia surge aquí porque anteriormente los efectos del sistema multirayecto estaban incluidos en las señales de referencia $x_{0q}(t)$ de (22) debidas a todos los caminos que originan la reverberación, y ahora se ha preferido aislar en (202) el retardo acústico de todos los demás efectos electroacústicos del fenómeno de captación en array de una señal reverberante. La razón de esta nueva forma de expresar la respuesta del array es que, a la hora de estudiar la reverberación para tratar de reducirla o filtrarla, el fenómeno que más influye es el de la función de transferencia multirayecto, generada por la suma de Q funciones temporales con diferentes retardos.

En el dominio de la frecuencia, la respuesta al impulso $h_{iq}(t)$ expresada en (202) quedaría:

$$H_{iq}(\omega) = B_{iq}(\omega) \exp(-j\omega\tau_q) \approx |H_{iq}(\omega)| \exp(-j\omega\tau_q) \quad (203)$$

siendo $B_{iq}(\omega)$ la función de transferencia en frecuencia del micrófono i respecto al camino q excluyendo los efectos del retardo τ_q . En (203) se ha asumido, simplificando al máximo, que toda la fase de cada función de transferencia correspondiente a un camino acústico, está contenida en el retardo τ_q . En la Figura 43 se hace una representación vectorial de la función de transferencia de cada uno de los caminos que componen un sistema multirayecto. La función de transferencia total del sistema multirayecto entre el emisor y el micrófono será la suma de las funciones de transferencia de cada uno de los Q caminos integrantes como se vio en (201). Como se aprecia en la Figura 43, cada uno de los caminos queda representado por un vector que, al aumentar la frecuencia, gira en sentido antihorario a diferentes velocidades angulares, dependiendo del retardo τ_q . El resultado de la suma será un vector o fasor con módulo y fase muy rápidamente variable con la posición relativa entre emisor y receptor (que determina los retardos) y con la frecuencia.

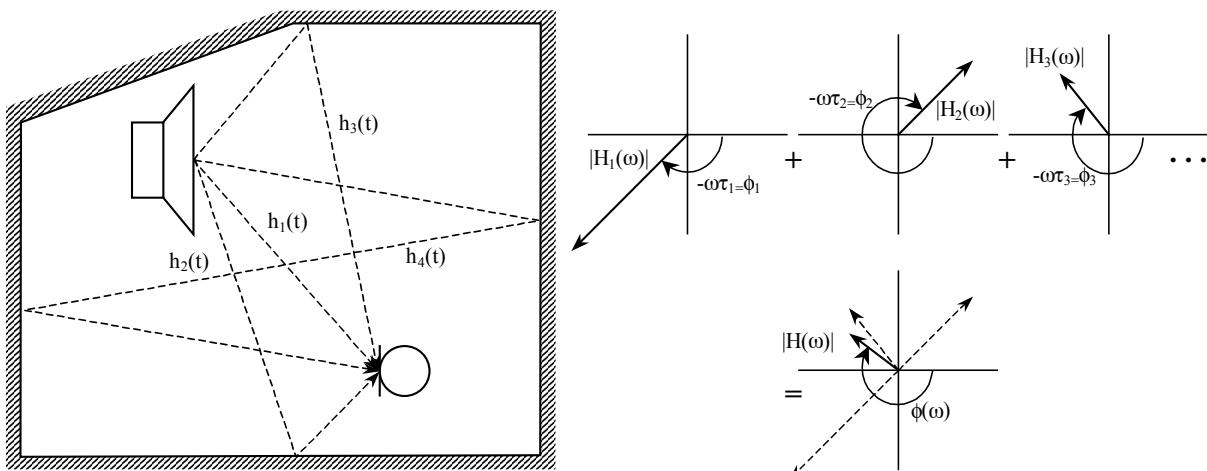


Figura 43. Sistema multirayecto. La función de transferencia electroacústica de cada camino viene representada por $|H_q(\omega)| \exp(-j\omega\tau_q)$. La suma de las Q funciones de transferencia individuales origina la función de transferencia global multirayecto, $|H(\omega)| \exp[j\phi(\omega)]$.

En el micrófono i, la respuesta al impulso del sistema multirayecto será:

$$h_i(t) = \sum_{q=1}^Q h_{iq}(t) \quad (204)$$

y la función de transferencia en frecuencia,

$$H_i(\omega) = \sum_{q=1}^Q H_{iq}(\omega) = \sum_{q=1}^Q B_{iq}(\omega) \exp(-j\omega\tau_q) = |H_i(\omega)| \exp[j\phi_i(\omega)] \quad (205)$$

donde $|H_i(\omega)|$ representa el módulo de la función de transferencia del sistema multirayecto para el micrófono i, incluyendo todos los efectos: respuesta en frecuencia de elementos electroacústicos, absorciones en las paredes, y por supuesto los retardos acústicos. De igual forma $\phi(\omega)$ es la fase total del sistema multirayecto, incluyendo los efectos anteriormente mencionados.

Por lo tanto, en el dominio de la frecuencia, la respuesta eléctrica $Y_i(\omega)$ del micrófono i ante una excitación de presión directa $p_1(\omega)$ de tal manera que incluya el efecto de la reverberación quedaría expresado por el siguiente producto de funciones en frecuencia:

$$Y_i(\omega) = p_1(\omega)H_i(\omega) \quad (206)$$

En lo sucesivo se evitará representar el subíndice i que identifica al micrófono del array para expresar la función de transferencia electroacústica del sistema multirayecto, que será $H(\omega)$.

La Figura 44 ilustra la respuesta al impulso típica de un sistema multirayecto acústico. Normalmente la respuesta al impulso $h(t)$ de (204) o la Figura 44 decrece exponencialmente en su energía. El parámetro básico que caracteriza la respuesta al impulso de un recinto se conoce como tiempo de reverberación T_{60} , que es el tiempo que tarda la energía acústica tomada en un punto del recinto, en decaer 60dB cuando la señal acústica incidente cesa bruscamente. El valor típico de T_{60} es del orden de 1s, para la clase de recintos en los que se usarán los arrays microfónicos estudiados en esta Tesis, pequeñas salas de conferencia, aulas, de no más de $1000m^3$ de volumen y con coeficientes medios de reflexión inferiores a $\beta = 0.8$.

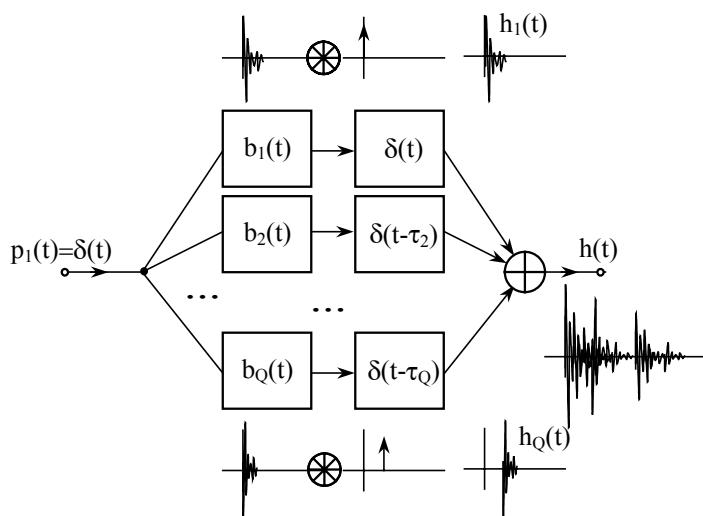


Figura 44. Respuesta al impulso de un multirayecto acústico. La señal $p_1(t)$ representa la presión directa, que está afectada de retardo $\tau_1=0$. Las funciones $h_q(t)$ representan las respuestas al impulso individuales de cada camino, con $h(t)$ la respuesta al impulso global.

En la Figura 45 se muestra una función de transferencia típica de un sistema acústico multirayecto, para una sala de dimensiones normales. Es muy característica la rápida variación con la frecuencia que experimentan tanto el módulo $|H(\omega)|$ como la fase $\phi(\omega)$. La existencia de múltiples caminos hace que al variar ligeramente la frecuencia varíen apreciablemente el módulo y la fase de la función de transferencia (véase la Figura 43 y la explicación en el pie de figura). Se ha estimado estadísticamente [Schroeder 64] que el intervalo promedio en frecuencia Δf , entre un máximo y un mínimo consecutivos de la función de transferencia, o entre dos puntos de cambio de fase 180° vale:

$$\Delta f = \frac{4}{T_{60}} \quad (207)$$

Esta variabilidad en frecuencia de la función de transferencia $H(\omega)$ también se manifiesta cuando el punto de captación (normalmente el micrófono) se mueve ligeramente de una posición a otra.

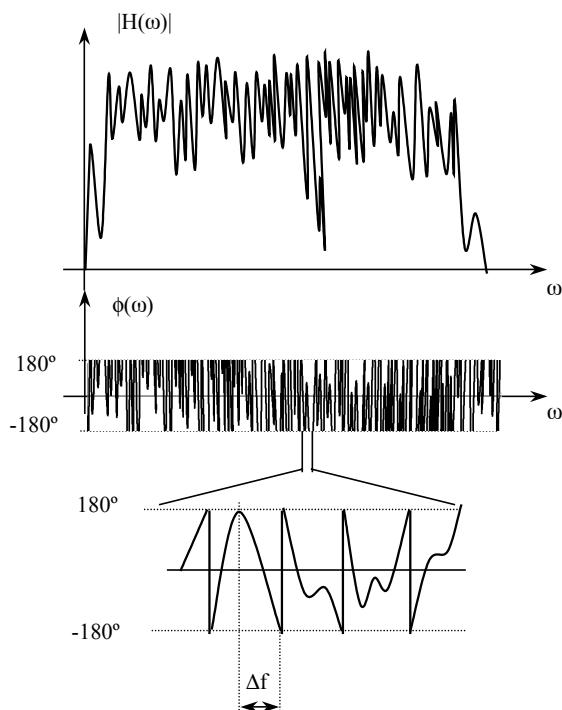


Figura 45. Función de transferencia $H(\omega)$ para un recinto de dimensiones típicas. Se destaca el Δf de Schroeder (207).

Los métodos de derreverberación que existen actualmente están basados en dos filosofías diferentes, a saber: seguimiento con filtrado inverso de la función de transferencia del sistema [Allen 77-a] [Jan 95] [Jan 96] y derreverberación ciega.

El primer método de filtrado inverso, consiste en calcular o estimar en cada punto de la sala donde se sitúen los micrófonos del array, la función de transferencia $H(\omega)$ –ó $h(t)$ –. Una vez invertida mediante $H^{-1}(\omega)$ (con la problemática de invertibilidad que conlleva) se puede aplicar al canal correspondiente del micrófono i –en (206) por ejemplo– para obtener una versión limpia (sin reverberación) de la señal sucia $Y(\omega)$. Aunque este tipo de filtrado inverso en teoría puede ofrecer buenos resultados, en la práctica es de muy difícil aplicación, por la

gran y rápida variabilidad ya comentada de $H(\omega)$ con la frecuencia y la posición del micrófono. Además, otro problema que surge a la hora de pretender la inversión de $H(\omega)$ es que esta función no es en general invertible [Neely 79] [Radlovic 00], debido a que puede tener ceros que se conviertan en polos inestables del filtro inverso, originando una función ni causal ni estable.

La derreverberación ciega consiste simplemente en filtrar o eliminar las componentes de la señal de habla que dan lugar a la sensación de reverberación. Un ejemplo de estas técnicas se muestra en [Cole 97]. La derreverberación ciega es menos efectiva sobre el papel que el filtrado inverso, pero puede ser mejorada de forma importante si es complementada con la derreverberación inherente que produce un sistema de captación en array. Existe un gran número de variantes en los métodos de derreverberación antes apuntados, cuyo detalle se puede encontrar reflejado en la literatura científica producida en las dos últimas décadas. Sin embargo, por la importancia que tiene para esta Tesis, se expone con mayor detalle el método de derreverberación ciega basado en descomposición cepstral de la señal de habla.

4.2.1 Derreverberación mediante descomposición en componentes fase mínima-paso todo

De entre todos los esquemas de derreverberación ciega existente en el ámbito del procesado de habla, merece especial interés para esta Tesis el basado en la separación de la señal de habla en sus componentes fase mínima y paso todo, mediante análisis cepstral [Liu 96]. La razón de ese interés es que con este método se hacen experimentos de derreverberación usando arrays microfónicos, según se desarrolla en la parte de pruebas y resultados de la Tesis. A continuación se explican brevemente los fundamentos de la derreverberación mediante descomposición fase mínima y paso todo.

Sea un sistema de captación mediante micrófonos (una fuente sonora y un micrófono, por ejemplo), dentro de un recinto con reverberación. Si el emisor produce una señal de presión $p(\omega)$, que puede ser la presión p_1 de la fuente origen o directa de (206), a la salida del micrófono se tendrá $Y(\omega)$:

$$Y(\omega) = H(\omega) p(\omega) \quad (208)$$

siendo $H(\omega)$ la función de transferencia de la captación electroacústica, incluyendo la reverberación del recinto y el micrófono del array.

La función de transferencia $H(\omega)$ se puede descomponer en dos componentes llamadas fase mínima $H_{\min}(\omega)$ y paso todo $H_{\text{all}}(\omega)$. En frecuencia:

$$H(\omega) = H_{\min}(\omega) H_{\text{all}}(\omega) \quad (209)$$

con

$$H_{\min}(\omega) = |H(\omega)| \exp[j\phi_{\min}(\omega)] \quad (210)$$

$$H_{\text{all}}(\omega) = \exp[j\phi(\omega) - j\phi_{\min}(\omega)] = \exp[j\phi_{\text{all}}(\omega)] \quad (211)$$

Se dice que una señal es fase mínima si, considerando el dominio transformado z, su transformada Z no contiene polos o ceros fuera de la circunferencia unidad. Una función de transferencia fase mínima es invertible, es decir, su inversa es causal y estable. La función de transferencia de la sala $H(\omega)$ no es en general una función de transferencia fase mínima o lo que es lo mismo no es invertible. Para superar este escollo, lo que se suele hacer es invertir

sólo la componente fase mínima de (210). Si $p(\omega)$ es fase mínima, se puede demostrar la siguiente expresión:

$$p(\omega) = H_{\min}^{-1}(\omega) Y_{\min}(\omega) \quad (212)$$

con lo cual es suficiente con invertir la parte fase mínima de $H(\omega)$ para recuperar $p(\omega)$, libre de reverberación. Esto último es un acercamiento bastante teórico ya que la función de transferencia $H(\omega)$ no se conoce y es muy difícil de calcular y por lo tanto de invertir, aunque como se ha dicho hay toda una rama del procesado acústico que se ocupa de la estimación e inversión de la función de transferencia $H(\omega)$, asunto que no será tratado aquí. Sin embargo se puede relacionar la reverberación de la señal captada con las características de cada una de las componentes de $H(\omega)$, como se muestra a continuación.

La descomposición de la función de transferencia del recinto en componentes fase mínima y paso todo puede aportar un conocimiento extra que ayude a derreverberar una señal de voz. Efectivamente la componente fase mínima se ve menos afectada por la reverberación que la componente paso todo. La componente paso todo contiene la información sobre el sistema multirayecto, es decir sobre la posición relativa fuente-micrófono y las reflexiones con las paredes de la sala. En este sentido, puede suponerse que la información de reverberación está fundamentalmente contenida en la componente paso todo. En la Figura 46 se representa un ejemplo de respuestas al impulso, con las componentes fase mínima y paso todo, en dos casos, para una fuente cercana (1m) y por tanto que representa poca reverberación, y para una fuente más alejada (7m) y por tanto con mayor reverberación. La componente paso todo varía en mayor cuantía cuando crece la reverberación, mientras que la componente fase mínima se ve menos afectada.

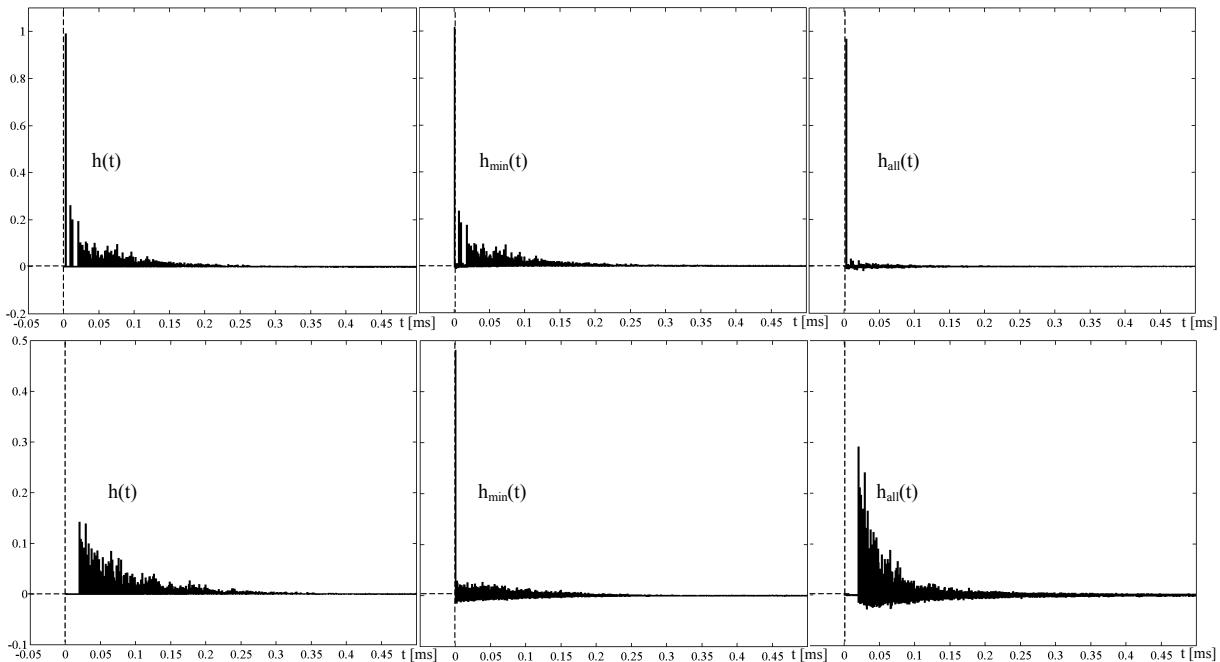


Figura 46. Respuestas al impulso en un punto dentro de un recinto, a partir de una simulación teórica. Las tres superiores corresponden a un emisor situado a $r_0=1\text{m}$ del micrófono. Las tres de abajo corresponden a un emisor situado a $r_0=7\text{m}$. El recinto tiene $T_{60}=1\text{s}$, según Sabine –ver [Kuttruff 91]–.

Se puede obtener la componente fase mínima de la función de transferencia del recinto mediante la transformada cepstrum. Sea una función temporal, por ejemplo la respuesta al

impulso del recinto, el cepstrum de $h(t)$ ó $\hat{h}(t)$ se expresa por la siguiente transformación temporal¹:

$$\hat{h}(t) = \mathcal{F}^{-1}\{\log[H(\omega)]\} \quad (213)$$

La descomposición de $H(\omega)$ en componentes fase mínima y paso todo sigue siendo válida en el dominio del cepstrum, pero ahora la contribución de cada parte es aditiva y no multiplicativa:

$$\begin{aligned} \hat{h}(t) &= \mathcal{F}^{-1}\{\log[|H(\omega)| \exp[j\phi_{\min}(\omega)] \exp[j\phi_{\text{all}}(\omega)]]\} = \\ &= \underbrace{\mathcal{F}^{-1}\{\log[|H(\omega)|]\}}_{\hat{h}_{\min}(t)} + j\mathcal{F}^{-1}\{\phi_{\min}(\omega)\} + j\mathcal{F}^{-1}\{\phi_{\text{all}}(\omega)\} \end{aligned} \quad (214)$$

Se llama cepstrum real de una función temporal $h(t)$ al cepstrum del módulo de su espectro, es decir:

$$\hat{h}_r(t) = \mathcal{F}^{-1}\{\log[|H(\omega)|]\} \quad (215)$$

Se puede demostrar [Furui 91] que la parte fase mínima de $h(t)$, $\hat{h}_{\min}(t)$ viene dada por:

$$\hat{h}_{\min}(t) = \begin{cases} \hat{h}_r(t), & t = 0 \\ 2\hat{h}_r(t), & t > 0 \\ 0, & t < 0 \end{cases} \quad (216)$$

o lo que es igual,

$$\hat{h}_{\min}(t) = \hat{h}_r(t)r(t) \quad (217)$$

con:

$$r(t) = \begin{cases} 1, & t = 0 \\ 2, & t > 0 \\ 0, & t < 0 \end{cases} \quad (218)$$

Los métodos para obtener una derreverberación ciega basados en descomposición fase mínima-paso todo, utilizan el hecho de que, para señales de habla, la parte fase mínima del cepstrum sin reverberación es más corta que la parte fase mínima del mismo cepstrum pero con reverberación. Por lo tanto, el método consiste en aplicar un filtro paso bajo en el dominio del cepstrum(o *liftering*) a la parte fase mínima de la señal que se pretende reconstruir, y posteriormente rehacer la señal global añadiendo la parte paso todo. Por supuesto, esta característica de la función de trasferencia $H(\omega)$ se puede extender a la salida eléctrica de cada micrófono que se ha llamado $y(t)$ ó $Y(\omega)$. Es decir se puede hablar de la componente fase mínima de la salida eléctrica de cada micrófono en tiempo $y_{\min}(t)$ o en frecuencia $Y_{\min}(\omega)$, o de forma equivalente se puede considerar la componente paso todo $y_{\text{all}}(t)$ o $Y_{\text{all}}(\omega)$, cumpliéndose

$$Y(\omega) = Y_{\min}(\omega) Y_{\text{all}}(\omega) \quad (219)$$

¹ Para adoptar la nomenclatura habitual usada en la literatura sobre el tema, en las definiciones de cepstrum la función “log” se refiere a logaritmo en base “e”, a diferencia de otras definiciones aparecidas en la Tesis en las que se sobreentiende el logaritmo en base 10.

Para hacer esta descomposición se usa el mismo método de descomposición cepstral que el propuesto para la función de transferencia electroacústica $H(\omega)$ de cada micrófono. La reverberación de $Y(\omega)$ se puede atenuar igualmente mediante un *liftering* de $y_{\min}(t)$, según se explica en [Liu 96].

La derreverberación monocanal mediante el *liftering* paso bajo de $y_{\min}(t)$ puede ser mejorada si se aplica a una señal multicanal procedente de un array microfónico, como se explica a continuación. La derreverberación ciega multicanal mediante descomposición cepstral ha sido propuesta en [Liu 96] y aplicada por el autor a un array anidado de 15 micrófonos en [González-Rodríguez 99-b] [González-Rodríguez 00] y [Sánchez-Bote 00]. A continuación se desarrolla con más detalle este esquema de derreverberación ciega basado en descomposición fase mínima-paso todo. En la Figura 47 se muestra el diagrama de bloques de este derreverberador.

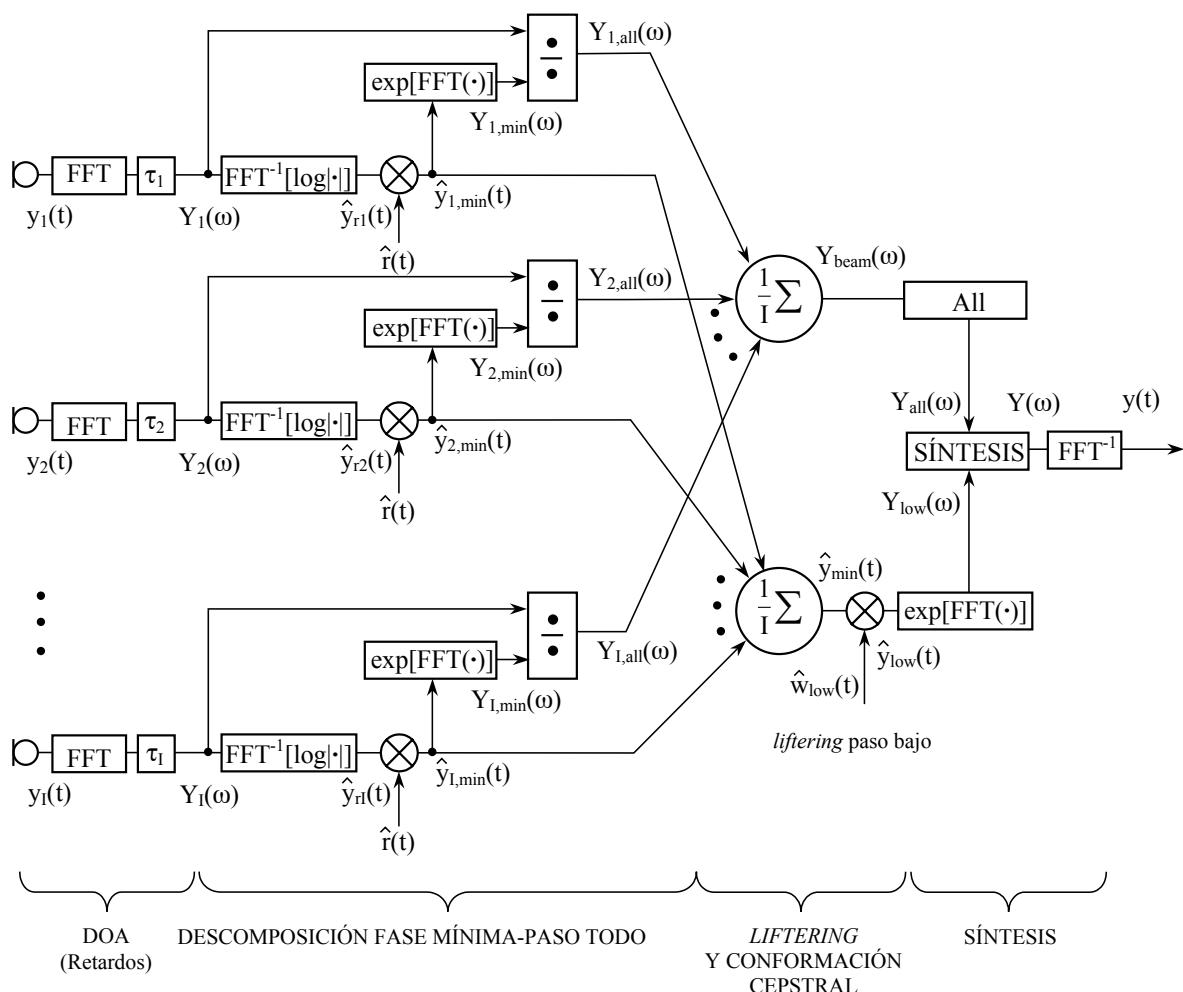


Figura 47. Derreverberador ciego multicanal implementado sobre un array lineal de I micrófonos y basado en descomposición de la señal en sus componentes fase mínima y paso todo.

La señal entregada por cada micrófono una vez realizada la conversión al dominio de la frecuencia (mediante la FFT) y alineada temporalmente mediante los retardos τ_i , se divide en sus componentes paso todo $Y_{i,all}(\omega)$ y fase mínima $Y_{i,min}(\omega)$. Para ello se calcula el cepstrum real $\hat{y}_{ri}(t)$ de cada canal utilizando (215). Despues se aplica la función $r(t)$ de (218) para obtener el cepstrum de la parte fase mínima $\hat{y}_{i,min}(t)$. Mediante la transformación inversa al

cepstrum de (213) se obtiene el espectro de la parte fase mínima buscada $Y_{i,\min}(\omega)$. Después se extrae la parte paso todo de cada canal mediante:

$$Y_{i,\text{all}}(\omega) = \frac{Y_i(\omega)}{Y_{i,\min}(\omega)} \quad (220)$$

Se hace una conformación convencional (conformación cepstral) de cada una de las dos partes halladas, el cepstrum de la parte fase mínima –obteniéndose $\hat{y}_{\min}(t)$ – y el espectro de la parte paso todo –obteniéndose $Y_{\text{beam}}(\omega)$ –. A la parte fase mínima $\hat{y}_{\min}(t)$ se le hace un *liftering* paso bajo, mediante el filtro $\hat{w}_{\text{low}}(t)$ obteniéndose a la salida $\hat{y}_{\text{low}}(t)$ que como se ha dicho presenta menor reverberación que el cepstrum de partida, a la salida de cada micrófono $\hat{y}_{i,\min}(t)$. El espectro $Y_{\text{beam}}(\omega)$ no tiene una característica paso todo ya que la operación suma que realiza el conformador no preserva las propiedades paso todo de cada una de las entradas del sumador. Para reducir distorsión se pasa $Y_{\text{beam}}(\omega)$ por el reconstructor “all” de la Figura 47 que extrae el espectro paso todo de la señal conformada mediante la expresión:

$$Y_{\text{all}}(\omega) = \frac{Y_{\text{beam}}(\omega)}{Y_{\text{beam,min}}(\omega)} \quad (221)$$

siendo $Y_{\text{beam,min}}(\omega)$ la parte fase mínima (espectro) de $Y_{\text{beam}}(\omega)$. Finalmente se reconstruye el espectro de salida mediante

$$Y(\omega) = Y_{\text{low}}(\omega) Y_{\text{all}}(\omega) \quad (222)$$

siendo $Y_{\text{low}}(\omega)$ el correspondiente cepstrum inverso de $\hat{y}_{\text{low}}(t)$, obteniéndose la señal $Y(\omega)$ a la salida del procesador, para ser finalmente convertida en $y(t)$, ya en el dominio temporal.

El esquema de la Figura 47 equivale a un conformador en el dominio cepstral con *liftering* paso bajo para eliminar las componentes de reverberación contenidas en la parte fase mínima de la señal captada por los micrófonos. La conformación del tipo filtrado y suma, tal y como se explicó en el punto 2.1.6 de esta Tesis, está contenida en los dos sumadores de la Figura 47, con los acondicionamientos necesarios en las señales intermedias para minimizar la distorsión inherente a la descomposición cepstral multicanal.

4.2.2 Otros métodos de derreverberación mediante postfiltrado

En un sistema de comunicación “manos libres”, donde existe un altavoz y un micrófono enfrentados a una fuente sonora, se produce el fenómeno de la realimentación acústica, o del “eco telefónico” según sea el caso, como se explica a continuación.

Los fenómenos de la realimentación acústica o del “eco telefónico” están producidos porque la señal captada por el micrófono es reproducida por el altavoz, y llega al mismo micrófono con cierto retardo. Si este retardo está ocasionado por el sistema acústico multirayecto, es de pequeña cuantía (algunos milisegundos) y se produce el fenómeno de la realimentación acústica. La realimentación acústica aumenta la reverberación natural del recinto acústico donde se produce la comunicación, pudiéndose llegar a la oscilación del sistema, porque en alguna frecuencia la ganancia del lazo de realimentación puede producir una amplificación excesiva. Si el retardo es muy grande (centenas de milisegundos) el efecto que se produce no es la oscilación, sino una molesta sensación de eco.

La filosofía más extendida usada para la cancelación de los ecos acústicos consiste en restar de la señal producida por el micrófono la señal del eco producido por la transmisión acústica (habitualmente sólo el asociado al camino directo y a las primeras reflexiones entre altavoz y micrófono). Esto se hace normalmente mediante un filtro adaptativo que minimiza en tiempo real la salida del micrófono. Esta metodología de funcionamiento, enlaza directamente con el concepto de filtrado inverso, introducido al principio del punto 4.2 sobre derreverberación genérica. Por supuesto, este esquema de cancelación de eco y filtrado adaptativo se puede usar no sólo para minimizar la realimentación y el eco acústico sino para reducir la reverberación. El problema suele venir siempre dado por la gran cantidad de coeficientes que debe tener el filtro adaptativo si se quiere atenuar la reverberación (la respuesta al impulso que hay que eliminar es muy larga) y por la variabilidad del sistema multirayecto que origina el eco acústico.

En cualquier caso, la cancelación de ecos acústicos en comunicaciones manos libres es un asunto que ha sido extensamente tratado en el ámbito de la acústica y la tecnología, existiendo en la literatura científica una amplia oferta de algoritmos y de implementaciones prácticas. En las partes I y II de [Gay 00] se puede encontrar una extensa exposición del estado del arte sobre este tema.

Recientemente se están desarrollando otros métodos para la derreverberación ciega, entre los se destacan los basados en la modificación de los residuos de codificación por predicción lineal (LP ó LPC). Estos métodos de reconstrucción LPC se pueden utilizar de forma general para la mejora de la señal de habla, en condiciones de baja SNR (véase el punto 4.1.5 de la Tesis), aunque en los últimos años están siendo utilizados para tratar específicamente la reverberación [Yegnanarayana 98] [Yegnanarayana 00] [Griebel 01]. La idea es modificar los parámetros LP residuales de acuerdo con la cantidad de reverberación, ya que ésta modifica de manera característica dichos residuos, aumentando su entropía temporal.

Para finalizar se hace una breve recapitulación. Como se ha visto existen básicamente dos grandes grupos de técnicas para minimizar la reverberación. El filtrado adaptativo y la derreverberación ciega mediante postfiltrado de reconstrucción. En el primer grupo se pueden incluir los esquemas de derreverberación basados en inversión de $H(\omega)$ y la cancelación de ecos acústicos. En el segundo la derreverberación ciega basada en descomposición cepstral y la reconstrucción de los residuos LP. Los primeros están limitados por la máxima longitud admisible para el filtro adaptativo y por la rapidez de adaptación, sujeta a la variabilidad del sistema acústico multirayecto. Los métodos de postprocesado ciego suelen estar limitados por la complejidad computacional que puede suponer un serio inconveniente para la implementación en tiempo real.

En cualquier caso, debe resaltarse que la derreverberación producida por un determinado algoritmo, siempre se ve mejorada en gran medida cuando se pasa de una implementación monocanal a una multicanal, mediante un array de micrófonos, debido a la atenuación de las señales laterales al eje de máxima captación que produce la conformación de haz.

5 EVALUACIÓN OBJETIVA DE CALIDAD DE LA SEÑAL DE HABLA

El principal propósito del trabajo propuesto en esta Tesis es producir una mejora de la señal de habla cuando ésta se capta en condiciones adversas. En el proceso de captura y transmisión de señal de voz existen muchas fuentes de distorsión y de ruido, como se ilustra en la Figura 48. Por supuesto una parte de esa distorsión está asociada al ruido y la reverberación, objetivo principal en la mejora producida por el array de micrófonos. No obstante existen otras fuentes de distorsión que degradan a la señal de voz, por ejemplo la falta de fidelidad de los elementos de amplificación (distorsión lineal y no lineal, ruido interno), las imperfecciones en la respuesta de los transductores electroacústicos, como el efecto proximidad en los micrófonos directivos [Sánchez-Bote 02-a], y la distorsión en la transducción mecánica, etc. Por otra parte, al utilizar en asociación con el array de micrófonos un sistema de procesado digital, la propia herramienta de limpieza de voz introduce distorsión, más específicamente origina el llamado ruido musical, como se ha explicado en apartados anteriores (capítulo 4.1), que se debe básicamente a que el procesado mediante la transformada STFT ocasiona variaciones rápidas y discontinuidades [Klabbers 01] en el espectro instantáneo a la salida del procesador. Normalmente, esta distorsión, inherente al procesado, es tanto mayor cuanto mayor sea la degeneración de la señal a limpiar (mayor la reverberación o menor la SNR) puesto que es en ese caso cuando el comportamiento del procesador se ve más influido por las perturbaciones acústicas. Lo que nunca debe ocurrir es que la distorsión introducida sea mayor que la distorsión que se ha pretendido eliminar.

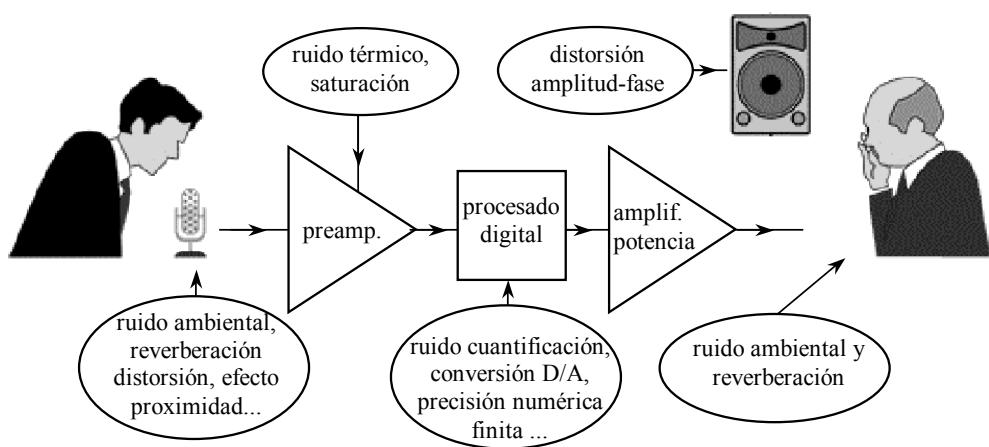


Figura 48. Comunicación mediante señal vocal. Fuentes de distorsión.

Sin embargo la cantidad de distorsión remanente en la señal procesada es muy difícil de evaluar, tanto objetivamente como de forma subjetiva. Objetivamente lo es porque los efectos de algunas fuentes de distorsión son difíciles de medir, por ejemplo la cantidad de reverberación existente en una señal. Pero también existe una dificultad en la evaluación

subjetiva, ya que las fuentes de distorsión son de naturaleza tan distinta (por ejemplo el ruido musical y la reverberación) que se hace difícil saber cuáles tienen un efecto más perjudicial y en qué grado a la hora de establecer un baremo que objetivice las impresiones subjetivas.

La valoración del grado de distorsión en la señal de habla se llamará aquí de forma general evaluación de la calidad de habla y tiene por tanto dos vertientes, la evaluación subjetiva y la evaluación objetiva. En cualquiera de los casos la evaluación de la calidad de habla es problemática por su gran dosis de subjetividad. La evaluación de la calidad de habla se utilizará en esta Tesis para calificar el comportamiento de un determinado procesador diseñado para la mejora de habla, para lo cual será necesario medir la calidad de la señal antes y después de dicho procesador.

En la evaluación subjetiva siempre se necesita contar con la opinión de un grupo de oyentes que comparan la señal limpia con la procesada. De estas pruebas subjetivas normalmente se infieren modelos más o menos objetivos que cuantifican en un determinado baremo la calidad de la señal vocal de salida, en función de los factores de distorsión. En este sentido podría hablarse por otro lado de las pruebas quasi objetivas, en las cuales para evaluar la calidad de la señal de habla, no se mide exactamente dicha señal sino las características del canal de transmisión (tiempo de reverberación, retardo o ruido), que previamente han sido evaluadas subjetivamente. La evaluación objetiva pura compara determinados parámetros de la señal de voz limpia con la señal de voz sucia y ofrece medidas de parecido o de distancia. Estas medidas casi nunca se realizan en el dominio temporal sino en algún dominio transformado.

Tradicionalmente, la evaluación de calidad de habla ha sido abordada por dos ámbitos de la ciencia y la tecnología: la telefonía por una parte y la acústica de recintos y electroacústica para el refuerzo sonoro por otra parte.

El campo de la telefonía tiene una problemática propia, por ejemplo, el efecto en una comunicación de los retardos de varios segundos o la influencia de los codificadores (ADPCM, GSM...), aspectos que para esta Tesis no tienen gran interés, porque el procesado en array carece de las peculiaridades de un sistema de transmisión telefónica. Sin embargo, tanto la telefonía como los sistemas de reproducción sonora están interesados en lo que genéricamente podría llamarse claridad de habla. La claridad de habla [Furui 91] [Denisowski 01] es el concepto que usualmente se conoce como inteligibilidad en el dominio de la acústica de recintos y la electroacústica.

Históricamente, la claridad de habla ha sido evaluada mediante técnicas subjetivas en el campo de la telefonía. La más utilizada hasta el momento ha sido la prueba MOS (*Mean Opinion Score*) en la que la calidad de la voz se califica con un baremo de 1 a 5.

Desde el punto de vista de evaluación subjetiva, se utilizan otros tests además del MOS, que pueden ser encontrados en la literatura científica sobre procesado de habla. Se pueden destacar el test DRT (*Diagnostics Rhyme Test*) usado para palabras o el SUS (*Semantically Unpredictable Sentences*) para frases, cuyas referencias originales puede encontrarse en [Tsoukalas 97].

Los evaluadores objetivos de claridad de habla consideradas como un estándar en la actualidad en el mundo telefónico son dos, la medida PSQM (*Perceptual Speech Quality Measurement*) y la PAMS (*Perceptual Analysis Measurement System*). La primera de ellas ha sido desarrollada por KPM Research en Holanda y la segunda por British Telecom en Gran Bretaña. Ambos esquemas de medida se basan en el análisis perceptual en bandas críticas de la señal vocal, como se ha estudiado en el punto 4.1.3 de esta Tesis. Los espectros de la señal

de referencia y la señal distorsionada de salida se descomponen en bandas críticas, se comparan y se obtiene una calificación numérica sobre su mayor o menor grado de parecido. La gran aceptación de ambos evaluadores se sustenta en que sus resultados coinciden de forma importante con otros métodos de evaluación subjetiva, por ejemplo la prueba MOS.

Las medidas de calidad objetiva más convencionales en el campo del procesado digital de habla se basan en la relación señal a ruido [Quackenbush 88], bien sea SNR simplemente, con algún tipo de ponderación como la red de valoración A (SNR_A) o aquellas que evalúan la cantidad de ruido audible, como la relación señal a enmascaradora (NMR o *Noise to Masking Ratio*) [Herré 92] según se definió en el apartado 4.1.3 de esta Tesis. Otros tipos de medidas objetivas de la calidad de habla utilizan una evaluación de distancias de parámetros característicos de la señal original (que tiene que estar disponible) con la señal antes y después de ser procesada. Estos parámetros característicos tienen que ser inherentes a la señal de habla. Son muy usados, por ejemplo, las distancias de componentes cepstrales y de parámetros LAR (*Log Area Ratio*).

Otro grupo importante de evaluadores de calidad de habla son los índices de inteligibilidad. El mundo de la acústica de recintos y la electroacústica ha estimado tradicionalmente la inteligibilidad de la señal vocal mediante lo que antes se ha llamado pruebas cuasi objetivas, generándose diferentes índices de inteligibilidad, entre los que destacan el índice de articulación –AI o *Articulation Index* [Kryter 62]– y la pérdida de articulación de consonantes –Alcons o *Articulation loss of consonants* [Peutz 71]–. En todas ellas se evalúa subjetivamente, mediante encuestas a oyentes, la calidad de la señal vocal recibida y se parametriza el resultado en función de características objetivas como el ruido, la reverberación y los ecos, inherentes al canal de transmisión acústica. Entonces, para evaluar finalmente la calidad o inteligibilidad de una señal de voz, sólo será necesario medir objetivamente las características del canal en el que se va a propagar.

Desde hace unos 20 años se viene utilizando con gran éxito un método puramente objetivo, y además bastante original, para evaluar la inteligibilidad de la señal de habla; se trata del índice de transmisión del habla, STI –*Speech Transmission Index* [Steeneken 80] [Houtgast 85]– que ha derivado en el índice STI rápido, el RASTI –*Rapid STI* [Steeneken 85]–. La gran ventaja del método STI es que es capaz de detectar reverberación, calidad de la que carecen otros métodos puramente objetivos como el PAMS, PSQM o AI. Esto es de gran interés aquí, puesto que una de las misiones más importantes de la mejora de habla usando un array de micrófonos será la derreverberación.

En el tratamiento de los resultados obtenidos con el procesado en array que se expondrán posteriormente en esta Tesis (en las partes 2 y 3) se han utilizado evaluadores objetivos de calidad de habla de tres tipos diferentes. Por una parte los basados en las distancias cepstrales y de parámetros LAR, por otra parte los que utilizan medidas más tradicionales basadas en la mejora de SNR y finalmente los índices de inteligibilidad de tipo acústico, como el índice de articulación AI y el RASTI. A continuación se exponen con mayor detalle los fundamentos de estos indicadores objetivos.

5.1 MEDIDAS DE LA RELACIÓN SEÑAL A RUIDO. SNR, SNR_A Y NMR

El parámetro más frecuentemente utilizado para la evaluación objetiva de la calidad de habla es la relación señal a ruido, SNR [Quackenbush 88]. La relación señal a ruido es una medida que establece la cantidad de ruido que se añade a la señal de habla limpia. En (28) se

estableció el concepto de ruido aditivo que se suma a la señal multicanal proporcionada por un array microfónico. En (157) se expone el mismo concepto de ruido aditivo pero considerando la señal monocanal única –señal conformada $y_{SD}(t)$ ó $Y_{SD}(\omega)$ – correspondiente a la salida del array funcionando como conformador superdirectivo. Efectivamente, considerando que en la salida procesada del array está presente la señal de referencia $x_0(t)$ ó $X_0(\omega)$ que es la señal deseada que se quiere recuperar –o una copia equivalente a la señal de presión $p(t)$ producida por la fuente de habla–, entonces, el ruido aditivo puede calcularse mediante la simple sustracción:

$$N(\omega) = Y(\omega) - X_0(\omega) \quad (223)$$

Téngase en cuenta que, puesto de esta manera, el ruido $N(\omega)$ ó $n(t)$ incluye también la reverberación y la presencia de otras fuentes de voz consideradas como ruido. Efectivamente, como se definió en su momento la señal de referencia $X_0(\omega)$ es la que la fuente principal produce en un micrófono hipotético situado en el centro del array, y por tanto corresponde a la voz seca, sin reverberación. La reverberación, debida al sistema multirayecto acústico, se suma –véase (25)– a la señal anecoica original, y por tanto $N(\omega)$ es algo más que el ruido aditivo en su versión más tradicional.

Para evaluar la cantidad de ruido según (223) tiene que estar disponible la señal de referencia $X_0(\omega)$, lo cual no es siempre posible. En la práctica lo que se suele hacer es captar, a la vez que se realiza el procesado, la fuente de voz de referencia con un micrófono igual a los que integran el array microfónico y situado en campo cercano de la fuente, de tal manera que el ruido y la reverberación sean mínimos en la adquisición, para después proceder al cálculo (223). Sin embargo, a la hora de aplicar el cálculo del ruido por (223), la señal $Y(\omega)$ debe alinearse en tiempo con la referencia $X_0(\omega)$, que por estar captada cerca de la fuente presentará un retardo diferente a la salida del array $Y(\omega)$. Cualquiera de los métodos de localización de fuente usados en el punto 3.2.2 puede ser útil para calcular los retardos necesarios en la alineación temporal. En especial suele utilizarse el método PHAT, descrito en el apartado 3.2.2, por su gran sencillez y efectividad.

La SNR suele considerarse en términos de potencia. Desde un punto de vista más formal habría que evaluar el autoespectro de ruido mediante la esperanza matemática (37):

$$\Phi_{NN}(\omega) = E\{[Y(\omega) - X_0(\omega)][Y^*(\omega) - X_0^*(\omega)]\} \quad (224)$$

En la práctica, por no estar disponible toda la información estadística de la señal, el cálculo de la potencia de ruido se limitará a una trama temporal de voz, k

$$|N(\omega, k)|^2 = |Y(\omega, k) - X_0(\omega, k)|^2 \approx |Y(\omega, k)|^2 - |X_0(\omega, k)|^2 \quad (225)$$

que posteriormente será tratada estadísticamente junto con otras tramas de tal manera que se tenga una muestra significativa del ruido presente en la señal de voz. En (225) el signo “aproximado” podrá ser aplicado cuando se presuponga incorrelación total entre voz y ruido. Esta presunción puede ser muy atrevida en la mayoría de los casos, ya que como se ha dicho $N(\omega)$ contiene la reverberación que no es totalmente incoherente con $X_0(\omega)$. Sin embargo si se dispone de la señal $X_0(\omega)$ en cada trama temporal y alineada en tiempo con $Y(\omega)$ no hará falta tal aproximación, ya que se puede calcular el segundo término de (225).

En la práctica el ruido se evalúa en una determinada banda b de frecuencias, como se hacía en (183) para el cálculo de la potencia por banda crítica auditiva:

$$|N(b, k)|^2 = \int_{\omega_{ib}}^{\omega_{fb}} |N(\omega, k)|^2 d\omega \quad (226)$$

con ω_{ib} y ω_{fb} las pulsaciones inicial y final de la banda b considerada.

En (226) hay que hacer alguna consideración, ya que es normal que cada trama temporal k se extraiga mediante la multiplicación de la señal de voz por una ventana ²v(t) de ponderación, necesaria para el análisis en frecuencia mediante la transformada de Fourier a corto plazo (STFT). Este proceso se llama enventanado temporal y sirve para atenuar el efecto borde en los extremos de la trama, que produce una distorsión excesiva del espectro que se pretende obtener cuando no se efectúa dicho enventanado. Esta operación de enventanado no conserva la potencia de la señal, por lo cual el cálculo (226) debe hacerse antes del enventanado, estimando la potencia de la señal mediante la suma cuadrática de las muestras temporales. En la práctica, las ventanas v(t) más utilizadas (la ventana Hanning, por ejemplo) distorsionan poco la potencia calculada. Considerando además que se va a calcular una relación SNR, hay que suponer que la desviación de potencia obtenida afectará de forma parecida a la señal y al ruido, con lo que se puede aplicar (226) después del enventanado sin mucho error. En [Gade 87-a] y [Gade 87-b] pueden encontrarse más detalles sobre la problemática del enventanado y la conservación de la potencia de la señal.

Una vez hechas estas consideraciones, la relación SNR como en (163), en la banda b y en la trama temporal k, vendrá dada por:

$$\text{SNR}(b, k) = \frac{|X_0(b, k)|^2}{|N(b, k)|^2} \quad (227)$$

Por otra parte, es muy frecuente en el ámbito de las tecnologías acústicas y de audio ponderar o filtrar el ruido, de tal manera que se considere la respuesta subjetiva del oído humano. Esto es, como el sistema auditivo humano es menos sensible a las altas y a las bajas frecuencias, se suele filtrar el ruido, para atenuar esas componentes del espectro. La curva de ponderación más usada para este propósito es la red A, aunque existen otras [IEC/CD 1672]. En la Figura 49 se representan las curvas de ponderación A, B y C.

De esa manera el ruido ponderado según la curva A sería:

$$|N_A(b, k)|^2 = \int_{\omega_{ib}}^{\omega_{fb}} |N(\omega, k)|^2 A^2(\omega) d\omega \quad (228)$$

siendo $|N_A(b, k)|^2 \leq |N(b, k)|^2$. Aquí cabrían las mismas consideraciones sobre el enventanado y el cálculo de la potencia hechas anteriormente. La relación señal a ruido A es entonces:

$$\text{SNR}_A(b, k) = \frac{|X_0(b, k)|^2}{|N_A(b, k)|^2} \quad (229)$$

cumpliéndose también que $\text{SNR}(b, k) \leq \text{SNR}_A(b, k)$.

² Nótese que se utiliza v(t) y no w(t), como es habitual en la literatura, para no crear confusiones con los coeficientes w de ponderación del array.

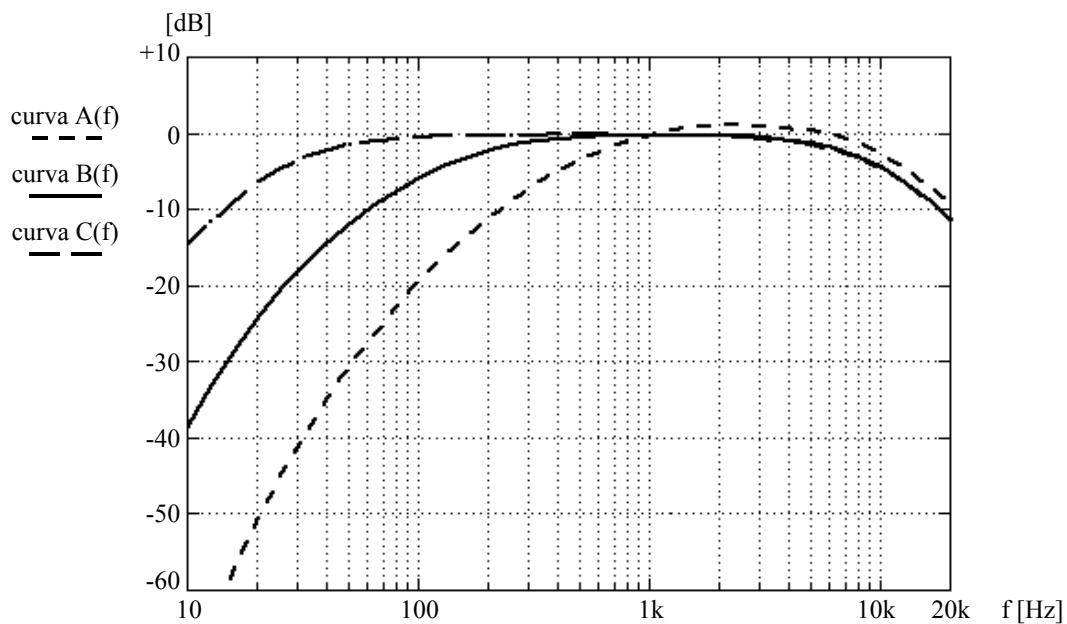


Figura 49. Curvas de ponderación A(f), B(f) y C(f) según [IEC/CD 1672].

Como se ha explicado, la relación SNR_A intenta dar un enfoque subjetivo al ruido captado. Una mejor aproximación a la sensación subjetiva de ruido la da el tratamiento auditivo, según se expuso en el punto 4.1.3 de esta Tesis. De esa manera la relación NMR (*Noise to Masking Ratio*) [Herre 92] ya vista en (182) también se utiliza para los mismos fines de evaluación objetiva. Una vez realizado un análisis en bandas críticas auditivas y calculados la potencia de ruido $|N(b, k)|^2$ y el umbral de enmascaramiento por banda crítica $T(b, k)$, para calcular NMR(b, k) se ha de aplicar la siguiente fórmula:

$$\text{NMR}(b, k) = \frac{|N(b, k)|^2}{(\omega_{fb} - \omega_{ib})T(b, k)} \quad (230)$$

donde se ha considerado el cociente $|N(b, k)|^2/(\omega_{fb}-\omega_{ib})$ que da la densidad espectral de potencia en la banda considerada, ya que el umbral de enmascaramiento $T(b, k)$ debe compararse con densidad espectral de potencia. Se aprecia que la relación NMR es un concepto equivalente a la SNR (227), pero hay que considerar que el ruido está en el numerador, por lo que a diferencia de esta última interesaría que la NMR sea pequeña. Si se consideran B bandas críticas en la totalidad del espectro bajo evaluación, la relación NMR en la banda B será

$$\text{NMR}(B, k) = \frac{1}{B} \sum_{b=1}^B \text{NMR}(b, k) \quad (231)$$

que como se ve se calcula como un promedio simple.

Las tres relaciones de calidad objetiva expuestas pueden ponerse en unidades logarítmicas de la siguiente forma.

$$\text{SNR}(b, k) [\text{dB}] = 10 \log \text{SNR}(b, k) \quad (232)$$

$$\text{SNR}_A(b, k) [\text{dB}] = 10 \log \text{SNR}_A(b, k) \quad (233)$$

$$\text{NMR}(B, k) [\text{dB}] = 10 \log \text{NMR}(B, k) \quad (234)$$

Por otra parte se ha de realizar algún tratamiento estadístico que considere las K tramas temporales del análisis. Normalmente, para el análisis en frecuencia cada trama k tiene una duración temporal igual a la de la ventana FFT, pero no tiene por qué ser así siempre. Un procedimiento frecuente de cálculo global es promediar las relaciones logarítmicas en las K tramas bajo estudio, de la siguiente manera:

$$\text{SNR}(b) [\text{dB}] = \frac{1}{K} \sum_{k=1}^K \text{SNR}(b, k) [\text{dB}] \quad (235)$$

$$\text{SNR}_A(b) [\text{dB}] = \frac{1}{K} \sum_{k=1}^K \text{SNR}_A(b, k) [\text{dB}] \quad (236)$$

$$\text{NMR}(B) [\text{dB}] = \frac{1}{K} \sum_{k=1}^K \text{NMR}(B, k) [\text{dB}] \quad (237)$$

En la práctica esas K tramas sólo incluyen los períodos temporales con actividad de habla. Por lo tanto se necesita una detección VAD para extraer los promedios anteriores.

Las expresiones (235), (236) y (237) consideran relaciones objetivas de calidad en términos absolutos. Las relaciones SNR, SNR_A y NMR serían unos buenos estimadores objetivos si la señal de referencia considerada $X_0(\omega)$ fuese la misma que está inmersa en la señal sucia $Y(\omega)$. Sin embargo esto no es del todo cierto, puesto que en el mejor de los casos la $X_0(\omega)$ utilizada para los cálculos será una versión atenuada o amplificada de la que realmente está presente mezclada con el ruido $N(\omega)$. Por ello en vez de calcular las relaciones absolutas anteriores se estudian las ganancias en SNR, SNR_A o NMR de la siguiente forma:

$$\text{GSNR}(b) [\text{dB}] = \text{SNR}_Y(b) [\text{dB}] - \text{SNR}_{Y_i}(b) [\text{dB}] \quad (238)$$

$$\text{GSNR}_A(b) [\text{dB}] = \text{SNR}_{AY}(b) [\text{dB}] - \text{SNR}_{AY_i}(b) [\text{dB}] \quad (239)$$

$$\text{GNMR}(B) [\text{dB}] = \text{NMR}_{Y_i}(B) [\text{dB}] - \text{NMR}_Y(B) [\text{dB}] \quad (240)$$

donde $\text{SNR}_Y(b)$, $\text{SNR}_{AY}(b)$ y $\text{NMR}_Y(B)$ son las mismas que en (235), (236) y (237) pero calculadas utilizando la señal $Y(\omega)$ de salida del procesador. Por otra parte $\text{SNR}_{Y_i}(b)$, $\text{SNR}_{AY_i}(b)$ y $\text{NMR}_{Y_i}(B)$ se calculan con la señal $Y_i(\omega)$ a la entrada del procesador, que es la salida de cualquier micrófono del array (con índice i genérico), normalmente el canal central del array. Evidentemente si el procesador funciona correctamente y produce una mejora efectiva de la calidad de habla, las tres ganancias GSNR, GSNR_A y GNMR deben ser positivas.

A pesar de que las medidas de relación señal a ruido apuntadas son las que mejor dan cuenta del ruido y la reverberación remanentes en la señal procesada, por la forma de calcular $N(\omega)$, hay que tener presente que la forma más típica de cuantificar en primera instancia la cantidad de ruido que presenta una señal de audio es mediante la relación señal a ruido a posteriori, que se calcula, sobre señales temporales, de la siguiente forma

$$\text{SNR}_{\text{post}} [\text{dB}] = 10 \log \frac{y_{\text{MS}}}{n_{\text{MS}}} \quad (241)$$

donde y_{MS} y n_{MS} representan el valor cuadrático medio (MS, *Mean Squared*) correspondiente a las señales $y(t)$ y $n(t)$ calculados mediante las conocidas expresiones:

$$y_{MS} = \frac{1}{T_y} \int_{T_y} y^2(t) dt \quad (242)$$

$$n_{MS} = \frac{1}{T_n} \int_{T_n} y^2(t) dt \quad (243)$$

El periodo temporal T_y corresponde a los lapsos de tiempo donde la señal de habla está presente y T_n a aquellos donde está ausente. Por lo tanto se necesita una detección de actividad de habla VAD como paso previo al cálculo de (241). Evidentemente la expresión (241) sobreestima la señal ya que en (242) se considera el ruido y la reverberación como señal e infraestima el ruido, ya que en (243) no se cuenta la reverberación como perturbación que no es sumada al ruido. Por lo tanto se cumplirá siempre que SNR_{post} [dB] >> $SNR(b)$ [dB] de (235) y su magnitud está en consonancia con la asignación tradicional de relación señal a ruido que se le da a una señal de habla con ruido. Por eso SNR_{post} [dB] (241) suele usarse como una forma de catalogación inicial de la calidad de la señal de habla, teniendo en cuenta la cantidad de ruido de fondo en los silencios de señal de voz y es lo que en los capítulos posteriores de esta Tesis se llamará a veces medida tradicional de la relación señal a ruido.

5.2 MEDIDAS BASADAS EN PREDICCIÓN LINEAL (LP)

Existe todo un campo de conocimiento dentro del procesado de voz que modela la señal de habla mediante el análisis por medio de predicción lineal (LP o *Linear Prediction*). La filosofía del método consiste básicamente en que la generación de voz humana puede equipararse a un modelo de emisión basado en una señal que atraviesa un filtro con un determinado número de polos y ceros –una profundización extensa sobre predicción lineal puede encontrarse en [Deller 01]–. Se puede demostrar que si se modela la voz con un filtro de sólo polos no se pierde una cantidad significativa de información. La ventaja de esta simplificación es que los coeficientes del filtro de sólo polos se pueden determinar mediante la resolución de una serie de ecuaciones lineales. Los coeficientes que implementan dicho filtro de sólo polos son conocidos como coeficientes LP. Existen diferentes tipos de coeficientes LP en función de la forma de construir el filtro generador de habla, pero el concepto de filtro de sólo polos se mantiene en todos ellos.

El análisis de la señal de voz mediante coeficientes LP no tiene en cuenta la información de fase de dicha señal, cosa que en principio es poco importante desde el punto de vista de percepción subjetiva, ya que muchas investigaciones han demostrado que el oído humano no considera de modo determinante la fase del sonido a la hora de producirse la percepción acústica. Esto puede ser muy interesante al implementar una evaluación objetiva basada en parámetros LP, ya que las posibles desalineaciones de fase en una trama de voz estacionaria no influirían en el análisis de la misma. Esta ventaja no la tienen las medidas de calidad objetivas basadas en SNR, y desarrolladas en el apartado anterior, ya que se realiza una comparación punto a punto entre dos señales y por consiguiente se necesita una perfecta alineación temporal previa a toda comparación. Además, cualquier efecto colateral poco importante a priori (al menos desde el punto de vista subjetivo), en cualquiera de las señales a analizar (por ejemplo que una locución a comparar haya sido captada por un micrófono o un preamplificador ligeramente distinto), puede variar de forma importante la fase de la señal captada y alterar de forma significativa las medidas objetivas de calidad.

Cuando se dispone de una estimación de la señal limpia de voz o referencia, dada por $X_0(\omega)$, la forma de estimar la calidad de habla es comparar el “parecido” de los parámetros LP

de la señal sucia (antes de procesar) con la referencia, con el “parecido” de la señal procesada con la referencia. Este parecido suele ser una medida de distancia de un juego de parámetros LAR correspondiente a dos tramas de voz. Esta distancia puede ser una distancia euclídea sin ponderar o ponderada. Una medida de distancia muy utilizada es la distancia de Itakura basada en las diferencias existentes en el modelo de sólo polos entre las dos tramas de voz a analizar –más detalles en [Deller 01]–. Un conjunto de medidas de calidad parecidas engloba las llamadas distancias paramétricas LPC (*Linear Predictive Coding* o Codificación por Predicción Lineal).

Se pueden usar diferentes tipos de coeficientes LP para hacer medidas de distancia (coeficientes de predicción de reflexión, parcor, etc.), aunque el juego de parámetros LP más usados para la estimación objetiva de la calidad de habla son los parámetros LAR. Los parámetros LAR fueron propuestos inicialmente para codificación de habla. Diversos autores, por ejemplo en [Quackenbush 88] han mostrado que los parámetros LAR tienen una mayor correlación con la impresión subjetiva de calidad que otros tipos de parámetros LP.

Sea $\mathbf{g}_{X_0}(k)$ el vector de parámetros LAR de la señal de referencia $X_0(\omega, k)$ que corresponde a la transformada STDFT de $x_0(t)$ en la trama temporal k y $\mathbf{g}_Y(k)$ el correspondiente a la salida procesada $Y(\omega, k)$ de un array de micrófonos. Ambos son vectores columna de longitud G , correspondiente al número de parámetros LAR considerado en cada trama de voz. Si se usa la norma euclídea, la distancia entre ambos conjuntos de parámetros LAR viene dada por:

$$dLAR_{X_0Y}(k) = \|\mathbf{g}_{X_0}(k) - \mathbf{g}_Y(k)\| \quad (244)$$

donde $\|\cdot\|$ representa dicha norma euclídea.

Si se promedian las K tramas temporales donde existe actividad de voz:

$$dLAR_{X_0Y} = \frac{1}{K} \sum_{k=1}^K dLAR_{X_0Y}(k) \quad (245)$$

se obtiene un valor único de distancia entre dos fragmentos de habla. Siguiendo la filosofía de cálculo adoptada en la evaluación de la mejora de relación señal a ruido, lo que se suele hacer aquí es estimar la mejora de la señal de voz producida por el procesador antes y después de ser procesada, y siempre tomando como referencia la señal $X_0(\omega, k)$. Es decir, si se quiere evaluar mediante $dLAR$ la mejora de la señal $y_i(t)$ (que corresponde a una de las salidas eléctricas de array, normalmente la correspondiente al micrófono central) con respecto a la salida $y(t)$ del array, y se conoce la señal de referencia $x_0(t)$, lo que se evalúa es:

$$GdLAR = \frac{dLAR_{X_0Y_i}}{dLAR_{X_0Y}} \quad (246)$$

y de forma logarítmica,

$$GdLAR [\text{dB}] = 20 \log GdLAR \quad (247)$$

de tal manera que si el resultado es positivo, el procesador ha producido mejora.

Sin abandonar los parámetros LP, el cepstrum de la señal de habla (véase el punto 4.2.1 de esta Tesis) también se usa frecuentemente en la determinación de la calidad objetiva de la voz. En concreto se usa el cepstrum real (215) en contraste con el cepstrum complejo (213) (aunque debe decirse que si la señal de voz es fase mínima no existe una diferencia sustancial

entre ambos tipos de medidas de cepstrum). Existe un gran paralelismo entre el modelo de predicción lineal para la voz y el concepto de cepstrum. De hecho los parámetros de cepstrum real de una señal de voz pueden obtenerse a partir de los parámetros LP [Deller 01]. Y como estos últimos, los parámetros de cepstrum también pueden ser indicativos de la calidad de una señal de voz, cuando se comparan con la señal de referencia. Puede demostrarse, que la distancia euclídea entre dos juegos de parámetros cepstrales es una medida de la similitud de los espectros de los que proceden. Así, dada una trama de voz con índice k, la distancia en el cepstrum real de la señal a la salida del procesador $Y(\omega, k)$ con respecto a la señal de referencia $X_0(\omega, k)$ viene dada por:

$$dRCEP_{X_0 Y}(k) = \|\mathbf{c}_{X_0}(k) - \mathbf{c}_Y(k)\| \quad (248)$$

con ³ $\mathbf{c}_{X_0}(k)$ y $\mathbf{c}_Y(k)$ los vectores de cepstrum real, de dimensión L (siendo L el número de puntos de análisis de la ventana temporal correspondiente a la trama k), asociados respectivamente a la señal de referencia $X_0(\omega, k)$ y a la salida del array $Y(\omega, k)$.

Hay que decir que para la estimación de la calidad de habla mediante el cepstrum real, normalmente sólo se consideran los L_1 primeros componentes cepstrales, es decir, se admite cierta cantidad de *liftering* en los vectores $\mathbf{c}_{X_0}(k)$ y $\mathbf{c}_Y(k)$. Además se excluye el primer parámetro RCEP de cepstrum real para independizar la distancia cepstral de la amplitud de la señal.

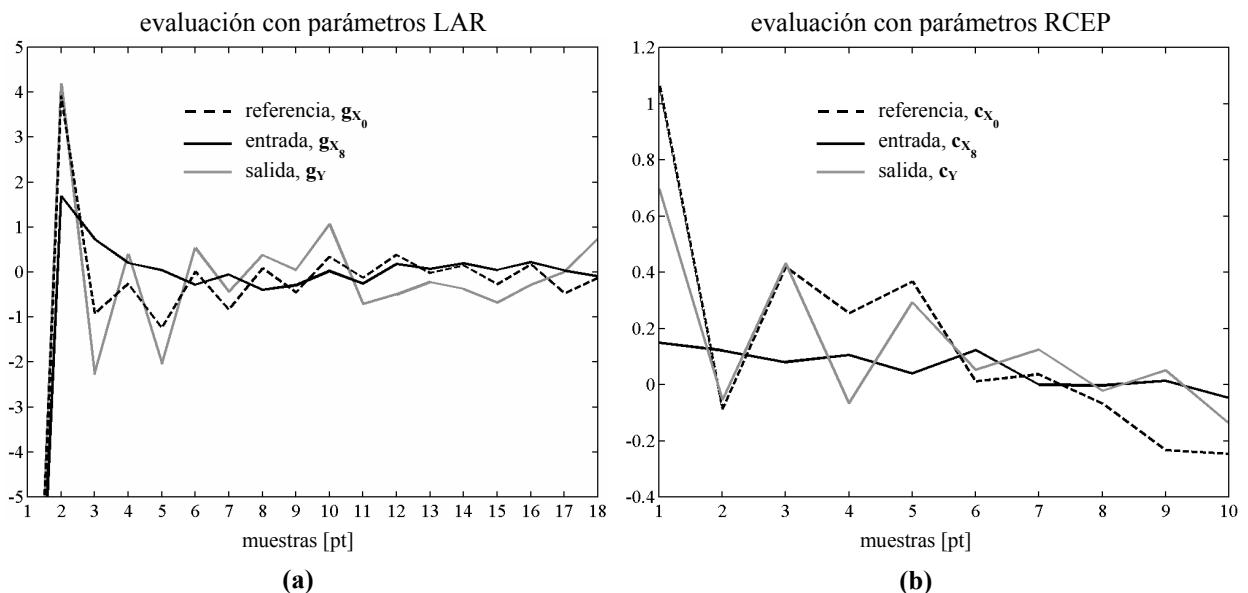


Figura 50. Muestra de una evaluación objetiva de mejora de señal de voz producida por un array de 15 micrófonos. Se considera una trama de voz de 512pt, con $f_s=16\text{kHz}$. **(a)** A partir de los parámetros LAR (orden 18). **(b)** A partir del cepstrum real, considerando los diez primeros componentes de RCEP.

De igual modo a como se hizo con los parámetros LAR, si se promedia en las tramas activas de voz,

$$dRCEP_{X_0 Y} = \frac{1}{K} \sum_{k=1}^K dRCEP_{X_0 Y}(k) \quad (249)$$

³ Aquí se ha optado por no emplear la tilde circunfleja “ $\hat{\cdot}$ ” para designar al cepstrum, en contra de lo que se hace en el capítulo 4.2.1 de la Tesis, para no crear confusión con la referencia a la estimación de una señal.

y si se considera la ganancia de distancia con respecto a la referencia:

$$GdRCEP = \frac{dRCEP_{X_0 Y_i}}{dRCEP_{X_0 Y}} \quad (250)$$

o en decibelios

$$GdRCEP [dB] = 20 \log GdRCEP \quad (251)$$

En la Figura 50 se representa una muestra de análisis de los parámetros LAR y RCEP en una trama de voz de 512 puntos ($f_s = 16\text{kHz}$) antes y después de pasar por un procesador en array con 15 micrófonos. Puede observarse cómo los parámetros de la señal procesada se asemejan más al original que la entrada sucia del array, representada por el canal octavo del mismo. Por lo tanto las medidas de distancia euclídea al original serán más pequeñas para la señal procesada $Y(\omega)$ que para la entrada del array $Y_8(\omega)$, y las ganancias (247) y (251) resultarán positivas.

5.3 ÍNDICE DE ARTICULACIÓN AI

El Índice de Articulación AI [Kryter 62] es un índice comprendido entre 0 y 1 que se utiliza para medir la inteligibilidad y por tanto la calidad de la señal de habla. Pertenece al grupo de los llamados índices de inteligibilidad y se ha usado desde la década de los 60 del siglo pasado para calificar la calidad acústica de recintos y sonorizaciones electroacústicas. El índice AI evalúa la relación SNR para diferentes bandas de frecuencia (considera normalmente 5 octavas). A continuación se desarrolla el método de cálculo del AI.

El Índice de Articulación AI en la trama de voz k se calcula con la siguiente fórmula:

$$AI(k) = \sum_{b=1}^5 c_b [\text{SNR}_T(b, k) [\text{dB}] + 12\text{dB}] \quad (252)$$

donde el índice b hace referencia a cada una de las cinco bandas de octava consideradas. El factor c_b es diferente para cada banda, como muestra la Tabla 4.

f_{0b} [Hz]	f_{ib} [Hz]	f_{fb} [Hz]	b	c_b
250	177	354	1	0.0024
500	354	707	2	0.0048
1000	707	1414	3	0.0074
2000	1414	2828	4	0.0109
4000	1414	5657	5	0.0078

Tabla 4. Coeficientes c_b para el cálculo del Índice de Articulación AI. Se indican las frecuencias centrales inicial y final de cada banda de octava.

El coeficiente c_b mayor corresponde a la octava de 2kHz, es decir se establece que el ruido es más perjudicial en esta banda a efectos de índice AI. $\text{SNR}_T(b, k)$ [dB] corresponde a la relación señal a ruido en dB en la banda b y en la trama k , como en (232), pero truncado entre -12dB y +18dB. Eso quiere decir que $\text{SNR} = -12\text{dB}$ y $\text{SNR} = +18\text{dB}$ son respectivamente los límites inferior y superior de calidad de habla considerados aquí. La suma de 12dB en (252) se hace para ajustar el valor obtenido a un número entre 0 y 1. Técnicamente, ese incremento

equivale a convertir los valores RMS de la señal de voz en valores de pico, ya que los coeficientes c_b de la Tabla 4 son aplicables a relaciones SNR de pico. El índice AI promediado en las K tramas donde se considera actividad de voz, se calcula por:

$$AI = \frac{1}{K} \sum_{k=1}^K AI(k) \quad (253)$$

y la ganancia en AI según,

$$GAI = AI_Y - AI_{Y_i} \quad (254)$$

Nótese que en este caso la ganancia del procesador se expresa como diferencia y no como cociente, tal y como ocurría con los evaluadores basados en SNR o en predicción lineal.

Para finalizar, hay que considerar que el Índice de Articulación en su formulación original considera la relación SNR de forma tradicional, como en (241), estimando la SNR a posteriori mediante la medida del ruido sólo en las tramas de ausencia de voz. Por lo tanto el método aquí propuesto mediante (252) no es muy adecuado para medir el índice AI absoluto según [Kryter 62], puesto que el ruido calculado por (223) sobreestima la medida tradicional de dicho ruido. Sin embargo, el método sí es adecuado para evaluar la ganancia GAI de (254), que será en la forma que dicho índice se utilizará en esta Tesis.

5.4 ÍNDICES DE INTELIGIBILIDAD STI y RASTI

Hasta ahora para medir de forma objetiva la calidad de la señal de habla se ha considerado de alguna u otra forma la relación señal a ruido. Es cierto, que si la señal de referencia $X_0(\omega)$ está perfectamente alineada en tiempo con la señal $Y(\omega)$ a la salida del procesador, es posible considerar la reverberación mediante la simple diferencia expresada en (223). Sin embargo, cuando esto no es así, es difícil evaluar la cantidad de reverberación mediante una observación a posteriori de la señal contaminada. El índice de transmisión del habla o STI (*Speech Transmission Index*) [Steeneken 80] [Houtgast 85] trata de salvar esta dificultad. El STI evalúa la disminución de inteligibilidad ocasionada en el proceso de transmisión electroacústica de la señal de habla a partir del análisis de las pérdidas de modulación del cuadrado de la señal de habla, equivalente a la intensidad acústica.

Efectivamente, la señal de intensidad acústica $I(t)$ de una señal de habla –que es el cuadrado de la presión acústica $p(t)$ – puede asociarse a una modulación de amplitud, como se muestra en la Figura 51.

La señal vocal típica (en su versión intensidad acústica) puede equipararse a una suma de señales moduladas en amplitud. Esta modulación consiste en una señal de banda ancha (caracterizada por la frecuencia moduladora f) multiplicada o modulada por una serie de tonos puros de muy baja frecuencia (caracterizados por la frecuencia de modulación F). La señal moduladora, por corresponder a la información de habla, es de espectro parecido al ruido rosa.

Como normalmente se hace un análisis en octavas, las frecuencias moduladoras f van desde 125Hz hasta 8kHz (en octavas). Las frecuencias de modulación F habitualmente consideradas son 14, desde 0.63Hz hasta 12.5Hz (en tercios de octava).

El espectro de muy baja frecuencia de la intensidad acústica $I(t)$, que contiene las frecuencias de modulación F constituye la envolvente de dicha señal de intensidad. Está asociado al efecto de modulación de nivel sobre una señal de banda ancha que se produce cuando se habla.

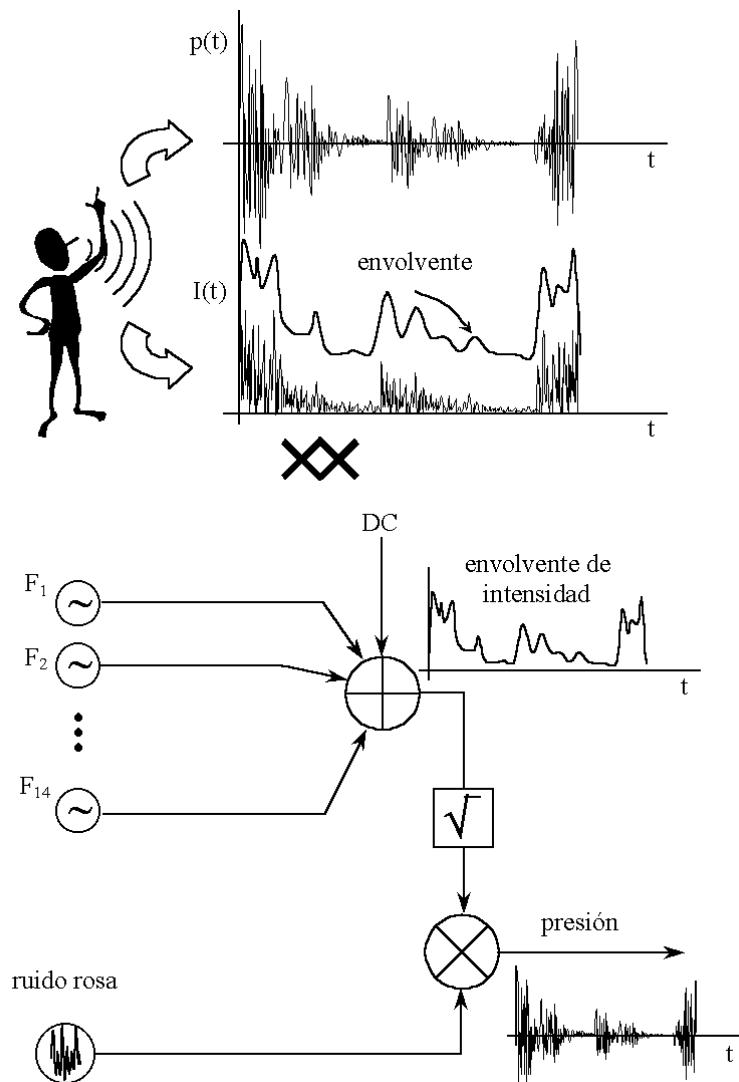


Figura 51. Envolvente de intensidad acústica.

5.4.1 Índice de modulación del habla, m

Para evaluar las pérdidas de inteligibilidad según el STI, se mide la degradación de la envolvente de la intensidad acústica $I(t)$. Dicha degradación equivale a pérdidas de modulación. Para explicar esto, considérese una sola frecuencia de modulación F . Sea $I_e(t)$ la envolvente de intensidad producida por un orador. La envolvente captada por el oyente será $I_s(t)$. Esta señal ha pasado por el medio de transmisión electroacústica, y en consecuencia ha sido contaminada por ruido aditivo y reverberación. Al degradarse ha perdido índice de modulación.

Se define índice de modulación del habla m :

$$m = \frac{I_{\max} - I_0}{I_0} \quad (255)$$

y se aplica a la señal envolvente de intensidad, siendo I_{\max} e I_0 respectivamente el valor máximo alcanzado y el valor medio correspondientes a dicha envolvente.

En la Figura 52 se ilustra el fenómeno de la modulación del habla. Se supone inicialmente que el índice de modulación en emisión es la unidad, y por tanto que el índice de modulación en recepción ha de ser menor que la unidad. Es decir, la degradación que sufre la señal vocal en su viaje por la cadena de reproducción electroacústica se traduce en una pérdida de modulación. Cuanto menor sea el m recibido mayor degradación habrá y menor inteligibilidad será percibida.

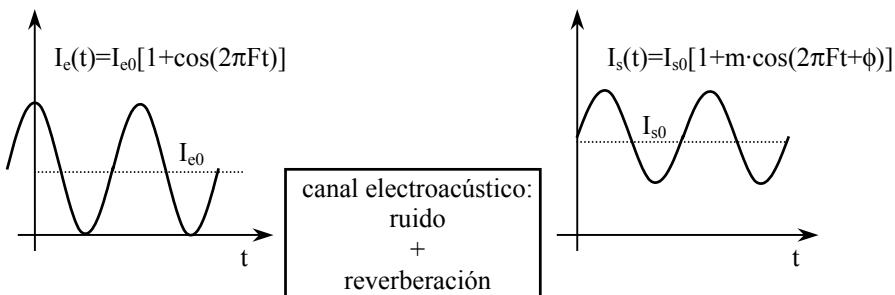


Figura 52. Efecto del canal de transmisión acústica en la pérdida de índice de modulación del habla, m .

En realidad, considerando la señal de voz, la modulación del habla en emisión no es la unidad sino algo menor. En cualquier caso el canal actúa como un reductor de la modulación de entrada, siendo la disminución relativa de modulación la que se define como índice de modulación del habla m . Este índice m es sólo achacable a dicho canal, e incluye los efectos del ruido, la reverberación y los ecos.

La pérdida de modulación m se puede medir fácilmente observando el espectro de baja frecuencia de la intensidad acústica o lo que es lo mismo el cuadrado de la señal conformada y procesada por el array microfónico, $y^2(t)$. En la práctica se consideran 98 valores m diferentes, 14 frecuencias de modulación por 7 moduladoras ($14F \times 7f$). Si se considera una determinada trama de habla, pueden calcularse las curvas MTF (*Modulation Transfer Function* o Función de Transferencia de Modulación), que consisten en expresar de forma gráfica los 98 valores de m . En la Figura 53 se representa un ejemplo de curvas MTF. Las curvas MTF tienen una gran utilidad, ya que permiten conocer no sólo que la inteligibilidad es baja, sino a qué frecuencia y por qué se produce este decremento de la calidad [Houtgast 85].

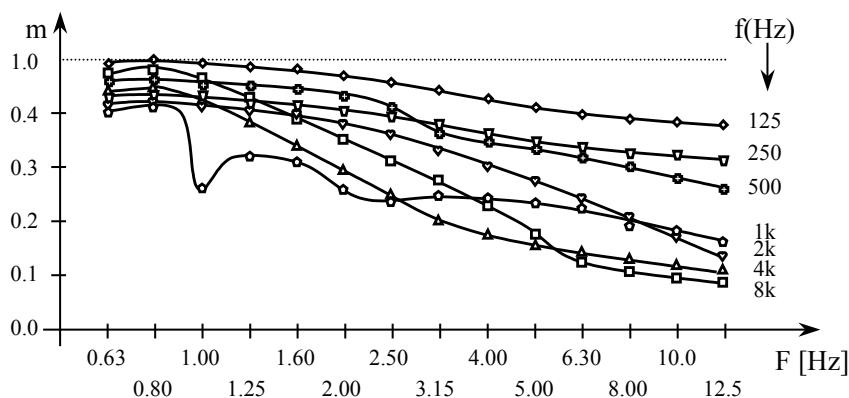


Figura 53. Ejemplo de curvas MTF con los 98 valores de m .

Aunque los 98 valores de m dan muchos detalles sobre la calidad de la señal de voz basada en las pérdidas de modulación, pueden constituir un exceso de información. En la

práctica todos esos índices se resumen en uno solo, llamado STI, y que está comprendido entre en el intervalo [0, 1]. El STI es un promedio ponderado (se da más valor a las frecuencias medias) de los 98 valores de m , adaptado al intervalo [0, 1]. A continuación se describe brevemente cómo se obtiene el índice STI.

5.4.2 Cálculo del STI

A partir de los 98 valores de m se obtiene el índice STI de la mediante el siguiente proceso.

- 1.- Cálculo de m en unidades logarítmicas, para obtener la llamada relación señal a ruido aparente: $\text{SNR}_{\text{ap}}(f_i, F_j)$, con $i = 1..7$ y $j = 1..14$. Para ello se emplea la siguiente fórmula:

$$\text{SNR}_{\text{ap}}(f_i, F_j) = 10 \log \frac{m}{1-m} \quad (256)$$

- 2.- Truncamiento de las 98 $\text{SNR}_{\text{ap}}(f_i, F_j)$ en el intervalo [+15dB, -15dB].

- 3.- Promediado lineal en F_i , obteniéndose 7 valores de $\text{SNR}_{\text{ap}}(f_i)$:

$$\text{SNR}_{\text{ap}}(f_i) = \frac{1}{14} \sum_{j=1}^{14} \text{SNR}_{\text{ap}}(f_i, F_j) \quad (257)$$

- 4.- Promediado ponderado en f_i , según unos coeficientes c_i de la Tabla 5.

$$\text{SNR}_{\text{ap}} = \sum_{i=1}^7 c_i \text{SNR}_{\text{ap}}(f_i) \quad (258)$$

$f_i(\text{Hz})$	i	c_i
125	1	0.13
250	2	0.14
500	3	0.11
1k	4	0.12
2k	5	0.19
4k	6	0.17
8k	7	0.14

Tabla 5. Coeficientes c_i para el cálculo del índice STI.

- 5.- La relación SNR_{ap} obtenida antes hay que escalarla en el intervalo [0, 1], mediante la siguiente expresión.

$$\text{STI} = \frac{\text{SNR}_{\text{ap}} + 15}{30} \quad (259)$$

que será finalmente el índice buscado.

En la práctica, para evaluaciones reales de inteligibilidad se emplea, más que el STI, el índice RASTI (*Rapid STI* o STI rápido) [Steeneken 85]. A continuación se describe brevemente el método de cálculo de dicho índice RASTI.

5.4.3 El índice RASTI

El índice STI necesita demasiados coeficientes para su cálculo, y en definitiva el resultado es un sólo valor que ha sido despojado de todos los detalles del origen y la causa de las pérdidas de inteligibilidad. El RASTI es una simplificación del STI que considera menos valores de m . Se tienen en cuenta menos frecuencias de modulación F y menos frecuencias de moduladora f , de tal manera que son seleccionadas aquéllas que más afectan a la inteligibilidad del habla. En total se consideran 5 moduladoras F para $f = 2\text{kHz}$ y 4 moduladoras F para $f = 500\text{Hz}$ (es decir 9 coeficientes m). En la Tabla 6 se indican cuáles son esas frecuencias.

f [Hz] \ F [Hz]	125	250	500	1k	2k	4k	8k
0.63							
0.80							
1.00							
1.25							
1.60							
2.00							
2.50							
3.15							
4.00							
5.00							
6.30							
8.00							
10.00							
12.50							

Tabla 6. Frecuencias elegidas para calcular el índice RASTI.

Ahora el proceso de cálculo es el mismo que para el STI, pero las 9 $\text{SNR}_{ap}(f_i, F_j)$ se promedian linealmente (sin ninguna ponderación dependiente de f_i) para obtener un índice que ahora es llamado RASTI.

Aunque el índice RASTI puede calcularse mediante el análisis de la señal $y^2(t)$ (equivalente a la intensidad acústica) a la salida del procesador, fue propuesto inicialmente para medirse “in situ” sobre la señal captada por un micrófono que previamente ha sido emitida en un determinado recinto con ruido y reverberación. En la práctica, el índice RASTI se obtiene entonces emitiendo en la sala bajo estudio una señal pregrabada, o señal RASTI, x_{RASTI} (véase la Figura 54) que contiene las 9 moduladoras F para 500Hz y 2kHz, y con una modulación m de partida ya conocida.

La señal y_{RASTI} captada en recepción, una vez ha sido modificada por el canal electroacústico (ruido + reverberación + altavoces) se filtra (véase la Figura 55) para obtener las dos bandas de 500Hz y 2kHz, se eleva al cuadrado, para obtener la intensidad y después de un filtrado paso bajo, se explora su espectro de baja frecuencia mediante una DFT o FFT. Los 9 índices de modulación m se calculan utilizando (255), mediante la comparación del valor de la intensidad $y^2(t)$ en cada una de las 9 frecuencias F , desde 0.7Hz hasta 11.2Hz, con el valor de continua I_0 , obtenido del espectro DFT.

El método RASTI de inteligibilidad consigue evaluaciones de la calidad de habla a la salida de un sistema electroacústico acordes con la apreciación subjetiva, cuando se utiliza como entrada del mismo la señal x_{RASTI} , que contiene las componentes de modulación del

habla a las frecuencias F precisas y en la cantidad adecuada para dicho método RASTI. Sin embargo, este método no se puede aplicar directamente sobre la señal de voz, ya que en general ésta no tendrá la misma composición espectral que presupone el procedimiento de cálculo del RASTI. En la parte 2 de esta Tesis (apartado 7.2), se propone un método RASTI alternativo para ser usado como evaluador objetivo, de aplicación sobre la señal de voz procesada por un array de micrófonos (método E-RASTI propuesto por el autor).

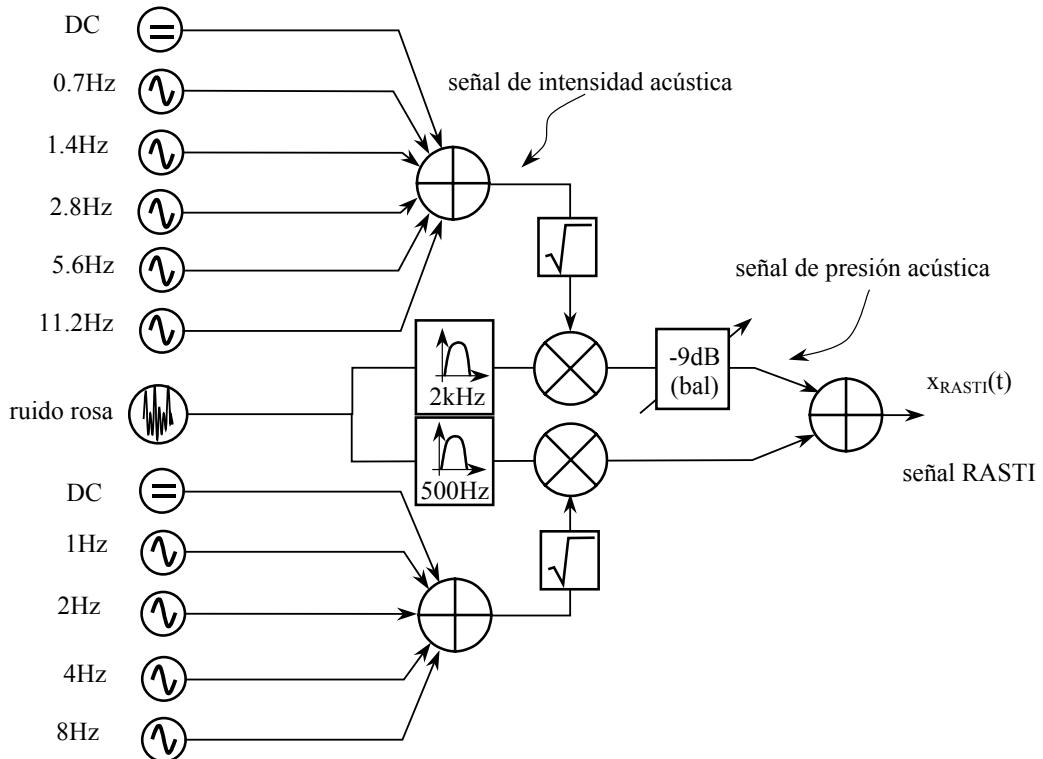


Figura 54. Generación de la señal RASTI.

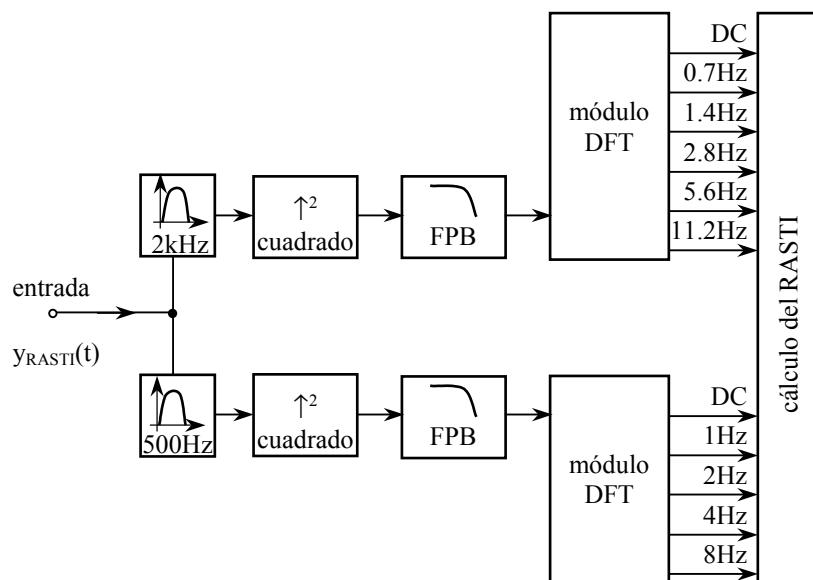


Figura 55. Cálculo del índice RASTI.

PARTE 2

***PROPUESTAS, EXPERIMENTOS Y
RESULTADOS PRELIMINARES***

6 ELEMENTOS PRELIMINARES DE TRABAJO

El punto de partida de los experimentos realizados en esta Tesis se sitúa en los trabajos con arrays microfónicos para reconocimiento robusto de locutor en malas condiciones acústicas, que son referidos en la Tesis Doctoral de Joaquín González Rodríguez [González-Rodríguez 99-a], director del presente trabajo.

A partir de ahí, antes de afrontar la implementación práctica de un array microfónico prototipo, como se tratará en la tercera parte de esta Tesis, es necesario realizar pruebas preliminares mediante simulaciones *software*. Esto implica necesariamente disponer de una base de datos con señal de voz multicanal que responda al modelo de array que se pretende realizar. A la hora de confeccionar una base de datos útil para este propósito se han considerado dos estrategias, por una parte utilizar una base de datos multicanal preexistente (“base de datos real”) y por otra simular una base de datos multicanal a partir de una grabación monocanal (“base de datos simulada”).

La base de datos real se ha confeccionado a partir de una base de datos más amplia –véase [CMU]–. Ésta ha sido realizada en la Universidad de Carnegie Mellon (EE.UU.) por Tom Sullivan y Richard Stern [Sullivan 96] y por eso se llama también aquí base de datos CMU (*Carnegie Mellon University*). Contiene grabaciones multicanal de señal de voz procedente de diferentes locutores y en diferentes situaciones acústicas de ruido y reverberación, y ha sido captada con un array microfónico de banda ancha de 15 canales, que responde al prototipo de array anidado expuesto en el punto 2.2.2 de esta Tesis.

La base de datos simulada, que recibirá el nombre de simCMU, se ha construido mediante la adición artificial, mediante un *software* propio de simulación acústica de recintos, de reverberación y ruido a un conjunto de archivos de voz monocanal en diferentes condiciones acústicas.

La elección de la base de datos real para la realización de los experimentos condiciona totalmente el tipo de array cuyo comportamiento va a ser investigado, ya que dicha base de datos está asociada a una configuración geométrica de partida de los micrófonos del array. En el caso de esta Tesis, siguiendo los trabajos expuestos en [González-Rodríguez 99-a] se decidió seguir trabajando con la propuesta de un array lineal anidado de [Sullivan 96], por lo que se utilizó CMU como base de datos real. Por otra parte hay que considerar que la base de datos de partida, tanto si es real como si es simulada, también condiciona el tipo de micrófono usado en el array, y sobre todo su directividad, que es la característica electroacústica que más importa al trabajo que aquí se refiere. Efectivamente, la directividad de cada micrófono influye de forma importante en la directividad total del array, que es un factor fundamental en los resultados de esta Tesis. Por lo tanto también habrá que considerar este aspecto.

A continuación se describen con más detalle los elementos iniciales con los que se contaba para realizar las pruebas preliminares reflejadas en esta Tesis. Éstos son básicamente las bases de datos utilizadas, tanto la real como la simulada, confeccionada ésta última por el autor, y adicionalmente el modelo de array que se deriva de las bases de datos anteriores. En

el caso del array, es necesario conocer su comportamiento teórico en cuanto a directividad y respuesta en frecuencia, para así poder extraer consecuencias más precisas sobre las propuestas de mejora de voz evaluadas con la configuración geométrica particular del array.

6.1 ARRAY ANIDADO DE 15 CANALES: CARACTERÍSTICAS

El array anidado con el que se grabó la base de datos real estaba dispuesto como se muestra en la Figura 56. Esta configuración será la utilizada en todas las pruebas preliminares y en la implementación del prototipo en tiempo real, como se verá más adelante. Existen tres grupos de siete micrófonos, cada uno de ellos constituye un array lineal uniforme que se encarga de captar una banda parcial del espectro total de voz. En la base de datos real se utilizaron micrófonos electret modelo Panasonic WD-063. Estos micrófonos son directivos [Sullivan 96] y están enfocados hacia la dirección *broadside* ($\theta_0 = 90^\circ$). Habrá que considerar la directividad conjunta de los micrófonos y el array (véase la Figura 66), ya que si los micrófonos son directivos hay una atenuación suplementaria a la proporcionada por el array ante sonidos incidentes lateralmente.

A continuación se especifican con más detalle las características directivas teóricas del array anidado de 15 canales equiponderados, utilizado en las pruebas preliminares. La banda de frecuencias de trabajo usada para dichas pruebas (y para todos los experimentos de esta Tesis) abarca desde 20Hz hasta 8kHz, que puede considerarse como de alta calidad para aplicaciones vocales. Esta banda viene determinada por la frecuencia de muestreo utilizada en la adquisición digital de señal, que será de $f_s = 16\text{kHz}$.

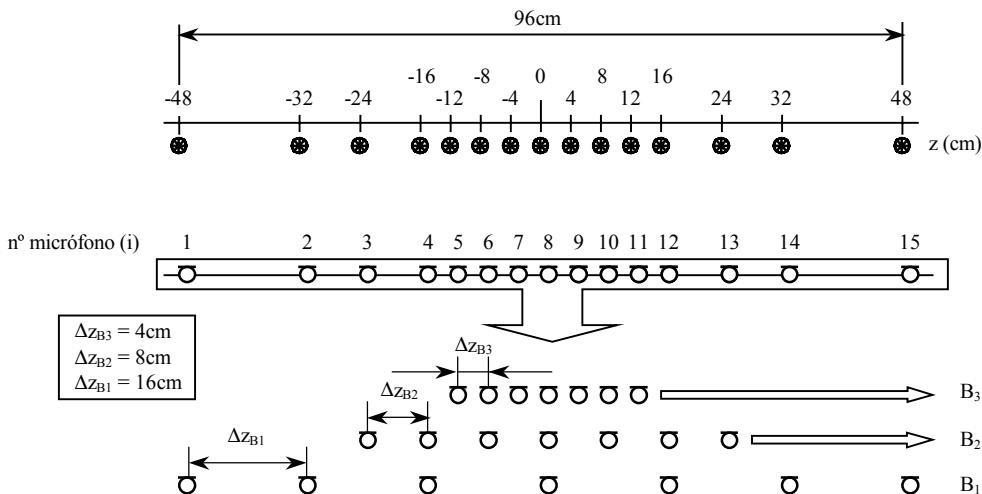


Figura 56. Array anidado. Posición de los micrófonos.

Se puede considerar por tanto que el array global con $I_T = 15$ micrófonos, se descompone en tres subarrays lineales uniformes, cada uno de ellos con $I = I_{B1} = I_{B2} = I_{B3} = 7$ micrófonos y correspondientes a tres subbandas de frecuencia que se llamarán respectivamente B_1 , B_2 y B_3 (Figura 57), determinadas por las dos frecuencias f_1 y f_2 . La banda de alta frecuencia es B_3 y está captada por el array menor, con $\Delta z_{B3} = 4\text{cm}$. Para la banda de frecuencias medias, B_2 se cumple que $\Delta z_{B2} = 8\text{cm}$ y para la de baja frecuencia $\Delta z_{B1} = 16\text{cm}$. La longitud total del array viene determinada por la subbanda de baja frecuencia siendo $\Delta z_{B1}(I_{B1}-1) = 96\text{cm}$. Esta configuración en octavas del espaciado intermicrofónico ofrece un ancho de banda teórico, en el cual la directividad puede ser considerada como

aproximadamente, constante de tres octavas, desde $f_1/2$ hasta $2f_2$, según se expuso en el capítulo 2.2.2 de esta Tesis.

Inicialmente la conformación de haz se hace mediante el método de filtrado y suma (o retardo y suma), como se muestra en la Figura 58, cuyos detalles ya se han expuesto en el punto 2.1.6 de esta Tesis. La red de filtros o retardos, representada por el vector $\mathbf{w} = \mathbf{w}_{RS}$, alinea temporalmente los canales microfónicos para enfocar el array a la dirección principal de apuntamiento o DOA dada por (r_0, θ_0) . Los filtros B_1 , B_2 y B_3 seleccionan las bandas de frecuencia de interés, para obtener finalmente mediante una suma, la señal conformada de salida.

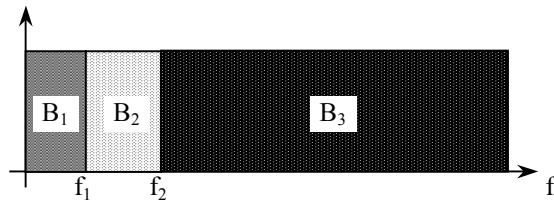


Figura 57. Subbandas consideradas en el array anidado de la Figura 56.

Atendiendo a la expresión (57), los retardos τ_i necesarios en cada canal para alinear temporalmente los micrófonos vendrán dados por:

$$\tau_i = \frac{r_0 - \sqrt{[r_0 \operatorname{sen}(\theta_0)]^2 + [r_0 \cos(\theta_0) - z_i]^2}}{c} \quad (260)$$

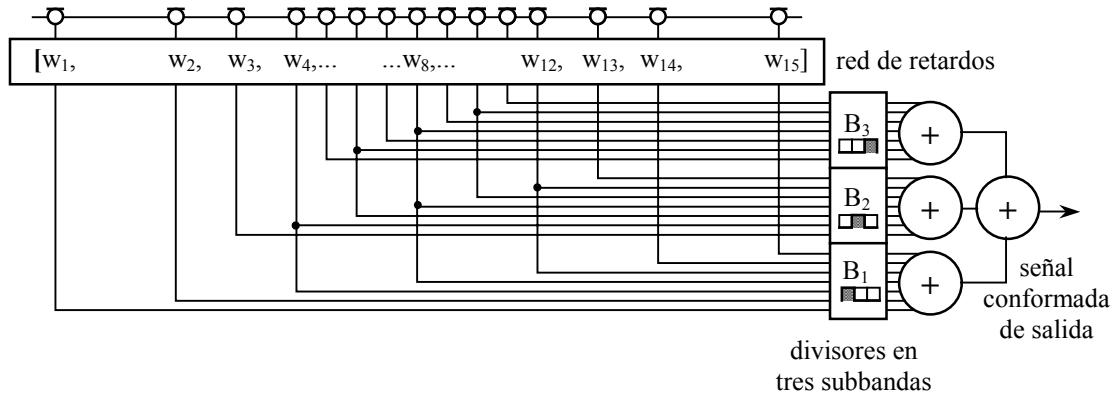


Figura 58. Estructura de retardo y suma para el array anidado de 15 micrófonos.

En la Tabla 7 se representan los retardos necesarios para apuntar al array a cuatro posiciones significativas. El retardo máximo a aplicar en la práctica corresponde a $|\tau_1| = |\tau_{15}| = 1.375\text{ms}$ para los micrófonos laterales en la configuración *endfire*. En la Tabla 8 se traducen esos retardos a unidades angulares y se calcula la corrección de fase $\Delta\phi_i [^\circ]$ necesaria para cada micrófono en dos frecuencias características, $f = 125\text{Hz}$ y $f = 4\text{kHz}$.

Si se admite que los micrófonos son omnidiireccionales (no es el caso pero simplifica los cálculos teóricos), la directividad del array vendrá determinada únicamente por la geometría del mismo, considerada en la Figura 56. Según se comentó en el punto 2.1.8 de esta Tesis, dos parámetros muy importantes determinan la directividad de cada una de las subbandas, por supuesto la posición de apuntamiento (r_0, θ_0) , que en la aproximación de campo lejano se

sustituye por θ_0 únicamente, y también el periodo de muestreo espacial Δz o de forma equivalente la pulsación de muestreo espacial Ω_s dada en (69).

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
z _i [cm]	-48	-32	-24	-16	-12	-8	-4	0	4	8	12	16	24	32	48	
r ₀ [m]	θ ₀ [°]	retardo, τ _i [ms]														
1	90	-0.313	-0.143	-0.081	-0.036	-0.021	-0.009	-0.002	0.000	-0.002	-0.009	-0.021	-0.036	-0.081	-0.143	-0.313
1	0	-1.375	-0.917	-0.688	-0.458	-0.344	-0.229	-0.115	0.000	0.115	0.229	0.344	0.458	0.688	0.917	1.375
10	90	-0.033	-0.015	-0.008	-0.004	-0.002	-0.001	0.000	0.000	0.000	-0.001	-0.002	-0.004	-0.008	-0.015	-0.033
10	0	-1.375	-0.917	-0.688	-0.458	-0.344	-0.229	-0.115	0.000	0.115	0.229	0.344	0.458	0.688	0.917	1.375

Tabla 7. Retardos que hay que aplicar a cada uno de los micrófonos del array para cuatro posiciones de apuntamiento, que comprenden el apuntamiento *broadside* y *endfire* a distancias de 1m y 10m.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
z _i [cm]	-48	-32	-24	-16	-12	-8	-4	0	4	8	12	16	24	32	48		
f [Hz]	r ₀ [m]	θ ₀ [°]	desfase, ΔΦ _i [°]														
125	10	90	-0.7	-0.3	-0.2	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.2	-0.3	-0.7	
	10	0	-30.9	-20.6	-15.5	-10.3	-7.7	-5.2	-2.6	0.0	2.6	5.2	7.7	10.3	15.5	20.6	30.9
4000	10	90	-23.8	-10.6	-5.9	-2.6	-1.5	-0.7	-0.2	0.0	-0.2	-0.7	-1.5	-2.6	-5.9	-10.6	-23.8
	10	0	-990.3	-660.2	-495.1	-330.1	-247.6	-165.0	-82.5	0.0	82.5	165.0	247.6	330.1	495.1	660.2	990.3

Tabla 8. Desfases ΔΦ_i [°] que hay que aplicar a cada uno de los micrófonos del array según dos posiciones de apuntamiento (*broadside* y *endfire* para r₀=10m) calculados en dos frecuencias, f=125Hz y f=4kHz. Se sombrean los desfases no aplicados por no utilizarse el micrófono i para la frecuencia correspondiente.

A la hora de configurar el array lineal anidado de 15 micrófonos interesa evitar el *aliasing* espacial. Según (71) se tendrá *aliasing* espacial para frecuencias mayores a f_a (frecuencia de *aliasing*) siendo:

$$f_a = \frac{c}{\Delta z} \Rightarrow f_{aB_1} = 2181\text{Hz}, f_{aB_2} = 4363\text{Hz}, f_{aB_3} = 8725\text{Hz} \text{ configuración } broadside \quad (261)$$

$$f_a = \frac{c}{2\Delta z} \Rightarrow f_{aB_1} = 1091\text{Hz}, f_{aB_2} = 2181\text{Hz}, f_{aB_3} = 4363\text{Hz} \text{ configuración } end-fire$$

(se ha supuesto c = 349ms⁻¹, para más detalles véase el punto 9.2.2 de la Tesis). Se aprecia que la configuración *endfire* tiene mayor tendencia al *aliasing* espacial. Hay que considerar que si el máximo objetivo perseguido a la hora de configurar el apuntamiento del array es conseguir una buena selectividad espacial (o directividad), la idoneidad del array viene determinada por varios factores, que se relacionan a continuación.

1.- El criterio de resolución de Rayleigh establece la existencia de al menos un mínimo de captación del array (nulo en un array ideal) en el intervalo $\theta \in [0, \pi]$ y ya fue establecido en (77) y (78). Si se llama f_{nulo} a la frecuencia del primer nulo de directividad, que puede ser asociada a aquélla a partir de la cual el array es suficientemente directivo, entonces

$$f_{nulo} = \frac{c}{I\Delta z} \Rightarrow f_{nuloB_1} = 306\text{Hz}, f_{nuloB_2} = 613\text{Hz}, f_{nuloB_3} = 1225\text{Hz} \text{ para } broadside \quad (262)$$

$$f_{nulo} = \frac{c}{2I\Delta z} \Rightarrow f_{nuloB_1} = 153\text{Hz}, f_{nuloB_2} = 306\text{Hz}, f_{nuloB_3} = 613\text{Hz} \text{ para } endfire$$

A la vista de (262) la configuración *broadside* es menos selectiva que la configuración *endfire* puesto que los primeros nulos de captación se sitúan a mayor frecuencia.

2.- La anchura angular del lóbulo principal determina la selectividad espacial del array. Para evaluarla en los dos casos particulares de apuntamiento de $\theta_0 = 90^\circ$ y $\theta_0 = 0^\circ$, se usan las expresiones (83) y (84). Así, según (83), el ancho del lóbulo principal a la frecuencia de 500Hz vale:

$$\Delta\theta_{\text{bsB}_1} (f = 500\text{Hz}) = 1.3\text{rad} \equiv 77^\circ \text{ para } \textit{broadside} (\theta_0 = 90^\circ) \quad (263)$$

y según (84)

$$\Delta\theta_{\text{efB}_1} (f = 500\text{Hz}) = 2.4\text{rad} \equiv 136^\circ \text{ para } \textit{endfire} (\theta_0 = 0^\circ) \quad (264)$$

Puede comprobarse que la configuración *endfire* produce lóbulos de captación más anchos que la configuración *broadside*, aunque por contra, y debido a la simetría de revolución del array alrededor de $\theta = 0^\circ$, esta última posee un lóbulo de captación gemelo al principal en $\theta = 180^\circ$, lo cual puede ser muy perjudicial a la hora de captar fuentes de sonido ajenas a la deseada. En la Figura 59 se representa el mapa de directividad de cada uno de los tres subarrays, en configuraciones *broadside* y *endfire*, donde se pueden comprobar todos los resultados anteriores.

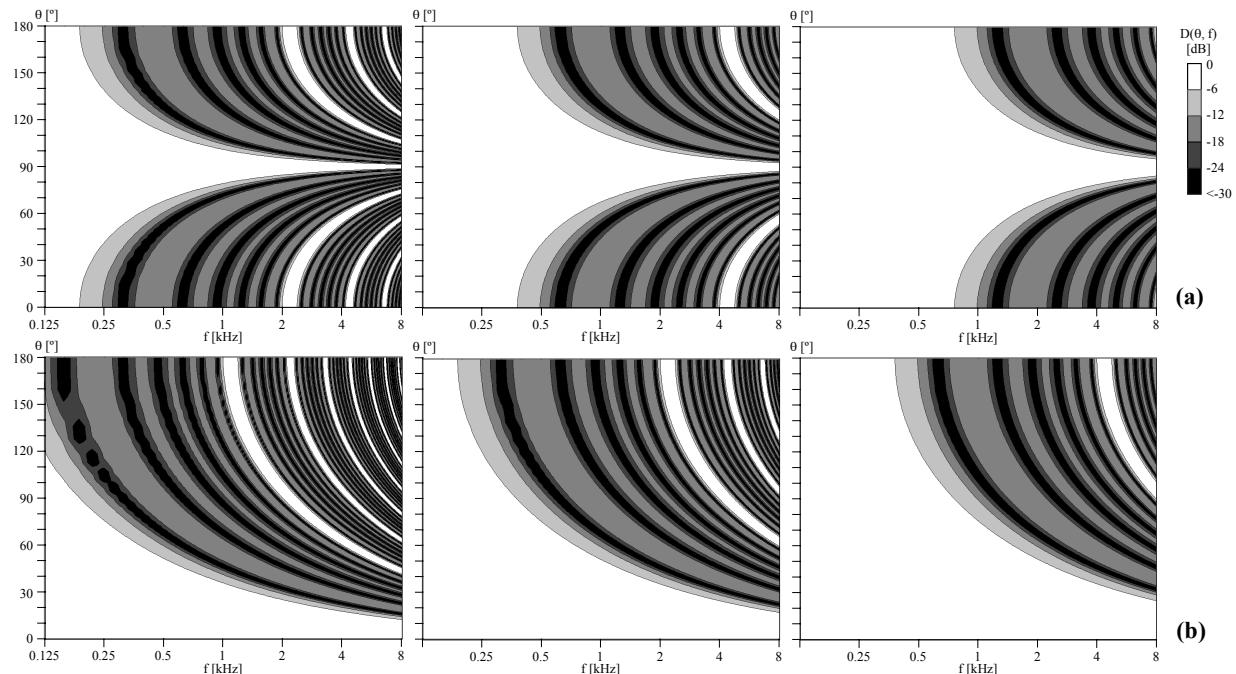


Figura 59. Mapa de directividad $D(\theta, f)$ [dB] de cada uno de los subarrays de 7 micrófonos correspondientes al array anidado de 15 micrófonos, en la aproximación de campo lejano. **(a)** Para apuntamiento *broadside* ($\theta_0=90^\circ$). De izquierda a derecha, bandas B_1 , B_2 y B_3 . **(b)** *Endfire* ($\theta_0=0^\circ$). Bandas B_1 , B_2 y B_3 .

En la práctica, los resultados obtenidos en esta Tesis, tanto en los experimentos preliminares de la Parte 2 como en los de la Parte 3 realizados con un prototipo real, han utilizado una división en tres bandas dada por $f_1 = 1\text{kHz}$ y $f_2 = 2\text{kHz}$. De esta manera, manteniendo los criterios de directividad en un array anidado, se puede considerar la banda de 500Hz a 4kHz como aquélla en la que la directividad se mantiene como aproximadamente constante. En la Figura 60 se representa el mapa de directividad global del array anidado de 15 canales para diferentes ángulos de apuntamiento, con las dos frecuencias de separación indicadas antes. Puede comprobarse cómo en la configuración *endfire* aparece *aliasing* espacial por encima de 4kHz pero mejora la selectividad espacial en baja frecuencia, por

debajo de 250Hz. Los saltos bruscos producidos a las frecuencias f_1 y f_2 se pueden corregir fácilmente utilizando filtros limitadores para las bandas B_1 , B_2 y B_3 que sean de transición suave. En las Figuras 61, 62 y 63 se representan detalladamente las curvas de directividad $D(\theta)$ [dB] en unidades logarítmicas y en bandas de 1/3 de octava, del array anidado para tres apuntamientos diferentes, $\theta_0 = 0^\circ$, $\theta_0 = 45^\circ$ y $\theta_0 = 90^\circ$. En la Figura 64 se representa esta misma directividad, en la banda 500Hz - 1kHz, considerada de directividad constante, para los tres ángulos de apuntamiento anteriores.

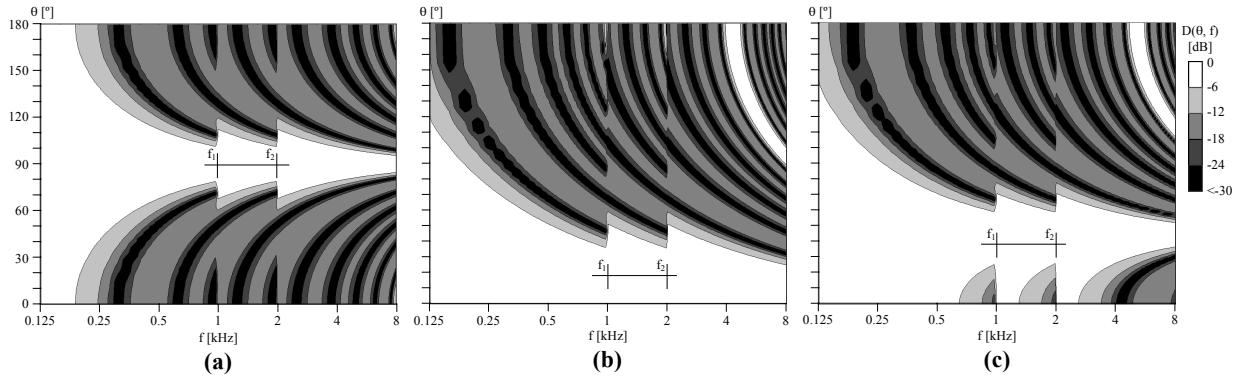


Figura 60. Mapa de directividad $D(\theta, f)$ [dB] del array anidado de 15 micrófonos en la aproximación de campo lejano con $f_1=1\text{kHz}$ y $f_2=2\text{kHz}$. **(a)** *Broadside* ($\theta_0=90^\circ$). **(b)** *Endfire* ($\theta_0=0^\circ$). **(c)** Apuntamiento lateral ($\theta_0=45^\circ$).

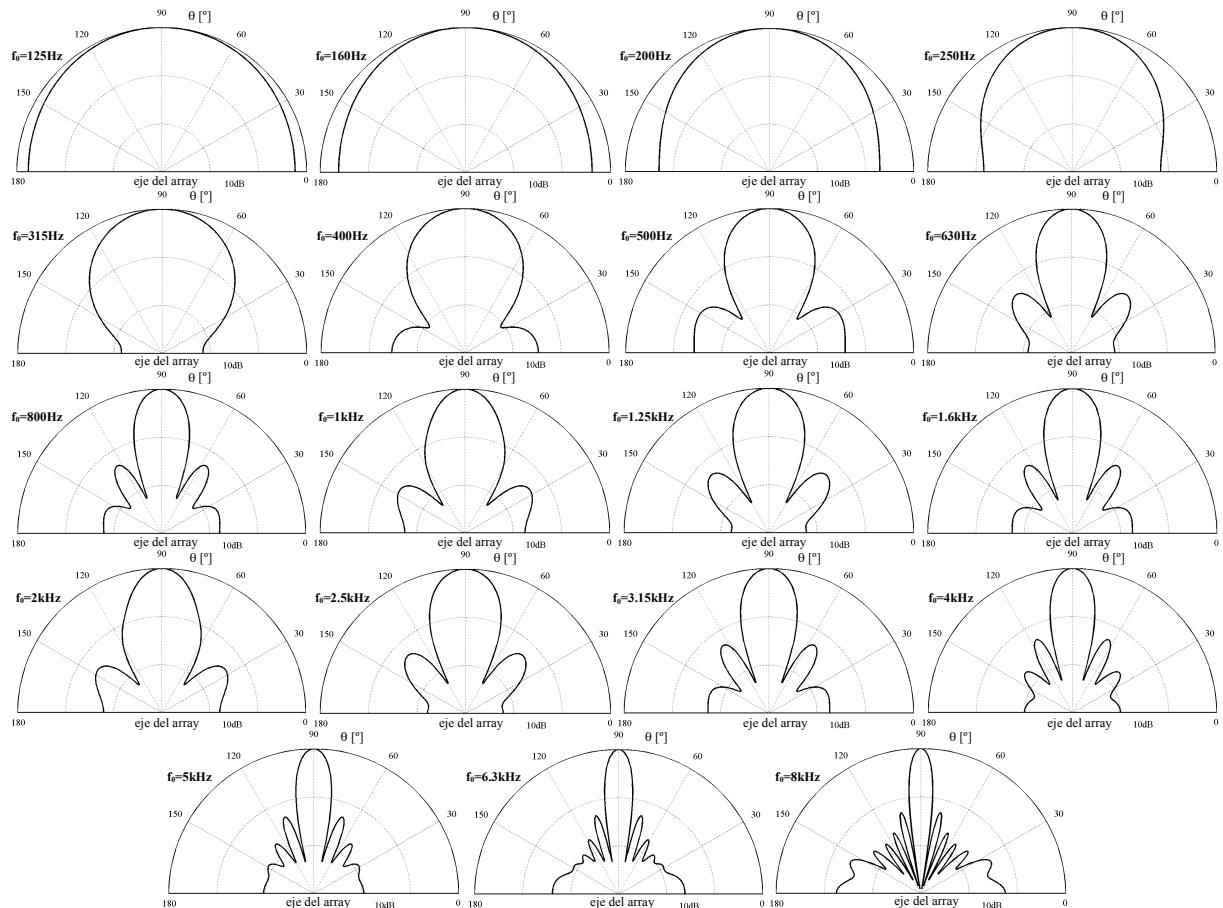


Figura 61. Curvas polares de directividad $D(\theta)$ [dB] para el array anidado de 15 micrófonos en bandas de 1/3 de octava. Apuntamiento *broadside* ($\theta_0=90^\circ$). Se ha usado la aproximación de campo lejano.

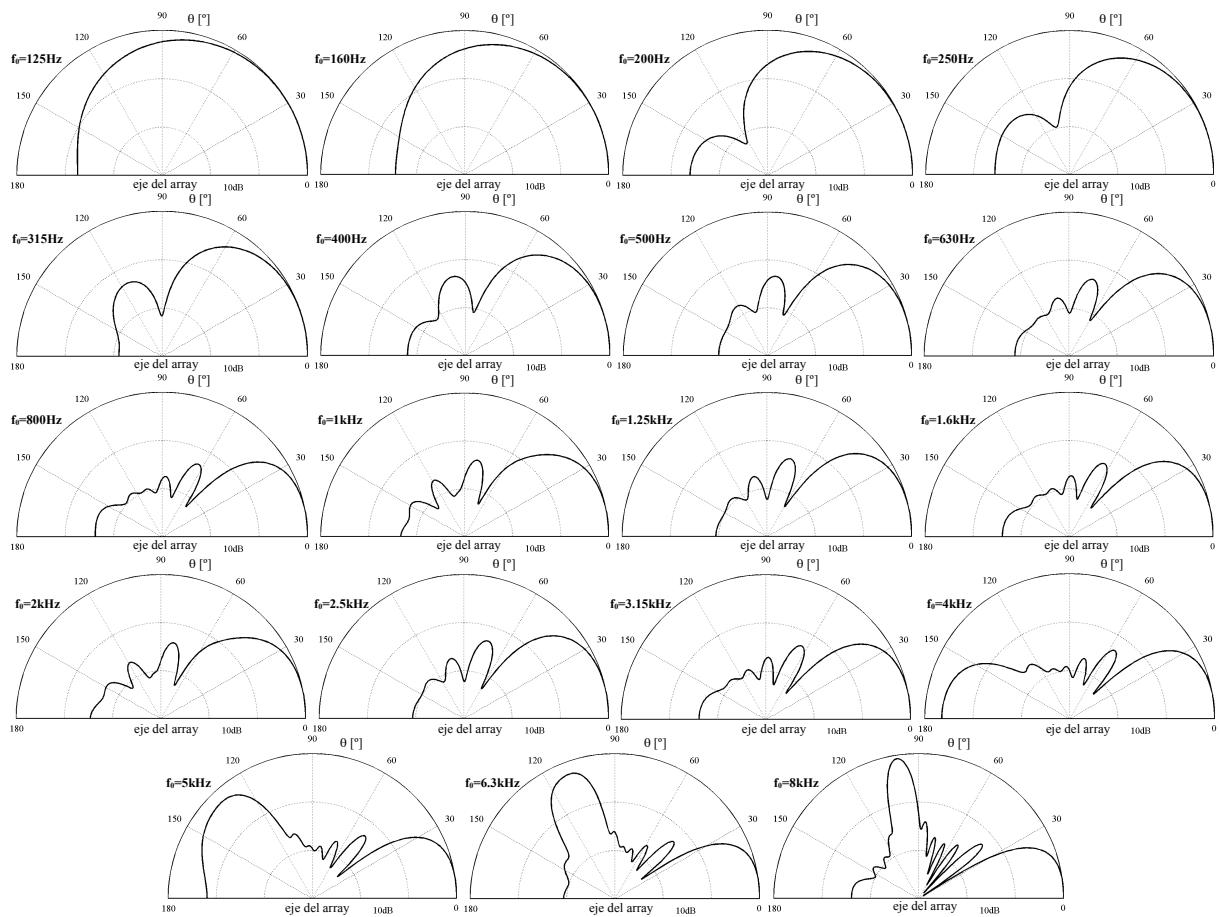


Figura 62. Curvas polares de directividad $D(\theta)$ [dB] para el array anidado de 15 micrófonos en bandas de 1/3 de octava. Apuntamiento *endfire* ($\theta_0=0^\circ$). Se ha usado la aproximación de campo lejano.

Por otra parte siempre se tiene que tener presente que la directividad del array anidado se representa por una superficie espacial, que en términos generales depende de las coordenadas angulares θ y φ , es decir $D = D(\theta, \varphi)$. Eso quiere decir que, aun en el caso de que haya simetría de revolución respecto al eje **Z** (eje del array) cuando los micrófonos sean omnidiireccionales, debe considerarse la captación tridimensional del array, puesto que dos curvas $D(\theta)$ muy similares pueden dar lugar a dos figuras de captación totalmente diferentes dependiendo del ángulo de apuntamiento θ_0 . En efecto, en la Figura 65 se representa la curva de directividad espacial para tres apuntamientos distintos con el array equiponderado de 15 micrófonos omnidiireccionales en la banda de directividad constante. Puede verse que los apuntamientos distintos al *endfire* producen figuras de captación poco convenientes, puesto que no atenúan en ninguna medida las fuentes sonoras procedentes de ciertas direcciones laterales. Este inconveniente se puede solucionar utilizando micrófonos directivos con los ejes de captación alineados hacia una dirección próxima al apuntamiento electrónico. En ese caso es de aplicación el teorema del producto de las directividades, según se expresó en (62), que establece que la directividad conjunta $D(\theta, \varphi)$ de un array compuesto por micrófonos iguales y alineados, es el producto de la directividad del array –también factor de array $D_{FA}(\theta, \varphi)$ – multiplicada por la directividad de los micrófonos $D_\mu(\theta, \varphi)$. Para mostrar el efecto de la directividad de los micrófonos sobre el array, en la Figura 66 se representa la directividad del array anidado apuntado hacia $\theta_0 = 90^\circ$ cuando los micrófonos son cardioideos y paralelos al eje **X**. Puede verse cómo la directividad de dichos micrófonos modula la directividad del array, de tal manera que se atenúa la captación lateral. Esto puede ser una ventaja, aunque en la

práctica hay que considerar la dificultad de encontrar micrófonos directivos con iguales características (micrófonos pareados), mantenidas en un amplio margen de frecuencias. Debido a esto último, si se opta por elegir micrófonos directivos para configurar el array, habrá que contar con que la directividad real medida con el array se apartará más de la predicción teórica. Este problema no existe con los micrófonos omnidireccionales ya que en la práctica es fácil que éstos tengan características similares en el margen de frecuencias de la voz.

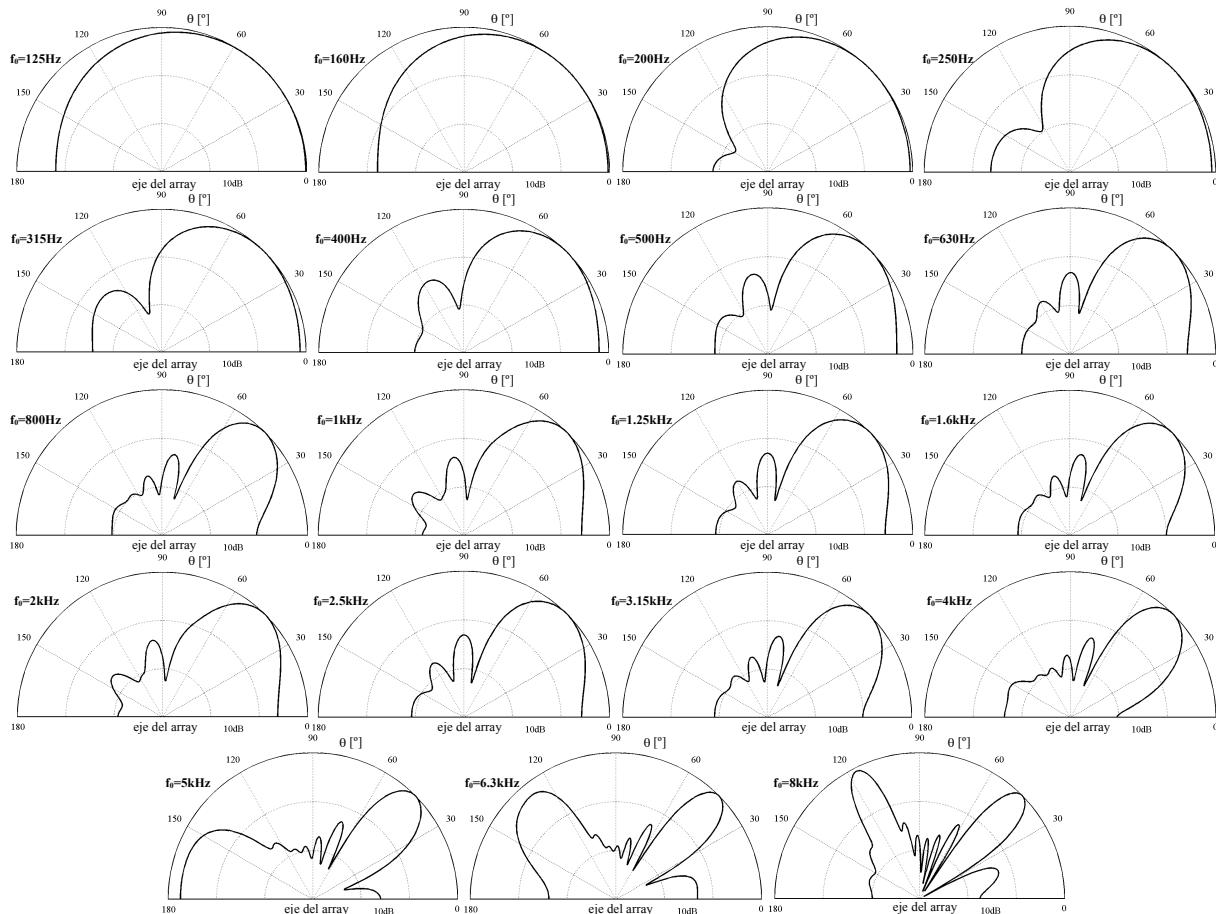


Figura 63. Curvas polares de directividad $D(\theta)$ [dB] para el array anidado de 15 micrófonos en bandas de $1/3$ de octava. Apuntamiento lateral ($\theta_0=45^\circ$). Se ha usado la aproximación de campo lejano.

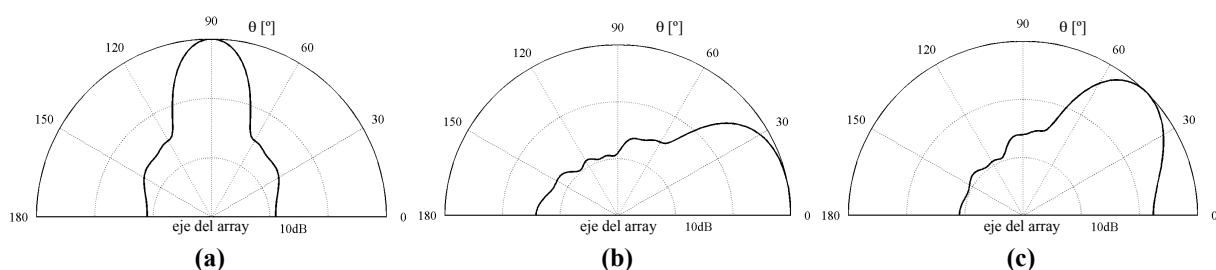


Figura 64. Curvas polares de directividad $D(\theta)$ [dB] del array anidado de 15 micrófonos, considerando el promedio de la banda 500Hz-4kHz, en la que se considera la directividad como aproximadamente constante. Compárese esta figura con el mapa de la Figura 60. Aproximación de campo lejano. (a) Broadside ($\theta_0=90^\circ$). (b) Endfire ($\theta_0=0^\circ$). (c) Lateral ($\theta_0=45^\circ$).

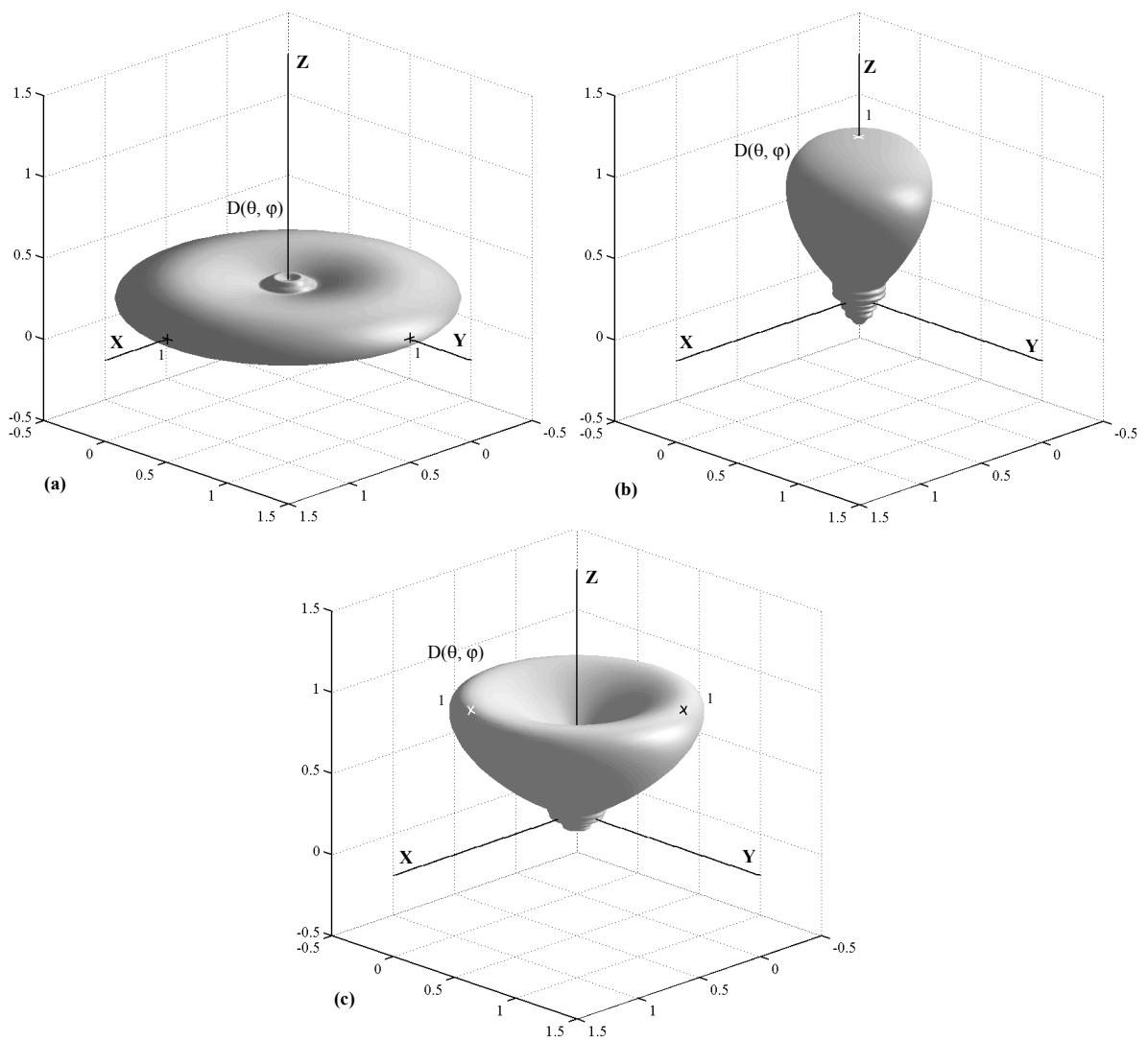


Figura 65. Directividad $D(\theta, \phi)$ (en unidades relativas) del array anidado de 15 micrófonos, mostrando el promedio de la banda 500Hz-4kHz, en la que se considera la directividad como aproximadamente constante. Aproximación de campo lejano. **(a)** Broadsire ($\theta_0=90^\circ$). **(b)** Endfire ($\theta_0=0^\circ$). **(c)** Lateral ($\theta_0=45^\circ$).

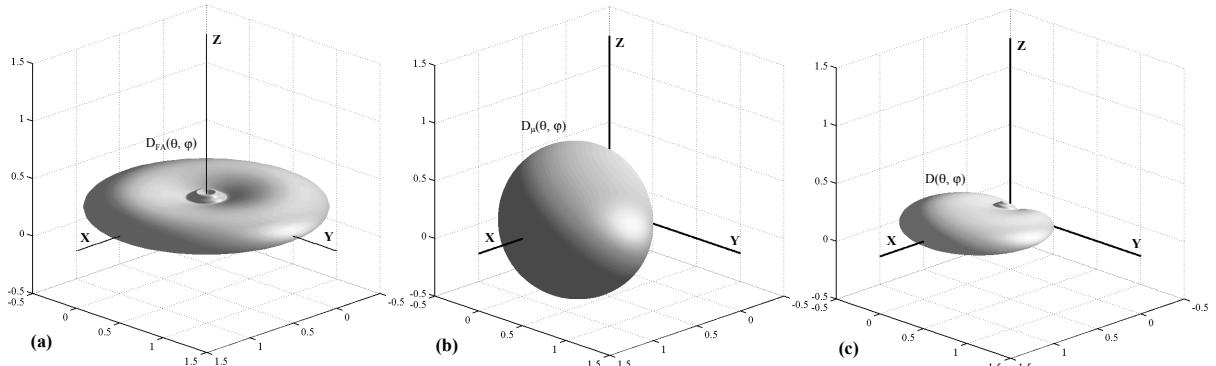


Figura 66. Directividad $D(\theta, \phi)$ (en unidades relativas) del array anidado de 15 micrófonos, promediado de la banda 500Hz-4kHz y apuntado en Broadsire ($\theta_0=90^\circ$). Se ha considerado que los micrófonos son cardioideos, con sus ejes paralelos y apuntando al eje X positivo. **(a)** Directividad del factor de array $D_{FA}(\theta, \phi)$. **(b)** Directividad de los micrófonos $D_\mu(\theta, \phi)$. **(c)** Directividad conjunta según el teorema del producto (62), $D(\theta, \phi)=D_\mu(\theta, \phi) D_{FA}(\theta, \phi)$.

3.- Factor de directividad máximo Q_{MAX} (53) –o índice de directividad máximo DI_{MAX} (55)–. Según se expone en el punto 2.1.7 de esta Tesis, el factor de directividad máximo da una idea global del rechazo de energía por parte del array en las direcciones no coincidentes con la DOA principal. Por lo tanto ofrece una buena valoración integral de la selectividad espacial del array, aunque no da ningún detalle de la forma de la curva $D(\theta, \varphi)$. Cuanto mayor sea DI_{MAX} más selectivo será el array y provocará más atenuación del ruido y la reverberación. En la Figura 67 se representa el índice de directividad máximo para los tres apuntamientos referidos anteriormente. Puede verse cómo la directividad se mantiene aproximadamente constante en el intervalo 500Hz - 4kHz, con los saltos de DI_{MAX} a f_1 y f_2 debidos al crecimiento de la directividad dentro de cada subarray. Estas discontinuidades se pueden suavizar mediante un filtrado adecuado de las tres subbandas B_1 , B_2 y B_3 , como de explicó en el apartado 2.2.2 de la Tesis (Figura 15) dedicado a los arrays anidados. Continuando con el análisis de la Figura 67, puede apreciarse la baja selectividad del array anidado en baja frecuencia y los defectos producidos en alta frecuencia por el *aliasing* espacial.

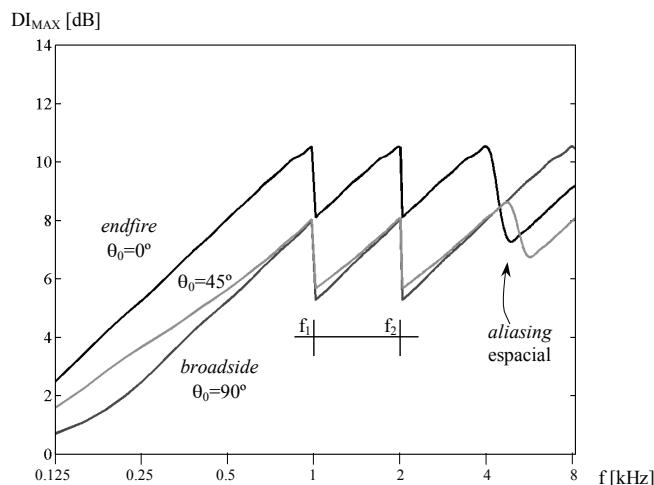


Figura 67. Índice de directividad máximo DI_{MAX} en función de la frecuencia para el array anidado de 15 micrófonos en las configuraciones *broadside* ($\theta_0=90^\circ$), *endfire* ($\theta_0=0^\circ$) y apuntamiento lateral ($\theta_0=45^\circ$). Se ha usado la aproximación de campo lejano.

La configuración *endfire* es la más directiva atendiendo al factor DI_{MAX} aunque tenga un lóbulo frontal más ancho, con lo que en principio se puede considerar como la más apropiada para implementar un array con micrófonos omnidiireccionales, ya que será la que menor captación global de ruido y reverberación tenga.

6.2 BASE DE DATOS REAL CMU

La base de datos real, utilizada en las pruebas preliminares pertenece como ya se ha dicho a una más extensa grabada por Tom Sullivan y Richard Stern en la Universidad de Carnegie Mellon. El subcorpus utilizado aquí consiste en una selección de 15 locuciones multicanal grabadas con un array anidado de 15 canales al que se ha hecho referencia en el punto anterior (Figura 56). Cada locución incorpora además la grabación limpia monocanal correspondiente a la señal multicanal. Ésta ha sido captada con un micrófono modelo

Sennheiser HMD-414, que corresponde a un micrófono de proximidad de “tipo casco” y directividad supercardioide. Esta señal se puede utilizar como referencia en cada una de las pruebas ya que contiene la voz libre de ruido y reverberación, aunque para ello es necesario alinearla temporalmente con la señal captada por el array.

La base de datos seleccionada contiene archivos de señal de audio con una frecuencia de muestreo $f_s = 16\text{kHz}$ y una resolución de 16bits. En la Tabla 9 se detallan las características más importantes de los fragmentos de voz multicanal utilizados como base de datos real. Aunque se ha mantenido en cierta medida la nomenclatura usada en [Sullivan 96] se ha resumido ésta en lo posible, puesto que aquí sólo se utilizará una parte pequeña de la base de datos multicanal total, mucho más extensa.

identificador de subcorpus	Nº de archivos usados	distancia, $r_0 [\text{m}]$	ruido	reverberación
arr4A	5	1	alto	baja
arrC1A	5	1	medio	media
arrC3A	5	3	medio	alta

Tabla 9. Base de datos real multicanal utilizada (CMU) y características más importantes de cada subcorpus.

Las locuciones de la Tabla 9 son de habla inglesa y pertenecen al mismo locutor, Tom Sullivan. Todas fueron grabadas en configuración *broadside*, es decir con $\theta_0 = 90^\circ$, con lo que la Figura 66 puede dar una idea gráfica de cuál es la directividad del array en la banda de 500Hz - 4kHz. Tomando como referencia la Tabla 9, a continuación se describen con más detalle las características de cada uno de los tres subcorpus utilizados.

- a.- Subcorpus **arr4A**: corresponde a cinco grabaciones con el array anidado de 15 micrófonos en un laboratorio ruidoso y con poca reverberación, con el orador situado a $r_0 = 1\text{m}$ del centro del array (micrófono 8 de la Figura 56).
- b.- Subcorpus **arrC1A**: como en el caso anterior, pero ahora el lugar de grabación corresponde a una sala de reuniones con mayor volumen y por lo tanto mayor reverberación. El ruido aditivo de fondo es menor aquí. El orador se sitúa también a $r_0 = 1\text{m}$ del centro del array.
- c.- Subcorpus **arrC3A**: las condiciones son idénticas al subcorpus **arrC1A** pero ahora el orador está a $r_0 = 3\text{m}$ del centro del array. Esto hace que la reverberación y ruido captados sean mayores que en el caso anterior.

A la vista de las características de la base de datos real usada, ésta se puede considerar como apropiada para las hacer pruebas preliminares de limpieza de voz en condiciones de ruido alto (subcorpus **arr4A**) y de reverberación media-alta (subcorpora **arrC1** y **arrC3A**). Lógicamente la reverberación en los fragmentos **arrC1** es menor que en **arrC3** (aunque la sala sea la misma) porque la distancia de la fuente al micrófono es menor. La base de datos real, sin embargo, está limitada porque todos sus componentes fueron grabados con el orador en posición *broadside*, por lo que el autor ha considerado la realización, a partir de la base CMU, de una nueva base de datos para las pruebas preliminares, que amplíe la casuística en cuanto a la geometría entre el array y el locutor, lo cual se describe a continuación.

6.3 BASE DE DATOS SIMULADA: simCMU-1 y simCMU-2

La base de datos CMU resultaba insuficiente para hacer pruebas exhaustivas de procesado en array. Por ello se implementaron dos bases de datos alternativas diseñadas a partir de la primera. Son las bases de datos simCMU-1 y simCMU-2, ambas simuladas a partir de una señal de voz monocanal. Para confeccionar las bases de datos simuladas se utilizaron los fragmentos de voz limpia o de referencia correspondientes a la base de datos real, y que como ya se ha dicho fueron grabados con un micrófono directivo de proximidad, situado cerca del orador.

La base de datos simCMU-1 es un extracto de una mayor (simCMU-Db) descrita en [González-Rodríguez 99-a] y en [Chamorro 99]. Contiene señales multicanal pertenecientes a al array anidado de 15 micrófonos situado en un recinto de $6m \times 5m \times 3m$ ($90m^3$) con diferentes tiempos de reverberación y a diferentes distancias de la fuente y considerando el orador en posición *broadside* con respecto al array. Se ha generado convolucionando las señales de referencia de la base CMU con las respuestas al impulso del recinto, obtenidas mediante el método de imágenes [Allen 77-b]. La siguiente tabla muestra las características de los subcorpora usados para la base de datos simCMU-1. Los cinco archivos de los que consta cada subcorpus han sido generados con la señal de referencia correspondiente a cada uno de los cinco integrantes de cada subcorpus de la base de datos CMU.

identificador de subcorpus	Nº de archivos usados	distancia, r_0 [m]	T_{60} [s]
A0tf1s1	5	1	0.1
A0tf1s3	5	1	0.3
A0tf1s8	5	1	0.8
A0tf3s1	5	3	0.1
A0tf3s3	5	3	0.3
A0tf3s8	5	3	0.8

Tabla 10. Base de datos simulada multicanal simCMU-1. Características más importantes de cada subcorpus.

La base simCMU-1 contiene archivos multicanal sin ruido añadido. Es decir sólo se considera contaminación por reverberación ya que la señal de referencia con la que fue generada estaba libre de ruido. Para añadir ruido se hizo un primer intento a partir de la base de datos anterior, generando la variante simCMU-1.1 que contiene los elementos de simCMU-1 con ruido $n(t)$ añadido en los micrófonos del array. Este ruido puede ser de dos tipos, ruido aleatorio o ruido determinista. El ruido aleatorio consiste en ruido blanco gaussiano y el ruido determinista en una señal diente de sierra de frecuencia $f_N = 500Hz$ (descrita en la Figura 72). El ruido se añadió directamente, sumándolo a cada uno de los canales del array, con un nivel tal que cumpla unas condiciones de relación señal a ruido determinadas en el canal central del array (SNR_8). En el caso de la señal determinista, lógicamente se aplicaron los retardos correspondientes desde el punto de emisión hasta cada uno de los micrófonos del array (Figura 68), simulando una fuente de ruido en una posición determinada del campo acústico. Por lo anterior, simCMU-1.1 no considera la reverberación del ruido, con lo que se quita realismo a la simulación. En la Tabla 11 se expresan las condiciones de ruido añadido en simCMU-1.1.

La única reverberación entonces considerada en simCMU-1.1 es la conseguida inicialmente por el recinto de $90m^3$ (como máximo 0.8s), y asociada sólo a la señal (no al

ruido). Este recinto es demasiado pequeño y por ello muy seco acústicamente. Por ello fue necesario ampliar la base de datos anterior añadiendo ruido de diferentes características y una reverberación mayor, mediante la simulación en un recinto de mayor volumen. La consecuencia es la base de datos simCMU-2, obtenida por el autor de esta Tesis.

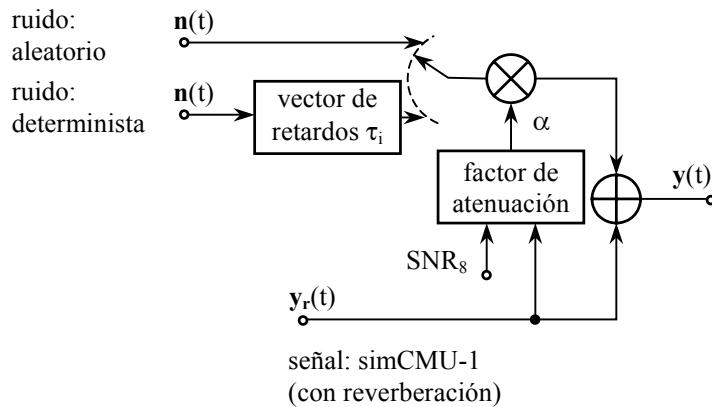


Figura 68. Proceso de generación de la base de datos simCMU-1.1 a partir de simCMU-1.

identificador de sucorpus	tipo de ruido	Nº archivos usados	r_N [m]	θ_N [°]	SNR_8 [dB]
dif 10	blanco	30	-	-	10
dif 15	blanco	30	-	-	15
dif 20	blanco	30	-	-	20
c 45 10 10	d. sierra	30	10	45	10
c 45 10 15	d. sierra	30	10	45	15
c 45 10 20	d. sierra	30	10	45	20

Tabla 11. Condiciones de ruido añadido para generar simCMU-1.1 a partir de simCMU-1. Los parámetros r_N y θ_N son respectivamente la distancia al centro y el ángulo respecto del array en los que se situó la fuente de ruido. Los 30 archivos de cada subcorpus corresponden al total de simCMU-1 referido en la Tabla 10.

La base de datos simCMU-2 se ha diseñado mediante simulación en un recinto sencillo, descrito en la Figura 69. El recinto tiene dimensiones 12m x 10m x 3.5m ($420m^3$). El array se sitúa a una altura de 2m del suelo, con su eje paralelo a la pared trasera del recinto, a 4m de la pared izquierda y a 4m de la pared trasera. Los coeficientes de reflexión sonora de todas las paredes son los mismos, con un valor de $\beta = 0.83$. Estas condiciones dan como resultado un tiempo de reverberación según Sabine de $T_{60} = 1s$ o según Eyring de $T_{60} = 0.92s$ –para más detalles sobre acústica de recintos se puede consultar [Kuttruff 91]–. Se han situado 8 puntos de emisión, del P₁ al P₈, colocados a la misma altura sobre el suelo que el array y a diferentes distancias y ángulos del mismo (Tabla 12).

Se ha obtenido la respuesta al impulso punto-array (15 respuestas al impulso, una por cada micrófono del array) para cada uno de los ocho puntos considerados – $h_{P1}(t), \dots, h_{P7}(t)$ –. En la Figura 70 se representan, a modo de ejemplo, dos respuestas al impulso utilizadas para la simulación de simCMU-2. Para el cálculo se ha usado el algoritmo de Allen [Allen 77-b] que implementa el método de fuentes imagen para hallar la respuesta al impulso en un recinto. Nótese, a la vista de la Figura 70, que el T_{60} medido sobre $h_{P7}(t)$, evaluando el tiempo en 60dB de atenuación, es de aproximadamente 0.7s. Seguidamente se ha situado una fuente de señal (voz) en uno de los puntos y una fuente de ruido en otro de los puntos, de tal manera que

se han generado diversas combinaciones de voz limpia y ruido, emitidos por fuentes situadas en diferentes puntos del recinto.

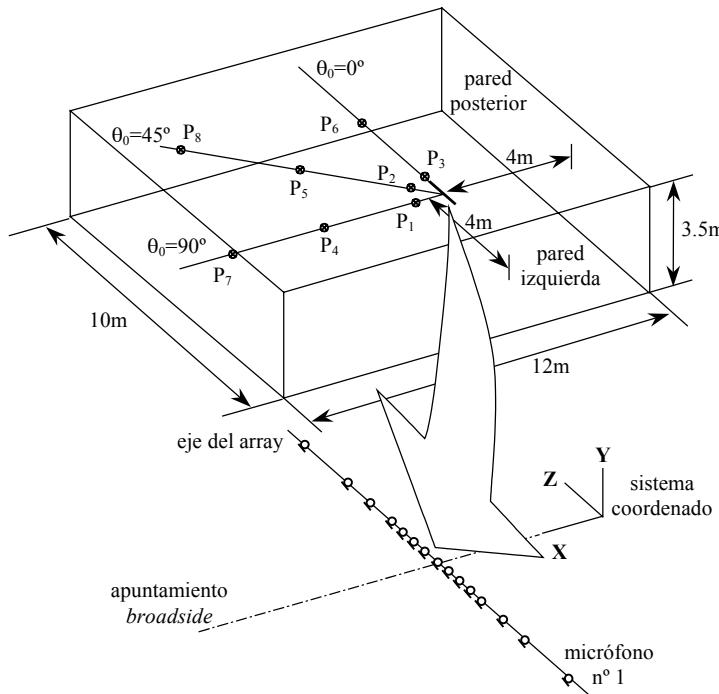


Figura 69. Escenario de generación de la base de datos simulada simCMU-2 a partir de las grabaciones limpias de la base de datos real CMU descrita en el punto 6.2 de esta Tesis.

punto	θ_0 [°]	ϕ_0 [°]	r_0 [m]
P ₁	90	0	1
P ₂	45	0	1
P ₃	0	0	1
P ₄	90	0	4
P ₅	45	0	4
P ₆	0	0	4
P ₇	90	0	7
P ₈	45	0	7

Tabla 12. Coordenadas de la fuente, para generar la base de datos simCMU-2, respecto al sistema coordinado centrado en el array, basado en el de la Figura 6.

La Figura 71 ilustra el proceso de generación de la base de datos simCMU-2. Tanto la señal original o de referencia $x_0(t)$ como el ruido $n(t)$ se convolucionan con el vector de respuestas al impulso correspondientes al punto donde se ha situado la señal o el ruido –vectores $\mathbf{h}_{pj}(t)$ y $\mathbf{h}_{pk}(t)$ de la Figura 71–. El vector de ruido $\mathbf{n}(t)$ resultante, que contiene reverberación, se multiplica por un factor α de atenuación, dependiente de la relación SNR₈ requerida (se considera el canal central del array, $i = 8$) y del nivel de la señal sin ruido en el vector $\mathbf{y}_r(t)$ –voz reverberante, como en (26)–. Las dos señales multicanal de salida se suman para generar la señal multicanal resultante, $\mathbf{y}(t)$, que será uno de los integrantes de la base de datos simulada. Como resultado, a diferencia de simCMU-1.1, tanto la señal como el ruido están afectados por la reverberación, para hacer más realista la simulación. El ruido $n(t)$

considerado ha sido elegido entre estos dos casos, ruido aleatorio o ruido determinista. El ruido aleatorio consiste en ruido blanco gaussiano y el ruido determinista en una señal diente de sierra de frecuencia $f_N = 500\text{Hz}$ (Figura 72).

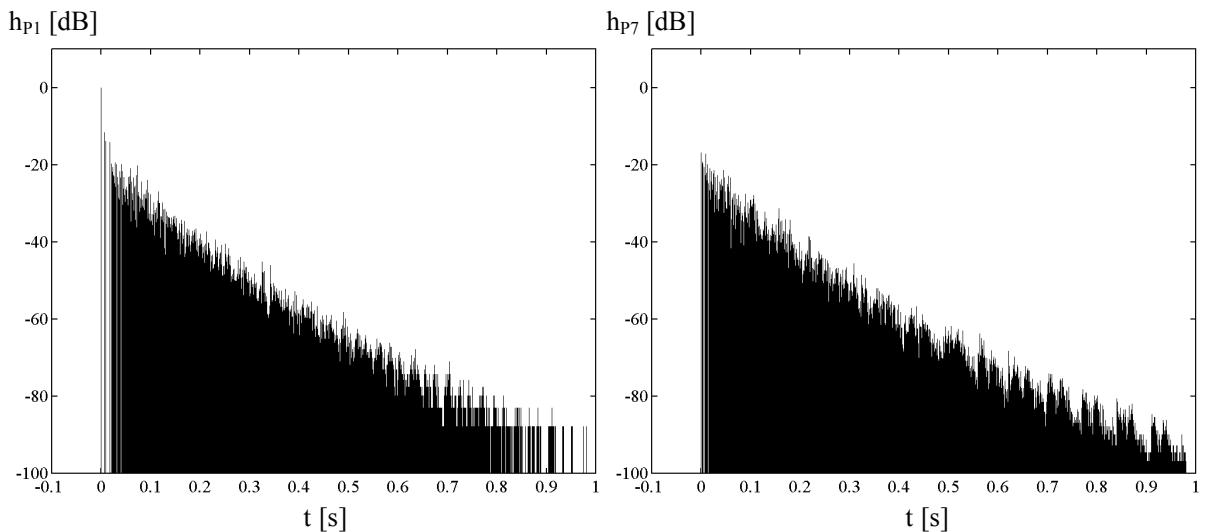


Figura 70. Ejemplo de respuesta al impulso h_{Pj} [dB] en los puntos P_1 (h_{P1}) y P_7 (h_{P7}) (considerando el elemento $i=8$ –micrófono central del array– del vector de respuestas al impulso) para la generación de simCMU-2.

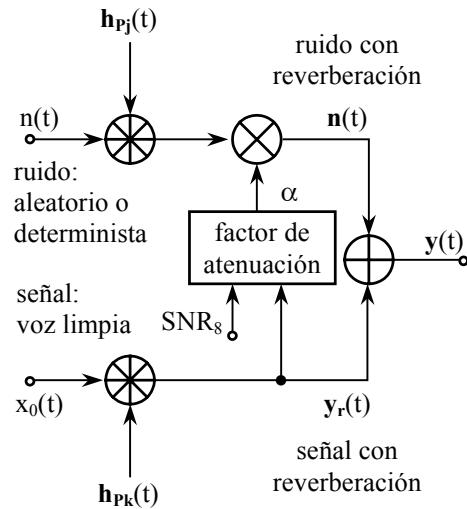


Figura 71. Proceso de generación de la base de datos simCMU-2.

En la Tabla 13 se exponen las características más importantes de la base simCMU-2, utilizada en muchas de las pruebas y experimentos del capítulo 7 de la Tesis.

El subcorpus **rev** contiene sólo señal reverberante (sin ruido) generada en el recinto de la Figura 69. Los subcorporus **dif**, **Mdif** y **Mcoh** contienen combinaciones de voz reverberante con ruido reverberante procedentes de diferentes puntos del recinto. Se ha limitado el número de combinaciones de posición y SNR₈, y aun así simCMU-2 contiene 207 archivos multicanal.

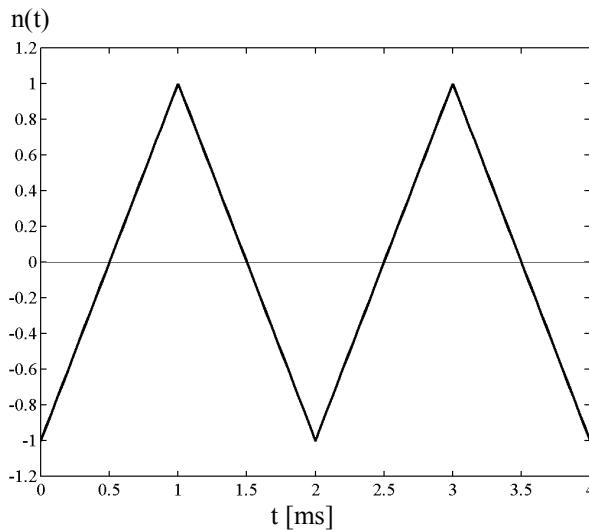


Figura 72. Señal diente de sierra de frecuencia 500Hz considerada para añadir ruido determinista a la señal de voz limpia (bases simCMU-2 y simCMU1.1).

identificador de sucopus	Nº de archivos	tipo de ruido	señal	ruido	SNR ₈ [dB]
rev	40 ⁽¹⁾	-	P ₁ a P ₈	-	-
dif	27 ⁽²⁾	blanco	P ₁ a P ₃	P ₄ , P ₆ y P ₇	10, 20 y 30
Mdif	20 ⁽¹⁾	blanco	P ₄ y P ₅	P ₄ y P ₆	20
	60 ⁽¹⁾		P ₁ , P ₂	P ₆ , P ₇ y P ₈	0 y 10
	40 ⁽¹⁾		P ₄ , P ₅ , P ₇ y P ₈	P ₇	
Mcoh	20 ⁽¹⁾	d. sierra	P ₄ y P ₅	P ₄ y P ₆	20

Tabla 13. Base de datos simulada multicanal utilizada (simCMU-2) y características más importantes de cada subcorpus. ⁽¹⁾ Contiene cinco locuciones diferentes por cada combinación de puntos. ⁽²⁾ Está generada con la misma locución de voz en todas las combinaciones de puntos.

7 PROPUESTAS DE MEJORA DE SEÑAL DE VOZ EN PRESENCIA DE RUIDO Y REVERBERACIÓN

Como se refiere en la introducción, esta Tesis es una continuación de los trabajos de investigación sobre procesado en array iniciados en el grupo ATVS. Las investigaciones iniciales se centran en desarrollar los algoritmos óptimos para reducir ruido y reverberación mediante captación en array. En [González-Rodríguez 99-a] se demuestra el buen comportamiento del algoritmo de Wiener con modificación de coherencia, que utiliza un postfiltrado de Wiener multicanal dependiente de la información de coherencia intercanal que puede obtenerse con el array microfónico. Es lo que aquí será llamado algoritmo MW (*Modified Wiener*) y del que posteriormente se darán más detalles. Además en [González-Rodríguez 99-a] se prueban otros algoritmos alternativos, demostrándose que la derreverberación cepstral basada en descomposición fase mínima-paso todo, en lo sucesivo algoritmo MPAP (*Minimum Phase-All Pass*) –ver el punto 4.2.1 de esta Tesis y [Liu 96]–, es bastante adecuada para reducir reverberación. Por eso, los primeros intentos del trabajo de investigación de esta Tesis, se centraron en proponer y probar un algoritmo que combinara ambas técnicas sobre el array anidado de 15 micrófonos. Paralelamente era necesario adaptarlo a una posible implementación de un prototipo en tiempo real, con lo que había que eliminar las partes menos eficientes del algoritmo, desde el punto de vista computacional. Estos primeros intentos dieron lugar al algoritmo combinado MW-MPAP. Seguidamente se adaptaron al caso multicanal algunas mejoras encontradas en la literatura, relacionadas con las mejoras subjetivas que produce el postfiltrado perceptual (punto 4.1.3), basado en el procesado dependiente del enmascaramiento auditivo [Tsoukalas 97] [Virag 99]. Con ello se propone posteriormente el algoritmo ANS-MW, que utiliza el procesado ANS (punto 4.1.3) combinándolo con el postfiltrado de Wiener que se ha mostrado tan eficiente.

Otra línea de investigación iniciada en esta Tesis es la evaluación objetiva de la calidad de habla. Como se explica en el capítulo 5, existen diversos métodos para obtener una calificación del procesador en array cuando se dispone de la señal de referencia X_0 . Los métodos basados en SNR son bastante adecuados para evaluar la cantidad de ruido eliminado, pero se muestran poco eficientes a la hora de sopesar la reverberación. Efectivamente, la evaluación de la cantidad de reverberación ha sido un caballo de batalla tanto en el mundo del procesado de voz como en el mundo de la acústica de recintos, dando lugar en este último caso a una gran variedad de índices de inteligibilidad. Actualmente el índice RASTI es una referencia para medir la inteligibilidad del habla en un recinto, y por ende para evaluar de alguna manera la reverberación, que interfiere negativamente en dicha inteligibilidad. Por ello se hacía necesario profundizar en la utilización del índice RASTI adaptado al procesado de señal vocal. En esta Tesis se propone el índice E-RASTI (*Emulated RASTI*) que no es más que una medida de la ganancia de modulación producida en una señal de voz por los procesadores propuestos. La novedad del método consiste en que se miden dichas pérdidas de modulación sobre la señal vocal a evaluar y no se usa la señal RASTI (Figura 54).

Una vez realizadas las pruebas preliminares referidas en este capítulo, se estaba en disposición de implementar un prototipo real utilizando los algoritmos propuestos, pero eso será el centro de atención de la Parte 3 de la Tesis. En el presente capítulo se muestran por tanto las propuestas, pruebas y resultados relacionados con cada uno de los tres siguientes elementos propuestos: algoritmo MW-MPAP, algoritmo MW-ANS y método E-RASTI.

7.1 PROCESADOR EN ARRAY BASADO EN FILTRADO DE WIENER MULTICANAL MODIFICADO POR COHERENCIA MÁS DERREVERBERADOR CIEGO BASADO EN DESCOMPOSICIÓN FASE MÍNIMA-PASO TODO (MW-MPAP)

La propuesta inicial para conseguir un procesador ciego de limpieza de voz basado en un array anidado de 15 micrófonos, se basa en la conjunción del postfiltrado de Wiener modificado por coherencia y la derreverberación mediante descomposición fase mínima-paso todo, es decir el procesador MW-MPAP. Los primeros experimentos sobre este procesador en array han sido publicados en [González-Rodríguez 99-b] y [González-Rodríguez 00] y contienen un acercamiento inicial al problema. Básicamente persiguen verificar que el procesador funciona correctamente, es decir, proporciona mejores resultados que los ofrecidos de forma individual por la parte MW y MPAP de dicho procesador. Seguidamente, como segunda fase de los experimentos, se hizo un ajuste fino de los parámetros del procesador, referidos sobre todo al enventanado y a los parámetros asociados a los estimadores de señal y de ruido. Estos resultados han sido publicados en [Sánchez-Bote 00]. A continuación se describen en detalle las propuestas aportadas.

7.1.1 Descripción del sistema

El diagrama de bloques del procesador MW-MPAP se muestra en la Figura 73. Básicamente consiste en un filtro de Wiener multicanal (véase el punto 4.1.1 de esta Tesis) mejorado con un derreverberador ciego basado en descomposición cepstral. El procesador consta de tres grandes bloques, que se describen a continuación.

Conformador de haz (*Beamformer*)

Realiza una conformación de haz en el dominio de la frecuencia. Es decir, previamente se debe realizar una transformada STFT mediante enventanado (normalmente se usan ventanas tipo Hanning) más FFT. El conformador utiliza el método convencional de retardo y suma aplicado al array anidado de 15 canales, con división en tres subbandas y descrito en el punto 6.1. El retardo necesario para alinear temporalmente los 15 canales puede estimarse de forma automática (véase el capítulo 3 sobre localización de fuente), pero en este procesador se introduce manualmente. Conocido el retardo, existen dos posibilidades para la alineación temporal, lo que se llamará compensación de retardo en frecuencia y la compensación de retardo en tiempo. La compensación de retardo en frecuencia se aplica sobre la señal en frecuencia $Y_i(\omega)$ mediante:

$$Y_{Ri}(\omega) = Y_i(\omega) e^{j\omega\tau_i} \quad (265)$$

siendo τ_i el retardo con el que cada canal $y_i(t)$ llega con respecto al canal central del array $y_8(t)$. Véase la Tabla 7 con los retardos necesarios para dos apuntamientos típicos. El método

de alineación temporal en frecuencia tiene algunos inconvenientes que se explicarán más adelante.

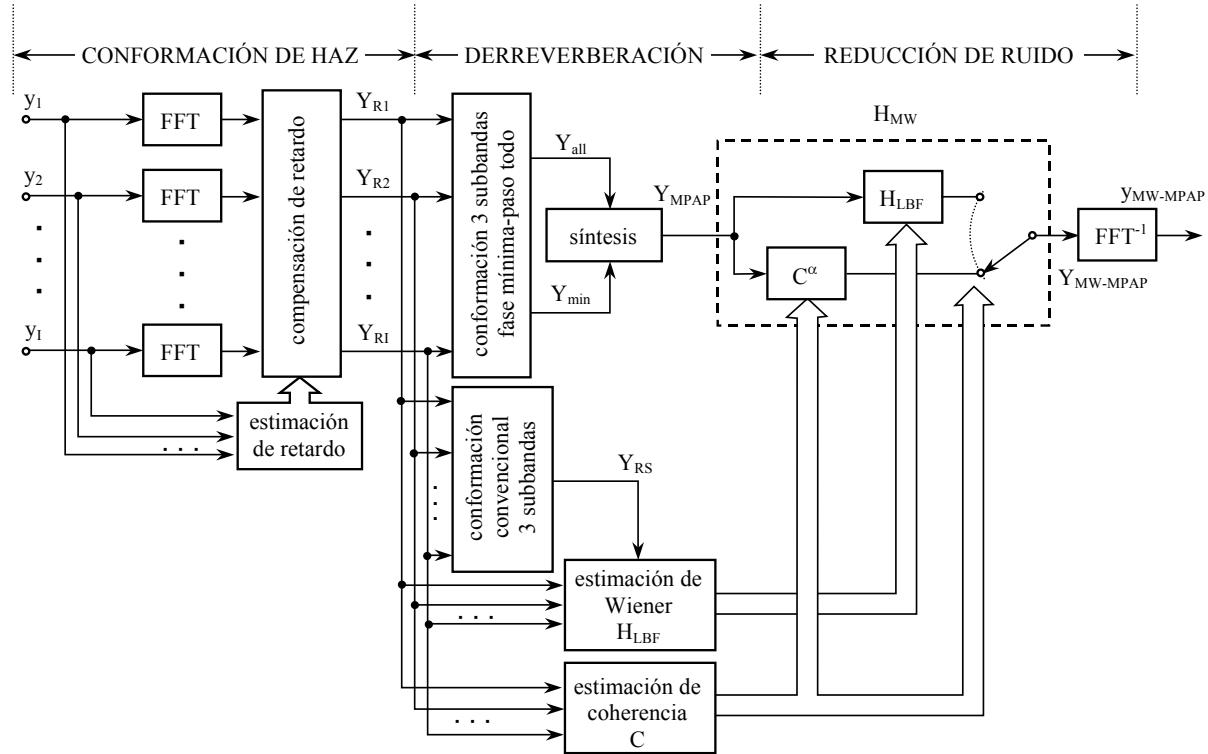


Figura 73. Diagrama de bloques del procesador MW-MPAP.

Por otra parte se puede hacer una compensación temporal de retardos. Es ese caso se realiza en el dominio del tiempo, antes del bloque FFT, retrasando el número de muestras necesario para obtener el retardo requerido. El inconveniente es que la precisión de retardo está limitada al periodo de muestreo, que es $T_s = 1/f_s = 1/16000 = 62.5\mu s$. Para aumentar esta precisión habrá que utilizar interpolación numérica o aumentar la frecuencia de muestreo, con la sobrecarga computacional que esto conlleva sobre aplicaciones en tiempo real.

A pesar de que en la Figura 73 se representa la compensación en frecuencia de retardos (puesto que se hace después del bloque FFT), todas las pruebas de este capítulo sobre el procesador MW-MPAP se han realizado aplicando la compensación temporal de retardos, utilizando cierta tasa de interpolación numérica, que se especificará más adelante, para aumentar la precisión en el dominio del tiempo. Las frecuencias de división de las tres subbandas fueron $f_1 = 1\text{kHz}$ y $f_2 = 2\text{kHz}$. A la salida del conformador convencional se obtiene la señal $Y_{RS}(\omega)$ a la cual ya se le ha restado cierta cantidad de reverberación y ruido debido al proceso de conformación de haz, según se desarrolló en el capítulo 2 de la Tesis.

Derreverberador ciego mediante descomposición cepstral

Se aplica el método de descomposición cepstral multicanal MPAP descrito en el punto 4.2.1 de esta Tesis y en [Liu 96] mediante el esquema de la Figura 47, pero aplicando la conformación en tres subbandas. Como parámetro más importante de esta parte del procesador está $W_C [\text{pt}]$ que es la “quefrecuencia” máxima (de *quefreny*, frecuencia cepstral) en puntos aplicada al *liftering* de la parte fase mínima, como se explicó con anterioridad. Es decir

W_C es la cuestión de corte, expresada en muestras, del filtro cepstral \hat{w}_{low} del punto 4.2.1. La salida $Y_{\text{MPAP}}(\omega)$ corresponde en realidad a una conformación en el dominio cepstral a la que se ha añadido la derreverberación producida por el *liftering*.

Como consta en el apartado 7.1.4 de resultados, en algunas pruebas se consideró la conformación cepstral en una sola banda mediante la suma directa de todas las salidas microfónicas, por lo que en éstas el diagrama de bloques de la Figura 73 variaría ligeramente.

Reducción de ruido mediante filtrado Wiener modificado por coherencia

La primera aproximación realizada para reducir el ruido añadido a la señal de voz consiste en la implementación de un filtro de Wiener multicanal. Ya se ha expuesto (punto 4.1.1) que una conformación de haz superdirectiva más un filtrado de Wiener monocanal (lo que comúnmente se llama filtrado Wiener multicanal) constituye el postfiltro óptimo en el sentido de que minimiza la SNR en banda ancha a la salida del array. En ese aspecto, el procesador MW-MPAP propuesto no es un filtro de Wiener multicanal óptimo ya que realiza una conformación convencional (retardo y suma) y no superdirectiva. La diferencia entre el conformador convencional y el superdirectivo sólo aparece en baja frecuencia, donde la coherencia intercanal es baja. En el array anidado utilizado esa diferencia sólo afectará a la subbanda B_1 . Sin embargo, se puede hacer un aprovechamiento óptimo de la información multicanal para estimar correctamente la señal limpia $-X_0(\omega)$ en la nomenclatura de la Tesis— a partir de la información espacial aportada por el array de micrófonos, como se explica a continuación. En este último aspecto sí puede considerarse que el filtrado Wiener modificado por coherencia utilizado aquí es óptimo.

La señal que llega a los micrófonos está contaminada por ruido y reverberación. Si se aplica un filtro de Wiener a la salida conformada del array, en el numerador de la función de transferencia correspondiente a dicho filtro hay que poner la estimación de la señal limpia. La ecuación (161) da la función de transferencia del filtro de Wiener monocanal óptimo. En el numerador está escrito el autoespectro o potencia de la señal de referencia $\Phi_{X_0 X_0}(\omega)$ que habrá que estimar. La forma más intuitiva de hacer esta estimación es simplemente restando la potencia estimada de ruido conformado, a modo de un sustractor espectral de potencia (véase el punto 4.1.2).

Ya se ha disertado anteriormente sobre la naturaleza del ruido aditivo que va a captar un array de micrófonos. Por una parte está el ruido coherente (en el sentido de coherencia intercanal), provocado por fuentes de ruido presentes o por reflexiones tempranas y muy localizadas temporalmente. Por otra parte está el ruido no coherente, que normalmente corresponde a la reverberación difusa o a fuentes de ruido lejanas o poco localizadas. También se puede incluir en este apartado al ruido inherente al procesador (micrófonos, preamplificadores, conversión analógica digital, etc). El ruido coherente se puede eliminar mediante una conformación óptima (recuérdese que el conformador óptimo para ruido difuso o isotrópico es el conformador superdirectivo). El ruido no coherente se puede estimar en los instantes de no-actividad de habla, si se dispone de un VAD eficiente. El estimador de señal limpia $\hat{\Phi}_{X_0 X_0}(\omega)$ tendrá que tener en cuenta esas consideraciones, como se explica a continuación.

Considérese una estimación de la potencia de la señal de referencia $\hat{\Phi}_{X_0 X_0}(\omega)$, calculada como:

$$\hat{\Phi}_{X_0 X_0}(\omega) = \hat{\Phi}_{YY}(\omega) - \hat{\Phi}_{NN}(\omega) \quad (266)$$

donde $\hat{\Phi}_{YY}(\omega)$ es una estimación de la señal de voz sin presencia de ruido no coherente y $\hat{\Phi}_{NN}(\omega)$ es una estimación de la potencia de ruido coherente hecha en los instantes de ausencia de voz. Dicha estimación, $\hat{\Phi}_{X_0 X_0}(\omega)$, estará libre de ruido no coherente –eliminado en $\hat{\Phi}_{YY}(\omega)$ – y de ruido coherente –eliminado mediante la sustracción de $\hat{\Phi}_{NN}(\omega)$ –. Las propuestas para hacer todas estas estimaciones son muy abundantes en la literatura científica. Aquí se ha considerado la propuesta de [Le Bouquin 97] y experimentada en [González-Rodríguez 99-a] con un array anidado de micrófonos usando la base de datos CMU.

En estos trabajos se proponen las siguientes estimaciones.

Para la señal sin ruido no coherente:

$$\hat{\Phi}_{YY}(\omega) = \left\langle \Phi_{Y_i Y_j}(\omega) \right\rangle_{i,j} \quad (267)$$

se hace un promedio en cada trama de voz de todos los espectros cruzados intermicrofónicos. El subíndice i, j indica que el promedio se hace considerando todas las parejas de micrófonos.

Para el ruido coherente:

$$\hat{\Phi}_{NN}(\omega) = \left\langle \Phi_{N_i N_j}(\omega) \right\rangle_{i,j} \quad (268)$$

se considera el mismo promedio activado sólo en los instantes de no-actividad de voz.

El inconveniente de este planteamiento es el gran coste computacional que requiere para ser adaptado al procesador propuesto. Efectivamente, si se tienen 15 micrófonos, el número de parejas de micrófonos a considerar es el número combinatorio $C(15, 2) = 105$ parejas de micrófonos. Es decir, en cada trama de voz hay que realizar 105 espectros cruzados lo cual es poco recomendable desde el punto de vista de un procesado en tiempo real.

Consecuentemente, la propuesta inicial ha sido adaptada por el autor para un comportamiento más eficiente desde el punto de vista computacional. Según dicha propuesta del autor, la estimación de señal sin ruido no coherente quedaría:

$$\hat{\Phi}_{YY}(\omega) = \left\langle \Phi_{Y_i Y_{ref}}(\omega) \right\rangle_i \quad (269)$$

que se calcula con el espectro cruzado promedio entre la señal salida de todos los canales del array y un canal de referencia, llamado canal “ref”. No hay que confundir este canal de referencia, que varía de trama en trama, con el canal central del array (canal 8), ni con el canal de referencia del array (canal 0). El canal “ref” de (269) se asigna en cada trama de tiempo, de tal manera que el espectro cruzado recorra todas las parejas de micrófonos en un tiempo de procesado suficientemente pequeño, durante el cual la señal de voz y el ruido puedan ser considerados como estacionarios. En (269) el subíndice i recorre todos los micrófonos del array menos uno ($I_T - 1$ valores), correspondiente al canal “ref” considerado en un momento dado.

Por otra parte, la estimación de ruido coherente:

$$\hat{\Phi}_{NN}(\omega) = \left\langle \Phi_{N_i N_{ref}}(\omega) \right\rangle_i \quad (270)$$

se realiza mediante el espectro cruzado intermicrofónico promedio entre la señal de ruido y el canal “ref”, estimado en las tramas sin actividad de voz, con la misma asignación de canal de referencia considerada anteriormente. Los promedios (269) y (270) excluyen los autoes-

pectros del canal de referencia $\Phi_{Y_{ref}Y_{ref}}(\omega)$ y $\Phi_{N_{ref}N_{ref}}(\omega)$ con lo que se contemplan $I_T - 1 = 15 - 1$ espectros cruzados por cada trama de voz. Como el canal de referencia va siendo asignado crecientemente en cada trama de voz desde $i = 1$ hasta $i = 15$, cuando se hayan recorrido 15 tramas de voz se habrán realizado $14 \times 15 = 210$ autoespectros que corresponden a las 105 parejas diferentes duplicadas, aunque en instantes distintos de tiempo. El valor típico de actualización de una trama de voz utilizado en el procesador propuesto está en torno a 10ms, con lo que en unos $15 \times 10\text{ms} = 150\text{ms}$ se habrán realizado los 210 espectros cruzados. Este tiempo es suficientemente corto para ser considerado aceptable. Recuérdese que sólo afecta a los estimadores del filtro de Wiener, que por otra parte van a tener que ser considerados con cierto grado de filtrado paso bajo o latencia temporal, en cuanto a la variación intertrama (véase el apartado 4.1.4 de esta Tesis).

La función de transferencia del filtro Wiener modificado propuesto, es entonces

$$H_{LBF}(\omega) = \frac{|\hat{\Phi}_{YY}(\omega) - \hat{\Phi}_{NN}(\omega)|}{\hat{\Phi}_{Y_{RS}Y_{RS}}(\omega)} = \frac{\left| \langle \hat{\Phi}_{Y_i Y_{ref}}(\omega) \rangle_i - \langle \hat{\Phi}_{N_i N_{ref}}(\omega) \rangle_i \right|}{\hat{\Phi}_{Y_{RS}Y_{RS}}(\omega)} \quad \text{con } 0 < H_{LBF}(\omega) < 1 \quad (271)$$

El subíndice LBF atiende a Jeannes R. le Bouquin y G. Faucon [Le Bouquin 97]. En el numerador se ha considerado el módulo para evitar que éste sea de naturaleza compleja o de valor negativo, ambos efectos causados por los errores de estimación, siempre presentes. Además el filtro LBF debe ser limitado en el intervalo $[0, 1]$, ya que la naturaleza de un filtro de tipo Wiener no permite otros valores. En el denominador figura la potencia de señal de voz a la salida del conformador, en este caso mediante el método de retardo y suma. Las tildes en $\hat{\Phi}$ indican que se hace una estimación recursiva, como se explica más adelante.

Las estimaciones incluidas en (271) pueden tener muchos errores que consecuentemente se traducen en un filtro de Wiener erróneo. Estas situaciones ocurrirán frecuentemente en situaciones de mucho ruido aditivo o de mucha reverberación. En estas ocasiones el filtro LBF resultará en un valor errático y la posibilidad de que produzca distorsión audible a la salida aumenta fuertemente. Los errores de estimación se darán entonces cuando la coherencia intercanal sea baja, debido a la presencia de ruido y reverberación. Por ello se hace recomendable adaptar un detector de coherencia [Carter 73] [Carter 87] [Mahmoudi 98] que interprete cuándo el filtro de Wiener es previsiblemente incorrecto, para en ese momento atenuar fuertemente la salida filtrada. Es lo que en esta Tesis se llama modificación de coherencia. El filtro resultante, que se llamará $H_{MW}(\omega)$ viene expresado por la ecuación:

$$H_{MW}(\omega) = \begin{cases} H_{LBF}(\omega) & \text{si } C(\omega) \geq UC \\ C^\alpha(\omega) & \text{si } C(\omega) < UC \end{cases} \quad (272)$$

con,

$$C(\omega) = \frac{|\hat{\Phi}_{Y_{RS}Y_8}(\omega)|}{\sqrt{\hat{\Phi}_{Y_{RS}Y_{RS}}(\omega) \hat{\Phi}_{Y_8Y_8}(\omega)}} \quad (273)$$

donde $C(\omega)$ es la función coherencia considerada. Las tildes en $\hat{\Phi}$ aquí también se refieren a la estimación recursiva. La función $C(\omega)$ sólo tiene en cuenta el grado de correlación entre la salida conformada del array $Y_{RS}(\omega)$ y el canal central del array $Y_8(\omega)$. Se podría calcular un promedio de las coherencias intercanal entre todos los pares de micrófonos, al modo de (269) y (270), pero, para la utilidad que va a dar a la función coherencia, se considera suficiente obtener tan solo $\hat{\Phi}_{Y_{RS}Y_8}(\omega)$, ya que si la coherencia intercanal es baja también lo será $C(\omega)$,

puesto que $Y_{RS}(\omega)$ participa de todos los canales. En (272), UC es un umbral de coherencia a partir del cual se considera aplicable la modificación de coherencia, y α es un factor de forma que se elige según el grado de coherencia intercanal de partida en la señal multimicrofónica.

Finalmente, la salida del procesador se consigue de la aplicación del filtro $H_{MW}(\omega)$ a la salida del derreverberador MPAP mediante:

$$Y_{MW-MPAP}(\omega) = H_{MW}(\omega) Y_{MPAP}(\omega) \quad (274)$$

Como se ha manifestado antes, $Y_{MPAP}(\omega)$ puede considerarse como la salida del conformador convencional en tres subbandas, aunque con un procesado cepstral adicional que consigue una derreverberación extra mediante *liftering*. El filtro de Wiener $H_{MW}(\omega)$ consigue en el conformador una eliminación de ruido adicional.

7.1.2 Estimación recursiva de los espectros de señal y de ruido

Todas las cantidades espectrales a estimar, es decir, las componentes de potencia de ruido y señal de (271) y (273), se han obtenido por el método recursivo descrito en el punto 4.1.4 de esta Tesis. Se ha utilizado una recursión de polo único en la forma expresada en (192). Así en la fórmula (271) para el filtro LBF se hace una estimación recursiva, con el parámetro de actualización λ_S para la señal, en los espectros cruzados del numerador y la señal conformada del denominador:

$$\hat{\Phi}_{Y_i Y_{ref}}(\omega, k) = \lambda_S \hat{\Phi}_{Y_i Y_{ref}}(\omega, k-1) + (1 - \lambda_S) \Phi_{Y_i Y_{ref}}(\omega, k) \quad (275)$$

$$\hat{\Phi}_{Y_{RS} Y_{RS}}(\omega, k) = \lambda_S \hat{\Phi}_{Y_{RS} Y_{RS}}(\omega, k-1) + (1 - \lambda_S) \Phi_{Y_{RS} Y_{RS}}(\omega, k) \quad (276)$$

mientras que λ_N se considera para la estimación de ruido del numerador,

$$\hat{\Phi}_{N_i N_{ref}}(\omega, k) = \lambda_N \hat{\Phi}_{N_i N_{ref}}(\omega, k-1) + (1 - \lambda_N) \Phi_{N_i N_{ref}}(\omega, k) \quad (277)$$

Por otra parte, mediante esta misma técnica se estima también la fórmula de coherencia (273) con el parámetro único λ_{coh} para sus tres componentes:

$$|\hat{\Phi}_{Y_{RS} Y_8}(\omega, k)| = \lambda_{coh} |\hat{\Phi}_{Y_{RS} Y_8}(\omega, k-1)| + (1 - \lambda_{coh}) |\Phi_{Y_{RS} Y_8}(\omega, k)| \quad (278)$$

$$\hat{\Phi}_{Y_{RS} Y_{RS}}(\omega, k) = \lambda_{coh} \hat{\Phi}_{Y_{RS} Y_{RS}}(\omega, k-1) + (1 - \lambda_{coh}) \Phi_{Y_{RS} Y_{RS}}(\omega, k) \quad (279)$$

$$\hat{\Phi}_{Y_8 Y_8}(\omega, k) = \lambda_{coh} \hat{\Phi}_{Y_8 Y_8}(\omega, k-1) + (1 - \lambda_{coh}) \Phi_{Y_8 Y_8}(\omega, k) \quad (280)$$

Como se verá en el transcurso de los experimentos, el uso de diferentes valores de los parámetros λ dará lugar a mayor o menor tasa de eliminación de ruido y consecuentemente mayor o menor cantidad de ruido musical remanente o distorsión, debido a la mayor o menor variación instantánea de la función de transferencia del filtro $H_{MW}(\omega)$ que produce la limpieza final de voz. Por otra parte, en el procesado MPAP no existe este tipo de distorsión, o al menos no es suficientemente apreciable, ya que el filtro que se le asocia, una vez determinado, mantiene sus parámetros constantes ventana a ventana, sin producir artefactos como el ruido musical.

7.1.3 Enventanado y reconstrucción de la señal temporal

El esquema del procesador basado en análisis y síntesis FFT mediante enventanado exige algunas reflexiones. Evidentemente se busca que la reconstrucción de la señal temporal sea lo más perfecta posible, es decir, que las ventanas $v(t)$ sintetizadas “encajen” perfectamente a la hora de regenerar la señal temporal de salida. En ese sentido se suele emplear la ventana Hanning (Figura 74). Existen dos posibilidades a la hora de diseñar una ventana Hanning:

$$v[n] = \frac{1}{2} \left(1 - \cos \frac{2\pi n}{L+1} \right) \quad n = 1, 2, \dots, L \quad \text{Hanning simétrica} \quad (281)$$

$$v[n] = \begin{cases} 0 & n = 1 \\ \frac{1}{2} \left(1 - \cos \frac{2\pi(n-1)}{L} \right) & n = 2, 3, \dots, L \end{cases} \quad \text{Hanning periódica} \quad (282)$$

siendo L el tamaño en puntos o muestras de la ventana. En (281) y (282) se considera la versión discreta de $v(t)$ con n el índice de tiempo discreto. La diferencia entre las dos versiones es que la ventana Hanning periódica no es simétrica, es decir el primer valor es cero y el último es distinto de cero.

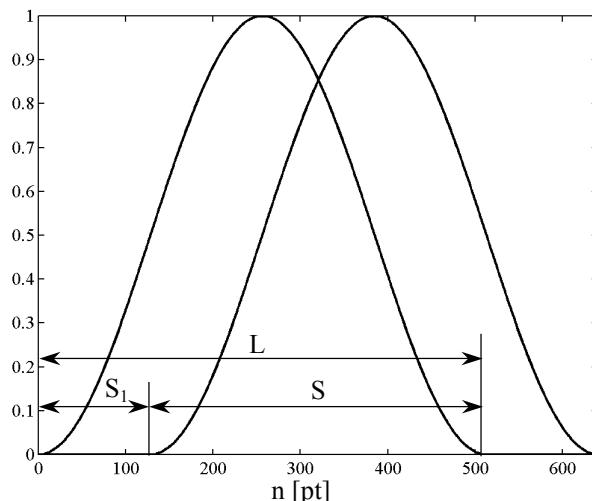


Figura 74. Ventana Hanning de longitud $L=512pt$ y solapamiento $S=384pt$, equivalente a $S=75\%$.

Parámetros importantes en el proceso de enventanado son:

- **S [pt] ó T_s [s]**: es el solapamiento interventana, es decir la cantidad de muestras comunes entre una trama o ventana y la siguiente, y
- **S_1 [pt] ó T_{s_1} [s]**: es el desplazamiento interventana, es decir cuánto se desplaza una ventana con respecto a la siguiente.

Se cumple además que:

$$L = S + S_1 \quad (283)$$

Si el enventanado no es perfecto, es decir no se reconstruye correctamente la señal de salida, puede aparecer un “ruido de enventanado”, que consiste en un impulso que se repite periódicamente con frecuencia S_1/f_s , y que puede llegar a ser perceptible. Parece evidente que

si en L cabe un número entero de periodos S_1 , la reconstrucción puede ser casi perfecta. En ese caso la suma de sucesivas ventanas desplazadas es una constante. Esto se comprueba en el siguiente experimento.

Se intenta reconstruir mediante el proceso enventanado-FFT-IFT-desenventanado una señal que consiste en una constante con un impulso de corta duración en una posición aleatoria (Figura 75). Se hacen medidas promediando 100 fragmentos de este tipo con diferentes tipos de enventanado. Los resultados se muestran en la Tabla 14. A la vista de dicha tabla, el proceso de enventanado cuando S_1 es un divisor de L proporciona unos resultados perfectos – $\text{error} < -200\text{dB}$!– cuando la ventana es periódica (sólo limitados por la precisión del procesador, que corresponde a 64bits en punto flotante para este experimento). Téngase en cuenta que los errores mostrados también incluyen los correspondientes al proceso FFT-IFT del impulso.

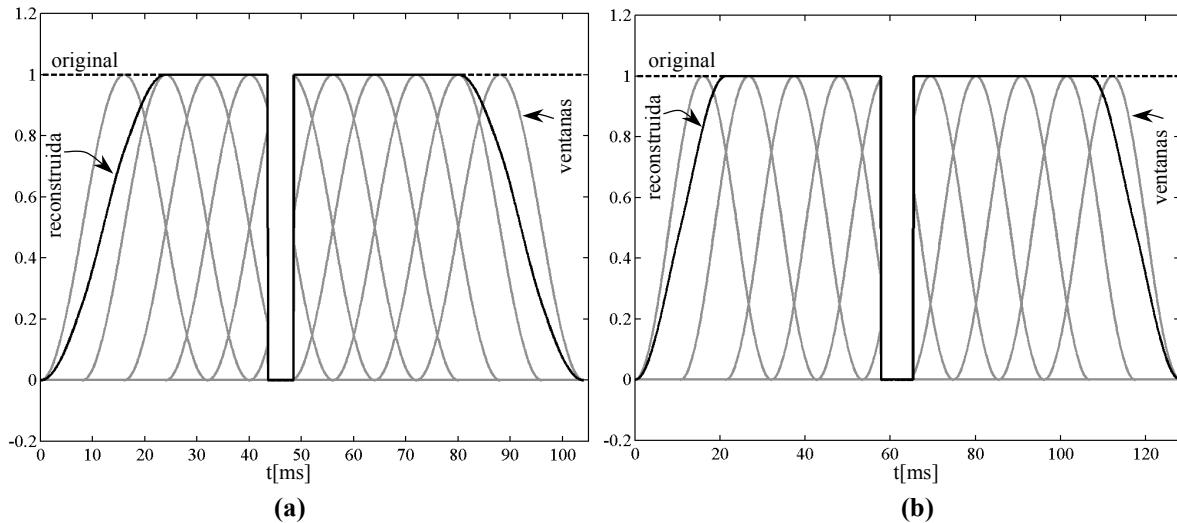


Figura 75. Detalle de reconstrucción de una señal con un transitorio mediante el proceso enventanado-FFT-IFT-desenventanado, con $L=512\text{pt}$ y $f_s=16\text{kHz}$, utilizando la ventana Hanning. (a) $S_1=128\text{pt}$ ($S=75\%$) ventana Hanning simétrica. La reconstrucción es perfecta. (b) $S_1=171\text{pt}$ ($S=67\%$) ventana Hanning periódica. La reconstrucción es perfecta.

L [pt]	S₁ [pt]/S[%]	Hanning	E[dB]	A
512	171/67	simétrica	-292	2/3
		periódica	-59	~2/3
512	128/75	simétrica	-64	~1/2
		periódica	-290	1/2
1024	341/67	simétrica	-59	~2/3
		periódica	-65	~2/3
1024	256/75	simétrica	-70	~1/2
		periódica	-283	1/2
2048	512/75	periódica	-284	1/2
4096	1024/75	periódica	-274	1/2

Tabla 14. Combinaciones de número de puntos L con diferentes solapamientos S , para la ventana Hanning simétrica y periódica. E es el error promedio en el proceso de reconstrucción enventanado-FFT-IFT-desenventanado cuando la entrada procesada es un transitorio del tipo al mostrado en la Figura 75. A es la amplificación necesaria en la señal reconstruida para que su valor RMS se iguale a la original.

Sin embargo, hay un resultado que será utilizado en el futuro. Si la longitud de la ventana es $L = 512\text{pt}$ y el desplazamiento es $S_1 = 171\text{pt}$ ($S = 66\%$), cuando la ventana es simétrica, la reconstrucción es perfecta. Este resultado es muy interesante ya que permite utilizar solapamientos intermedios entre $S = 75\%$ y $S = 50\%$ que son los más frecuentemente usados en la práctica. Hay que reseñar que, desde el punto de vista del procesador MW-MPAP conviene utilizar grandes solapamientos, ya que en ese caso el cambio en el valor instantáneo del postfiltro utilizado es menos perceptible desde el punto de vista subjetivo. Sin embargo, un solapamiento muy elevado aumenta el coste computacional en un procesador en tiempo real, ya que el número de FFT's por segundo que hay que realizar crece linealmente con S . Por eso $S_1 = 171\text{pt}$ puede ser una solución de compromiso para el procesado en tiempo real.

Otra cuestión que hay que considerar en cuanto al enventanado es la aplicación de retardo en el dominio de la frecuencia. Efectivamente, el conformador de haz que utiliza el array microfónico realiza una alineación en tiempo en el dominio de la frecuencia y ventana a ventana. Este procedimiento tiene algunas limitaciones, que se pueden resumir en que el tiempo de retardo debe ser bastante inferior a la duración de las L muestras que representan a una ventana. Efectivamente, en la Figura 76 se representa el efecto que tiene un retardo sobre una trama de señal enventanada (para ofrecer más claridad se ha enventanado un impulso tipo agujero). Como puede comprobarse la aplicación del retardo, ventana a ventana, en el dominio de la frecuencia –multiplicando por $\exp(-j\omega\tau)$ –, introduce cierta distorsión ya que las muestras finales regresan al principio de la ventana. Si el retardo es pequeño y el solapamiento S es grande, esta distorsión será poco apreciable, ya que las muestras temporales “adelantadas” estarán muy atenuadas por la ventana y quedarán enmascaradas por la siguiente ventana que estará muy solapada con la anterior.

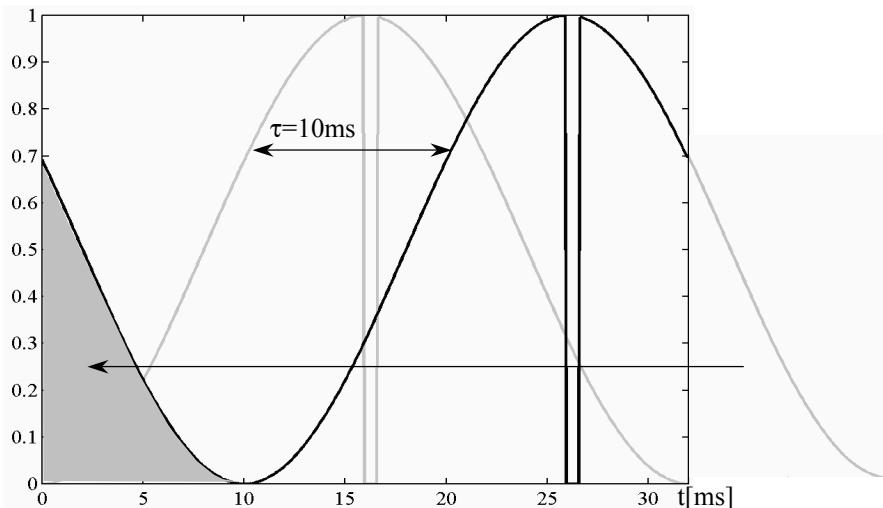


Figura 76. Impulso tipo ranura (trazo gris) enventanado con una ventana Hanning de 32ms ($f_s=16\text{kHz}$ y $L=512\text{pt}$) y esta misma función (trazo negro) aplicando un retardo de 10ms. La aplicación del retardo se ha hecho en el dominio de la frecuencia, multiplicando la FFT correspondiente por $\exp(-j\omega\tau)$. Los 10ms finales de la ventana “retroceden” al principio de dicha ventana.

Si se considera el peor de los casos tratados en los experimentos de esta Tesis, la ventana más corta corresponde a una duración de $T_L = 32\text{ms}$, y el retardo mayor aplicado (ver la Tabla 7) es de $\tau_1 = \tau_{15} = 1.375\text{ms}$, que sólo afecta a 2 de los 15 micrófonos. Este tiempo representa un 4% del tamaño de la ventana, valor que se puede considerar suficientemente pequeño.

Por otra parte, también hay que contemplar el efecto que tiene la compensación de retardo en la reconstrucción perfecta de la señal temporal después del enventanado. En la Figura 77 se reproduce el ejemplo de un experimento como el de la Figura 75, pero en este caso aplicando el retardo máximo de compensación para el conformador, $\tau = 1.375\text{ms}$. Ante una señal de entrada uniforme con un impulso negativo, se producen dos fenómenos reseñables, una reconstrucción imperfecta sólo en el transitorio –Figura 77(a)– o una reconstrucción imperfecta durante todo el tiempo –Figura 77(b)–. El segundo de los defectos es peor ya que puede percibirse continuamente, independientemente de que la señal a enventanar varíe rápidamente cuando hay un transitorio.

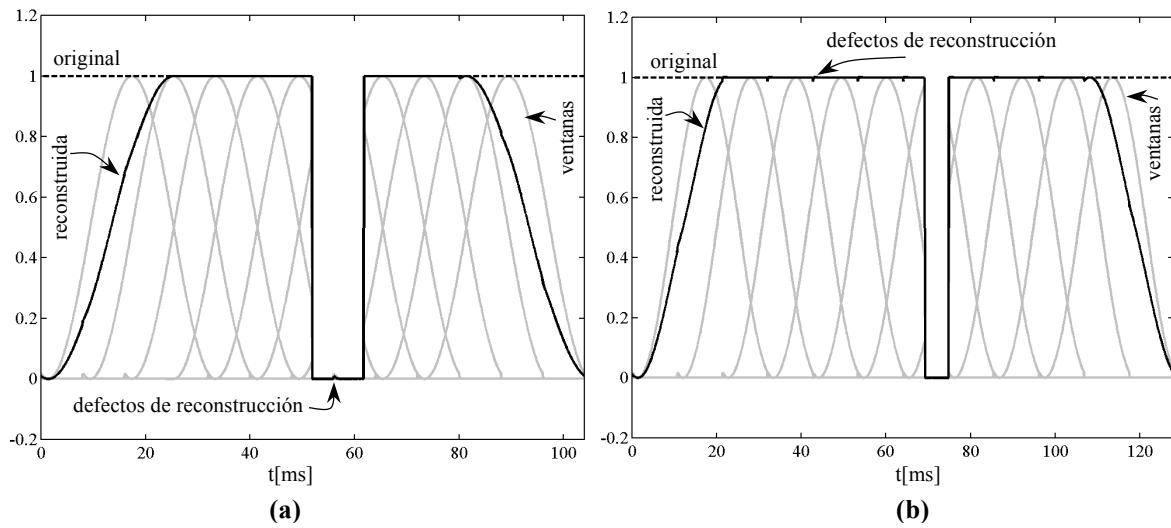


Figura 77. Detalle de reconstrucción de una señal uniforme con un transitorio mediante el proceso enventanado-FFT-IFFT-desenventanado aplicando el retardo máximo del procesador ($\tau=1.375\text{ms}$), con $L=512\text{pt}$ y $f_s=16\text{kHz}$ y utilizando la ventana Hanning. **(a)** $S_1=128\text{pt}$ ($S=75\%$) ventana Hanning simétrica. **(b)** $S_1=171\text{pt}$ ($S=67\%$) ventana Hanning periódica.

En la Tabla 15 se cuantifica el error cometido en el proceso de reconstrucción, para el retardo máximo. Puede concluirse que la reconstrucción con $S = 67\%$ es la peor, con la posibilidad de un sobreimpulso situado a -34dB respecto al original, en el peor de los casos. Por otra parte el solapamiento $S = 75\%$ reconstruye de forma perfecta la señal si ésta es uniforme. Si no lo es, los problemas sólo aparecen en los alrededores del transitorio.

señal original	$S_1 [\text{pt}]/S[\%]$	Hanning	E[dB]
uniforme	171/67	simétrica	-38
transitorio	171/67	simétrica	-34
uniforme	128/75	periódica	-307
transitorio	128/75	periódica	-95

Tabla 15. Efecto promedio de la compensación de retardo máxima del conformador ($\tau=1.375\text{ms}$) sobre la reconstrucción de la señal temporal. ($L=152\text{pt}$, $f_s=16\text{kHz}$). La señal original tipo uniforme es una constante de valor unidad, con un impulso negativo situado en posición aleatoria, como el mostrado en la Figura 77.

Como conclusión a este apartado, se puede decir que los parámetros del enventanado se tienen que elegir de una forma cuidadosa, si se quiere reproducir la señal de salida con la menor distorsión posible. Los mejores enventanados son aquellos en los que el solapamiento S [pt] o el desplazamiento S_1 [pt] es un divisor del tamaño de la ventana L [pt] (por ejemplo $S_1 = 128\text{pt}$ – $S = 75\%$ – ó $S_1 = 256\text{pt}$ – $S = 50\%$ –) aunque se pueden usar otros ($S_1 = 171\text{pt}$ – $S = 67\%$ –) dependiendo de la calidad que se deseé para la señal conformada de salida.

7.1.4 Experimentos y resultados

A continuación se describen con detalle los experimentos realizados mediante simulaciones *software* en esta fase preliminar, y los resultados más relevantes obtenidos con los mismos, en todos los casos llevados a cabo utilizando las propuestas realizadas por el autor.

Pruebas iniciales aplicando MW-MPAP a CMU y simCMU-1.1

Los primeros experimentos llevados a cabo que se reflejan en esta Tesis, consistieron en someter las bases de datos disponibles al procesador multicanal MW-MPAP. Aunque se hicieron pruebas inicialmente con las bases CMU y simCMU-1 conjuntamente, esta última se desechó, ya que por no contener una contaminación de ruido, la eficacia de la parte MW en el resultado era muy pequeña, por lo que se decidió convertir la base simCMU-1 en simCMU-1.1, añadiendo ruido aleatorio y determinista, sin considerar la reverberación del ruido, como se explicó en el punto 6.3.

Después de muchas pruebas informales de escucha subjetiva, se utilizaron finalmente los parámetros del procesador MW-MPAP indicados en la Tabla 16.

enventanado			MW-MPAP									
			MW						MPAP			
			LBF			Coherencia						
L[pt]	N[pt]	S_1 [pt]/S[%]	λ_s	$t_{\Delta s}$ [ms]	λ_N	$t_{\Delta N}$ [ms]	λ_{coh}	$t_{\Delta coh}$ [ms]	α	UC	W_c [pt]	W_c/N
2048	4096	512/75	0.7	90	0.5	46	0.9	304	50	0.85	256	1/16

Tabla 16. Parámetros seleccionados para el procesador MW-MPAP en las pruebas iniciales sobre las bases CMU y simCMU-1.1.

El número de puntos en tiempo L es muy grande (corresponde a 128ms con $f_s = 16\text{kHz}$) para optimizar la reducción de reverberación con la parte MPAP. La razón es que se necesita analizar tramas largas de voz si se quiere reducir suficientemente la reverberación. El parámetro N corresponde al número de puntos en frecuencia a la hora de implementar las FFT's. Como se ve se ha seleccionado un factor de *zero padding* de x 2 para aumentar la resolución en frecuencia. Los parámetros λ de estimación recursiva se han elegido de tal manera que las actualizaciones sean relativamente lentas ($\lambda \approx 1$) con $t_{\Delta s} = 90\text{ms}$ y $t_{\Delta coh} = 304\text{ms}$, excepto para el ruido, con $t_{\Delta N} = 46\text{ms}$ –véase (197)–, el cual necesita una variación más rápida en su estimación. El factor de forma se ha establecido en $\alpha = 50$ y el umbral de coherencia en UC = 0.85. La “cuefrencia” para el *liftering cepstral* corresponde a 1/16 de la cuefrencia máxima, a través del parámetro W_c .

Todos los experimentos descritos en este apartado se han realizado utilizando para el conformador una alineación temporal en el dominio del tiempo. Para conseguir una precisión suficiente en el dominio temporal se ha necesitado una interpolación previa de “x 4”, con el consiguiente diezmado complementario después de la alineación temporal, antes de que la señal entrara en el bloque FFT.

Por otra parte, la segmentación de habla o detección VAD se ha hecho utilizando la señal original de referencia $x_0(t)$ que estaba disponible en todos los fragmentos de prueba. Para ello se ha implementado un sencillo detector de actividad por energía. Este detector asigna a una determinada trama como voz si su potencia supera el 1% del valor de potencia máxima alcanzada en cualquiera de las tramas de todo el fragmento considerado. El funcionamiento del VAD así implementado se ha comprobado como suficientemente correcto, a pesar de su trivialidad. Esto se debe sobre todo a que la señal $x_0(t)$, sobre la que se realiza la detección de actividad de habla, es de muy buena calidad.

En la Figura 78 se representa un fragmento de $y_8(t)$, la señal temporal antes de entrar al procesador, confrontada con la señal de salida del procesador $y_{MW-MPAP}(t)$, correspondientes ambas a un archivo de la base de datos real, CMU. Puede comprobarse cómo la eliminación de ruido es bastante drástica, propiedad que también queda muy patente en las escuchas subjetivas.

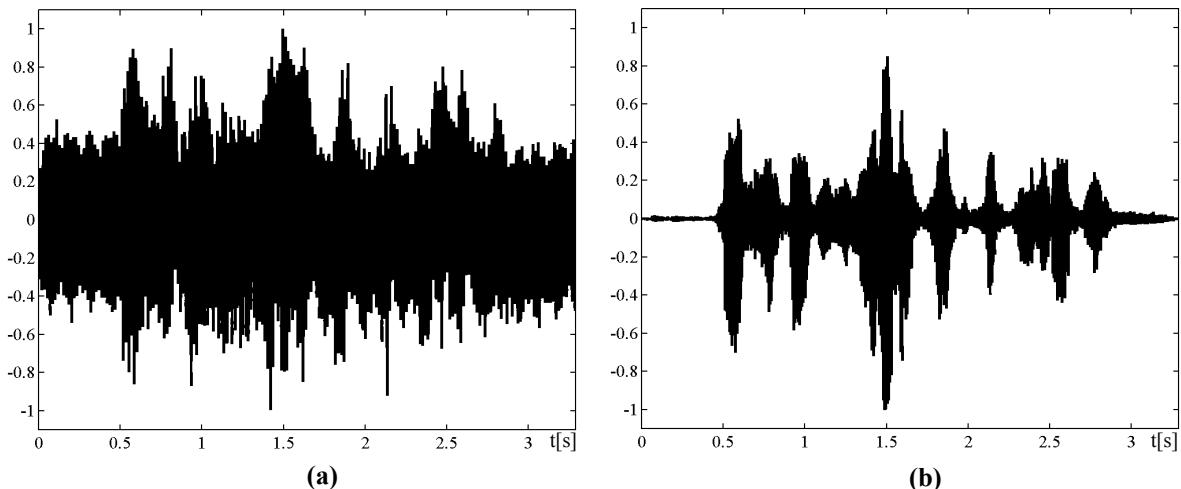


Figura 78. Muestra de señal antes y después del procesador en array MW-MPAP, perteneciente a la base de datos CMU (subcorpus **arr4A**). **(a)** Señal sucia del canal central $y_8(t)$, antes del procesador. **(b)** Señal de salida $y_{MW-MPAP}(t)$, después del procesador.

Efectivamente, se hizo una escucha cuidadosa para una calificación subjetiva de todos los fragmentos analizados, comprobándose cómo la presencia de ruido tipo diente de sierra disminuye muy apreciablemente en la señal procesada. El ruido blanco también queda atenuado, aunque en menor proporción desde un punto de vista subjetivo, debido a su naturaleza aleatoria y su amplio ancho de banda. La señal a la salida del array es más seca en todos los casos analizados, con lo que se pone de manifiesto la reducción de la reverberación producida por el procesador.

En la Figura 79(a) se muestra un ejemplo de la distancia $dLAR_{X,Y_8}(k)$ por trama, entre la señal original y la entrada al procesador comparada con la distancia $dLAR_{X_0,Y_{MW-MPAP}}(k)$, entre la original y la salida del procesador. Si el procesador produce mejora de señal de voz se debe obtener un decremento en estas distancias (GdLAR positiva), como así se manifiesta en la

Figura. En la Figura 79(b) se representa el equivalente en ganancia de distancia dRCEP para el mismo fragmento de voz.

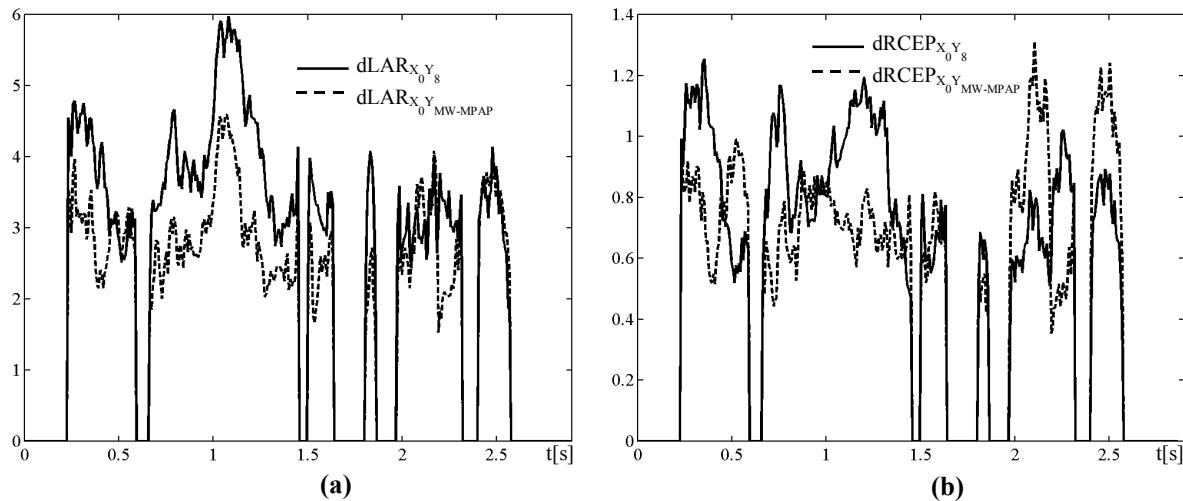


Figura 79. Ejemplo de evaluación de distancias entre la señal del array antes y después del procesador MW-MPAP, con la señal original. Pertenece a la base de datos simCMU-1.1 ($r_0=1m$, $T_{60}=0.8s$, $SNR_8=10dB$, contaminación con ruido blanco). **(a)** Comparación entre la distancia $dLAR(k)$ antes de procesar (entre la señal Y_8 y la original X_0) con la distancia $dLAR(k)$ después de procesar (entre $Y_{MW-MPAP}$ y la original X_0). **(b)** Ídem pero considerando distancias $dRCEP(k)$.

En la Figura 80 se representan los resultados de evaluación de 120 archivos de la base simCMU-1.1. Se promedian fragmentos distintos con iguales parámetros característicos, correspondientes sólo a $SNR_8 = 10dB$ y a $SNR_8 = 20dB$ (es decir, no se ha usado la totalidad de la base de datos). Se ha evaluado la ganancia que produce el procesador en distancias de parámetros LAR, cepstrum real y relación señal a ruido $-GdLAR$ [dB] (247), $GdRCEP$ [dB] (251) y $GSNR$ [dB] (238) (en la banda 20Hz - 20kHz) respectivamente— promediando tan solo las tramas de actividad de voz. Para la estimación de estos parámetros objetivos de mejora se ha usado un enventanado de $L = N = 512pt$ con solapamiento de $S = 75\%$ (compruébese que estos valores de los parámetros utilizados para el análisis de los resultados son diferentes a los usados para el procesador). Se han empleado además 18 parámetros LAR y 18 puntos (excluyendo el primero) de cepstrum real. La relación señal a ruido del canal central del array (SNR_8) y la ganancia en relación señal a ruido $GSNR$ se ha evaluado a posteriori usando (241) y no (227), es decir midiendo la potencia de la señal en las tramas de voz y del ruido en las tramas de no-actividad de voz, a partir del canal central del array, $y_8(t)$. Esa forma de determinación de la SNR es lo que se denominó evaluación tradicional en el capítulo 5.1 de esta Tesis y tiene la gran desventaja de que sólo considera el ruido presente en los blancos de voz, sin importar lo que sucede en las tramas con presencia de señal de habla. Las ganancias en parámetros objetivos han sido calculadas utilizando como referencia la señal limpia, $x_0(t)$, que siempre está disponible en las bases de datos.

En la Figura 81 se proporcionan los resultados del mismo tipo de evaluación para la base de datos real CMU. En ella se muestran los promedios de 15 archivos, cinco para cada subcorpus.

Los resultados de las Figuras 80 y 81 muestran, a partir de las medidas de $GdLAR$ y $GdRCEP$, un empeoramiento por parte del procesador sobre la señal de salida del array (valores negativos en las ganancias). Estos resultados no coinciden con la valoración subjetiva de cada uno de los fragmentos de CMU y simCMU-1.1 sometidos a prueba, por lo que los

estimadores basados en distancias de parámetros LAR y cepstrales serán reconsiderados más adelante.

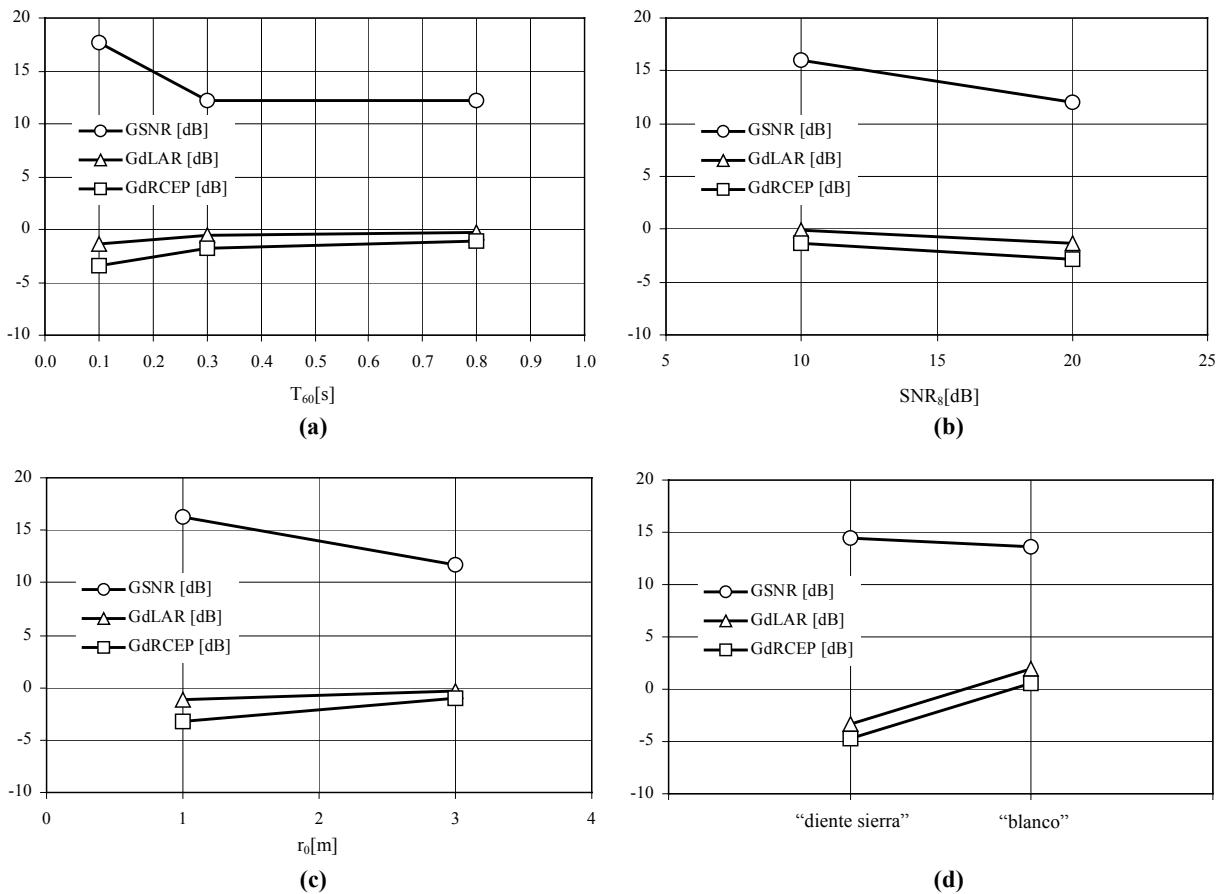


Figura 80. Resultados de mejora de mejora de habla utilizando las ganancias, GSNR [dB], GdLAR [dB] y GdRCEP [dB] a partir de la base de datos simCMU-1.1. **(a)** En función del tiempo de reverberación, T_{60} . **(b)** En función de la SNR_8 de entrada, en el canal 8 del array. **(c)** En función de la distancia r_0 de la fuente al array. **(d)** En función del tipo de ruido aditivo.

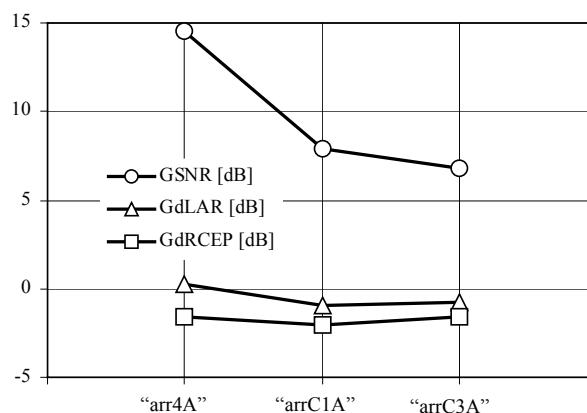


Figura 81. Resultados de mejora, GSNR [dB], GdLAR [dB] y GdRCEP [dB] utilizando la base de datos real CMU.

Por otra parte la ganancia GSNR es siempre positiva y bastante elevada, resultado más de acuerdo con la impresión subjetiva. Sin embargo la forma de obtener SNR en GSNR es un poco burda –recuérdese, a partir de (241)–, ya que sólo considera la disminución del ruido en los blancos de voz y no mide la derreverberación o la mejora obtenida en las tramas con actividad de voz.

Echando una mirada rápida a las figuras, parece que las mejoras en GdLAR y GdRCEP crecen cuando las condiciones de la señal se hacen peores: aumenta T_{60} –Figura 80(a)–, disminuye SNR_8 –Figura 80(b)–, aumenta r_0 –Figura 80(c)–, o se añade ruido blanco –Figura 80(d)–. En la Figura 81, para la base CMU, puede verse que las mejoras mayores son también para los subcorpus con menor calidad a priori, bien sea por excesiva cantidad de ruido (subcorpus **arr4A**) o por una reverberación mayor (subcorpus **arrC3A**). La ganancia GSNR también mejora cuando disminuye la relación SNR_8 –Figura 80(b) y Figura 81– aunque muestra tendencias contrarias a GdLAR y GdRCEP en los demás casos de la Figura 80.

A la vista de los resultados de la Figura 80(d), donde se consiguen ganancias positivas (aunque muy ligeras) en GdLAR y GdRCEP cuando el ruido aditivo suministrado era de naturaleza aleatoria, que quizás corresponda más con las condiciones reales (ruido y reverberación difusos), se decidió repetir las pruebas de medidas de parámetros objetivos eliminando los elementos de simCMU-1.1 portadores de contaminación con ruido coherente de tipo diente de sierra, consigiéndose los resultados que se muestran en las Tablas 17, 18 y 19. Los experimentos que dan origen a estos resultados han sido publicados en [González-Rodríguez 00]. Los resultados aparecidos en [González-Rodríguez 00] han sido adaptados para considerar las medidas de ganancias de forma logarítmica –GdLAR [dB] (247), GdRCEP [dB] (251) y GSNR [dB] (238)–.

En las Tablas 17, 18 y 19 se ha intentado no mezclar condiciones distintas de contaminación. Así la variación con T_{60} se considera sólo con $SNR_8 = 10\text{dB}$ (Tabla 17), la variación con SNR_8 (Tabla 18) y con r_0 (Tabla 19) se considera sólo un $T_{60} = 0.8\text{s}$. Ahora todas las ganancias son positivas, aunque GdLAR y GdRCEP son pequeñas. La tendencia de GdLAR ahora se manifiesta paralela a GSNR, no siendo así lo que sucede con GdRCEP en todas las ocasiones.

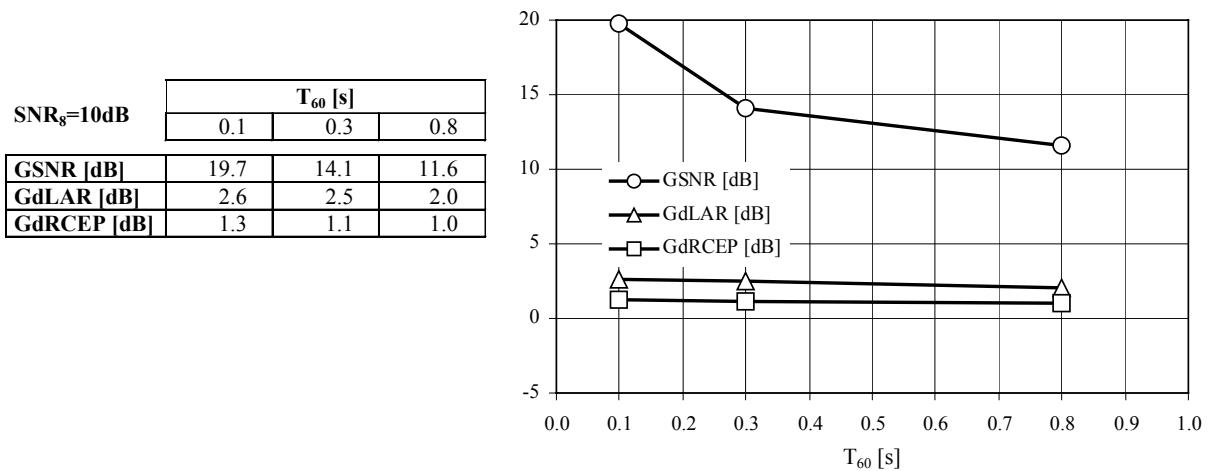


Tabla 17. Resultados parciales del procesador MW-MPAP en función del tiempo de reverberación T_{60} , con $SNR_8=10\text{dB}$, adaptados de [González-Rodríguez 00]. Sólo se considera contaminación con ruido blanco. Se mide la mejora de la relación señal a ruido SNR –GSNR [dB]–, distancia en parámetros LAR con la señal original –GdLAR [dB]– y distancia en coeficientes de cepstrum real con la señal original –GdRCEP [dB]–.

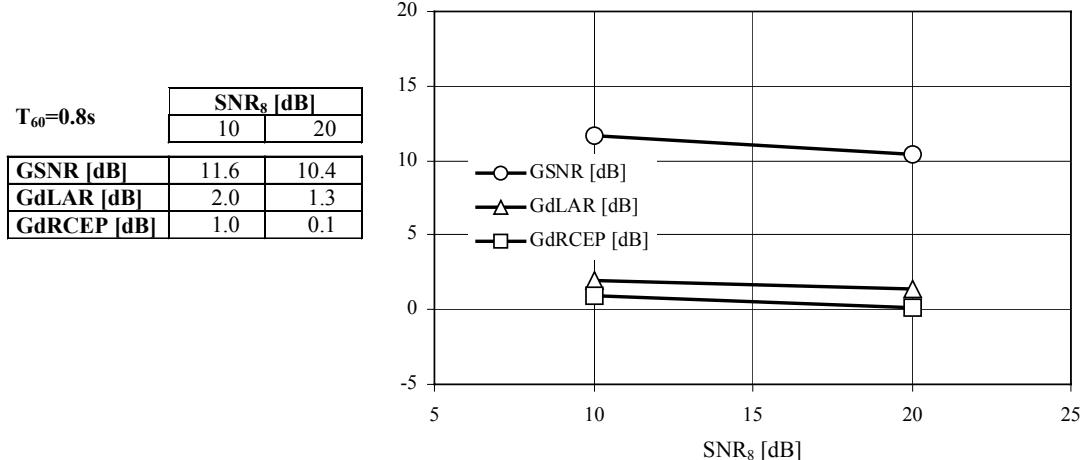


Tabla 18. Resultados parciales del procesador MW-MPAP en función de SNR_8 a la entrada ($T_{60}=0.8\text{s}$), medida en el canal 8 del array, adaptados de [González-Rodríguez 00] (sólo contaminación con ruido blanco).

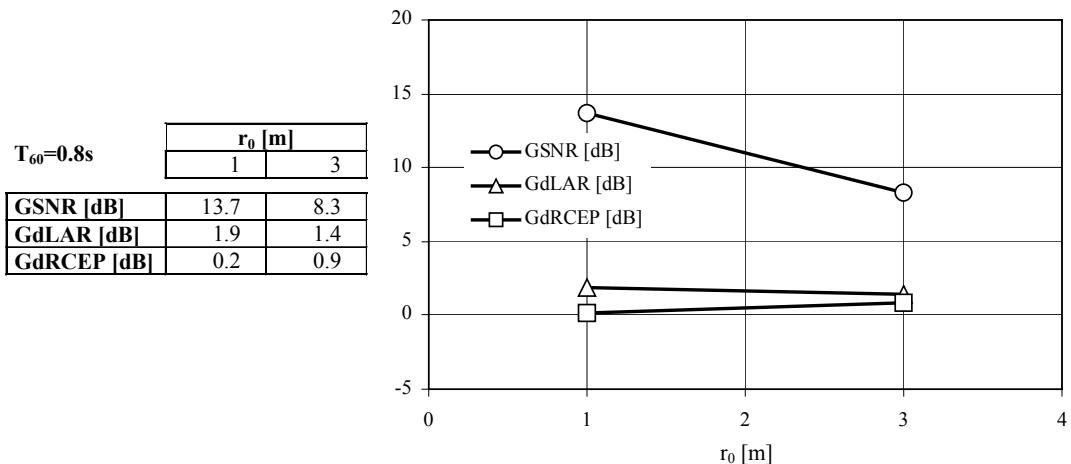


Tabla 19. Resultados parciales en función de la distancia fuente-array microfónico, adaptados de [González-Rodríguez 00] para $T_{60}=0.8\text{s}$ (sólo contaminación con ruido blanco).

Como conclusión de todo lo expresado antes, parece inferirse que los estimadores basados en parámetros LAR y RCEP son un poco erráticos a la hora de evaluar las habilidades del procesador en array. Desde luego, en las condiciones de trabajo utilizadas, no miden bien la reducción de ruido en las pruebas, hecho muy evidente en una apreciación subjetiva si el ruido es determinista y de banda estrecha. En este caso parece que consideran más la propia distorsión introducida por el procesador que la reducción de ruido. Los resultados obtenidos con GdLAR y GdRCEP sí parece que mejoran en caso de evaluar un ruido más realista de banda ancha. Sin embargo, tampoco se manifiestan muy eficaces en la detección de la reverberación. En esta última consideración hay que tener presente que en la base de datos simCMU1.1 el ruido se ha añadido sin reverberación.

Pruebas iniciales aplicando MW-MPAP a CMU y simCMU-2

Como se ha mostrado en el punto anterior, las medidas de reducción de ruido con la base CMU a partir de la mejora de las distancias de parámetros objetivos han dado resultados que no se corresponden en todos los casos con la apreciación subjetiva observada. Por ello se consideró crear otra base de datos más realista (simCMU-2) donde el ruido añadido (el ruido aleatorio y determinista ya considerados en simCMU-1.1) también sufran los efectos de la reverberación del recinto. Según se expone en el punto 6.3, la base de datos simCMU-2 está generada desde 8 puntos (P_1 a P_8 , véase la Figura 69) en un recinto de dimensiones y reverberación típicas, que pueden ser las de una sala de conferencias estándar. En los experimentos desarrollados a continuación, se refinaron aun más los parámetros asociados al procesador MW-MPAP de la Figura 73, para conseguir mejores resultados, así como los estimadores de parámetros objetivos para evaluar la calidad de la señal a la salida del array microfónico, de tal manera que se consiguiese un mayor paralelismo con la apreciación subjetiva de mejora de habla. Un extracto de los experimentos y resultados reflejados en este punto ha sido publicado en [Sánchez-Bote 00].

Experimentos con el tamaño de la ventana temporal

Se probó exclusivamente el conformador cepstral aplicando la parte MPAP del procesador en array a señales multimicrófono con sólo reverberación. Para ello se utilizó el subcorpus **rev** de la Tabla 13. Se varió el tamaño de la ventana L , aplicándose diferentes “cuerfencias” de corte cepstral, determinadas por el parámetro W_C , como se manifiesta en la Tabla 20. Aquí se sigue aplicando un número de puntos en frecuencia de $N = 2 \times L$.

L [pt]	N [pt]	W_C [pt]				
		N/16	N/12	N/8	N/6	N/4
512	1024	64	85	128	171	256
1024	2048	128	171	256	341	512
2048	4096	256	341	512	683	1024
4096	8192	512	683	1024	1365	2048
8192	16384	1024	1365	2048	2731	4096
16384	32768	2048	2731	4096	5461	8192
32768	65536	4096	5461	8192	10923	16384

(a)

L [pt]	N [pt]	W_C [pt]			
		256	512	768	1024
		N/W_C			
512	1024	4	2	4/3	1
1024	2048	8	4	8/3	2
2048	4096	16	8	16/3	4
4096	8192	32	16	32/3	8
8192	16384	64	32	64/3	16
16384	32768	128	64	128/3	32
32768	65536	256	128	256/3	64

(b)

Tabla 20. Variación del tamaño de la ventana temporal L , de la ventana en frecuencia N y del corte cepstral W_C para el *liftering* paso bajo en las pruebas del procesador MPAP solo, sobre el subcorpus **rev** de simCMU-2. (a) W_C porcentual. (b) W_C fijo.

Una vez procesado todo el subcorpus **rev** con el derreverberador, aplicando las combinaciones de la Tabla 20, se ha llegado a las siguientes conclusiones:

- un número de puntos L grande produce una mayor sensación subjetiva de derreverberación,
- el parámetro de corte cepstral W_C tiene que ser lo menor posible para aumentar la derreverberación pero no puede disminuir fuertemente sin introducir distorsión apreciable en la señal conformada de salida.

En este sentido, se encontró que las combinaciones más adecuadas en los parámetros de tamaño de trama para el procesador MPAP son de $L = 4096\text{pt}$ y $W_C = 256\text{pt}$ y de $L = 8192\text{pt}$ y $W_C = 768\text{pt}$.

Comparación del conformador cepstral MPAP en tres bandas con el conformador convencional en tres bandas

Ya se sabe que el conformador MPAP realiza una conformación de haz en el dominio cepstral en tres bandas, e incluye el *liftering* paso bajo para aumentar la derreverberación. Para estimar la cantidad extra de derreverberación que produce el procesador MPAP, se comparó éste con el conformador convencional en tres bandas mediante retardo y suma en el dominio de la frecuencia, utilizando el subcorpus **rev**. A partir de apreciaciones subjetivas de los resultados del array, se concluye que el derreverberador MPAP produce una mejora ligera, en cuanto a derreverberación que el conformador convencional, aunque con una distorsión (percibida de forma subjetiva) ligeramente superior.

Comparación del procesador MPAP 1-banda con el MPAP 3-bandas

Se ha comparado el conformador cepstral MPAP en una sola banda $B_T = [20\text{Hz}, 8\text{kHz}]$ con el mismo conformador operando en tres subbandas B_1 , B_2 y B_3 . Para ello se ha realizado la medida de parámetros objetivos de mejora (GdLAR, GdRCEP y GSNR) sobre el subcorpus **rev** de la base simCMU-2 (sólo reverberación con señal desde los 8 puntos de test). En estas pruebas, al igual que las realizadas con simCMU-1.1, la ganancia en relación señal a ruido, GSNR, se ha obtenido mediante la estimación tradicional de SNR de (241), con lo que la reverberación del subcorpus **rev** se considera como ruido cuando ésta se introduce en las tramas temporales de ausencia de actividad de voz, calificadas así mediante el análisis VAD en la señal de referencia $x_0(t)$.

Los resultados promedio se muestran en las Figuras 82 y 83. Han sido obtenidos con $L = 4096\text{pt}$ y $W_C = 300\text{pt}$ para 3-bandas y con $L = 8192\text{pt}$ y $W_C = 800\text{pt}$ para 1-banda. Puede comprobarse cómo (a diferencia de las pruebas hechas con simCMU-1.1) se obtienen mejoras en cualquiera de los casos evaluados, siendo mayores las correspondientes al conformador en tres bandas. Esto tiene sentido, ya que el conformador en tres bandas evita los problemas de conformación relatados en el capítulo dedicado a arrays lineales (Parte 1 de esta Tesis, punto 2.1.8), como el *aliasing* espacial en alta frecuencia y la baja directividad en baja frecuencia. Consecuentemente parece que hay que desechar la conformación monobanda, a costa de una mayor complejidad del procesador, para aprovechar las ventajas del array anidado en tres subbandas.

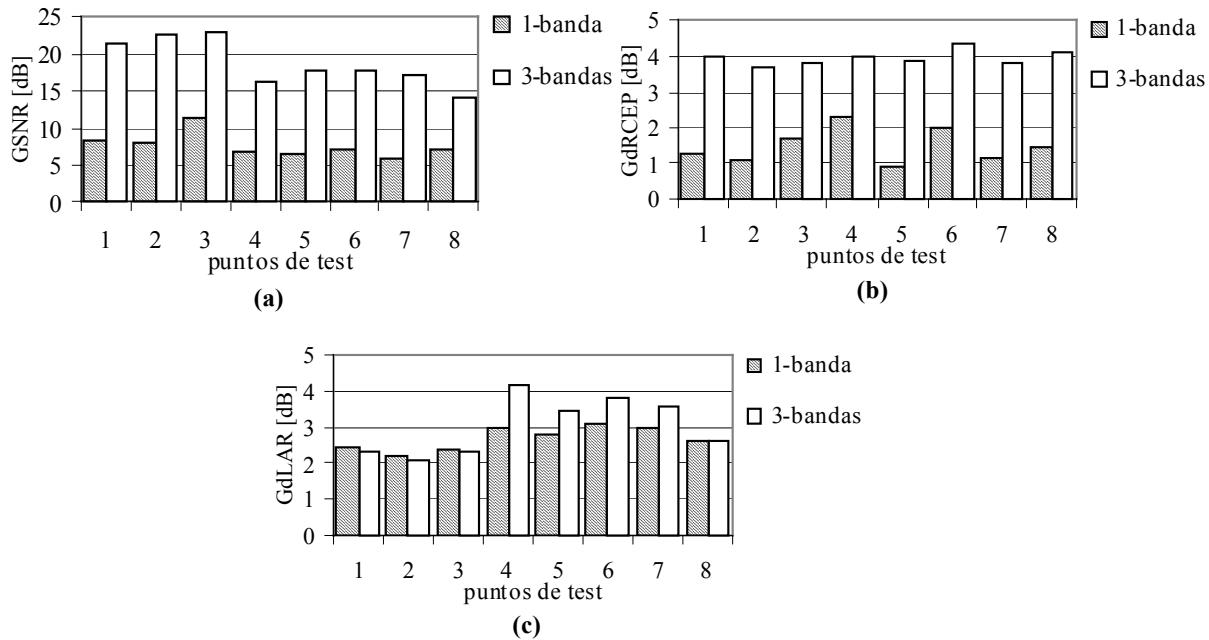


Figura 82. Medida de parámetros objetivos de mejora sobre el subcorpus **rev** de la base simCMU-2 (sólo reverberación con señal desde los 8 puntos de test) según aparece en [Sánchez-Bote 00]. Se comparan el conformador cepstral MPAP en una sola banda $B_T=[20\text{Hz}-8\text{kHz}]$ con el mismo procesador aplicando una descomposición en tres subbandas B_1 , B_2 y B_3 . **(a)** GdLAR [dB]. **(b)** GdRCEP [dB]. **(c)** GSNR [dB].

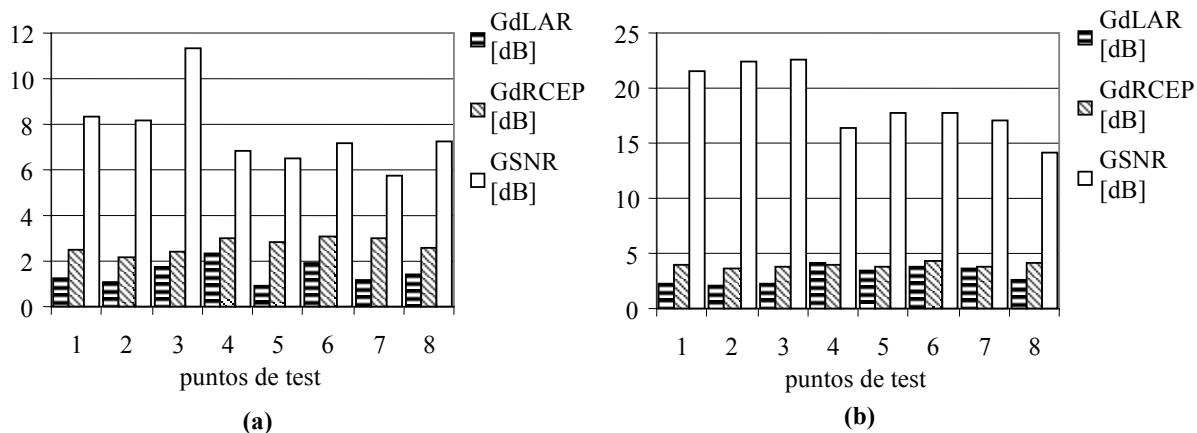


Figura 83. Medida de parámetros objetivos de mejora sobre el subcorpus **rev** de la base simCMU-2 (sólo reverberación con señal desde los 8 puntos de test). **(a)** Conformador cepstral MPAP en una sola banda $B_T=[20\text{Hz}-8\text{kHz}]$. **(b)** Conformador cepstral MPAP en tres subbandas B_1 , B_2 y B_3 .

Medidas objetivas sobre los subcorporus **dif**, **Mdif** y **Mcoh** de la base simCMU-2 y sobre CMU con el procesador final MW-MPAP, utilizando conformación en tres bandas

Una vez verificada la configuración más conveniente del derreverberador cepstral se hicieron medidas sobre la base de datos simCMU-2 (subcorpus **dif**, **Mdif** y **Mcoh**, Tabla 13), incluyendo todos los fragmentos de voz multicanal con ruido y reverberación. Para estas pruebas se ha utilizado el procesador completo MW-MPAP de la Figura 73, que incluye la parte de reducción de ruido mediante Wiener modificado por coherencia.

Como la ventana seleccionada para el procesador cepstral es muy grande (4096pt), la actualización del filtro de Wiener trama a trama es muy lenta, con lo que se ha decidido calcular dos tramas de filtrado de Wiener por cada trama del procesador MPAP, obteniéndose dos ventanas temporales de diferente tamaño, $L_1 = 2048\text{pt}$ (128ms) para el procesador MW y $L_2 = 4096\text{pt}$ (256ms) para el procesador MPAP. En ambos casos se ha considerado un solapamiento de $S = 75\%$.

Los parámetros seleccionados para el procesador ajustado finalmente se muestran en la Tabla 21.

enventanado				MW-MPAP									
MW		MPAP		MW						MPAP			
				LBF			Coherencia						
$L_1[\text{pt}]$	$N_1[\text{pt}]$	$L_2[\text{pt}]$	$N_2[\text{pt}]$	λ_s	$t_{\Delta s} [\text{ms}]$	λ_N	$t_{\Delta N} [\text{ms}]$	λ_{coh}	$t_{\Delta coh} [\text{ms}]$	α	UC	$W_c [\text{pt}]$	W_c/N_2
2048	4096	4096	8192	0.8	143	0.8	143	0.8	143	50	0.7	200	0.02

Tabla 21. Parámetros seleccionados en el procesador MW-MPAP para las pruebas iniciales sobre las bases CMU y simCMU-2. Se indican los dos enventanados diferentes para la parte MW y MPAP del procesador.

Los resultados obtenidos, promediados sobre simCMU-2 se muestran en las Figuras 84 y 85 y los obtenidos sobre CMU se representan en la Figura 86. Hay que resaltar que todos los resultados del procesador han sido evaluados también subjetivamente, obteniéndose una apreciación muy buena de la salida procesada del array.

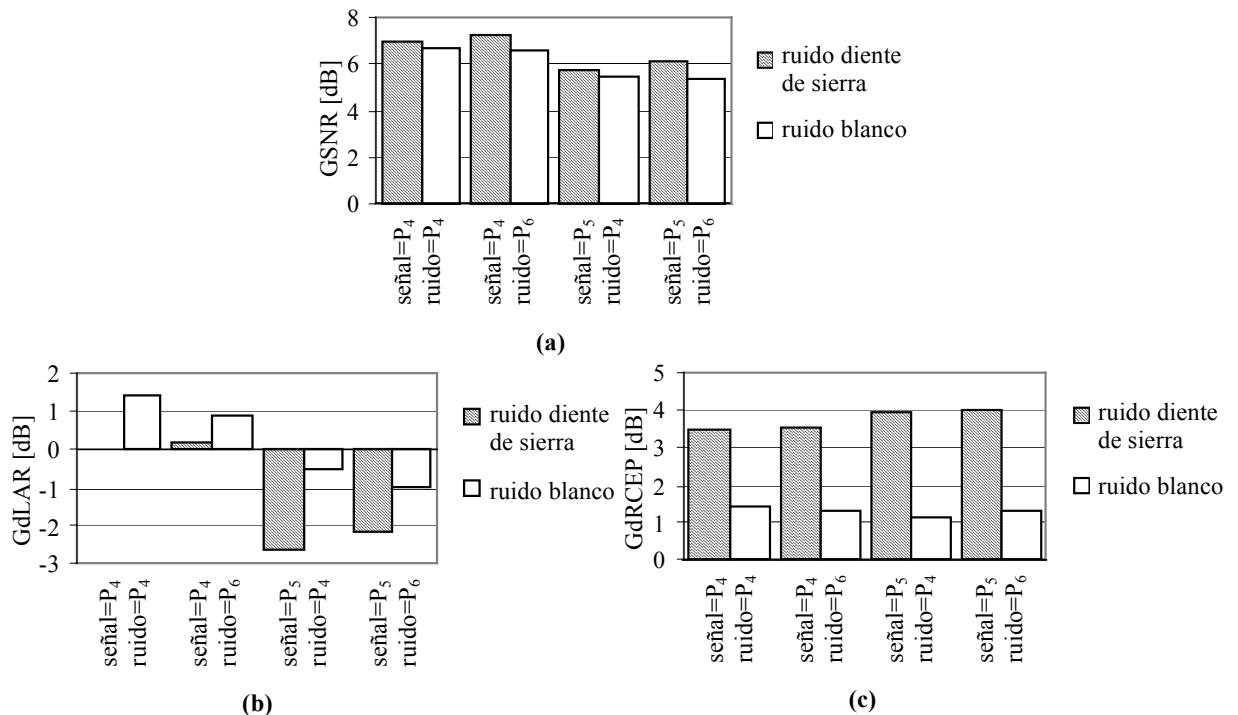


Figura 84. Resultado de la medida de parámetros objetivos sobre la salida del procesador MW-MPAP utilizando la base de datos simCMU-2. Se muestra el promedio de una selección de experimentos, como aparece en [Sánchez-Bote 00]. (a) GSNR [dB]. (b) GdLAR [dB]. (c) GdRCEP [dB].

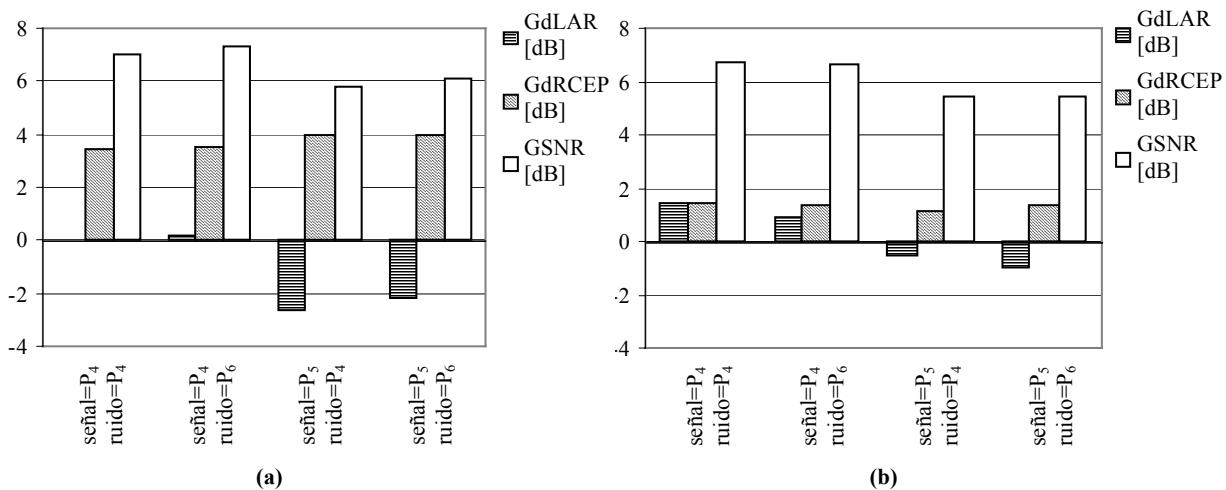


Figura 85. Resultado de la medida de parámetros objetivos sobre el procesador MW-MPAP utilizando la base simCMU-2, en función del tipo de ruido añadido. **(a)** Ruido diente de sierra. **(b)** Ruido blanco.

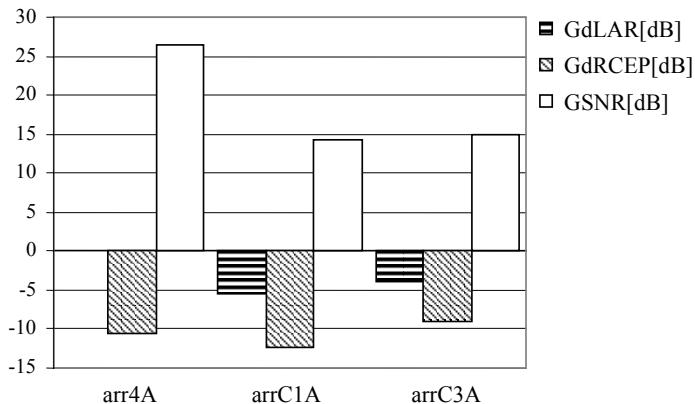


Figura 86. Resultado de la medida de parámetros objetivos sobre el procesador MW-MPAP utilizando la base de datos real CMU.

A la vista de los resultados mostrados se pueden obtener las siguientes conclusiones:

- Los resultados de mejora son superiores cuando la señal está contaminada por ruido de tipo “diente de sierra”, excepto cuando se mide GdLAR, ya que este ruido conduce a resultados no acordes con la apreciación subjetiva. Esto ya se había notado anteriormente y se debe a que para una señal de banda estrecha los estimadores dLAR, en este caso, atienden más a la distorsión propia del conformador, en las frecuencias en las que el ruido está ausente, que a la mejora de señal en las bandas donde el ruido está presente. Entonces, las medidas objetivas con ruido de banda ancha son más adecuadas.
- Las mejoras estimadas mediante GdLAR y GdRCEP en la base de datos CMU son negativas (resultado que no acorde con la apreciación subjetiva), con lo cual se pone en duda la eficacia de los estimadores GdLAR y GdRCEP en señales reales, ya que reflejan más la distorsión del procesador que la apreciación subjetiva de mejora.
- Los mejores resultados (Figura 85) corresponden a la fuente situada frontalmente (posición *broadside*) y el ruido situado lateralmente (posición *endfire*). Esta conclusión es muy razonable ya que todas las bases de datos evaluadas aquí están captadas con micrófonos directivos apuntando hacia $\theta = 90^\circ$. A pesar de ello, no se puede establecer ninguna conclusión relevante acerca de cuáles son las condiciones geométricas de situación de la señal o del ruido que producen unos resultados más favorables.

7.2 ESTIMACIÓN OBJETIVA DE MEJORA DE SEÑAL DE HABLA EN PRESENCIA DE RUIDO Y REVERBERACIÓN MEDIANTE EL MÉTODO E-RASTI (RASTI EMULADO)

Hasta ahora se ha evaluado la calidad de señal de voz utilizando una estimación de parámetros objetivos basados en relación señal a ruido y en parámetros LAR y cepstrales. Se ha visto que en algunos casos, el empleo de estos métodos sobre voz contaminada con ruido y reverberación no ofrece una calificación acorde a la impresión subjetiva de la mejora producida. Es decir, fragmentos de voz mejorados con el procesador son a veces considerados de peor calidad que la señal sucia a la entrada del mismo. El estimador usado hasta ahora basado en la mejora de relación señal a ruido es demasiado burdo –por estimar el ruido mediante (243)– y solo califica la limpieza en las tramas de no-actividad de voz. Es necesario idear un método de evaluación objetiva más acorde con la impresión subjetiva.

Como ya se explicó en el punto 5.4, el índice STI, y por extensión el índice RASTI, se ha utilizado con éxito en el mundo de la acústica de salas y la electroacústica para evaluar las pérdidas de inteligibilidad en señales de voz cuando están contaminadas por ruido y reverberación. Aunque el STI se ideó inicialmente estudiando la degradación mensurable de las características de la señal vocal, pronto derivó a métodos de cálculo que utilizaban señales sustitutas de la voz con características más controlables (la señal RASTI, x_{RASTI} , de la Figura 54), u otros basados en la estimación directa de las pérdidas de modulación mediante el análisis de la función de transferencia o la respuesta al impulso de la transmisión electroacústica (sala + ruido + altavoces + micrófonos). Estos métodos tradicionales de cálculo del RASTI no son aplicables generalmente para evaluar la voz procesada como se hace en las propuestas de esta Tesis. Por una parte la señal RASTI no sería tratada convenientemente por la mayoría de los procesadores en array propuestos en esta Tesis (sobre todo la sección de postfiltrado MW que necesita tramas de ausencia de voz para estimar el ruido) y por otra la función de transferencia del procesador en array es en su mayor parte no lineal, con lo que no se pueden usar los métodos tradicionales de cálculo de STI o RASTI que utilizan dicha función de transferencia. Por tanto ha sido necesario volver a los orígenes del STI para que pueda ser aplicado directamente sobre señal de voz y no sobre la señal RASTI. Es lo que aquí se propone como método E-RASTI y que ha sido desarrollado en [Sánchez-Bote 01-a].

7.2.1 Descripción del método E-RASTI propuesto

La idea básica del método E-RASTI consiste en calcular las pérdidas de modulación del habla sobre la señal de voz y a partir de este punto aplicar el método RASTI directamente. En la Figura 87 se representa el esquema de operación con el que se calcula el índice E-RASTI asociado a la mejora existente entre dos fragmentos de voz. En esa figura, se llama $x(t)$ a la señal de voz mejorada (entrada para el procesador E-RASTI) e $y(t)$ a la señal de voz sucia (salida para el procesador E-RASTI). La señal $x(t)$ será normalmente la salida del procesador en array, aunque también puede ser la señal de referencia original, sin ruido ni reverberación, que se ha llamado $x_0(t)$ a lo largo de la Tesis. Por tanto, el índice E-RASTI proporciona la tasa de mejora, en un baremo entre 0 y 1 de la entrada $x(t)$ con respecto a la salida $y(t)$. Si existe mucha diferencia de modulación entre ambas señales, el índice E-RASTI será pequeño y viceversa. Es decir, utilizado para evaluar un determinado procesador de voz, es beneficioso que el índice E-RASTI asociado al proceso entrada-salida del mismo, sea bajo (próximo a 0).

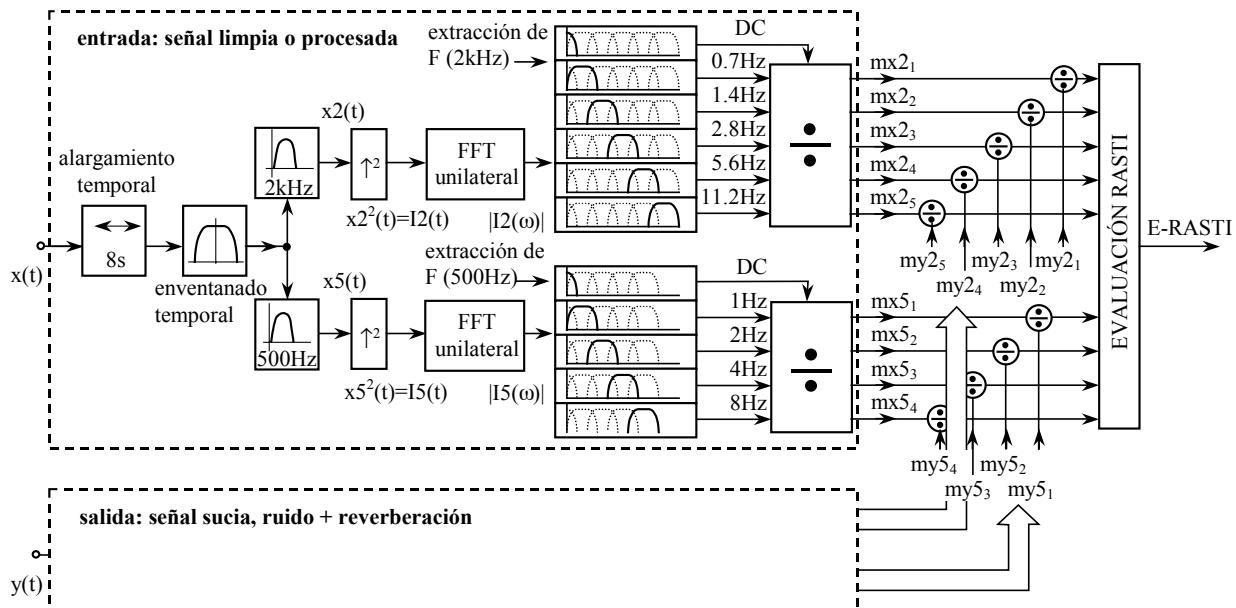


Figura 87. Diagrama de bloques del método E-RASTI para estimar la mejora de una señal de entrada $x(t)$ con respecto a una señal de salida $y(t)$.

A continuación se detalla cada uno de los bloques que componen el evaluador E-RASTI de la Figura 87.

Alargamiento temporal

Este bloque evalúa la longitud del fragmento de voz de entrada. Si es menor de 8s lo replica hasta obtener la duración anterior. Es necesario disponer de al menos 8s de voz para apreciar con suficiente resolución espacial el espectro de intensidad de voz –dado por el cuadrado $x^2(t)$ – hasta una frecuencia de unos 15Hz.

Enventanado temporal

Se utiliza una ventana Hanning para suavizar el efecto borde del fragmento de señal adquirido.

Filtrado paso banda

Se usan filtros de octava para seleccionar dos fragmentos del espectro, uno centrado en 500Hz y otro en 2kHz. Estas dos bandas representan las frecuencias más características asociadas a la inteligibilidad del habla. A su salida son obtenidas por separado las señales $x2(t)$ –2kHz– y $x5(t)$ –500Hz–.

Potenciación

Las señales temporales se elevan al cuadrado para convertirlas en un equivalente de la intensidad sonora $I_2(t)$ –2kHz– e $I_5(t)$ –500Hz–. Estas intensidades sonoras ya contienen el espectro de muy baja frecuencia donde se pueden apreciar las pérdidas de modulación.

FFT unilateral

Se realiza una FFT del fragmento de intensidad de 8s y se calcula su espectro unilateral, es decir se duplica la amplitud de las frecuencias mayores que 0Hz. Véase un ejemplo en la Figura 88.

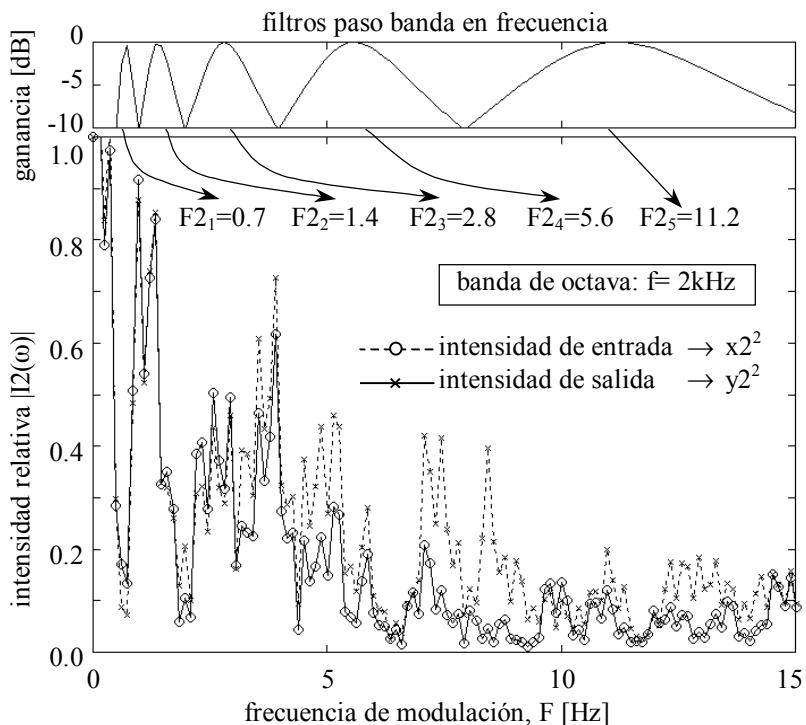


Figura 88. Ejemplo del espectro de intensidad de baja frecuencia $|I_2(\omega)|$ en la banda de octava de 2kHz con los filtros paso banda utilizados para extraer los índices de modulación m . Se compara la intensidad de entrada correspondiente a la señal procesada por el array de micrófonos con la señal de salida, que en este caso corresponde a la voz sucia extraída del canal 8 del array.

banda de octava: $f=500\text{Hz}$						
Frec. modulación. [Hz]	$F_{5_0}=0$	$F_{5_1}=1$	$F_{5_2}=2$	$F_{5_3}=4$	$F_{5_4}=8$	
banda de octava: $f=2\text{kHz}$						
Frec. modulación. [Hz]	$F_{2_0}=0$	$F_{2_1}=0.7$	$F_{2_2}=1.4$	$F_{2_3}=2.8$	$F_{2_4}=5.6$	$F_{2_5}=11.2$

Tabla 22. Frecuencias de modulación F para el índice RASTI.

Extracción de las frecuencias de modulación F

Se hace un filtrado paso banda en el dominio de la frecuencia para extraer las componentes espectrales de la intensidad, centradas en las frecuencias de modulación F y

definidas en el método RASTI (Tabla 22). También se extrae la intensidad alrededor de 0Hz (continua o DC). En total se calculan 5 + 1 filtros (5 F's + 1DC) para 2kHz y 4 + 1 filtros para 500Hz.

Cálculo de los índices de modulación m

Se compara el nivel de señal de intensidad de cada banda F con el nivel de continua de cada banda f, para obtener 9 índices de modulación m, 5 para la banda de 2kHz ($mx_{2_1} - mx_{2_5}$) y 4 para la de 500Hz ($mx_{5_1} - mx_{5_4}$). Cada uno de los índices m (comprendido entre 0 y 1) se divide por el homólogo correspondiente a la señal sucia y(t) ($my_{2_1} - my_{2_5}$) y ($my_{5_1} - my_{5_4}$), resultando los 9 índices de modulación del RASTI: ($m_{2_1} - m_{2_5}$) para la banda de 2kHz y ($m_{5_1} - m_{5_4}$) para la de 500Hz. Con estos coeficientes, utilizando el método descrito en el apartado 5.4 de esta Tesis, se calcula finalmente el índice E-RASTI.

7.2.2 Experimentos de validación del método E-RASTI

Aunque más adelante, en el capítulo 9 de la Tesis, se harán pruebas de verificación de un nuevo procesador utilizando el método E-RASTI, ha sido necesario comprobar la validez de dicho método en condiciones controladas de contaminación acústica, cuando se aplica sobre una señal RASTI o cuando se hace sobre fragmentos de voz real. Para ello se han implementado dos tipos de experimentos, que se describen a continuación.

Experimentos con señal RASTI

La idea básica consiste en comprobar si los resultados conocidos que produce el RASTI (ó STI) en condiciones conocidas de ruido y reverberación se reproducen cuando se utiliza el método propuesto. Recuérdese que el método RASTI se aplica en la actualidad para calificar la inteligibilidad de una sala excitando acústicamente con una señal prefabricada, que se ha llamado señal RASTI y cuya generación se describió anteriormente (x_{RASTI} en la Figura 54). Para validar el método, éste se ha probado inicialmente sobre la señal x_{RASTI} contaminada con ruido y reverberación, de tal manera que se pueda verificar que se confirman los resultados previstos por la teoría del STI. Para ello se ha generado, según el procedimiento descrito en la Figura 54, una señal RASTI de aproximadamente 8s, que ha sido contaminada con ruido y reverberación. El primero ha sido añadido sumando un ruido rosa a dicha señal RASTI, de tal manera que se consiga determinada SNR. Esta SNR ha sido medida considerando la energía total de la señal RASTI limpia y del ruido añadido. La reverberación se ha producido convolucionando la señal con las funciones de transferencia $h_{P1}(t) - h_{P8}(t)$ del recinto con el que se generó simCMU-2 (Figura 69). Como se sabe, este recinto tiene un tiempo de reverberación estimado (a través de la respuesta al impulso) de $T_{60} \approx 0.7s$.

Con este tipo de señal RASTI contaminada por ruido y reverberación, se han hecho dos tipos de experimentos. En primer lugar se ha aplicado el método E-RASTI propuesto para comparar la señal original y la contaminada por ruido o reverberación. En segundo lugar se ha probado la efectividad del derreverberador cepstral (procesador MPAP descrito en el apartado 7.1 de esta Tesis) aplicado también sobre la señal E-RASTI, de tal manera que puedan verificarse las mejoras subjetivas que produce el array de micrófonos. Se ha evaluado sólo la reverberación porque en principio es la más difícil de detectar objetivamente por los métodos utilizados con anterioridad, sobre todo los basados en distancias LAR y cepstrales.

En la Figura 89 re representa una muestra de funcionamiento del método E-RASTI ante la presencia de reverberación. Aparece el espectro de intensidad sonora ($f = 2\text{kHz}$) de la señal RASTI original y de la señal RASTI con reverberación (sin ruido añadido), después de haber sido convolucionada con la respuesta al impulso de una sala con $T_{60} \approx 0.7\text{s}$. Puede apreciarse cómo el resultado corresponde a lo que predice la teoría del STI. Es decir, ante la presencia de reverberación, la señal vocal (en este caso la señal RASTI) pierde modulación, tanto más cuanto mayor sea la frecuencia de modulación F . Efectivamente, en la Figura 89 se aprecia cómo las componentes espectrales de baja frecuencia van reduciéndose progresivamente a medida que crece la frecuencia de modulación.

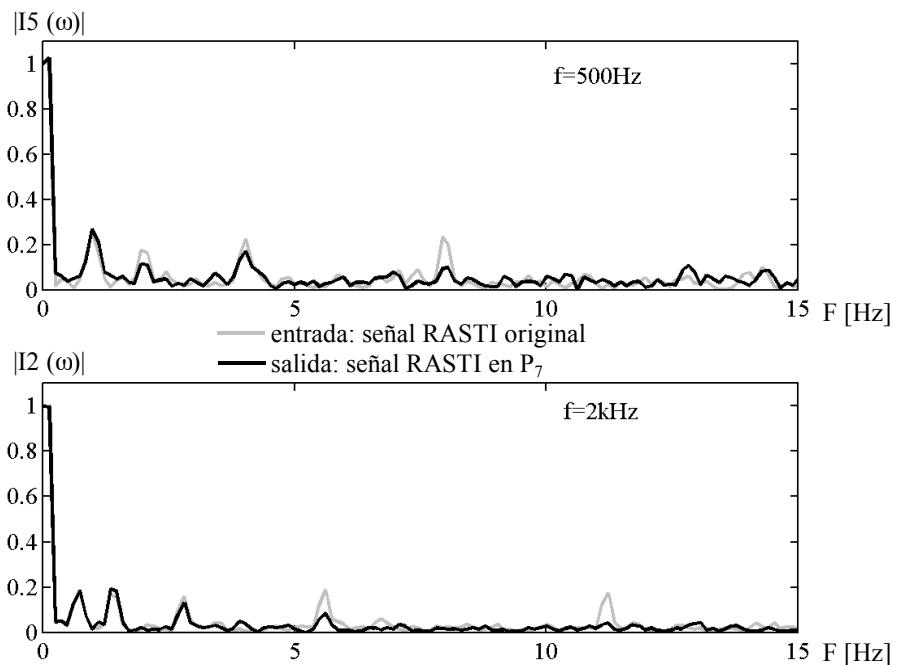


Figura 89. Intensidad sonora relativa obtenida cuando se excita al punto central del array con una señal RASTI desde el punto P_7 de la Figura 69 (no existe ruido añadido). Se compara con la intensidad sonora relativa de la señal RASTI original, sin reverberación añadida.

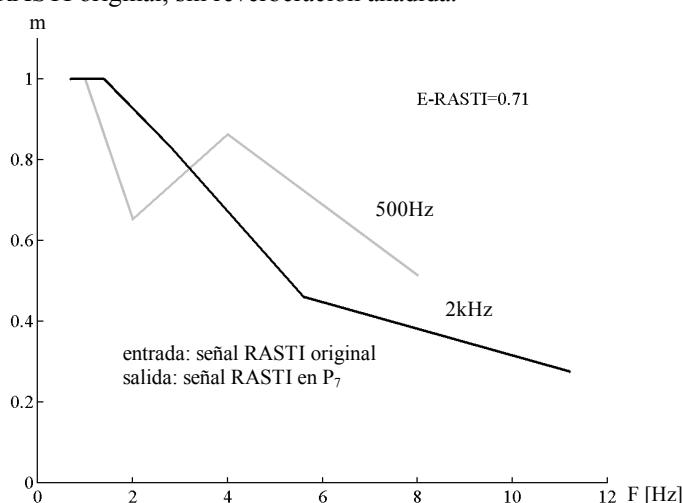


Figura 90. Pérdidas de modulación del habla (índices m) obtenidas con el método E-RASTI sobre el espectro de intensidad sonora de la Figura 89 para evaluar la degradación que produce la reverberación del recinto de la Figura 69 cuando la fuente de voz está situada en el punto P_7 .

En la Figura 90 se representan los índices m de modulación calculados con el esquema E-RASTI presentado en la Figura 87, cuando se considera como entrada la señal RASTI limpia y como salida la misma señal con reverberación añadida. Además se proporciona el índice E-RASTI calculado. Puede apreciarse la misma tendencia comentada anteriormente, por la que la modulación se reduce al aumentar F . Además lo hace de manera similar en las dos octavas de frecuencia centradas en $f = 500\text{Hz}$ y $f = 2\text{kHz}$. Esto tiene sentido porque en el recinto simulado no se ha incorporado ninguna variación con la frecuencia del coeficiente β de reflexión de las paredes.

En la Figura 91 se representa el efecto del ruido aditivo sobre el espectro de baja frecuencia de la intensidad sonora. En este caso no se ha añadido reverberación mediante la convolución con h_{P7} . Se aprecia cómo la amplitud de todas las frecuencias de modulación decrece en cuantía similar, independientemente de en qué punto del espectro estén situadas.

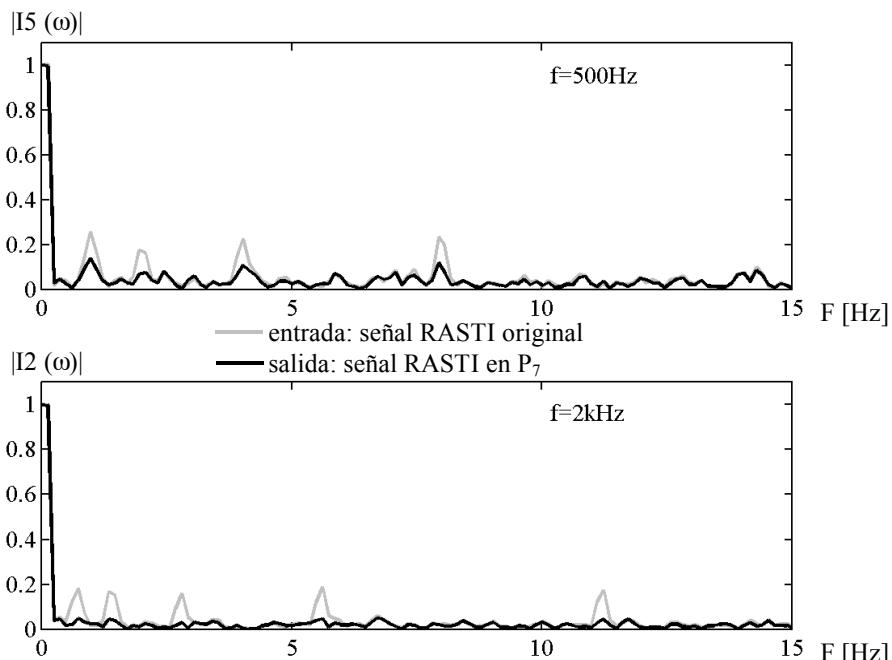


Figura 91. Intensidad sonora relativa obtenida cuando se suma a la señal E-RASTI un ruido rosa con SNR=-10dB. Se compara con la intensidad sonora relativa de la señal RASTI original, sin ruido añadido.

En la Figura 92 se proporcionan los resultados obtenidos en la evaluación de la reverberación por el método E-RASTI, promediando las respuestas en los ocho puntos de la sala de la Figura 69. En esa figura se comparan los resultados de E-RASTI con el RASTI predicho por la teoría [Sánchez-Bote 99], considerando las condiciones de reverberación existente, con $T_{60} = 0.7\text{s}$. Para obtener el E-RASTI se ha considerado como entrada la señal RASTI original, y como salida la misma señal con reverberación, considerándose que procede de alguno de los 8 puntos de test de la sala de simulación. Puede comprobarse cómo el E-RASTI tiende a sobreestimar el RASTI teórico. Pero esa sobreestimación es previsible, ya que los filtros en frecuencia para calcular el nivel de intensidad en las frecuencias de modulación (véase la Figura 88) tienden a sobreestimar los niveles de modulación, debido a su insuficiente selectividad. Sin embargo es necesario que su selectividad no sea muy grande, ya que cuando se analice una señal de voz real (y no la señal RASTI), las frecuencias de modulación no coincidirán en general con las predichas por la teoría del STI.

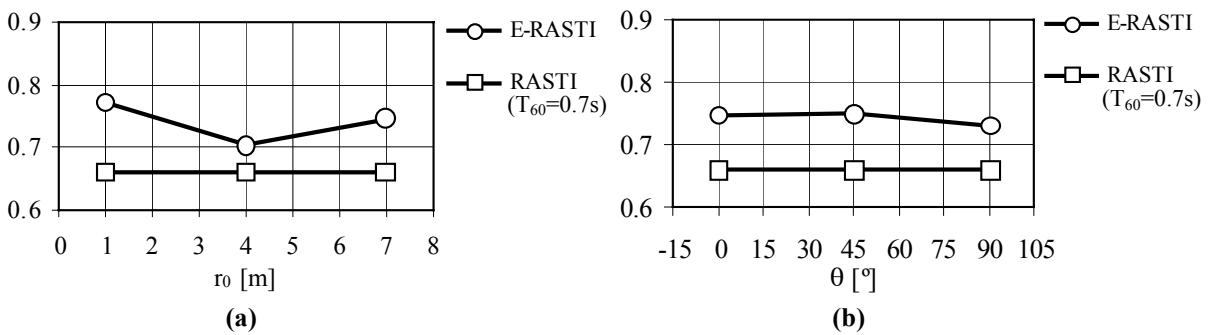


Figura 92. Valores promedio de los índices E-RASTI calculados sobre la señal RASTI emitida desde los puntos P_1 a P_7 del recinto de la Figura 69. La señal RASTI original se considera como entrada. El índice RASTI considerado en las gráficas es el calculado de forma teórica [Sánchez-Bote 99] para $T_{60}=0.7s$. **(a)** Promedio de los 7 puntos agrupados por distancia al array. **(b)** Ídem por ángulo θ de desviación respecto al eje del array (ver la Figura 69).

En la Figura 93 se representa el E-RASTI calculado sobre una señal RASTI contaminada con ruido rosa. Puede comprobarse que existe también sobreestimación del índice E-RASTI sobre el RASTI teórico, pero los resultados se van acercando a medida que la SNR mejora, porque en ese caso la energía de la intensidad sonora estará tan concentrada en las frecuencias de modulación que la baja selectividad del filtro F no influirá en el resultado. Por otra parte la tendencia ascendente del E-RASTI con la relación SNR es totalmente compatible con las predicciones teóricas.

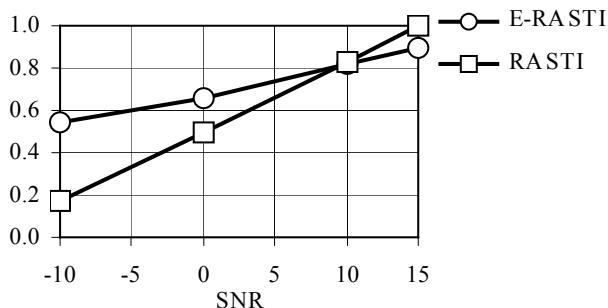


Figura 93. Índices E-RASTI calculados para la señal RASTI contaminada por ruido rosa, considerando diferentes relaciones SNR. El RASTI representado en las gráficas es el calculado de forma teórica [Sánchez-Bote 99] según las SNR especificadas.

Los experimentos mostrados de la Figura 89 a la Figura 93 acreditan que el método E-RASTI, en la forma que ha sido propuesto, obtiene unos resultados similares a los que se obtendrían en la verificación de inteligibilidad de una señal RASTI monocanal por los métodos convencionales. No obstante, manifiestan en algunos casos de baja calidad de la señal, una sobreestimación en la ganancia de modulación (índice E-RASTI menor que el teórico), cuyo origen ya ha sido explicado. De este modo, se está en disposición de realizar diferentes calificaciones con el método E-RASTI del procesador propuesto, lo cual se describe a continuación.

Como segundo grupo de experimentos se ha utilizado el índice E-RASTI para la evaluación de los resultados de derreverberación del procesador cepstral MPAP, con $L = 4096pt$, $N = 8192pt$ y $W_C = 200pt$. Se ha introducido en el procesador la señal RASTI

multicanal, resultado de convolucionar la señal RASTI original con la función de transferencia multicanal de cada uno de los 8 puntos de test y se ha procesado con el array de micrófonos, obteniéndose a la salida una señal monocanal mejorada $y_{MPAP}(t)$ reconstruida a partir de $Y_{MPAP}(\omega)$ de la Figura 73. Como entrada al evaluador E-RASTI se ha utilizado la señal a la salida del procesador en array –señal temporal $y_{MPAP}(t)$ – y como salida se ha utilizado la señal E-RASTI reverberada $y_8(t)$, extraída del canal central del array (canal 8). Los resultados se muestran en la Figura 94.

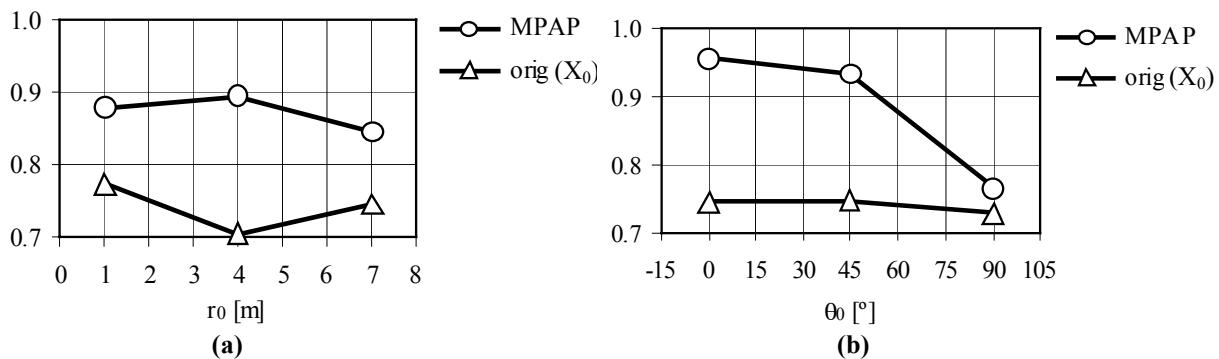


Figura 94. Resultados de los experimentos en los que se utiliza el índice E-RASTI para comprobar la mejora producida por el procesador cepstral MPAP descrito en el apartado 7.1 de esta Tesis. Señal de entrada: $y_{MPAP}(t)$. Señal de salida: $y_8(t)$ (canal central del array) que corresponde a la señal RASTI con reverberación, equivalente al subcorpus **rev** de simCMU-2. Se compara con los índices E-RASTI calculados para la señal original (Figura 92). En este caso, señal de entrada: $x_0(t)$ (RASTI original), señal de salida: $y_8(t)$ (canal central del array). **(a)** Resultados promedios sobre los 8 puntos del recinto simulado de la Figura 69 y agrupados en función de la distancia al array. **(b)** Ídem en función del ángulo θ_0 .

Un vistazo rápido a la Figura 94 da cuenta de que el derreverberador cepstral produce mejoras en la señal RASTI (índice E-RASTI menor que la unidad) y que la calidad de la señal procesada es, como cabía esperarse, peor a la de la señal original. Por eso las curvas correspondientes a la señal RASTI original tienen unos índices E-RASTI todavía menores. Como resultado reseñable, en la Figura 94(b) el índice E-RASTI revela una mayor mejora para el apuntamiento *broadside* ($\theta_0 = 90^\circ$) comparado con las otras direcciones de apuntamiento.

En definitiva, parece que los resultados ofrecidos mediante la aplicación del método E-RASTI a la señal RASTI producen unos resultados coherentes con la cantidad objetiva de contaminación por ruido y reverberación introducida en la señal RASTI de partida, y también son capaces de calificar la mejora de un procesador de forma congruente con la apreciación subjetiva.

Experimentos con señal de voz

Se han realizado experimentos para evaluar la derreverberación producida por el procesador en array de la Figura 73, utilizando las bases de datos simCMU-2 (sólo subcorpus **rev**) y CMU. Se considera como señal de entrada la voz procesada $y_{MPAP}(t)$ (procesador MPAP, con $L = 4096pt$, $N = 8192pt$ y $W_C = 200pt$, Tabla 23) y como salida el canal central del array, $y_8(t)$. Paralelamente se comparan los resultados con la señal de voz original, $x_0(t)$. En las Figuras 95 y 96 se representan los resultados obtenidos.

procesador MPAP			
enventanado			
L[pt]	N[pt]	W _C [pt]	W _C /N
4096	8192	200	0.02

Tabla 23. Parámetros seleccionados para evaluar mediante el método E-RASTI al procesador en array MPAP de la Figura 73. Pruebas E-RASTI sobre señal de voz perteneciente a las bases de datos CMU y simCMU-2 (sólo subcorpus rev).

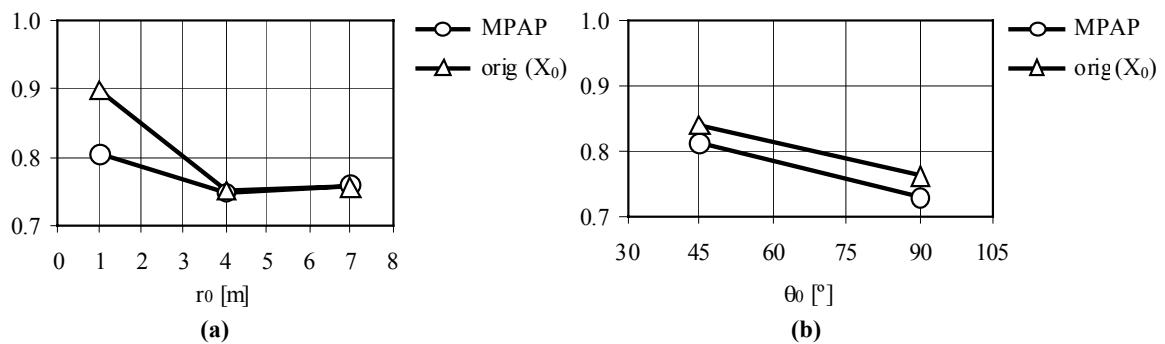


Figura 95. Resultados de evaluación del método E-RASTI con señal de voz, sobre la base de datos simCMU-2 (subcorpus rev). Se consideran como señales de entrada en cada caso la señal de voz procesada $y_{MPAP}(t)$ o la señal original $x_0(t)$. Como salida siempre se considera al canal central del array, $y_8(t)$. (a) Promedios sobre simCMU-2 (subcorpus rev) agrupados por distancias r_0 de la fuente al array. (b) Ídem, agrupados por ángulo θ_0 de la fuente con respecto al array.

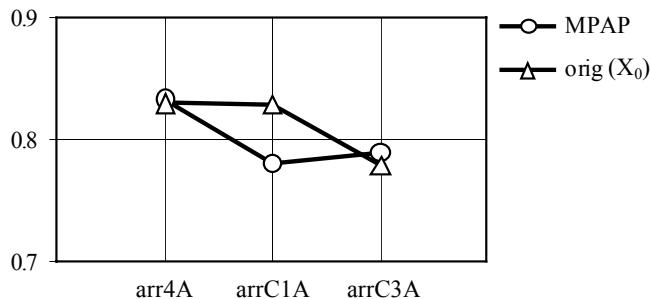


Figura 96. Resultados de evaluación del método E-RASTI con señal de voz, sobre la base de datos CMU. Se consideran como señales de entrada en cada caso la señal de voz procesada $y_{MPAP}(t)$ o la señal original $x_0(t)$. Como salida siempre se considera el canal central del array, $y_8(t)$.

En una inspección rápida de ambas figuras aparece claramente un resultado anómalo, que la señal original tiene mayor índice E-RASTI que la señal procesada, es decir, que se considera a ésta como de menor calidad. Esto evidentemente no se corresponde con la apreciación subjetiva, y quiere decir que habrá que tomar al índice E-RASTI con ciertas reservas. Lo que funcionaba muy bien para señal RASTI funciona un poco peor para señal de voz. Las razones pueden ser múltiples. Por una parte es evidente que las frecuencias teóricas F de modulación del RASTI no se encuentran tan claramente en el espectro de intensidad de la

señal de voz. Por otra parte puede ser que el procesado aumente la modulación de la señal de salida (incluso produciendo una modulación mayor que en la señal original) a base de introducir distorsión. No obstante los resultados son mucho mejores que los ofrecidos por las distancias LAR y cepstrales para evaluar derreverberación y obtenidos en el apartado 7.1.4 de la Tesis. Las diferencias mostradas en las Figuras 95 y 96 entre el E-RASTI de la señal original y la procesada son mínimas y siguen la misma tendencia, con lo que esta anomalía observada puede desaparecer cuando se analicen y comparan, utilizando el índice E-RASTI, diferentes resultados con diferentes procesadores en array. Es decir, la clave puede estar en utilizar el E-RASTI para comparar la mejora producida por varios procesadores en array del mismo tipo y no para comparar dos señales de voz muy diferentes, como son en este caso la señal original $x_0(t)$ y la salida del array, $y_{MPAP}(t)$.

Considerando globalmente los resultados de los experimentos realizados, como conclusión puede decirse que el índice E-RASTI es compatible con el cálculo tradicional del RASTI, en el sentido de que proporciona resultados parecidos a éste con la señal RASTI, y que es capaz de detectar la contaminación por ruido y reverberación de una señal de voz, con lo cual puede ser perfectamente utilizado en la evaluación de fragmentos de habla en condiciones acústicas adversas y por ende para evaluar las capacidades de los procesadores en array propuestos en esta Tesis.

7.3 PROCESADOR EN ARRAY BASADO EN SUPRESIÓN DE RUIDO AUDIBLE COMBINADA CON UN FILTRO DE WIENER MODIFICADO POR COHERENCIA (PROCESADOR ANS-MW)

El procesador MW-MPAP propuesto en el punto 7.1 de esta Tesis ha proporcionado muy buenos resultados objetivos y subjetivos. Se ha demostrado que la mayor parte de la derreverberación es aportada por la conformación de haz que proporciona el array de 15 micrófonos, mientras que el procesado cepstral incrementa ligeramente la sensación de “sequedad” en la señal procesada. Sin embargo, éste ocasiona cierta cantidad de distorsión, a lo que hay que añadir una complejidad de procesado bastante elevada, que lo hacen ineficiente para una implementación en tiempo real en un DSP convencional. Por otra parte, aunque el reductor de ruido MW es sencillo, eficiente y produce muy buenos resultados subjetivos en cuanto a cancelación de ruido, origina cierta cantidad de ruido musical, muy dependiente de la relación señal a ruido, del ancho de banda de dicho ruido y del ajuste de los parámetros λ de los estimadores recursivos de señal y ruido, que tiene que ser muy cuidadoso en la práctica. Por todo ello, como paso consecuente en los experimentos y propuestas que aquí se desarrollan, se propone eliminar la conformación cepstral (MPAP) del procesador en array planteado y añadir un supresor perceptual de ruido (ANS, *Audible Noise Suppression*) según los principios explicados en el apartado 4.1.3 de esta Tesis.

7.3.1 Descripción del procesador ANS-MW

Los métodos de postfiltrado auditivo, basados en las propiedades de enmascaramiento del oído humano fueron descritos en el punto 4.1.3 de esta Tesis. El método ANS (*Audible Noise Suppression*) consiste en filtrar la señal de voz contaminada, en función el umbral de enmascaramiento $T(\omega)$, o más específicamente en función de la relación ruido a umbral de enmascaramiento $NMR(\omega)$. Dos son los pasos que hay que dar para emplear el método ANS monocanal dentro del captador en array de 15 micrófonos. En primer lugar se debe conseguir

un filtrado multicanal óptimo, utilizando dicho método ANS. En segundo lugar se debe usar una estimación adecuada de la señal de voz limpia de ruido y reverberación, premisa necesaria ya que se necesitan los umbrales de enmascaramiento $T(\omega)$ correspondientes al habla sin perturbación acústica. El filtrado multicanal óptimo mediante el método ANS se puede conseguir filtrando la salida conformada del array, siguiendo el mismo razonamiento que se expuso en el punto 4.1.1 para el postfiltro multicanal de Wiener. Según éste, se demostraba que un filtro multicanal óptimo es equivalente a una conformación óptima (conformador superdirectivo) más un filtro monocanal óptimo aplicado a la salida conformada del array. La estimación óptima “a priori” de la señal limpia de voz puede hacerse según el método MW que como se sabe aprovecha la información multicanal proporcionada por el array, y ha dado muy buenos resultados de mejora de voz actuando por sí solo en el array anidado de 15 micrófonos. Ajustándose a lo anterior, se propone usar el método ANS con el array de micrófonos en la forma representada en la Figura 97.

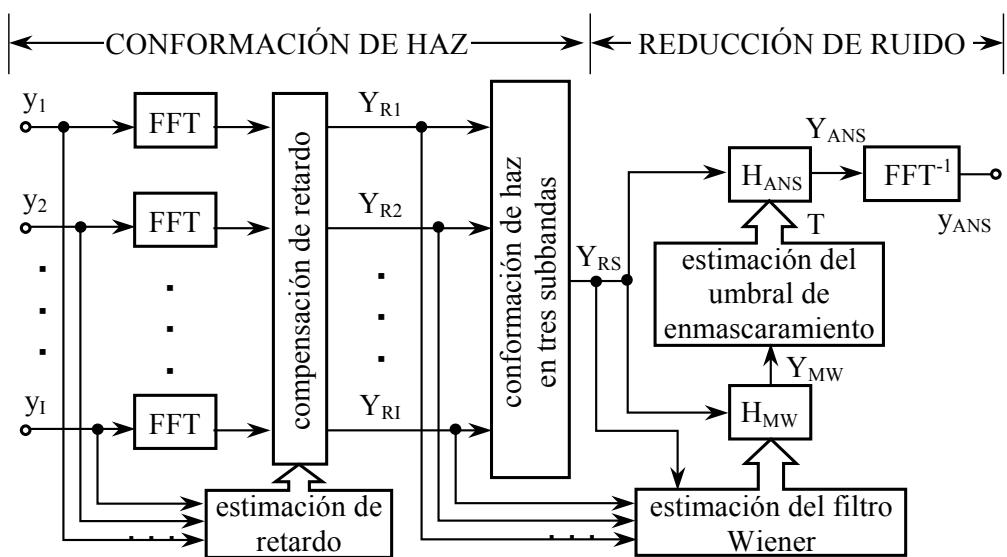


Figura 97. Diagrama de bloques del procesador ANS-MW.

Lo que hace el procesador ANS-MW es por tanto filtrar con el método ANS la salida conformada del array, utilizando los umbrales de enmascaramiento $T(\omega)$ calculados a partir de la señal de voz mejorada utilizando el método Wiener modificado por coherencia (MW).

A continuación se describen cada una de las etapas del procesador ANS-MW según aparecen en la Figura 97.

Conformación de haz

La conformación de haz en este procesador se realiza inicialmente mediante el método convencional de retardo y suma usando la anidación de las tres subbandas B_1 , B_2 y B_3 . En estas pruebas, los retardos necesarios para cada canal son introducidos manualmente a partir de la posición estimada de la fuente de voz. El retardo es compensado en el dominio del tiempo (en la Figura 97 se muestra la compensación en el dominio de la frecuencia, que también es posible), adelantando o retrasando el número de muestras necesarias (con una precisión temporal de $1/4f_s$ ya que la señal temporal se interpola “x 4”) para alinear temporalmente a los canales del array. La señal se enventana y se procesa por tramas de corta duración proporcionándose $Y_{RS}(\omega)$ a la salida del conformador.

Estimación a priori de la voz limpia mediante el método MW

Se utiliza el método MW para proporcionar la salida $Y_{MW}(\omega)$, que a su vez se usa como entrada en el bloque de estimación del umbral de enmascaramiento $T(\omega)$. El filtro MW, como ya se sabe, cancela ruido atendiendo a las componentes coherentes y no coherentes del mismo y sólo actúa en las frecuencias donde la coherencia intercanal estimada supera cierto umbral. Para más detalles sobre el funcionamiento de este bloque debe consultarse el apartado 7.1.1 de esta Tesis.

Cálculo de los umbrales de enmascaramiento $T(\omega)$

El cálculo de los umbrales de enmascaramiento se hace según J. D. Johnston [Johnston 88], como se explicó en el punto 4.1.3. A continuación se exponen las particularidades introducidas en el procesador ANS-MW.

Como se ha dicho antes, los umbrales de enmascaramiento se calculan sobre la potencia de la señal de salida del postfiltro de Wiener:

$$|Y_{MW}(\omega)|^2 = \Phi_{Y_{MW}Y_{MW}}(\omega) \quad (284)$$

Sin embargo, en previsión de que se desee incorporar cierta latencia en la estimación de los umbrales, para que el postfiltro no varíe instantáneamente (que induciría la producción de ruido musical), se utiliza una estimación recursiva según:

$$\hat{\Phi}_{Y_{MW}Y_{MW}}(\omega, k) = \lambda_S \hat{\Phi}_{Y_{MW}Y_{MW}}(\omega, k-1) + (1 - \lambda_S) \Phi_{Y_{MW}Y_{MW}}(\omega, k) \quad (285)$$

con un coeficiente de actualización λ_S igual al utilizado en las estimaciones de señal del procesador MW-MPAP.

Se consideran $B = 22$ bandas críticas auditivas, ya que la frecuencia superior de análisis es de 8kHz (ver la Tabla 1). La función de ensanchamiento SF(b) para considerar la influencia de bandas críticas adyacentes es la mostrada en la Figura 36 y la Tabla 2. El umbral de enmascaramiento por bandas críticas $T(b)$, se calcula con la expresión (186) y un factor de renormalización según la Tabla 3, un factor de *offset* según (187) y una naturaleza tonal según (188). La planitud espectral SMF se calcula por (189). Finalmente se pasa del valor de $T(b)$ por bandas críticas al valor $T(\omega)$ por frecuencia individual, considerándose que $T(\omega)$ se mantiene constante dentro de la misma banda crítica.

Supresión auditiva de ruido (ANS)

El paso final del procesador ANS consiste en filtrar la salida conformada del array $Y_{RS}(\omega)$ con el postfiltro $H_{ANS}(\omega)$. La función de transferencia $H_{ANS}(\omega)$ propuesta es similar, con algunas variaciones, a la genérica (180), donde se ha utilizado un factor de forma de valor $\varepsilon = 1$:

$$H_{ANS}(\omega) = \left[\frac{\Phi_{Y_{RS}Y_{RS}}(\omega)}{\delta(\omega)\eta\hat{\Phi}_{NN}(\omega) + \Phi_{Y_{RS}Y_{RS}}(\omega)} \right]^{\frac{1}{2}} \quad (286)$$

con el factor $\delta(\omega)$ como el definido en (181) y $\hat{\Phi}_{NN}(\omega)$ la estimación de ruido presente,

$$\delta(\omega) = \left[1 + \frac{T(\omega)}{\hat{\Phi}_{NN}(\omega)} \right] \left[\frac{\hat{\Phi}_{NN}(\omega)}{T(\omega)} \right] = \frac{\hat{\Phi}_{NN}(\omega)}{T(\omega)} + 1 = NMR(\omega) + 1 \quad (287)$$

Compruébese que (286) y (287) corresponden a una rescritura de (180) y (181) para $\epsilon = 1$ y aunque son ligeramente diferentes a las fórmulas de supresión de [Tsoukalas 97] están directamente adaptadas de allí, por lo que básicamente son las mismas. El ruido se estima de forma recursiva en los instantes de ausencia de actividad de habla:

$$\hat{\Phi}_{NN}(\omega, k) = \lambda_N \hat{\Phi}_{NN}(\omega, k-1) + (1 - \lambda_N) \Phi_{NN}(\omega, k) \quad (288)$$

siendo la potencia $\Phi_{NN}(\omega)$ de (288),

$$\Phi_{NN}(\omega, k) = \Phi_{Y_{RS}Y_{RS}}(\omega, k) \quad \text{en ausencia de señal} \quad (289)$$

el ruido conformado por el array, obtenido del cálculo de la potencia de salida del array en las tramas consideradas de ruido (detectadas por un VAD adecuado). El coeficiente de actualización λ_N es similar al usado en las estimaciones de ruido del procesador MW-MPAP.

El factor $\delta(\omega)$ es una medida de la relación ruido a umbral de enmascaramiento, verificándose que si $\hat{\Phi}_{NN}(\omega) > T(\omega)$ entonces $\delta(\omega)$ tendrá un valor elevado y viceversa. El factor $\eta \geq 1$ de (286) es una constante que se introduce para ocasionalmente, si se desea, una sobresustracción adicional de ruido y que se debe determinar de forma empírica, según la cantidad de perturbación presente en la señal de voz. Inicialmente debe ser la unidad ($\eta = 1$), como se propone originalmente en [Tsoukalas 97], pero si se admite que el supresor de ruido audible no cancela la perturbación acústica en una cantidad suficiente, se puede hacer $\eta > 1$ para conseguir una mayor eliminación de ruido, aunque a riesgo de aumentar la distorsión.

La función de transferencia del postfiltro utilizado, $H_{ANS}(\omega)$ en (286), es real por definición. Por lo tanto dicho filtro de mejora modifica únicamente el módulo de la señal a la entrada, es decir la fase de la señal $Y_{RS}(\omega)$ se deja pasar sin alteración. Como se sabe, esto es habitual en los procesadores que aplican postfiltrado (Wiener, sustracción espectral, etc.), y es la estrategia que se utiliza en el array microfónico aquí presentado.

7.3.2 Experimentos y resultados

Una vez desechada la parte MPAP de la propuesta inicial (ver el punto 7.1) para una futura implementación en tiempo real, el objetivo que se persigue ahora es probar la eficacia del nuevo procesador auditivo ANS-MW, con experimentos de escucha subjetiva y medidas objetivas, utilizando los evaluadores tradicionales basados en la relación SNR y el nuevo evaluador E-RASTI propuesto anteriormente. Por otra parte se hace necesario comparar los resultados del nuevo procesador con el cancelador de ruido MW funcionando en solitario, para indagar si merece la pena añadir la parte auditiva al cancelador MW probado con éxito anteriormente, es decir, si realmente el método auditivo reduce la sensación de distorsión y por tanto aumenta la calidad de la señal de salida obtenida.

Se han realizado múltiples pruebas sobre las bases de datos CMU y simCMU-2. Después de algunos ajustes y pruebas subjetivas iniciales, los parámetros del procesador ANS-MW han sido fijados como se indica en la Tabla 24. Puede comprobarse cómo una vez desecharo el procesador cepstral, se han usado ventanas de longitud L más pequeñas ($L = 512pt$, $T_L = 32ms$), que permiten un procesador de respuesta rápida o lenta, según se desee, con el ajuste adecuado de los parámetros λ de la estimación recursiva.

ANS-MW											
enventanado			MW								ANS
			LBF				Coherencia				
L[pt]	N[pt]	S ₁ [pt]/S[%]	λ _S	t _{ΔS} [ms]	λ _N	t _{ΔN} [ms]	λ _{coh}	t _{Δcoh} [ms]	α	UC	η
512	1024	128/75	0.8	143	0.8	143	0.8	143	50	0.7	1

Tabla 24. Parámetros seleccionados en el procesador ANS-MW para las pruebas sobre las bases CMU y simCMU-2.

Como prueba intermedia, antes de llegar a los resultados finales, se ha considerado alimentar al estimador de umbrales $T(\omega)$ con el numerador del filtro LBF de (271):

$$|Y_{\text{LBF}}(\omega)|^2 = \left| \langle \hat{\Phi}_{Y_i Y_{\text{ref}}}(\omega) \rangle_i - \langle \hat{\Phi}_{N_i N_{\text{ref}}}(\omega) \rangle_i \right| \quad (290)$$

que es una estimación de la señal limpia de ruido coherente y no coherente. Sin embargo, finalmente se decidió utilizar $Y_{\text{MW}}(\omega)$ (salida del filtro de Wiener) como se muestra en el procesador final de la Figura 97, ya que el array producía una salida con mayor calidad.

Los estimadores objetivos basados en SNR que se utilizarán serán la ganancia en relación señal a ruido con ponderación A, GSNR_A (239), la ganancia en relación ruido a umbral de enmascaramiento, GNMR (240) (ambos vistos en el punto 5.1) y la ganancia en índice de articulación GAI (254) (punto 5.3). Contrariamente a cómo se hizo para GNMR en los experimentos del apartado 7.1.4 sobre el anterior procesador propuesto, estos estimadores están basados ahora en el cálculo del ruido $N(\omega)$ mediante la diferencia (223) (después de una alineación temporal mediante el método PHAT descrito en el punto 3.2.2) entre la señal de salida del procesador $-Y_{\text{ANS}}(\omega)$ o $Y_{\text{MW}}(\omega)$ según sea el caso— y la señal original $X_0(\omega)$, haciendo comparaciones con la señal sucia $Y_8(\omega)$, correspondiente al canal central del array. Recuérdese que esto equivale a que se considera como perturbación no sólo al ruido aditivo, sino también a la reverberación, que no está presente en la voz original $X_0(\omega)$. Por tanto, aquí las estimaciones de SNR son más sólidas que las consideradas en el apartado 7.1.4 para el procesador MW-MPAP, donde se medía el ruido sólo en las tramas de ausencia de actividad de voz.

Las medidas de GSNR_A, GNMR y GAI se han obtenido promediando sólo las tramas temporales con presencia de voz. Además, las medidas de relación ruido a umbral de enmascaramiento NMR, se han hecho calculando los umbrales $T(\omega)$ a partir de la señal original $X_0(\omega)$ y no estimándolas con la salida $y_{\text{MW}}(t)$, como lo hace el procesador ANS-MW propuesto, ya que para él, $X_0(\omega)$ es desconocida.

Las medidas utilizando el E-RASTI se han realizado con el evaluador presentado en la Figura 87 sobre fragmentos de voz de 8s. Los resultados obtenidos han sido publicados en [Sánchez-Bote 01-a] según se muestran en las Tablas 25 y 26, pero a continuación se desarrollan con más detalle.

En la Figura 98 se muestra un ejemplo de la señal temporal perteneciente a la base de datos simCMU-2 (subcorpus **Mdif** con $\text{SNR}_8 = 10\text{dB}$) antes y después de pasar por el procesador ANS-MW o por el procesador MW. En la Figura 99 se representan las mejoras objetivas de tipo SNR correspondientes a la muestra de la Figura 98. Se representan GSNR_A(k), GAI(k) y GNMR(k) por trama temporal, para los procesadores ANS-MW y MW.

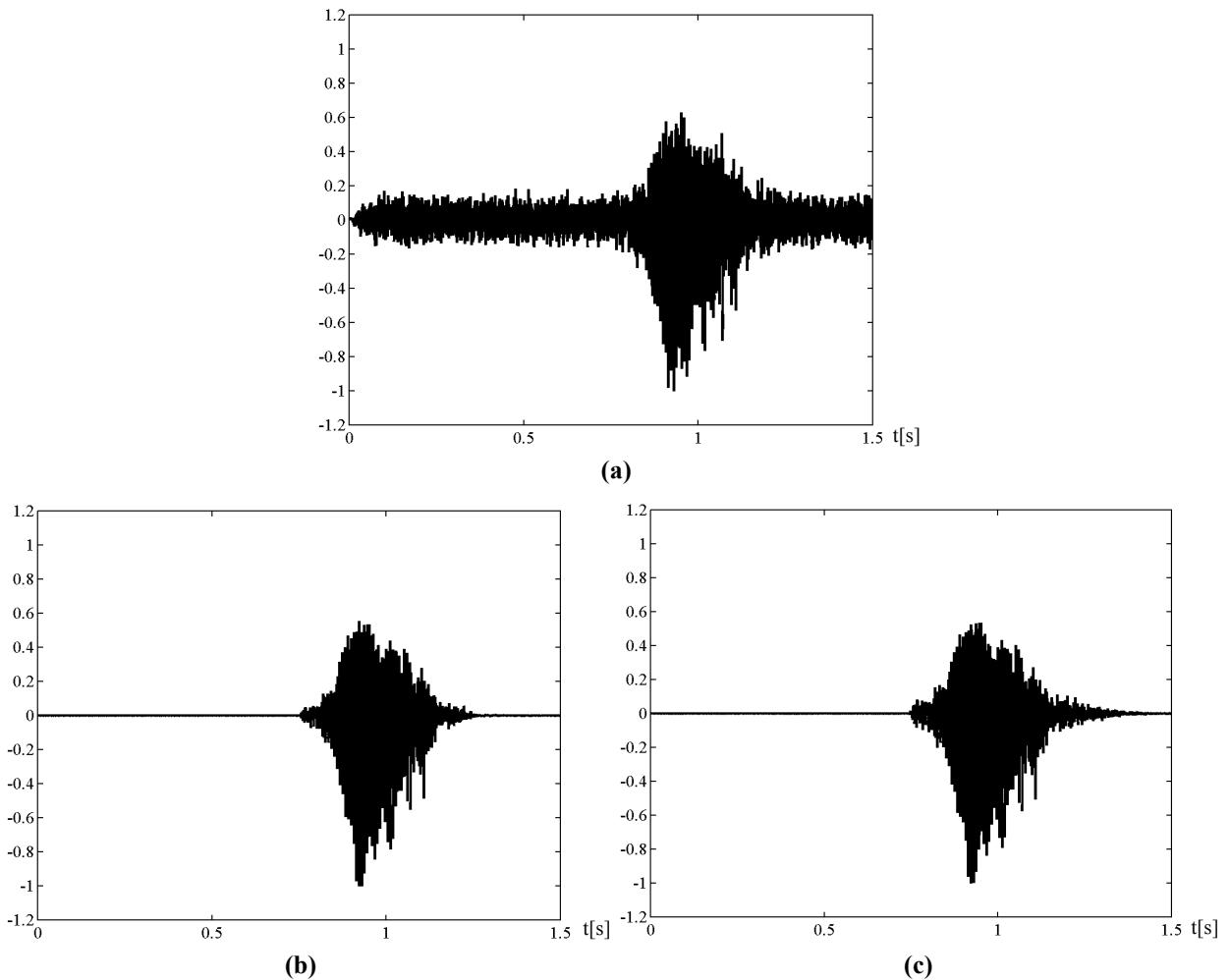


Figura 98. Muestra de señal perteneciente a la base de datos simCMU-2 (subcorpus **Mdif** con $\text{SNR}_s=10\text{dB}$) antes y después del procesador ANS-MW o el procesador MW. **(a)** Señal sucia del canal central $y_8(t)$, antes del procesador. **(b)** Señal de salida $y_{ANS}(t)$, después del procesador ANS-MW. **(c)** Señal de salida $y_{MW}(t)$, después del procesador MW.

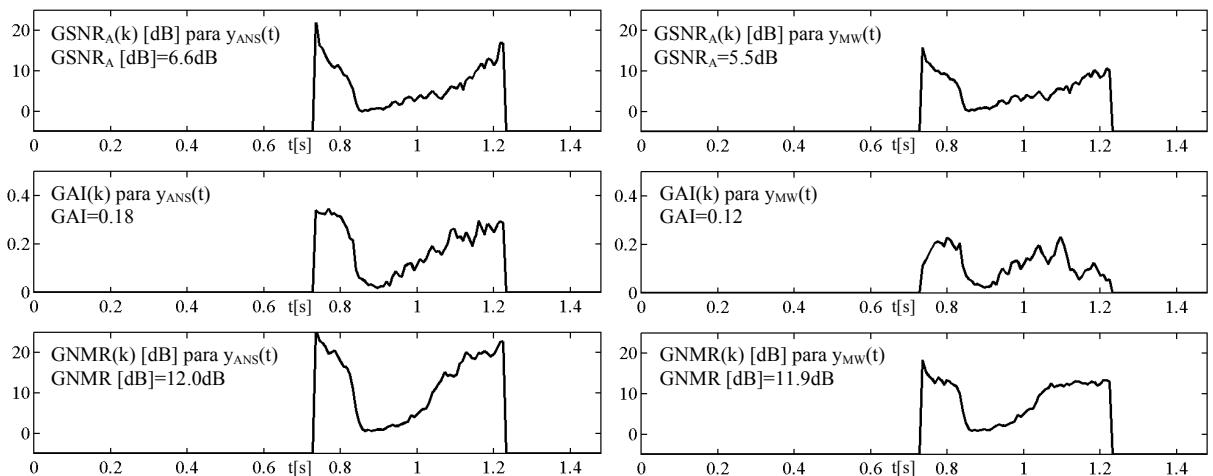


Figura 99. Ejemplo de cálculo de mejora objetiva basada en SNR correspondiente a la muestra de la Figura 98. Se representan las mejoras por trama $\text{GSNR}_A(k)$ [dB], $\text{GAI}(k)$ y $\text{GNMR}(k)$ [dB] para los procesadores ANS-MW y MW.

En las Figuras 100 y 101 se representan los resultados del procesador ANS-MW en función de la relación señal a ruido introducida en el canal central del array, SNR_8 y de la distancia r_0 de la fuente al centro del array. Se ha utilizado la base de datos simCMU-2, correspondiente a señal de voz contaminada por ruido blanco (subcorpus **Mdif**). Se comparan los resultados anteriores con el procesador MW actuando independientemente.

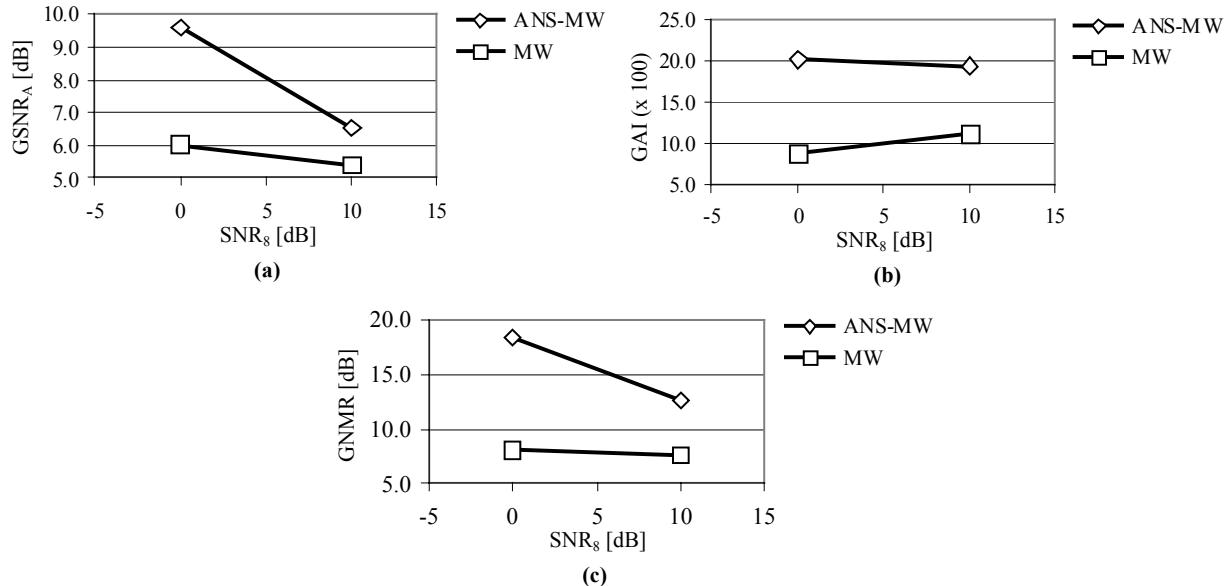


Figura 100. Resultados del procesador ANS-MW con la base de datos simCMU-2, correspondientes a señal de voz contaminada por ruido blanco (subcorpus **Mdif**). Se compara con el procesador MW actuando independientemente. Se representan los promedios de los resultados agrupados según la relación señal a ruido introducida en el canal central del array, SNR_8 . (a) Ganancia en relación señal a ruido con ponderación A, GSNR_A [dB]. (b) Ganancia en índice de articulación, GAI . (c) Ganancia en relación ruido a umbral de enmascaramiento, GNMR [dB].

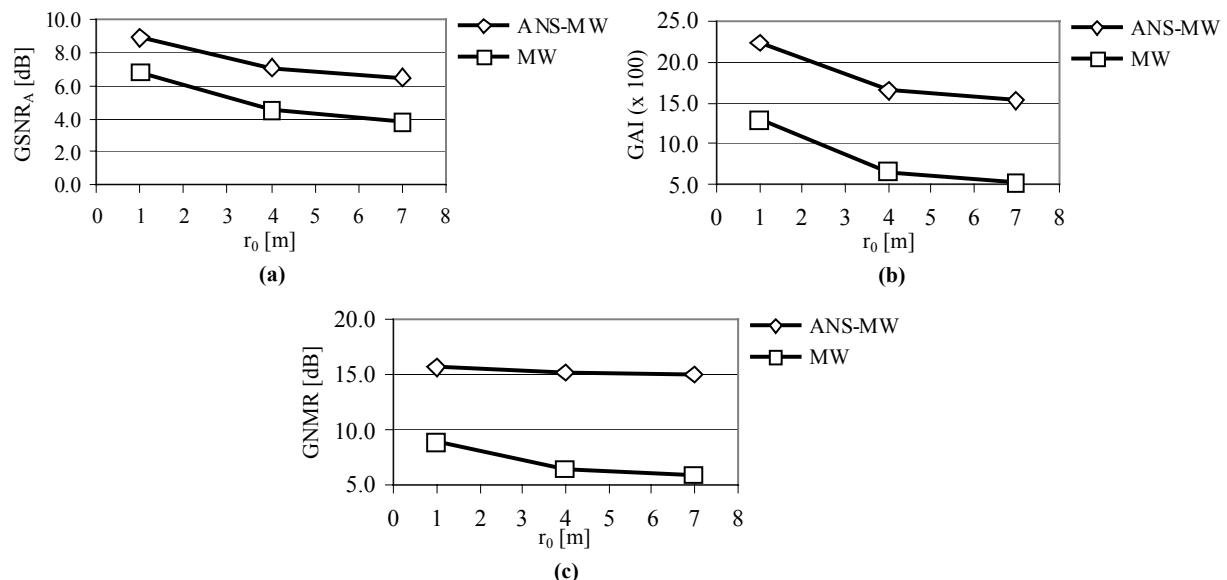


Figura 101. Resultados del procesador ANS-MW con la base de datos simCMU-2, correspondientes a señal de voz contaminada por ruido blanco (subcorpus **Mdif**). Se compara con el procesador MW actuando independientemente. Se representan los promedios de resultados agrupados según la distancia r_0 de la fuente al centro del array. (a) Ganancia en relación señal a ruido con ponderación A, GSNR_A [dB]. (b) Ganancia en índice de articulación, GAI . (c) Ganancia en relación ruido a umbral de enmascaramiento, GNMR [dB].

A la vista de las Figuras 100 y 101, puede comprobarse cómo en todos los análisis de parámetros objetivos realizados, el procesador auditivo es superior al procesador Wiener ya que las mejoras introducidas en la señal de salida son superiores. Esta superioridad se manifiesta más para bajas relaciones SNR₈ y para mayores distancias r₀ de la fuente, es decir, cuando la calidad de la señal de voz contaminada por ruido y reverberación empeora. Se debe hacer constar también que estos resultados objetivos corroboran la apreciación subjetiva, ya que las salidas y_{ANS(t)} gozaban de mayor calidad subjetiva que las salidas y_{MW(t)}. Las primeras poseen menor distorsión por ruido musical cuando se iguala la cantidad de ruido subjetivo remanente. También parece (excepto con la ganancia GAI) que las mejoras son mayores a medida que el fragmento de voz de entrada tiene menor SNR₈ y la distancia de la distancia r₀ de la fuente decrece.

Por otra parte, en la Figura 102 se representan, los índices E-RASTI en función de SNR₈, obtenidos por el método propuesto en el punto 7.2, para los dos procesadores anteriores, y comparados con el correspondiente E-RASTI de la señal original. En este caso, aunque los valores son muy parejos, el procesador auditivo también parece ser superior al procesador MW, ya que manifiesta índices E-RASTI menores, es decir existe mayor diferencia de modulación entre la señal procesada y la sucia. El E-RASTI de la señal original x_{0(t)} también presenta cierta tendencia a la anomalía comentada en el punto 7.2.2, donde se reflejaban los resultados preliminares del método E-RASTI. Esto es, la señal original parece tener una modulación parecida o incluso menor que la salida de los procesadores, anomalía que ya fue justificada con anterioridad, y que indica que con el E-RASTI sólo pueden compararse fragmentos de habla con el mismo tipo de distorsión. Las salidas y_{ANS(t)} e y_{MW(t)} lo son, ya que proceden de procesadores con formas de actuar muy parecidas.

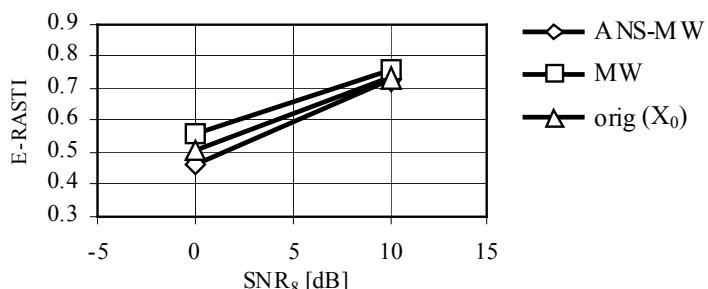


Figura 102. Resultados del análisis E-RASTI aplicado al procesador ANS-MW con la base de datos simCMU-2, correspondiente a señal de voz contaminada por ruido blanco (subcorpus **Mdif**). Se compara con el procesador MW actuando independientemente y con los índices E-RASTI de la señal original. Siempre la señal de entrada al evaluador E-RASTI es la señal procesada (u original) y la salida la señal sucia, correspondiente al canal central del array. Se representan los promedios de los resultados agrupados según la relación señal a ruido introducida en el canal central del array, SNR₈.

En la Figura 103 se representa el índice E-RASTI promedio de todos los procesadores probados, usados con la base de datos simCMU-2 (subcorpus **Mdif** para el procesador ANS-MW y MW y subcorpus **rev** para el procesador MPAP en solitario). Como se ve, se han incluido los resultados anteriores de derreverberación con el procesador cepstral MPAP (ver 7.2.2). Los resultados presentados corroboran lo dicho antes: el procesador ANS es superior al basado en filtrado de Wiener, y además con el evaluador E-RASTI se obtienen mejoras cualitativamente parecidas a las mostradas por las medidas objetivas basadas en SNR, con la

ventaja de que éste índice es más útil para medir derreverberación y no necesita tener como entrada tramas de voz perfectamente alineadas en tiempo.

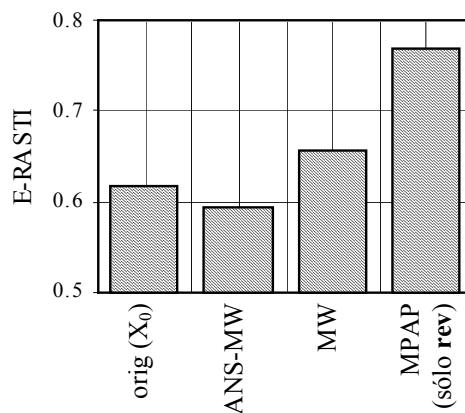


Figura 103. Índice E-RASTI promedio de todos los procesadores, probados con la base de datos simCMU-2. Los procesadores ANS-MW y MW con el subcorpus **Mdif** y el procesador MPAP con el subcorpus **rev**.

En las Figuras 104 y 105 se representa el mismo tipo de resultados, pero ahora obtenidos con la base de datos real CMU. Se muestran gráficamente tanto las mejoras usando evaluadores de tipo SNR como empleando el método E-RASTI. En el caso de la base de datos CMU, puede observarse que las mejoras son menores con todos los índices, ya que, ante señales procedentes de la realidad, con ruido y reverberación difusos, los procesadores funcionan peor. En cualquier caso se demuestra que todos los evaluadores estudiados obtienen resultados acordes con la impresión subjetiva de mejora, a excepción de GAI en algunos casos –Figura 104(b)–, que da valores levemente negativos. Este resultado positivo en las evaluaciones contrasta con lo obtenido sobre CMU con mediciones anteriores basadas en distancias LAR y RCEP (ver la Figura 86).

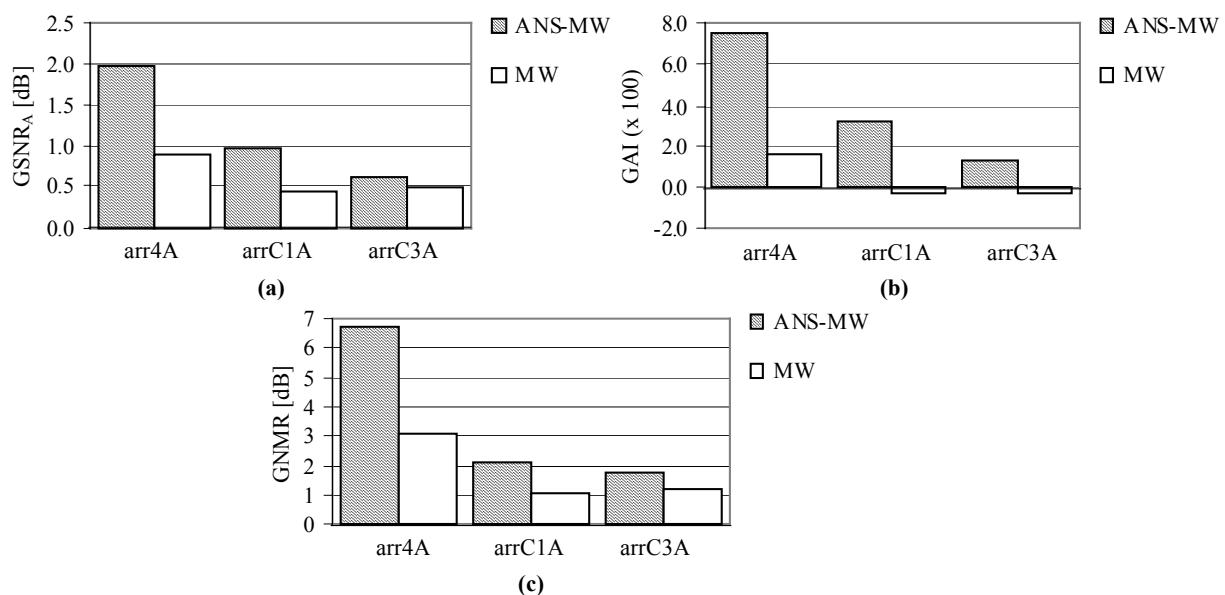


Figura 104. Resultados del procesador ANS-MW con la base de datos CMU. Se compara con el procesador MW actuando independientemente. (a) Ganancia en relación señal a ruido con ponderación A, $GSNR_A$ [dB]. (b) Ganancia en índice de articulación, GAI. (c) Ganancia en relación ruido a umbral de enmascaramiento, $GNMR$ [dB].

En la Figura 105 se resalta cómo los resultados de E-RASTI son muy similares en los dos procesadores analizados, mejorando cuando la señal de entrada es de menor calidad (subcorpus **arr4A**). También en esta figura se muestra la anomalía conocida con los resultados del procesador cepstral, ya que aun siendo la señal $y_{MPAP}(t)$ de peor calidad que $x_0(t)$ manifiesta un índice E-RASTI más bajo, lo cual ya ha sido justificado suficientemente.

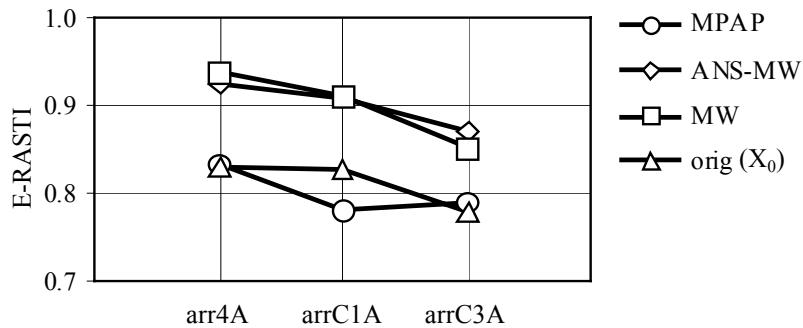


Figura 105. Resultados del análisis E-RASTI aplicado al procesador ANS-MW con la base de datos CMU. Se compara con el procesador MW actuando independientemente, el procesador cepstral MPAP actuando también solo y con los índices E-RASTI de la señal original. Siempre la señal de entrada al evaluador E-RASTI es la señal procesada (u original) y la salida la señal sucia, correspondiente al canal central del array.

Finalmente en las Tablas 25 y 26 se extraen los resultados más significativos tal y como fueron publicados en [Sánchez-Bote 01-a] y que resumen los mostrados anteriormente.

subcorpus	GSNR _A [dB]		GAI		GNMR [dB]	
	ANS-MW	MW	ANS-MW	MW	ANS-MW	MW
arr4A	2.0	0.9	0.07	0.020	6.7	3.1
arrC1A	1.0	0.4	0.03	-0.002	2.1	1.1
arrC3A	0.6	0.5	0.01	-0.002	1.8	1.3

(a)

SNR _s (dB)	GSNR _A [dB]		GAI		GNMR [dB]	
	ANS-MW	MW	ANS-MW	MW	ANS-MW	MW
0	9.6	6.0	0.20	0.09	18.3	8.1
10	6.5	5.4	0.19	0.11	12.6	7.6

(b)

Tabla 25. Extracto de resultados publicados en [Sánchez-Bote 01-a] correspondientes a las Figuras 100, 101 y 104. Se muestra la evaluación mediante parámetros objetivos basados en SNR. **(a)** Base de datos CMU. **(b)** Base de datos simCMU-2.

subcorpus	entrada: $y_{MPAP}(t)$ salida: $y_8(t)$	entrada: $y_{ANS}(t)$ salida: $y_8(t)$	entrada: $y_{MW}(t)$ salida: $y_8(t)$	entrada: $x_0(t)$ salida: $y_8(t)$
arr4A	0.83	0.92	0.94	0.83
arrC1A	0.78	0.91	0.91	0.83
arrC3A	0.79	0.87	0.85	0.78

(a)

SNR ₈ [dB]	entrada: $y_{MPAP}(t)$ salida: $y_8(t)$	entrada: $y_{ANS}(t)$ salida: $y_8(t)$	entrada: $y_{MW}(t)$ salida: $y_8(t)$	entrada: $x_0(t)$ salida: $y_8(t)$
0	-	0.46	0.56	0.51
10	-	0.73	0.75	0.73
sólo rev	0.77	-	-	0.80

(b)

Tabla 26. Extracto de resultados publicados en [Sánchez-Bote 01-a] correspondientes a las Figuras 102 y 105. Se muestran la evaluación mediante el índice E-RASTI propuesto. (a) Base de datos CMU. (b) Base de datos simCMU-2.

7.4 RECAPITULACIÓN SOBRE LOS RESULTADOS OBTENIDOS EN LAS PRUEBAS PRELIMINARES CON LAS APORTACIONES PROPUESTAS

En la parte 2 de esta Tesis se han hecho las propuestas preliminares que servirán para la implementación del prototipo de array en tiempo real y su evaluación posterior. Fundamentalmente se han propuesto varios algoritmos de mejora de voz, adaptados al trabajo multicanal con señal de voz, y un método de evaluación objetiva de la mejora producida en el habla por estos algoritmos. Resumiendo, las propuestas hechas en esta segunda parte han sido los procesadores combinados MW-MPAP y ANS-MW, y el método E-RASTI de evaluación objetiva. El procesador MW-MPAP combina el postfiltrado de Wiener multicanal óptimo con modificación de coherencia, con la derreverberación ciega basada en descomposición cepstral. El procesador ANS-MW utiliza la reducción de ruido audible, y por tanto dependiente de los umbrales de enmascaramiento auditivos, en conjunción con el método de Wiener mencionado anteriormente. El método de evaluación objetiva E-RASTI mide la pérdida/ganancia de modulación observada entre la señal vocal de entrada al procesador en array y la señal procesada de salida, obteniendo un índice de mejora comprendido entre 0 y 1.

Puede decirse que los resultados obtenidos en esta primera fase han sido satisfactorios. El procesador ANS-MW se ha mostrado como el más adecuado, por ser el que mejores resultados tanto subjetivos como objetivos proporciona, y también por ser el que más se adapta a la implementación en tiempo real. El procesador cepstral (MPAP) se ha mostrado poco eficiente puesto que no ocasiona una derreverberación especialmente apreciable y por contra consume muchos recursos computacionales, por lo que puede considerarse poco adecuado para la implementación del prototipo de array. Como se verá en la Parte 3, el prototipo de array reducirá en un paso más la reverberación de baja frecuencia, incorporando un conformador superdirectivo (SD), sin por ello complicar excesivamente la operación en tiempo real.

Los algoritmos propuestos se han adaptado al trabajo en el dominio de la frecuencia mediante el procedimiento FFT + enventanado. Se ha estudiado el efecto del enventanado y la alineación temporal de los canales del array en el dominio de la frecuencia, concluyéndose que los solapamientos del 50% y el 67% con ponderación Hanning son adecuados para la aplicación propuesta (señales muy contaminadas), aunque sean subóptimos, porque producen distorsión en la señal reconstruida, más audible en el caso del valor de 67%.

En todos los procesadores evaluados ha sido necesario un ajuste cuidadoso de los parámetros de control, especialmente el de las constantes λ de actualización de las estimaciones recursivas, ya que influyen considerablemente en la cantidad de ruido cancelado y de distorsión presente en la salida.

Otra tarea diferente ha sido la evaluación objetiva de los resultados obtenidos. Inicialmente se usaron estimadores de mejora basados en la SNR y en parámetros LAR y cepstrales (apartado 7.1). Los resultados de las evaluaciones no fueron suficientemente satisfactorios y en algunos casos se mostraron poco acordes con la apreciación subjetiva, especialmente los que utilizan las distancias dLAR y dRCEP, que parecen medir más la distorsión del procesador que la mejora subjetiva ocasionada sobre la señal de voz. Por eso, con el segundo procesador propuesto (ANS-MW en el punto 7.3) se han utilizado medidas objetivas basadas en la respuesta sujettiva del oído humano, y en el apartado 7.2 se ha propuesto el método E-RASTI de evaluación, que ofrece un enfoque diferente a los métodos basados en SNR y distancias, a la hora de estimar la calidad objetiva de la voz. Después de muchos experimentos, parece que la evaluación basada en los umbrales de enmascaramiento a través de la relación NMR es la más adecuada si se elige el grupo de los evaluadores objetivos tipo SNR. Los resultados con el E-RASTI han sido buenos igualmente, lográndose pautas de mejora de habla similares a las obtenidas por la ganancia GNMR y acordes también con la impresión subjetiva de la voz de salida del procesador.

Así, las principales ventajas del E-RASTI son, por una parte que no necesita un alineamiento temporal perfecto de las señales a comparar y por otra que evalúa mejor que los métodos de tipo SNR la derreverberación producida por el array. El principal defecto mostrado por el índice E-RASTI es que en algunas ocasiones puede valorar una mejora excesiva de la señal, probablemente por la sobresustracción ocasionada por el algoritmo de mejora. Este defecto, que ha aparecido cuando se calcula el índice E-RASTI sobre la señal $x_0(t)$ de referencia, no se ha mostrado sin embargo cuando se aplica a las señales de salida de los procesadores evaluados.

Una vez realizados con éxito los experimentos preliminares con una simulación *software*, se está en disposición de pasar a la fase más comprometida de implementación de un prototipo de array en tiempo real, lo cual se describe a continuación, en la Parte 3 de la Tesis.

PARTE 3

IMPLEMENTACIÓN SOBRE DSP DE

UN PROTOTIPO EN TIEMPO REAL

8 DESCRIPCIÓN DEL PROTOTIPO EN TIEMPO REAL IMPLEMENTADO

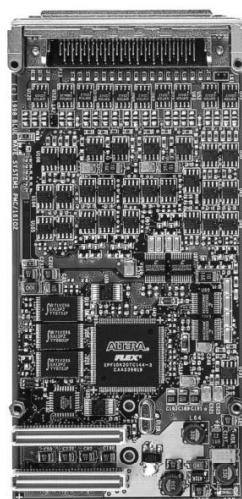
Después de los experimentos realizados y descritos en apartados anteriores, utilizando procesadores emulados por *software* (con señal multicanal real y contaminación acústica tanto real como simulada) es necesario encaminarse a uno de los objetivos más importantes de este trabajo de Tesis, es decir la implementación de un prototipo de array microfónico que aplique las ideas de mejora de señal de voz en condiciones acústicas adversas de ruido y reverberación, con las que se ha trabajado hasta el momento. La descripción del prototipo y los experimentos presentados en esta Parte 3 se expone de forma muy resumida en [Sánchez Bote 02-b] y [Sánchez-Bote 03-a].

8.1 ELEMENTOS *HARDWARE*

Para el proceso de implementación se necesita en primer lugar disponer de un procesador digital de señal (DSP) suficientemente rápido, que pueda albergar las propuestas algorítmicas con las que se ha experimentado hasta el momento, para una aplicación en tiempo real. Además se debe poseer la tecnología de adquisición adecuada para la conversión analógica digital síncrona de al menos 15 canales con calidad de audio.

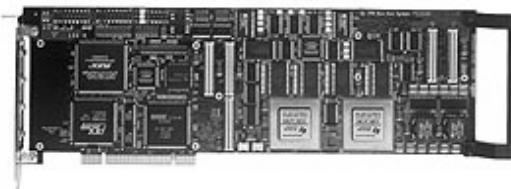
Cuando se inició la implementación del prototipo en tiempo real, el equipo hardware que mejor cumplía los requisitos propuestos era el conjunto BLUE WAVE PMC/16IO2 [BWS 99-b] + BLUE WAVE PCI/C6600 [BWS 99-a]. Ambos equipos son dispositivos de tipo tarjeta que se conectan al bus PCI de un ordenador personal, el primero de ellos mediante el protocolo PMC (*PCI Mezzanine Card*). En lo sucesivo el ordenador personal que soporta las dos tarjetas se llamará *Host PC*.

La tarjeta BWS-PMC/16IO2, que se muestra en la Figura 106, comprende un conversor analógico digital para 16 entradas y un conversor digital analógico para 2 salidas, que se comunica a una elevada velocidad de transferencia con la tarjeta que contiene el procesador DSP. Por tanto, este equipo tiene capacidad suficiente para ser alimentado por los 15 canales analógicos del array anidado propuesto, con la ventaja de que aun se dispone de un canal adicional que puede ser utilizado en los experimentos. La resolución de las entradas es de 20bits mientras que las salidas tienen 24bits, pudiendo llegar la frecuencia de muestreo hasta 48kHz. El terminal analógico de las 16 entradas y de las dos salidas es de tipo diferencial, con una excursión máxima de la tensión eléctrica dada por $\pm 10V$ (valor de pico). El acceso externo a las salidas y entradas analógicas se hace mediante un conector ULTRA SCSI de 68 pines. Cada uno de los 16 + 2 canales analógicos dispone un *buffer* de 1024 muestras en una memoria FIFO. El tamaño de la FIFO es un parámetro muy importante a la hora de implementar el sistema, ya que va a dictar el tamaño máximo del bloque de señal de voz que se puede procesar en cada instante de tiempo.



Analog I/O	■	Analog Inputs (differential)	Analog Outputs
Number of Channels	■	16	2
Conversion	■	Crystal Delta Sigma	Crystal Delta Sigma
	■	20 bit ADCs	24 bit DACs
Buffers	■	1K FIFO buffer per channel	1K FIFO buffer per channel
		± 10 V (pk)	± 10 V (pk)
Analog Voltage Range	■	Up to 48 kHz	Up to 48 kHz
Sample Rates	■	On-module 12.228 MHz	On-module 12.228 MHz
Clock Sources	■	crystal oscillator	crystal oscillator
		External clock input	External clock input
Digital I/O	■		
Number of Inputs /Outputs	■	4 lines, each configurable as input or output	
Voltage Levels	■	TTL levels	
Carrier Board Interface	■	32 bit slave PCI interface, 48M bytes/sec sustained	
Physical	■	Standard single width PMC module with 3.3 V or 5 V signalling	
Warranty	■	PMC/16IO2 comes with a 12 month worldwide warranty from the date of invoice.	

Figura 106. Principales características de la tarjeta de adquisición y conversión analógica-digital-analógica tipo PMC y modelo BLUE WAVE PMC/16IO2.



Processors	■	Two TMS320C6201/TMS320C6701 DSPs providing 3200 MIPS/2 GFLOPS peak performance.
External Memory	■	External Memory Interface (EMIF) on each processor provides access to 512K bytes Flash Memory and 4, 16 or 32M bytes of 1 ws SDRAM. Additionally the EMIF on each processor enables interprocessor communication and gives each DSP access to shared SRAM and the on-board PCI bus.
Shared Bus	■	Two banks of 256K or 1M byte of SRAM available to both C6000 processors via the shared buses.
PMC Site	■	A single width IEEE 1386.1 compliant PMC site includes Blue Wave's Direct-Connect interface.
On-Board PCI Bus	■	The on-board PCI bus provides access to the shared bus resources, PMC site, Test Bus Controller and the host PC.
Expansion Interface Site	■	A plug-in I/O interface module can be fitted to the board.
Serial Ports	■	Two bidirectional, synchronous Enhanced Buffered Serial Ports on each C6000 are connected to the PMC site Direct-Connect interface and to the Interface Module site.
Host Port Interface	■	The host port on each processor is connected together and provides an alternative data path from/to each DSP.
Warranty	■	PCI/C6600 comes with a 12 month worldwide warranty from the date of invoice.

Figura 107. Principales características de la tarjeta DSP modelo BLUE WAVE PCI/C6600.

La tarjeta BWS-PCI/C6600, cuyas características más importantes se destacan en la Figura 107, contiene el corazón del sistema, es decir el procesador DSP. Se conecta al *Host PC* mediante el bus PCI y a la tarjeta BWS-PMC/16IO2 también mediante el bus PCI, en este caso con el protocolo PMC. Tiene una memoria SRAM (*Shared RAM*) que contiene dos

bloques de 256K posiciones de 32 bits (en total 2MB) y cuyo acceso es muy rápido, a través del *Shared BUS*. Además se tiene una memoria SDRAM con dos bloques de 4M posiciones de 32 bits totales disponibles (en suma 16MB) y una memoria *Flash* de 1MB.

El procesador DSP es de modelo TMS320C6701 [TI 99]. Es un procesador en punto flotante con un reloj de 167MHz capaz de realizar hasta 1GFLOPS (*Floating-point Operations Per Second*) (el procesador puede realizar hasta ocho FLOPS de 32 bits por ciclo de reloj). Además el TMS320C6701 tiene dos bloques de 64KB cada uno, para memoria de programa y datos internos.

Los micrófonos seleccionados para implementar el array anidado son del modelo AKG-C417 (ver la Figura 108) cuyas características más importantes se muestran en la Figura 109. Se trata de una cápsula prepolarizada de tipo condensador-electret –para más detalles sobre este tipo de transducción véase [Sánchez-Bote 02-a]–. El receptor es de presión, por lo que ofrece una respuesta polar omnidireccional $-D_i(\theta, \varphi) = 1$ tal y como aparece en (59)– y una respuesta en frecuencia relativamente plana, tanto en módulo como en fase. Se conecta mediante una salida balanceada de tipo XLR compatible con la alimentación *PHANTOM*, necesaria para proveer de tensión continua al diminuto preamplificador integrado en la cápsula del micrófono. Los micrófonos han sido adquiridos con números de serie consecutivos, para facilitar en la mayor medida posible que sus características sean similares (micrófonos pareados).



Figura 108. Micrófono prepolarizado modelo AKG-C417.

Existe por tanto una característica diferenciadora con respecto a los micrófonos usados en las pruebas preliminares (base CMU, descrita en el punto 6.2), y muy relacionada con las cualidades finales del procesador implementado. Esta característica es la directividad de cada micrófono. Recuérdese que los micrófonos de la base CMU eran directivos, lo que confería al array final una directividad adicional. En este caso se optó por los micrófonos omnidireccionales por varias razones que se detallan a continuación.

- Los micrófonos de presión (omnidireccionales) tienen unas características electroacústicas más fáciles de reproducir de un ejemplar a otro. Según esto, dos micrófonos diferentes tendrán una respuesta en frecuencia y directividad similares hasta una determinada frecuencia, sobre todo si el micrófono es pequeño con relación a la

longitud de onda asociada a esa frecuencia (en la práctica esto se cumple por debajo de los 8kHz). Esta peculiaridad tiene un interés práctico inmediato, ya que será más fácil alinear la respuesta de los 15 micrófonos del array, mediante simples correcciones de nivel global, una vez implementado el prototipo final.

Datos técnicos:

Funcionamiento:	transductor de condensador con carga permanente
Característica direccional:	omnidireccional
Gama de frecuencia:	20–20.000 Hz
Sensibilidad a 1000 Hz:	10 mV/Pa Δ – 40 dBV re 1 V/Pa
Impedancia eléctrica a 1000 Hz:	200 Ω
Impedancia de carga recomendada:	\geq 1000 Ω
Presión acústica límite por 1 %/3 % de distorsión:	118 dB SPL/126 dB SPL
Nivel de presión acústica equivalente:	34 dB (DIN 45412)
Toma de corriente (con B 9 o MPA II):	2,2 mA
Condiciones climáticas aceptables:	<ul style="list-style-type: none"> – Gama de temperatura: -20° C ... + 60° C – Humedad relativa del aire: a + 20° C, 99 %
Tipo de conector:	C 417/B-lock; jack mono de 3,5 mm C 417: conector XLR de tres polos
Modo de conexión:	C 417/B-lock: punta del jack: conductor de sonido (inphase) mango: masa C 417: XLR: espiga 1: masa espiga 2: conductor de sonido (inphase) espiga 3: conductor de sonido
C 417 WL 900: WL: punta del jack: conductor de sonido (inphase) mango: masa	
Material de la caja:	parte sup.: plástico parte inf.: latón
Superficie:	negro opaco
Dimensiones:	7,5 Ø x 15 mm
Longitud de cable:	C 417: 3 m C 417/B-lock, C 417 WL: 1,5 m
Peso neto (sin cable)/bruto:	C 417/B-lock: 8 g/160 g C 417: 8 g/220 g

Volumen de suministros:	W 407 pantalla antiviento H 40/1 pinza-prendedor H 41/1 alfiler-prendedor
Accesorios recomendados	C 417/B-lock: B 29 alimentador de batería MPA II MicroMic II Phantom Power Adapter

Respuesta de frecuencia:

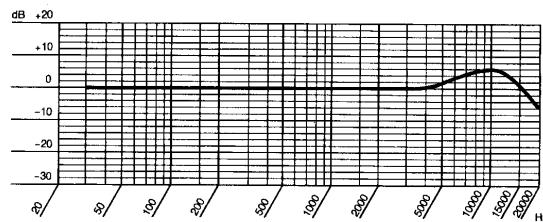
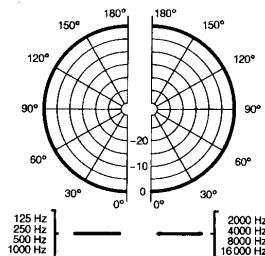


Diagrama polar:



Este producto cumple con la norma EN 50 082-1.

Figura 109. Características del micrófono prepolarizado modelo AKG-C417.

- Los micrófonos de presión (omnidireccionales) tienen una directividad constante con la frecuencia. Por contra, los micrófonos de gradiente de presión (directivos) suelen tener una característica polar muy variable con la frecuencia. Sólo los modelos de más alta calidad y precio mantienen, a duras penas, la directividad con la frecuencia. Eso es muy perjudicial para el prototipo que se propone implementar, ya que una respuesta lateral desigual de los micrófonos haría imposible el objetivo de una directividad controlada para el array de 15 micrófonos.
- Los micrófonos de gradiente de presión (directivos) suelen ser muy sensibles a los obstáculos cercanos, que pueden modificar apreciablemente su respuesta polar. En el prototipo final, los micrófonos deberán estar situados en un soporte rígido, que deberá situarse cerca de dichos micrófonos, a modo de estructura de soporte. Si este soporte obstruye o modifica acústicamente las diminutas aberturas laterales que confieren a los micrófonos directivos sus características polares, se estará modificando la directividad de cada cápsula microfónica y por tanto será muy difícil controlar la directividad del array.
- Se desea que el array tenga un eje principal de apuntamiento variable, determinado en teoría por el par (r_0, θ_0). Esta característica se obtiene como es sabido mediante la alineación temporal de los 15 micrófonos del array para esa dirección, pudiendo abarcar el apuntamiento al menos desde $\theta_0 = 0^\circ$ (*endfire*) hasta $\theta_0 = 90^\circ$ (*broadside*). Si los

micrófonos son directivos será imposible configurar todos los apuntamientos variables del array, ya que la dirección de máxima captación de cada micrófono no coincidirá en general con la DOA seleccionada. Consecuentemente, si se eligen micrófonos directivos, el array debería tener un apuntamiento fijo o al menos con una variación angular que difiriese poco del eje principal de captación de cada micrófono.

Por todo ello finalmente se han elegido los micrófonos C417, que tienen prestaciones de nivel profesional, y debido a su pequeño tamaño son adecuados para los fines que se persiguen.

Además, se ha previsto la incorporación al sistema de un micrófono adicional, llamado de referencia (véase la Figura 112). Será el encargado de captar en cada una de las pruebas una muestra de voz limpia –recuérdese que la referencia se denomina $x_0(t)$ a lo largo de la Tesis– que sirva de comparación en los experimentos. Por una parte es conveniente que este micrófono sea similar a cada uno de los 15 restantes que integran el array, ya que así las diferencias de respuesta entre micrófonos no se considerarán como distorsión en los experimentos. Sin embargo esto presenta un inconveniente y es que al ser el micrófono de referencia omnidireccional tiene bastante sensibilidad a la captación de ruido y reverberación. Por eso, cuando dicho micrófono se use en condiciones de mucho ruido y reverberación, será necesario que el orador lo acerque mucho a su boca, para que la señal captada pueda considerarse como libre de perturbación acústica. Existe la posibilidad de hacer captación no simultánea de la señal de referencia, siempre que la señal vocal a captar sea pregrabada (y por tanto repetible) y posteriormente, cuando se procesen los resultados de los experimentos, se realice una alineación temporal con los micrófonos del array.

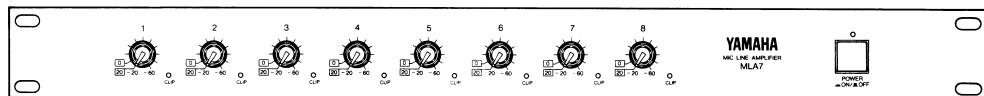
Los micrófonos necesitan ser amplificados por un preamplificador multicanal de al menos 15 entradas, que convierta el nivel de tensión microfónico, de unos pocos milivoltios, al nivel de línea necesario para alimentar a la tarjeta BWS-PMC/16IO2. Este elemento representa por tanto una interfaz analógica-analógica. Además, este equipo necesita disponer de alimentador tipo *PHANTOM* en cada una de sus entradas. Por eso, para este fin se ha elegido el modelo YAMAHA-MLA7, cuyas especificaciones se muestran en la Figura 110. Se trata de un preamplificador para 8 micrófonos (se necesitan por tanto dos de estas unidades) con *PHANTOM* de 48V en cada uno de los canales. Las entradas a las que se conectan los micrófonos son de tipo XLR balanceadas mientras que las salidas son de tipo *jack* no balanceadas, que pueden proporcionar hasta $\pm 7.75V$ de pico a la salida sin saturación, lo que hace al MLA7 muy apropiado para los $\pm 10V$ de pico admisibles por la tarjeta PMC. Cada uno de los canales tiene un selector de amplificación que permite ajustar independientemente el nivel entregado a la salida por cada uno de los canales.

Los micrófonos se conectan al preamplificador MLA7 mediante una manguera de 16 canales de calidad microfónica profesional, con una longitud total de 15m. Las 16 salidas del MLA7 se conectan a la entrada ULTRA SCSI de la tarjeta PMC mediante un cable multipar adecuado.

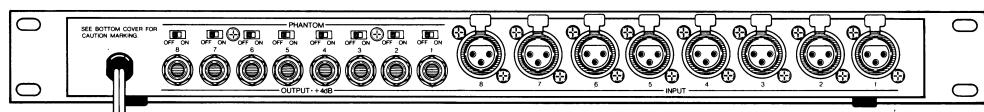
La colocación de los micrófonos en un soporte ha tenido algunas alternativas durante el proceso de implementación del prototipo. Inicialmente se situaron los micrófonos en posición elevada sobre un soporte cilíndrico rígido –ver la Figura 111(a)–. Sin embargo se comprobó que esta configuración producía alteraciones significativas en la respuesta en frecuencia del micrófono alrededor de 7kHz, debidas probablemente a la reflexión de la onda acústica en este soporte. Por tanto se optó finalmente por enrasar los micrófonos como muestra la Figura 111(b), de tal manera que desapareciese la posibilidad de que el micrófono captase la

reflexión cercana del cilindro de soporte. Se comprobó que esta disposición producía mejor respuesta en frecuencia de cada uno de los micrófonos en el array.

FRONT PANEL



REAR PANEL



GENERAL SPECIFICATIONS

Total Harmonic Distortion	Less than 0.1%, 20Hz ~ 20kHz @ +4dB into 10 k-ohms.
Frequency Response	+1, -3dB, 20Hz ~ 20kHz @ +4dB into 10 k-ohms.
Hum and Noise (20Hz ~ 20kHz, 150-ohm termination)	-128dBu equivalent input noise, PAD 0 GAIN control max. -87dBu equivalent input noise, PAD 20 GAIN control min.
Maximum Voltage Gain	64dB CH IN to CH OUT.
Crosstalk	-70dB at 1kHz/10kHz, adjacent channels

Power Requirements	U.S. & Canadian models 120V (105 – 130V) AC, 60Hz General model 110 – 120/220 – 240V AC, 50/60Hz	
Power Consumption	U.S. & Canadian models 20W General model 20W	
Dimensions (W x H x D)	480 mm x 45.5 mm x 231.6 mm (18-7/8" x 1-3/4" x 9-1/8")	
Weight	3.25 kg (7.2 lbs)	
<i>*0dB is referenced to 0.775V RMS. *Specifications subject to change without notice.</i>		

● INPUT SPECIFICATIONS

CONNECTION	ACTUAL LOAD IMPEDANCE		FOR USE WITH NOMINAL	SENSITIVITY** (AT MAX. GAIN)	INPUT LEVEL		CONNECTOR	
	PAD	GAIN			NOMINAL	MAX. BEFORE CLIP		
INPUT	OFF (0dB)	-60dB	4k ohms	50 ~ 250 ohm Microphones or 600 ohm Lines	-60dB μ (0.775mV)	-60dB μ (0.775mV)	-44dB μ (4.88mV)	XLR-3-31 type (Balanced)
		-20dB		600 ohm Lines	-20dB μ (77.5mV)	-20dB μ (77.5mV)	-4dB μ (488mV)	
	ON (20dB)			600 ohm Lines	0dB μ (775mV)	0dB μ (775mV)	+16dB μ (4.88V)	

● OUTPUT SPECIFICATIONS

CONNECTION	ACTUAL SOURCE IMPEDANCE	FOR USE WITH NOMINAL	OUTPUT LEVEL		CONNECTOR
			NOMINAL	MAX. BEFORE CLIP	
OUTPUT	150 ohms	10k ohm Lines	+4dB μ (1.23V)	+20dB μ (7.75V)	Phone Jack (Unbalanced)

* : In these specifications, when dB represent a specific Voltage, 0dB μ is referenced to 0.775V.

** : Sensitivity is the level required to produce an output of +4dB (1.23V).

Figura 110. Características principales del preamplificador microfónico multicanal modelo YAMAHA-MLA7.

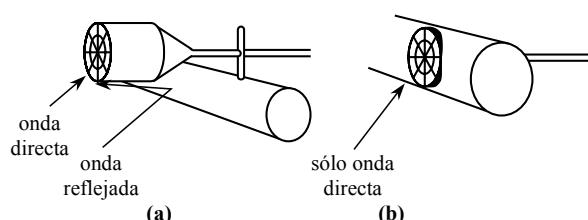


Figura 111. Colocación del micrófono sobre el soporte. (a) Elevado. (b) Enrasado.

En las primeras fases de la implementación del prototipo, dentro del primer nivel de experimentos, la posición de la fuente (r_0, θ_0) se introducía manualmente por un operador mediante el teclado del ordenador *Host PC*. Sin embargo, finalmente se diseñó un método de apuntamiento más sofisticado. Éste consiste en montar sobre el prototipo una *Web Cam* que proporcione una imagen frontal del array microfónico, de tal manera que el operador, mediante un clic del ratón, indique el punto de la escena al que se quiere dirigir el apuntamiento del prototipo de array, mediante una compensación de retardos.

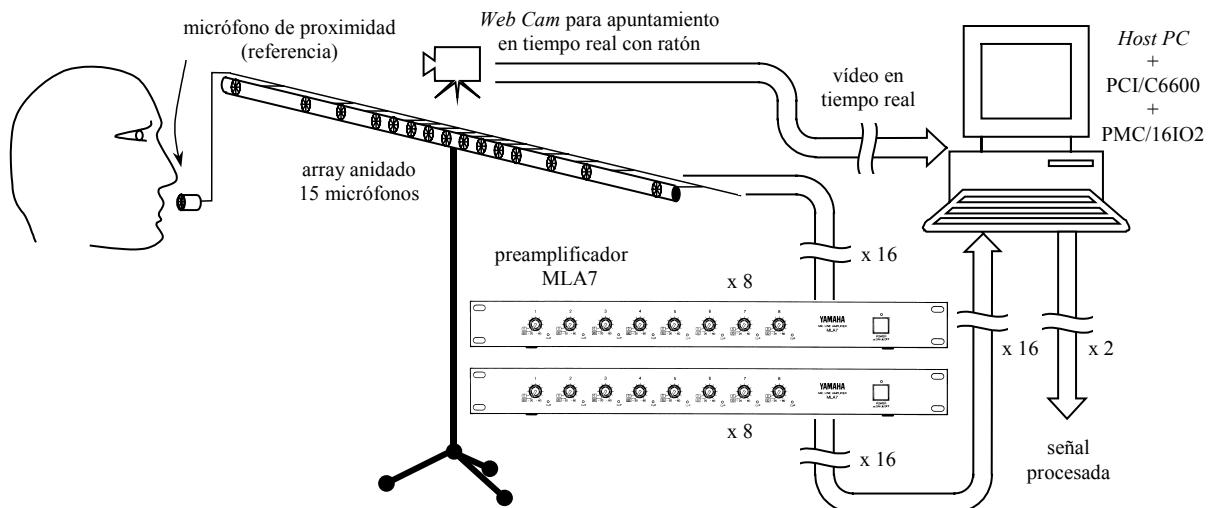


Figura 112. Sistema completo de array anidado de 15 micrófonos basado en DSP.

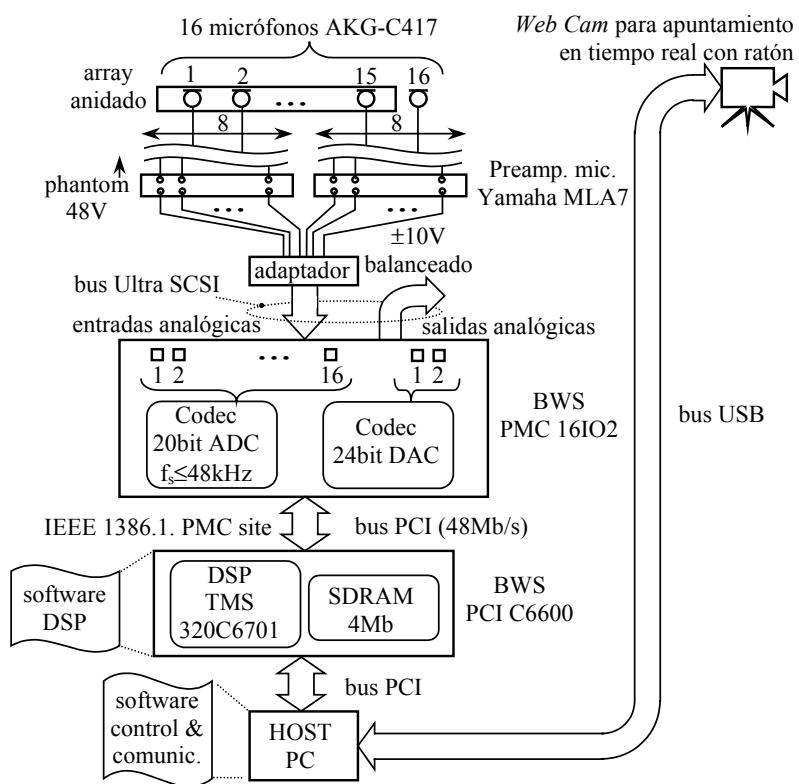


Figura 113. Esquema detallado de la implementación del prototipo propuesto de array microfónico de 15 canales anidados.

Por supuesto, este método de apuntamiento elimina la posibilidad de controlar la distancia r_0 a la que se sitúa la fuente, ya que la cámara sólo proporciona el ángulo θ_0 respecto al eje del array. No obstante, se ha comprobado que a partir de $r_0 > 5\text{m}$ la distancia no influye en el resultado del conformador, y ése es un alejamiento más que razonable a partir del cual se colocarán normalmente las fuentes de voz para ser mejoradas. Por tanto, cuando se use la *Web Cam* siempre se utilizará la aproximación de campo lejano, considerada en el punto 2.1.8 de esta Tesis. La *Web Cam* tiene una óptica de $\pm 45^\circ$ suficiente para la mayoría de los cambios finos de DOA, aunque no para pasar de apuntamiento *endfire* a *broadside*, y viceversa. Por tanto, se deberá fijar la cámara según un ángulo inicial de apuntamiento, previsto inicialmente. La *Web Cam* se conecta al *Host PC* mediante el protocolo USB. En la Figura 112 se representa el diagrama de bloques del prototipo implementado finalmente, con cada uno de los elementos más importantes ya comentados.

En la Figura 113 se representa un esquema detallado de implementación del prototipo de array microfónico propuesto.

Las Figuras 114, 115, 116 y 117 corresponden a galerías fotográficas del prototipo final implementado. En la Figura 114 se muestran los detalles constructivos del array microfónico. Los micrófonos omnidireccionales van enrasados en una estructura metálica tubular de soporte. Dicha estructura está constituida por una barra cilíndrica hueca de 10mm de diámetro sujetada a un pie regulable en altura. Las cápsulas microfónicas traspasan este soporte.



Figura 114. Galería fotográfica en la que se muestran los detalles constructivos del array microfónico prototipo implementado. Los micrófonos omnidireccionales van enrasados en una estructura tubular de soporte. Esta disposición aminora los efectos acústicos negativos del soporte sobre la respuesta en frecuencia de los micrófonos.

En la Figura 115 aparece el array de micrófonos finalmente implementado. También el micrófono de referencia para la captación de voz limpia. La imagen inferior derecha muestra la conexión de cada uno de los micrófonos al cable de 16 canales y 15m de longitud.



Figura 115. Galería de fotografías en la que se aprecia el array de micrófonos finalmente implementado. También al autor portando el micrófono de referencia para la captación de voz limpia. La imagen inferior derecha muestra la conexión de cada micrófono con el cable de 16 canales y 15m de longitud.

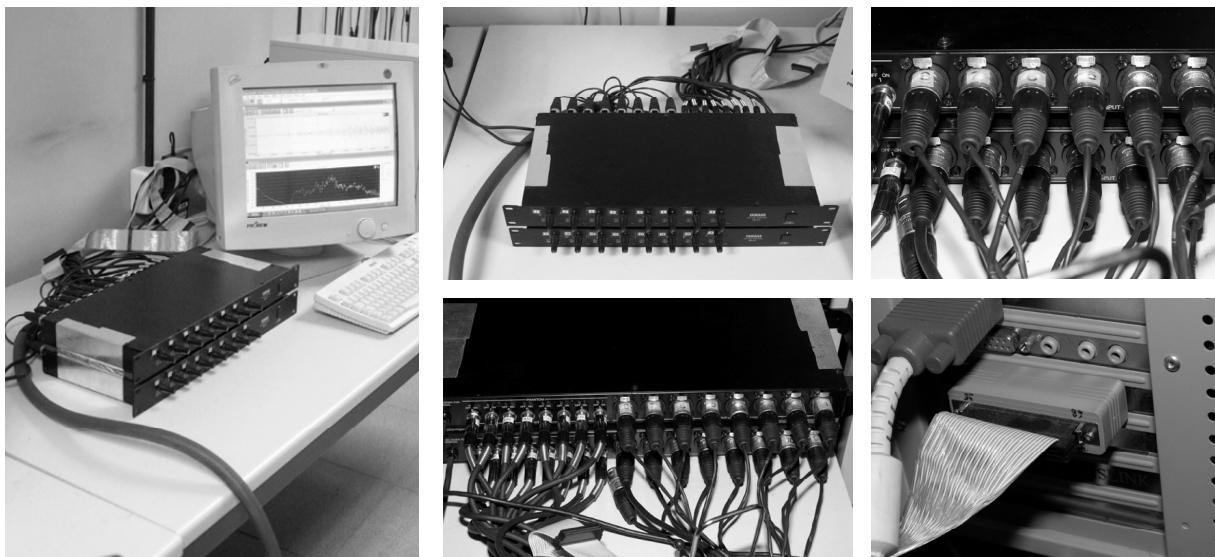


Figura 116. Galería de fotografías en la que aparece el preamplificador MLA7 conectado al *Host PC*. Se muestran detalles del panel posterior de conexiones con las entradas de micrófono (conectores XLR) y las salidas del preamplificador (conectores jack). En la imagen inferior izquierda se muestra la conexión del cable adaptador a la entrada de la tarjeta PMC mediante el conector ULTRA SCSI.

En la Figura 116 se visualiza el preamplificador MLA7 conectado al *Host PC*. Se muestra en detalle el panel posterior de conexiones, en el que se aprecian las entradas XLR para los micrófonos y las salidas jack del preamplificador. También la conexión a la entrada de la tarjeta PMC mediante el cable adaptador jack-ULTRA SCSI.

En la Figura 117 se visualiza la *Web Cam* usada para el apuntamiento del array. La fotografía de la derecha muestra en segundo plano la pantalla del *Host PC* en la que se reproduce la imagen entregada por la cámara y con la que el usuario puede apuntar la dirección de máxima captación del array mediante un clic del ratón.



Figura 117. Fotografías que muestran detalles de la Web Cam utilizada como soporte de apuntamiento del array de micrófonos. En la fotografía de la derecha aparece en segundo plano la pantalla del *Host PC* en la que se reproduce la imagen tomada por la cámara y con la que el usuario puede apuntar la dirección de máxima captación del array mediante un clic del ratón.

8.2 MEDIDAS ELECTROACÚSTICAS DE LOS MICRÓFONOS DEL ARRAY

La adecuada respuesta de cada uno de los micrófonos del array es un factor muy importante para que el array microfónico funcione correctamente a la hora de la conformación de haz. Recuérdese –ecuación (3)– que la respuesta en frecuencia de cada micrófono i del array viene determinada fuertemente por su sensibilidad \mathbf{S}_i [$V \cdot Pa^{-1}$].

No importa mucho que el módulo de la sensibilidad $|\mathbf{S}_i|$ varíe con la frecuencia, puesto que se puede corregir con facilidad. Sí puede ser importante, sin embargo, que haya diferencias considerables de módulo de un micrófono con respecto a otro. A la hora de implementar un conformador convencional basado en la estructura de retardo y suma, incluso esta diferencia del módulo de la sensibilidad entre dos micrófonos del array influye de forma poco importante, siempre que no sea excesiva. Al sumar los 15 canales alineados en fase, las diferencias de nivel sólo modifican (muy levemente) la amplitud de los lóbulos del diagrama polar de directividad, pero esto no afecta a la dirección de apuntamiento y a otras características importantes del array.

Al conformador superdirective sí le afectan más ese tipo de despareamientos en módulo. Recuérdese que el conformador superdirective (punto 2.2.3) basa la cancelación del sonido desde direcciones laterales en la resta de pares de micrófonos. Si se restan dos micrófonos de respuestas diferentes, no se va a obtener un nulo en la dirección deseada, como sería de esperar, y por tanto la conformación superdirective va a ser defectuosa.

Otra cuestión es la fase de la respuesta de cada micrófono, $\phi_i = \arg(\mathbf{S}_i)$. Ya se vio en la Tabla 8 cuáles eran los desfases $\Delta\phi_i$ que se necesitan aplicar a cada micrófono para conseguir la alineación temporal de todos los canales del array, para apuntamiento *broadside* y *endfire* y considerando dos frecuencias características ($f = 125Hz$ y $f = 4kHz$). Obviamente, los desfases más pequeños a utilizar siempre corresponden a las frecuencias más bajas. Es en esta

zona del espectro donde una desalineación de fase en la respuesta de los micrófonos puede producir los mayores problemas. De esa manera la banda B_1 es muy conflictiva, ya que, por ejemplo, para la alineación temporal en configuración *endfire* $f = 125\text{Hz}$ se necesita una corrección de fase para los micrófonos 1 y 15 de $|\Delta\phi_{1, 15}| = 31^\circ$. Cualquier diferencia en la respuesta en fase de los micrófonos que supere este valor puede hacer que la conformación no responda como se prevé. De nuevo, este problema afecta en mayor medida a la conformación superdirectiva. Haciendo un símil, se comete menos error relativo al sumar con cierto error de fase dos vectores de igual módulo y fase (conformación convencional) que cuando se restan esos dos vectores con el mismo error de fase. En definitiva, los errores de fase y módulo en las sensibilidades de los micrófonos afectan sobre todo a la conformación superdirectiva y especialmente en las bajas frecuencias.

Se ha medido la sensibilidad \mathbf{S}_i en función de la frecuencia de todos los micrófonos del array (montados en su posición final en el mismo) en una cámara anecoica, en la disposición mostrada por la Figura 118, con el array conectado al preamplificador MLA7 y todos los canales con la misma ganancia. Para ello se han comparado las respuestas de cada micrófono con un micrófono patrón de respuesta plana. Con el objeto de evitar en lo posible la influencia de la cámara anecoica (ondas estacionarias en baja frecuencia), se ha colocado cada micrófono del array cerca de la fuente $r_0 = 0.5\text{m}$ y siempre en el mismo punto. Es decir, para medir dos micrófonos diferentes se ha desplazado al array lateralmente de tal manera que en ambas medidas los dos micrófonos ocupen la misma posición respecto al altavoz.

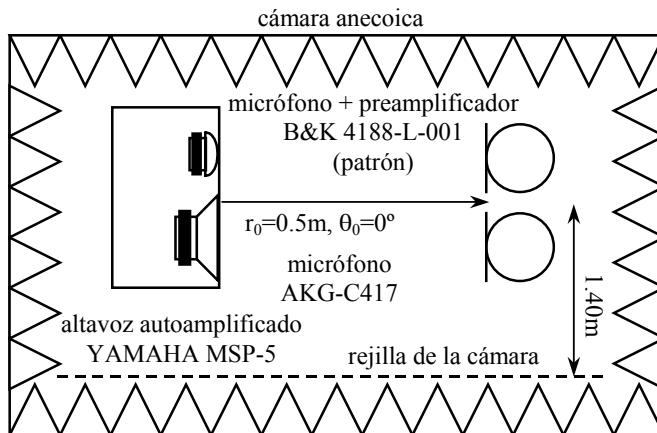


Figura 118. Disposición utilizada para medir la sensibilidad \mathbf{S}_i de los micrófonos del array. El sistema de medida es el PULSE™ 3560C de B&K.

Los resultados de las medidas se muestran en las Figuras 119 y 120 (módulo $|\mathbf{S}_i|$) y en las Figuras 121 y 122 (fase ϕ_i). Con estas medidas se manifiesta cómo los micrófonos están pareados hasta lo que se puede considerar razonable, como corresponde a unos transductores con calidad de audio profesional. Téngase en cuenta que se han medido montados en el array, y que esta disposición puede influir levemente en la respuesta de alta frecuencia.

Las respuestas en módulo $|\mathbf{S}_i|$ son bastante parejas. Todos los micrófonos tienen una sensibilidad a 1kHz en torno a $|\mathbf{S}_i| [\text{dB}] = -40\text{dB re } 1\text{V}\cdot\text{Pa}^{-1}$ que corresponde a $10\text{mV}\cdot\text{Pa}^{-1}$, como informa el fabricante en la Figura 109. Las diferencias globales de nivel no son importantes puesto que pueden ser compensadas globalmente en el procesador, mediante un simple control de ganancia (si se desease, se podría ecualizar también la respuesta en frecuencia de cada micrófono, puesto que los procesadores propuestos trabajan en el dominio de la frecuencia mediante la transformada FFT y no resultaría complicado).

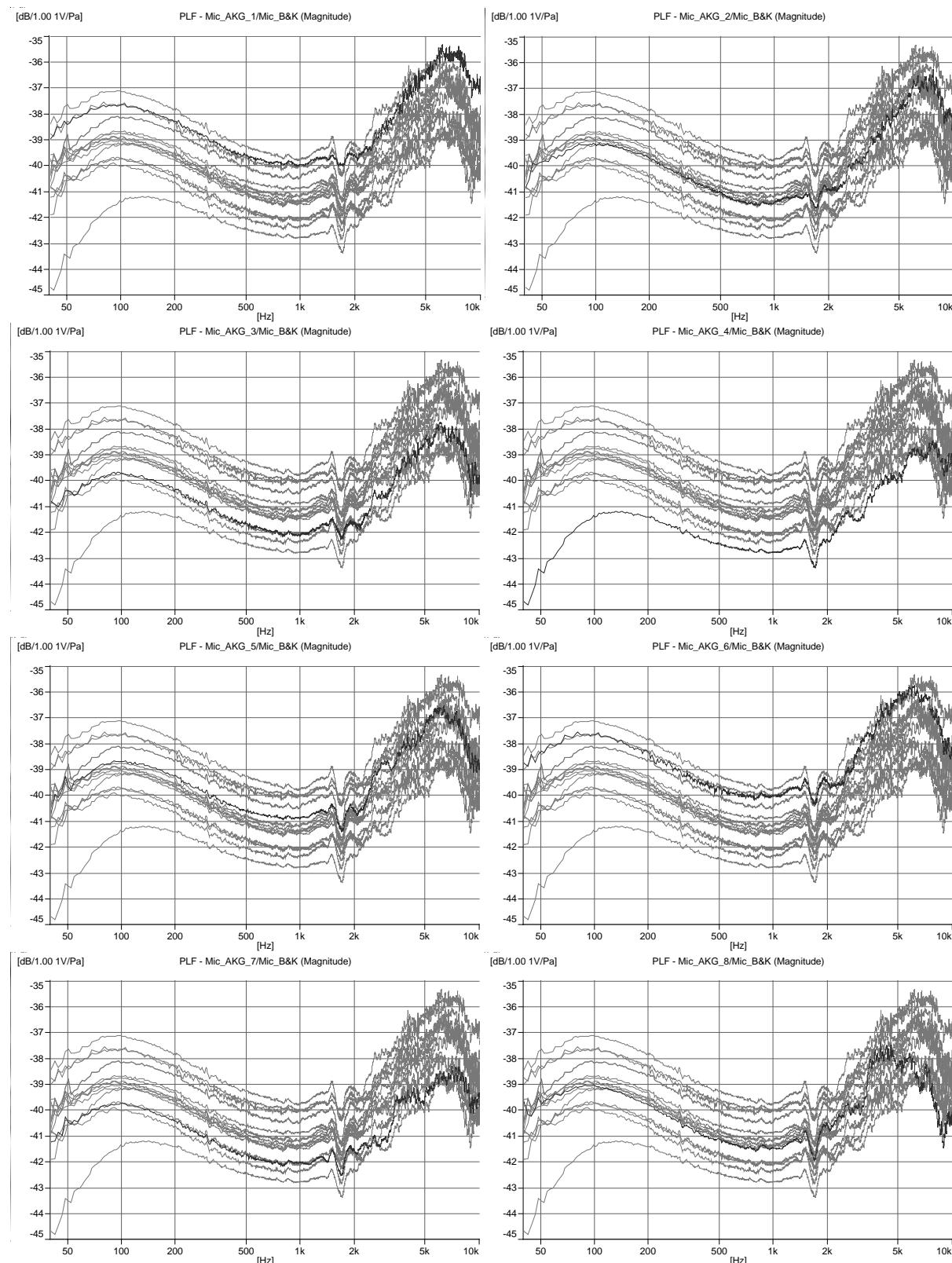


Figura 119. Módulo de la sensibilidad $|S_i|$ [dB] re $1V \cdot Pa^{-1}$ de los micrófonos del array ($i=1$ hasta $i=8$), medida en la configuración de la Figura 118. Se resalta en un tono más oscuro el micrófono al que se refiere el título de la gráfica.

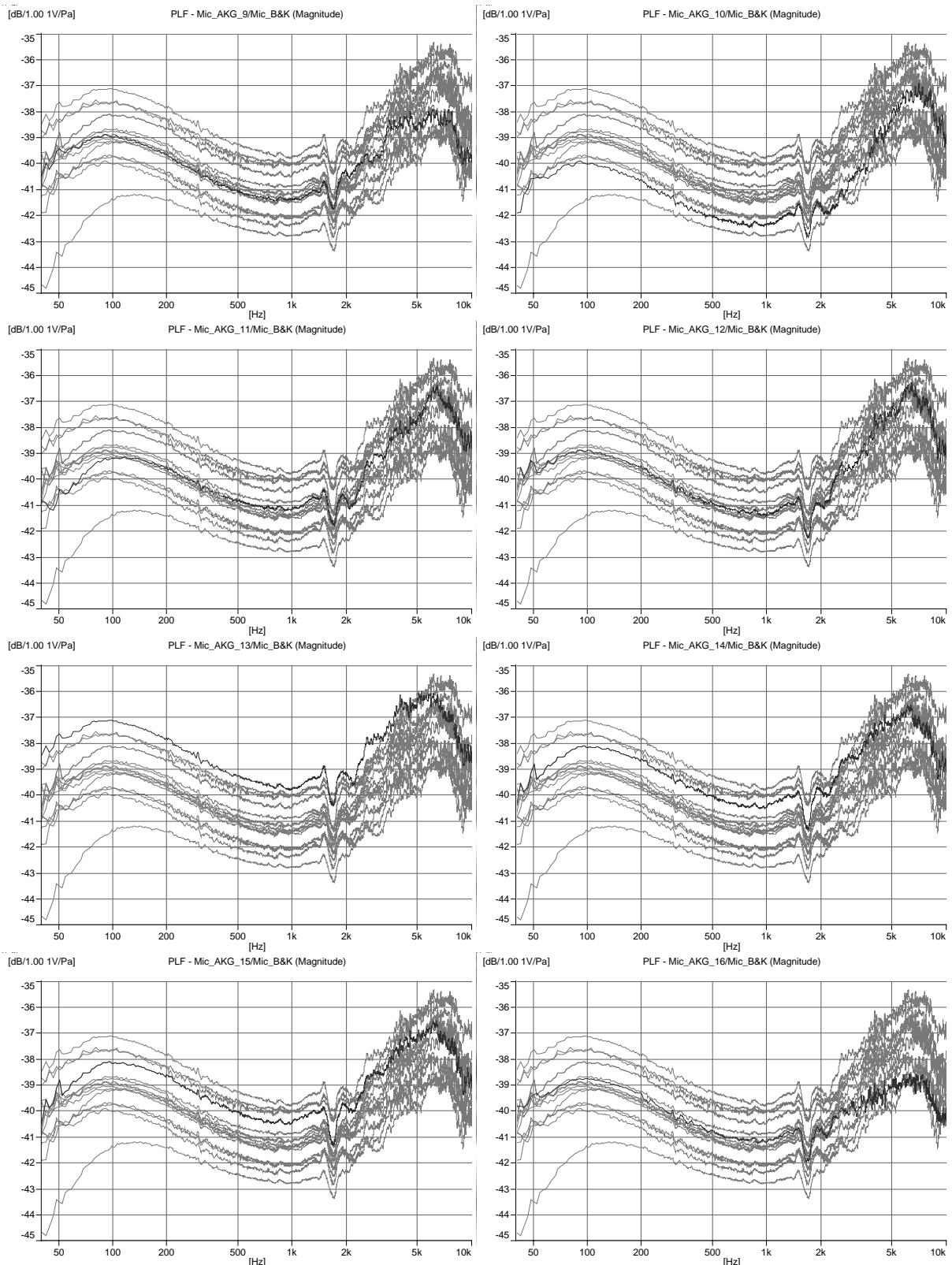


Figura 120. Módulo de la sensibilidad $|S_i|$ [dB] re $1V \cdot Pa^{-1}$ de los micrófonos del array ($i=9$ hasta $i=16$), medida en la configuración de la Figura 118. Se resalta en un tono más oscuro el micrófono al que se refiere el título de la gráfica.

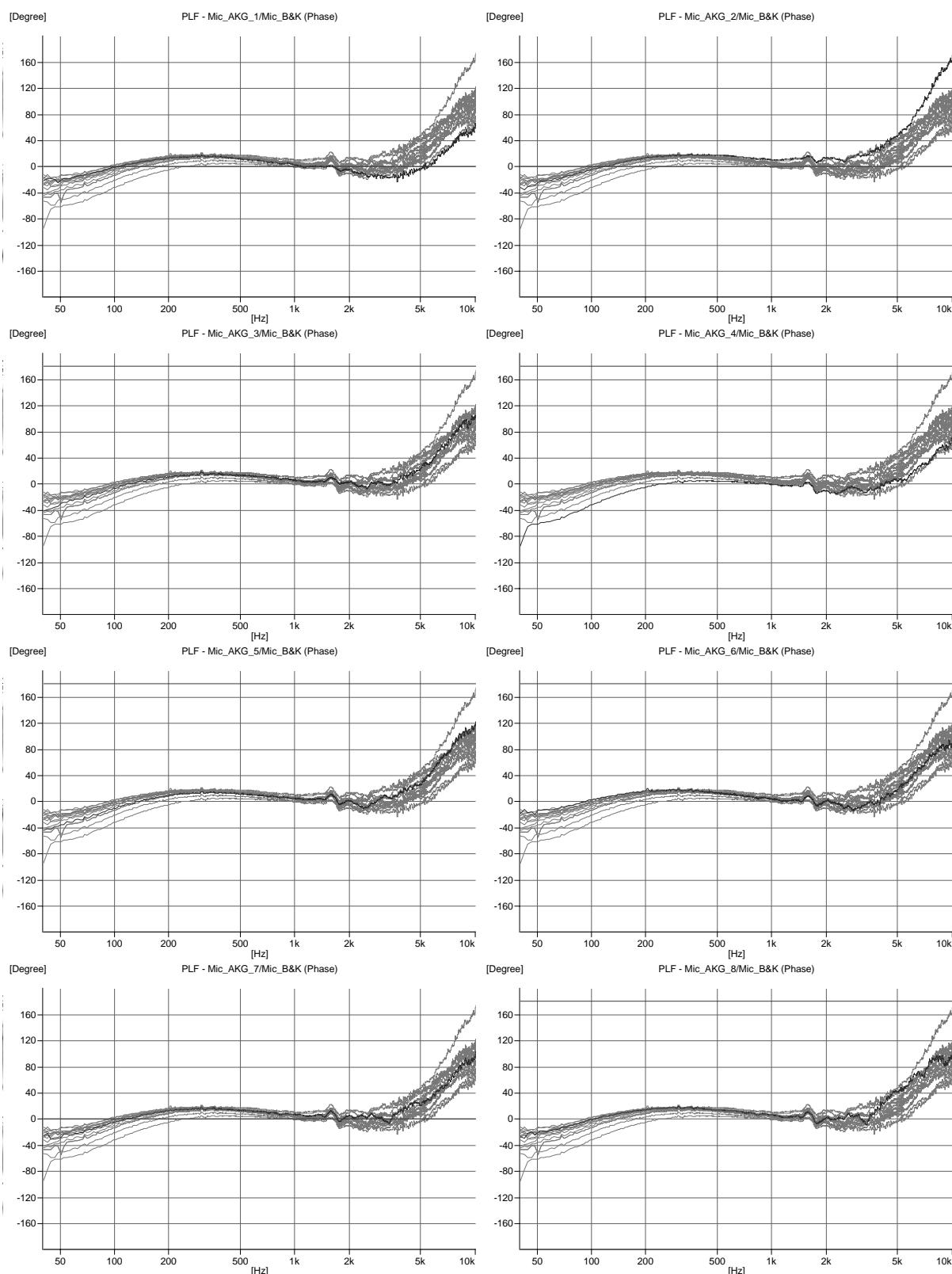


Figura 121. Fase ϕ_i [°] de la sensibilidad de los micrófonos del array ($i=1$ hasta $i=8$), medida en la configuración de la Figura 118. Se resalta en un tono más oscuro el micrófono al que se refiere el título de la gráfica.

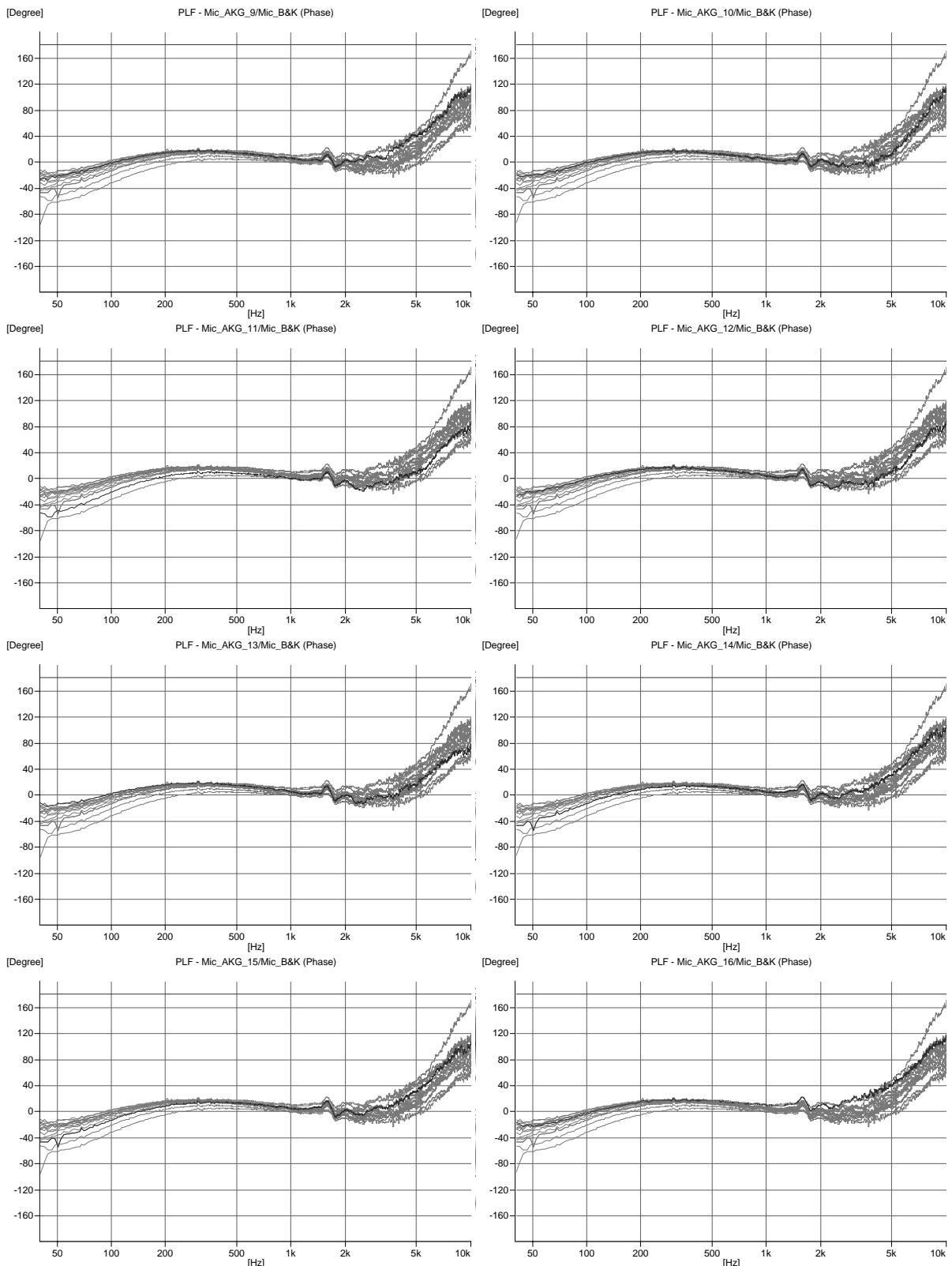


Figura 122. Fase ϕ_i [°] de la sensibilidad de los micrófonos del array ($i=9$ hasta $i=16$), medida en la configuración de la Figura 118. Se resalta en un tono más oscuro el micrófono al que se refiere el título de la gráfica.

Por todo lo anterior, lo que interesa conocer más bien son las diferencias de respuesta global. Esto se representa en la Figura 123(a). En esta figura se visualizan las diferencias de módulo de la sensibilidad $|\mathbf{S}_i| [\text{dB}] - |\mathbf{S}_8| [\text{dB}]$ entre cada uno de los micrófonos y el central del array (se ha compensado el valor de las curvas para que todas pasen por 0dB a la frecuencia de 1kHz). Como se ve, la diferencia mayor se produce en baja frecuencia y, a excepción de los micrófonos $i = 4$ e $i = 11$, no supera 1dB, lo que es bastante favorable.

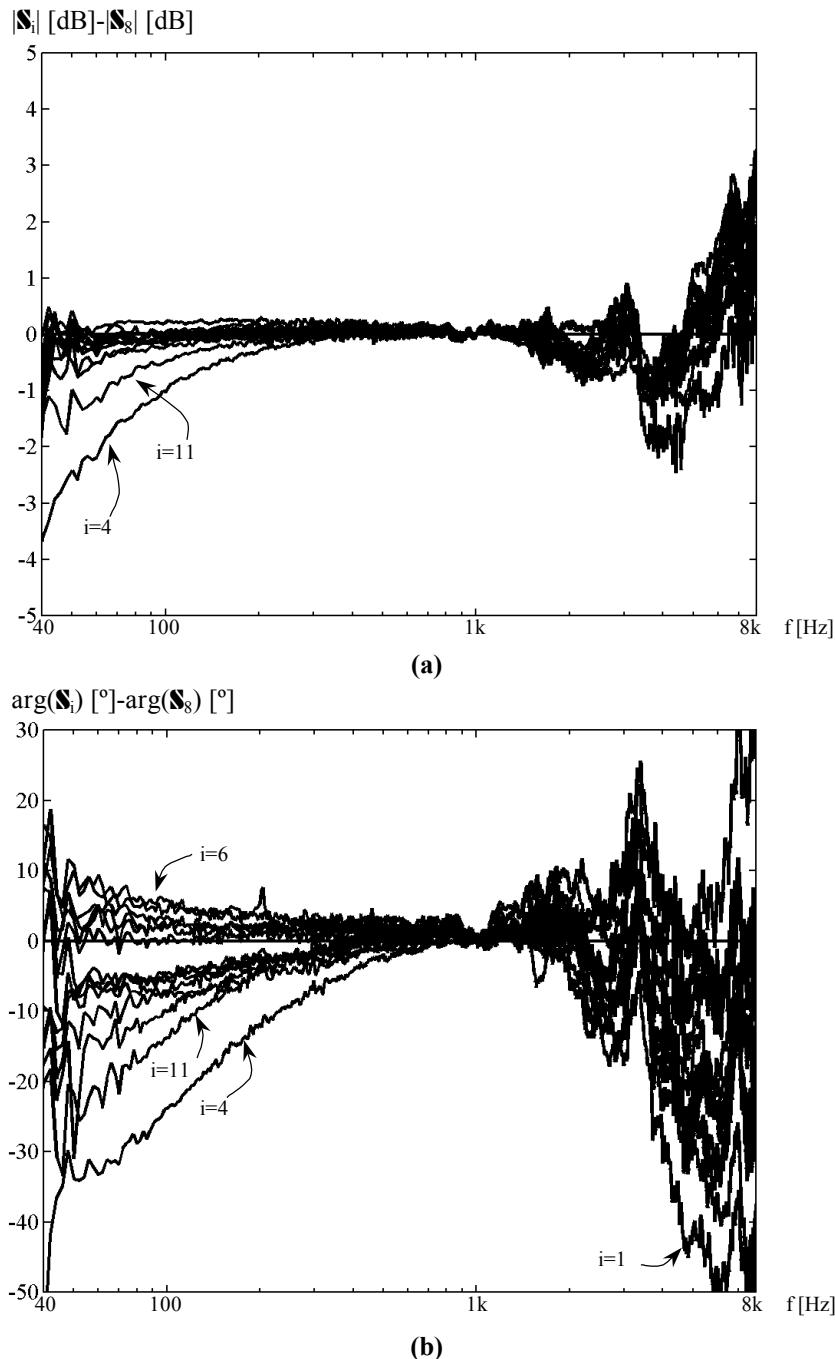


Figura 123. Diferencias de respuesta en frecuencia de los micrófonos del array, con respecto al canal central del mismo $i=8$, a partir de las Figuras 119, 120, 121 y 122. Se ha compensado el valor de las curvas para que todas ellas las pasen por 0 a la frecuencia de 1kHz. Se marcan los micrófonos más dispares. **(a)** Diferencia de módulo $|\mathbf{S}_i| [\text{dB}] - |\mathbf{S}_8| [\text{dB}]$. **(b)** Diferencia de fase $\phi_i [\text{°}] - \phi_8 [\text{°}]$.

En lo que respecta a la fase, en la Figura 123(b) aparece la respuesta en fase relativa al canal central del array, $\phi_i [^\circ] - \phi_8 [^\circ]$. La diferencia en este caso es más preocupante, sobre todo en baja frecuencia. Alrededor de 100Hz la mayor diferencia de fase está en torno a $\pm 10^\circ$, si se exceptúan los micrófonos $i = 4$ e $i = 11$, para los que esta diferencia es mayor.

En conclusión y para finalizar; el análisis de la respuesta en frecuencia de los micrófonos del array indica que es en baja frecuencia donde se pueden tener problemas de pareamiento, sobre todo si se utiliza la conformación superdirective que es la más sensible a la fase. También es posible que la falta de respuesta homogénea por encima de $f = 4\text{kHz}$, mostrada en las figuras anteriores, influya negativamente. Pero esa no es una banda que perjudique en exceso a la señal vocal y por tanto ese defecto tendrá poca importancia práctica.

8.3 ELEMENTOS SOFTWARE DEL SISTEMA BASADO EN DSP PCI/C6600 + PMC/16I02

Una vez finalizadas con éxito las pruebas preliminares desarrolladas en el capítulo 7, en las que se demuestra la eficacia del procesador en array para atenuar el ruido y la reverberación en la señal de habla, el paso siguiente consiste en adaptar los algoritmos probados anteriormente para operar en tiempo real en el DSP. Se considera el sistema basado en DSP descrito en el punto 8.1, que incluye la tarjeta de adquisición BWS-PMC/16I02 y la tarjeta de procesado BWS-PCI/C6600. Por otra parte es necesario diseñar un *software* de comunicaciones entre el *Host PC* y el DSP para que la persona que cumpla la función de operador del array, pueda manejar e introducir los parámetros que le sean necesarios al programa residente en el DSP, o para gobernar la grabación de archivos de habla, si se quieren almacenar las señales de audio producidas por el sistema. Esta funcionalidad se representa esquemáticamente en la Figura 124.

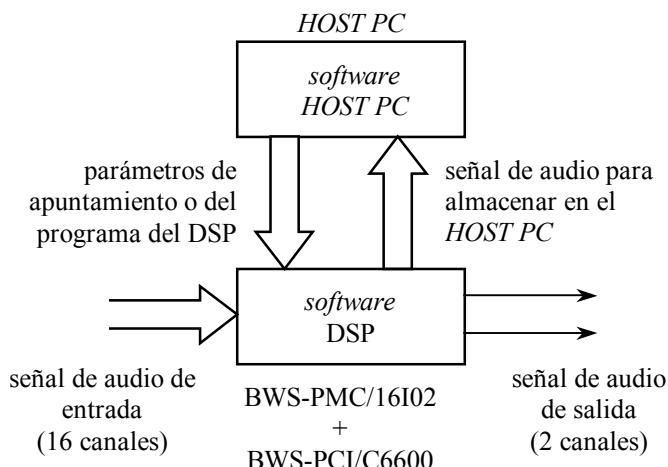


Figura 124. Diagrama de bloques con los componentes *software* utilizados en el procesador prototipo en tiempo real.

El problema del diseño del *hardware* no carece de importancia. Los sistemas simulados propuestos en el capítulo 7 han sido probados sin importar el tiempo de procesado. Será necesario hacer alguna adaptación de los mismos para que puedan ser incorporados al procesador en tiempo real. Por otra parte, el envío de las órdenes que el operador del array tiene que proporcionar al procesador se tiene que hacer sin interrumpir el normal funcionamiento del programa residente en el DSP. Todos estos detalles de implementación

han consumido una gran cantidad de esfuerzos, dirigidos al objetivo de un procesador en array en tiempo real totalmente operativo. A continuación se describen los detalles de este proceso que tengan suficiente interés científico a la hora de analizar e interpretar los resultados finales conseguidos con el prototipo implementado.

8.3.1 Programa residente en DSP

La programación del DSP se ha realizado utilizando el entorno *software* Code Composer Studio™ [TI 00] y las librerías específicas de procesado de señal de Texas Instruments [TI 02-a] [TI 02-b], además de las correspondientes a la tarjeta PCI/C6600 [BWS 00-a]. En la Figura 125 se representa un diagrama de bloques operativo que muestra el funcionamiento del programa residente en el DSP.

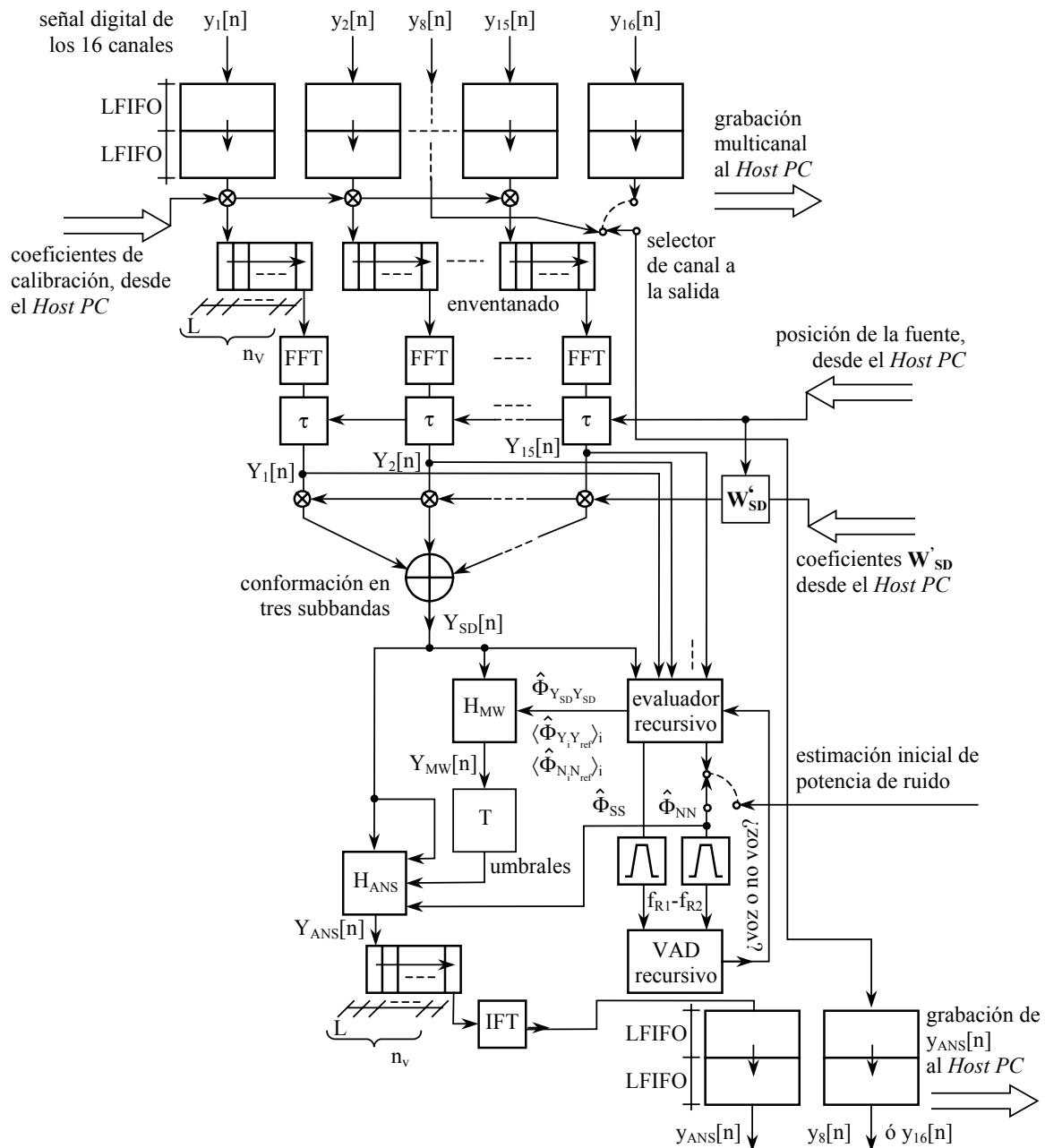


Figura 125. Diagrama de bloques operativo del software residente en el DSP.

El esquema representado en la Figura 125 básicamente consiste en el último procesador auditivo propuesto en las pruebas preliminares (procesador ANS-MW del punto 7.3). Sin embargo existen algunas particularidades y simplificaciones que ha sido necesario considerar, para adaptar este procesador al trabajo en tiempo real. A continuación se explican las peculiaridades de este *software* y su implicación con el procesado en tiempo real.

Procesado en doble *buffer*

La señal multicanal, digitalizada por el conversor ADC se adquiere en bloques de LFIFO puntos. Estos bloques deben contener un número entero de fragmentos S_1 (véase el punto 7.1.3 sobre enventanado temporal), parámetro que representa el desplazamiento en muestras entre dos ventanas consecutivas. Al número de tramas temporales (o ventanas) de tamaño L que caben en un bloque LFIFO se le llama aquí n_v . Así, si se decide considerar $L = 512\text{pt}$ y $S_1 = 171\text{pt}$ ($S = 67\%$), el tamaño del *buffer* de entrada recomendado será de $\text{LFIFO} = 855\text{pt}$, que representa $n_v = 5$ veces el tamaño de S_1 . Téngase en cuenta que LFIFO no puede superar el tamaño máximo de la memoria FIFO asignada a cada canal del ADC que es de 1024pt . Si se elige la combinación $L = 512\text{pt}$ y $S_1 = 256\text{pt}$ ($S = 50\%$) se recomendaría $\text{LFIFO} = 768\text{pt}$, que equivale a $n_v = 3$ veces el tamaño de S_1 . Otras combinaciones de mayor solapamiento exigirían demasiada capacidad de procesado al DSP, ya que el número de FFT's a realizar por segundo crece linealmente con la cantidad de solapamiento. Por otra parte, la segunda de las combinaciones elegidas no es óptima, ya que al ser el *buffer* LFIFO de menor tamaño, se producen más interrupciones por unidad de tiempo en el procesador y esto reduce su eficacia de cálculo. Por eso, parece que la combinación óptima en cuanto a velocidad de procesado es la primera, con un solapamiento de $S = 66\%$, aunque como ya se expuso en el punto 7.1.3, la elección de este valor puede producir una ligera distorsión a la hora de reconstruir temporalmente la señal, si la alineación en tiempo de canales se hace en el dominio de la frecuencia.

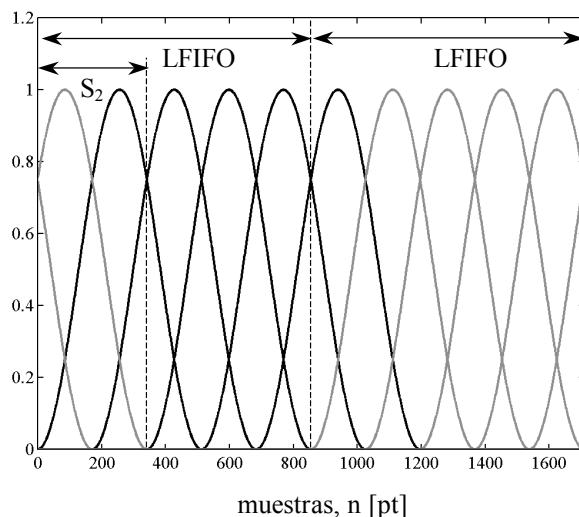


Figura 126. Distribución de $2 \times n_v$ ventanas en el doble *buffer* de tamaño $2 \times \text{LFIFO}$ para el caso $S_1=171\text{pt}$. Se necesitan S_2 puntos del bloque LFIFO anterior para conseguir que los bloques consecutivos reconstruyan la señal temporal sin pérdidas.

La señal así adquirida pasa a un *buffer* doble de longitud LFIFO, también de características “FIFO” (*First Input First Output*), de tal manera que el primer bloque de LFIFO puntos que entra es el primero en ser procesado. Esta disposición de doble *buffer* es necesaria tanto en la entrada como en la salida (véase la Figura 125). La adquisición en doble *buffer* va a influir en el retardo global de la señal procesada a la salida del conversor analógico, que será de 2 LFIFO’s, es decir de 107ms si se elige $S_1 = 171\text{pt}$ (LFIFO = 855pt) o de 96ms para $S_1 = 256\text{pt}$ (LFIFO = 768pt).

La necesidad de la configuración de adquisición anterior se explica porque, al finalizar el procesado de un bloque LFIFO, éste no está listo para ser enviado a la salida. Como se muestra en la Figura 126, cuando se acaba de procesar uno de estos bloques LFIFO se necesitan las S_2 primeras muestras temporales del siguiente bloque para poder reconstruir adecuadamente la señal temporal. Por lo tanto hay que esperar a que llegue el siguiente bloque. Si no fuese así, a la salida se reproduciría la caída final del enventanado correspondiente a las n_v ventanas por bloque.

Calibración

Como se muestra en la Figura 125, una vez realizada la adquisición digital por bloques, se realiza un ajuste de nivel en todos los canales del array, mediante unos coeficientes de calibración, que han sido previamente calculados y son enviados desde el *Host PC*. La misión de la calibración es corregir las diferencias en amplitud que puedan existir entre los canales del array, debido principalmente a las diferencias de sensibilidad de los micrófonos (según se explicó en el punto 8.2) y a los desajustes de amplificación del preamplificador MLA7.

Los coeficientes de calibración se calculan previamente, mediante un subprograma del DSP, obteniendo el nivel promedio captado por el procesador en cada uno de los canales. Dicho nivel se mide en una banda particular de frecuencias, que puede elegir según las circunstancias el operador de la calibración, y se determina por el ancho de banda (número de puntos de la FFT correspondiente) y la frecuencia central. Se recomienda hacer el cálculo de estos coeficientes en una cámara anecoica y con una fuente sonora patrón relativamente cerca del array (a un metro o menos), en posición *broadside*. Si no se hace así, la reverberación del recinto puede influir negativamente en la calibración al proporcionar niveles diferentes en cada uno de los micrófonos del array y falsear el cálculo de coeficientes. Las diferencias de nivel por distancia fuente-micrófono están corregidas en el programa de cálculo de coeficientes de calibración, por lo que dicha calibración se puede hacer, en teoría, con la fuente a cualquier distancia del array, incluso en posiciones muy cercanas a éste. Los coeficientes de calibración se almacenan en el *Host PC* y son recuperados durante la operación normal del array.

Bloque FFT

El bloque FFT es la parte más exigente en cuanto a recursos de procesado del sistema. Téngase en cuenta que para procesar $I_T = 15$ canales en el dominio de la frecuencia, es necesario realizar FFT’s en un número igual a I_T , más una FFT adicional para la reconstrucción temporal. Todo esto debe realizarse en el tiempo correspondiente al desplazamiento interventana S_1 . Es decir, la tasa t_F de FFT’s por segundo viene dado por la siguiente expresión:

$$t_F = \frac{(I+1)f_s}{S_1} [\text{FFT/s}] \quad (291)$$

Así para $S_1 = 171\text{pt}$ se tiene $t_F = 1497 \text{ FFT/s}$, mientras que para $S_1 = 256\text{pt}$ se tiene $t_F = 1000 \text{ FFT/s}$. Esto hace que se haya dedicado un gran esfuerzo en la optimización del bloque FFT.

En primer lugar se plantea la disyuntiva de si conviene realizar las FFT's en punto fijo o en punto flotante. A primera vista parece que, puesto que se utiliza un procesador en punto flotante que dispone de una ALU dedicada para operar en este formato, será más eficiente operar en punto flotante. Nada más lejos de la realidad ya que es manifiestamente más rápido operar en punto fijo. El DSP TMS320C6701 [TI 99] [BWS 99-a] dispone de 4 ALU's en punto flotante pero también de dos ALU's en punto fijo donde se realizan este tipo de operaciones y es mucho más rápido utilizar estas últimas para realizar las FFT's.

Una vez decidido el cálculo de las FFT's en punto fijo, cobra importancia el problema de la saturación y la precisión numérica, cuestiones que operando en punto flotante son casi irrelevantes. Si se opera en punto fijo la operación FFT es muy sensible a la saturación, sobre todo si la señal a analizar tiene pocas componentes espectrales. El caso peor corresponde al análisis de un tono puro que contenga un número entero de periodos en el tamaño L de la ventana, ya que es entonces cuando mayor concentración de energía existe en el espectro resultante. Si ese tono ha sido capturado con el nivel máximo permitido por la tarjeta de adquisición, le corresponden 20bits significativos, según la hoja de características de la Figura 106. Entonces, la FFT resultante será una barrapectral que necesitará para su digitalización 20bits x el número de puntos L. Puesto que la longitud de la trama en el procesador implementado es $L = 512\text{pt}$, valor que puede ser representado con 9bits, serían necesarios $20\text{bits} + 9\text{bits} = 29\text{bits}$ significativos en el peor caso, para operar sin saturación con la mayor precisión disponible por el conversor ADC. Eso supone en la práctica la capacidad de realizar FFT's de 32bits en punto fijo. Aunque esta posibilidad ha sido experimentada por el autor, siguiendo [Ahnoff 00], no se consigue la eficiencia computacional necesaria para la tasa de FFT's exigida. Por eso ha habido que implementar FFT's en punto fijo de 16bits. Eso ofrece una precisión numérica equivalente de 16bits - 9bits = 7bits. Como mucho, se podría subir un par de bits más para llegar a 9 bits, pensando que el caso del tono puro no va a darse en la realidad (además el enventanado reduce la concentración espectral). Sin embargo 9 bits es una precisión numérica muy pobre, ya que ofrece una SNR de unos 48dB en el mejor de los casos. Por todo ello ha sido necesario utilizar técnicas de autoescalado en la realización de las FFT's [Cheng 00] [TI 02-a] para conseguir una precisión numérica máxima de 13bits, lo que equivale a una SNR de cuantificación de 13 bits, que es totalmente admisible para el análisis y procesado de voz en condiciones de alto ruido y reverberación, que son las que se tratan en esta Tesis.

Coefficientes W'_{SD}

Como se aprecia en la Figura 125, después del bloque FFT y previo a la conformación de haz, es necesario aplicar unos coeficientes almacenados en el *Host PC*. Éstos son los coeficientes de ponderación superdirectivos –véase (112) en el apartado 2.2.3 sobre conformación superdirectiva–. El valor introducido en el DSP depende del ángulo θ_0 de posición de la fuente y su utilidad se detalla en el capítulo 9, sobre la propuesta de un conformador superdirectivo en la banda B_1 .

Operaciones matemáticas no lineales en punto flotante

Otro de los escollos importantes encontrados, desde el punto de vista de la eficacia computacional a la hora de la implementación en tiempo real, ha sido la lentitud de las operaciones no lineales (trigonometrías, potencia y raíz cuadrada) necesarias para el funcionamiento del procesador propuesto. Se han tenido que incorporar librerías matemáticas rápidas [TI 02-b] especialmente diseñadas para trabajar con la máxima rapidez en el procesador específico TMS320C6701.

Evaluador recursivo de potencias, VAD recursivo y estimación del filtro H_{MW}

Una gran parte de los elementos del procesador dependen de estimaciones de potencia de señal y de ruido. Por ejemplo, el filtro de Wiener modificado por coherencia H_{MW} propuesto en las ecuaciones (271) y (272) del punto 7.1, necesita de la estimación de los espectros cruzados intercanal de la señal y del ruido.

Las estimaciones para el cálculo de H_{MW} se hacen por el método recursivo de polo simple –(192) y (275) a (280)–, explicado en el punto 4.1.4 de la Parte 1 de esta Tesis. Además, aquí se aplican las consideraciones prácticas para operación en tiempo real expresadas en (269) y (270), que establecen un canal de referencia “ref” flotante, de tal manera que en un intervalo de tiempo suficientemente corto se obtengan los espectros cruzados de todas las parejas de micrófonos. Por tanto, el filtro H_{MW} necesita conocer la estimación de la potencia de señal conformada $\hat{\Phi}_{Y_{SD}Y_{SD}}(\omega)$ (el subíndice SD atiende a que, como se explica más adelante, se implementa una conformación superdirectiva), el promedio de espectros cruzados intercanal de señal cuando ésta se considera voz $\langle \hat{\Phi}_{Y_i Y_{ref}}(\omega) \rangle_i$ y el promedio de espectros cruzados cuando se considera ruido $\langle \hat{\Phi}_{N_i N_{ref}}(\omega) \rangle_i$. Es necesario, por tanto, distinguir qué tramas de señal son ruido y cuáles son señal, es decir se necesita complementar el estimador recursivo con un detector de actividad de voz (VAD) eficiente y sencillo, para ser incorporado al procesador.

El algoritmo de VAD, seleccionado por su sencillez, corresponde al estimador recursivo de polo simple y dos lados de (199), que decide que una trama de señal es de voz si su nivel se incrementa sobre la anterior y viceversa. Para disminuir la posibilidad de error del VAD, se ha considerado la estimación recursiva sólo en una banda de frecuencias, características de la señal de voz, para establecer el cálculo de potencia que precede a la decisión. Las frecuencias consideradas que limitan esta banda han sido $f_{R1} = 500\text{Hz}$ y $f_{R2} = 3\text{kHz}$. Es decir, para que el VAD proporcione la decisión final, sólo se realiza un cómputo energético entre estas dos frecuencias. Por eso $\hat{\Phi}_{SS}(\omega)$ y $\hat{\Phi}_{NN}(\omega)$ en la Figura 125, que representan las estimaciones de potencia de señal o de ruido, son filtradas en esa banda de análisis. La salida del VAD expresa la decisión de si la trama analizada es de voz o de ausencia de voz (véase la Figura 125) y a su vez alimenta al evaluador recursivo de señal y de ruido. Con esta información el evaluador recursivo distingue todas las estimaciones necesarias, es decir: $\hat{\Phi}_{Y_{SD}Y_{SD}}(\omega)$, $\langle \hat{\Phi}_{Y_i Y_{ref}}(\omega) \rangle_i$, $\langle \hat{\Phi}_{N_i N_{ref}}(\omega) \rangle_i$, $\hat{\Phi}_{SS}(\omega)$ y $\hat{\Phi}_{NN}(\omega)$. La potencia de ruido $\hat{\Phi}_{NN}(\omega)$, salida del evaluador recursivo, también es utilizada para el cálculo del filtro auditivo H_{ANS} .

Nótese también, que el VAD necesita para su arranque, cuando comienza a funcionar el procesador, una estimación inicial de la potencia de ruido $\hat{\Phi}_{NN}(\omega)$. Existe la posibilidad de dejar esa estimación inicial de ruido a cero, con lo que el arranque del VAD es más lento, hasta que detecta correctamente los espectros de voz y de ruido. También se puede dejar que las tramas iniciales captadas por el procesador sean exclusivamente de ruido (con un tiempo

de estimación de 1s es suficiente), con lo que el “enganche” del VAD en la estimación correcta del espectro de ruido es más rápido. Si se elige la segunda opción, debe cuidarse especialmente que en ese tiempo inicial de estimación de ruido no se capte nada de voz, que sería interpretada como ruido y cuyas componentes espectrales se cancelarían en instantes posteriores, con el aumento consecuente de la distorsión.

Estimación del filtro H_{ANS}

El filtro H_{ANS} , implementado en el procesador en tiempo real, cumple básicamente las mismas pautas que fueron consideradas en los experimentos preliminares del punto 7.3.1 de esta Tesis para el procesador allí propuesto –ecuaciones (286) y (287)–. Se consideran $B = 22$ bandas críticas auditivas. Los umbrales de enmascaramiento $T(b)$ o $T(\omega)$ se calculan por (186) a partir de la salida Y_{MW} del filtro Wiener con la función de ensanchamiento SF de la Tabla 2. Se utiliza el factor de renormalización $D(b)$ de la Tabla 3 y el factor de *offset* $O(b)$ de (187). La única diferencia con la propuesta preliminar anterior es que para economizar cálculos se considera fija la naturaleza tonal de la señal incidente, estableciéndose $\text{ton} = 0.8$ (véase la Figura 37) con lo que se evita el cálculo de la planitud espectral SMF de (189) en cada una de las tramas de voz incidentes. Se ha comprobado que esta simplificación no modifica sensiblemente el cálculo de los umbrales auditivos y sí reduce notablemente el número de operaciones realizadas en tiempo real.

8.3.2 Software de comunicación y pruebas

El programa residente en el DSP debe ser controlado desde el *Host PC* para comunicarle diversos parámetros de operación que necesita conocer el procesador en array. Esta comunicación se debe hacer a veces en tiempo real, sin interrumpir el correcto funcionamiento del procesador. Por ejemplo, las coordenadas de apuntamiento (r_0, θ_0) se han de poder introducir desde el *Host PC* cuando se quiera cambiar instantáneamente de directividad en el array. Otras veces no será necesaria la comunicación en tiempo real, como por ejemplo la calibración automática de los 15 canales del array o la grabación de los resultados en una base de datos.

Este conjunto de programas, diseñado mediante las herramientas adecuadas al *hardware* disponible [BWS 00-b], es lo que aquí se ha llamado *software* de comunicación y pruebas, y su descripción detallada puede encontrarse en [Alonso 03].

La misión del *software* de comunicación y pruebas es, por tanto, permitir la interacción, en tiempo real o no, entre el DSP y el *Host PC*. A continuación se describen brevemente las posibilidades más relevantes que ofrece este *software* de control (véase la Figura 127).

Calibración

Consiste en calcular los coeficientes C de calibración del array de 15 micrófonos (Figura 127), que sirven para igualar la sensibilidad de la conversión acústica-eléctrica en todos los canales del procesador. La calibración se implementa mediante un subprograma y debe ejecutarse en condiciones geométricas y acústicas controladas.

En el proceso de calibración se deben especificar las coordenadas r_0 y θ_0 de la fuente utilizada para la calibración y la frecuencia central f_0 y ancho de banda ΔB (expresado por el

número de líneas correspondiente a una FFT de 512pt) considerados para dicha calibración. También debe introducirse el tiempo de promediado T que se va a usar para estimar los coeficientes de salida. La salida del subprograma es un archivo de calibración que contiene los coeficientes de los 15 canales, necesarios para la igualación de niveles de los mismos. Se proporciona también un archivo multicanal con la señal multimicrófono a la que se ha aplicado la calibración, para que el operador pueda verificar que se consigue una igualación adecuada de niveles. Se recomienda hacer la calibración en una cámara anechoica, con una fuente de señal senoidal de baja frecuencia situada a una distancia cercana al array y en posición *broadside*, $\theta_0 = 90^\circ$.

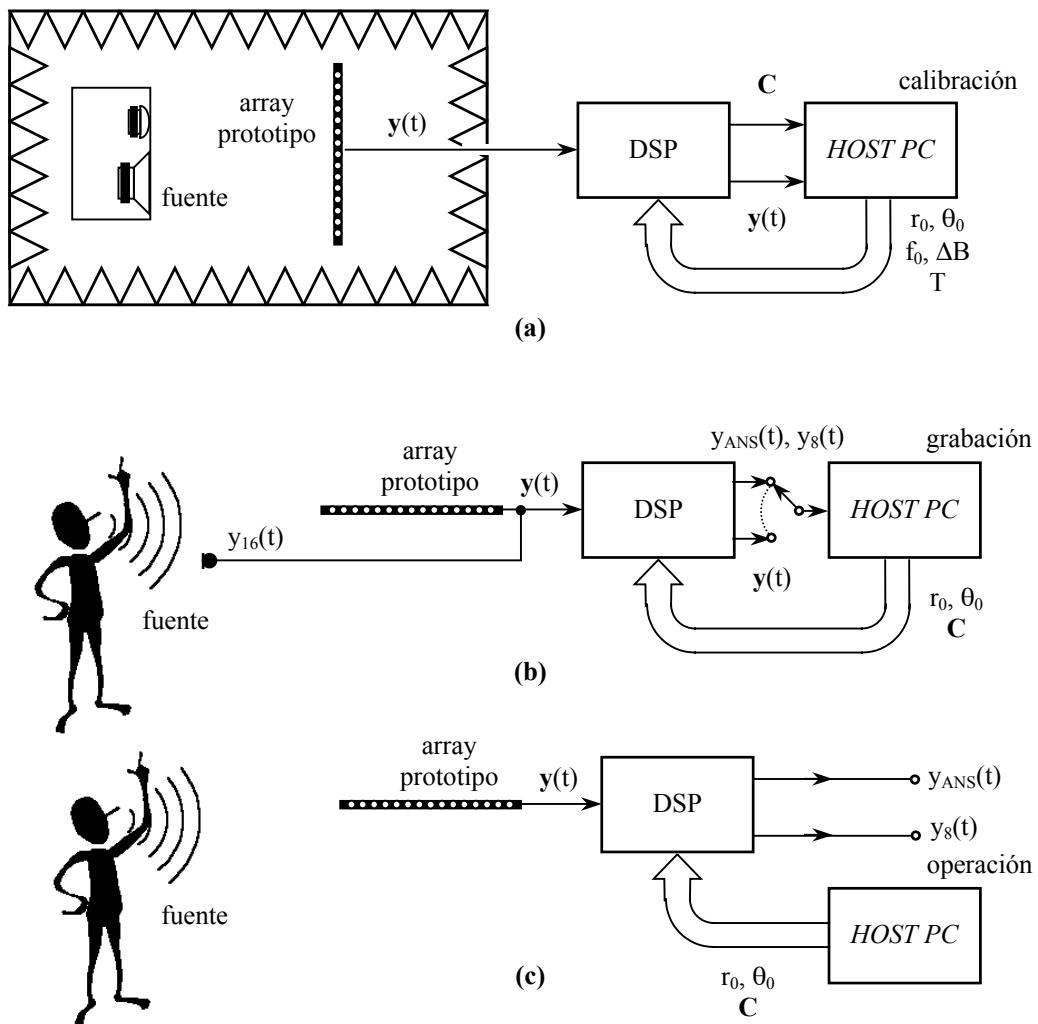


Figura 127. Modos de funcionamiento del *software* de control y comunicaciones.
(a) Calibración. **(b)** Grabación. **(c)** Operación.

Grabación

El subprograma de grabación permite almacenar en el disco duro del *Host PC* las señales más interesantes que puedan servir para analizar los resultados o para generar bases de datos multicanal. Se han implementado dos modos de funcionamiento. En “modo base de datos” se almacena el vector $y(t)$ que corresponde a los 15 canales del array sin procesar, más el canal de referencia $y_{16}(t)$. En “modo filtrado” se graba la salida del procesador –por

ejemplo $y_{ANS}(t)$ para el postfiltro auditivo— junto con cualquiera de los canales de entrada al array, que puede ser elegido por el operador —normalmente se elige $y_8(t)$, el canal central del array o el de referencia $y_{16}(t)$ —. El sistema no tiene recursos suficientes para almacenar simultáneamente los 16 canales del array y a la vez procesarlos.

Con el programa de grabación se puede seleccionar además qué programa se carga en el DSP, el archivo de coeficientes de calibración y si se utiliza sólo la conformación de haz en tres subbandas —salidas $y_{RS}(t)$ o $y_{SD}(t)$ — o se usa el procesador completo, con el postfiltro de mejora —salida $y_{ANS}(t)$ —. También si el apuntamiento se hace de modo manual (introduciendo la distancia r_0 y el ángulo θ_0 de la fuente) o con el apoyo de la *Web Cam*, en cuyo caso aparece en pantalla la imagen de la escena frente al array y se puede elegir con el ratón un punto determinado de la misma.

Operación

El subprograma de operación es similar al de grabación con la diferencia de que se obtiene la señal procesada por las dos salidas analógicas del procesador. Aunque la salida estéreo del procesador puede ser configurada (véase la Figura 125), normalmente por uno de esos canales se selecciona salida procesada —por ejemplo $y_{ANS}(t)$ — y por el otro la salida sin procesar, correspondiente a uno de los canales del array, normalmente el central $y_8(t)$ o, según se desee, la señal del micrófono de referencia $y_{16}(t)$.

9 PROTOTIPO DE ARRAY SUPERDIRECTIVO PERCEPTUAL (SD-ANS-MW) EN TIEMPO REAL

Las pruebas iniciales de conformación de haz con el prototipo final en tiempo real muestran una cierta falta de selectividad espacial en baja frecuencia, característica que ya se había previsto en el capítulo 6, donde se mostraban las curvas de directividad esperadas para el conformador convencional por el método de retardo y suma. En la Figura 60 se muestran los mapas de directividad para diferentes apuntamientos del array y se puede apreciar que por debajo de 500Hz el conformador se vuelve casi omnidireccional. Esto se destaca con más detalle en la Figura 61 (apuntamiento *broadside*), la Figura 62 (*endfire*) y la Figura 63 (apuntamiento lateral, $\theta_0 = 45^\circ$). El método más conocido y ampliamente utilizado para aumentar la directividad es el de conformación superdirectiva, descrito en el punto 2.2.3 de la Parte 1 de la Tesis. Puesto que no se plantea el objetivo de implementar un conformador adaptativo, la forma más eficiente de configurar un array superdirectivo para atenuar en baja frecuencia el mayor ruido y reverberación posibles, es hacerlo para la cancelación de ruido isotrópico ideal (o ruido difuso). Por todo ello se plantea la implementación de un conformador superdirectivo en la banda de baja frecuencia, que complemente al procesador ANS-MW descrito en el apartado 7.3. Ésta es la última propuesta realizada en el marco de esta Tesis y será aquí designada como procesador SD-ANS-MW. A continuación se describe dicho procesador que ha sido implementado finalmente en el sistema en tiempo real y con el que se han realizado los experimentos y se han obtenido los resultados finales de esta Tesis.

9.1 DESCRIPCIÓN DEL CONFORMADOR SUPERDIRECTIVO EN LA BANDA DE BAJA FRECUENCIA

En la Figura 128 se representa el diagrama de bloques correspondiente al conformador SD-ANS-MW implementado. Los elementos de procesador correspondientes a la supresión auditiva de ruido son los que se explicaron en el punto 7.2.2. También hay que considerar los detalles relativos a la implementación práctica en tiempo real, que han sido expuestos en el punto 8.3.1 de esta Tesis. En este sentido, los cambios más importantes introducidos han sido los siguientes. Para el filtro $H_{MW}(\omega)$ ha sido implementado un VAD recursivo según (192), (198) y (199), basado en un estimador de energía de polo único y dos lados (con frecuencias límite $f_{R1} = 500\text{Hz}$ y $f_{R2} = 3\text{kHz}$ –ver el punto 8.3.1–). Para el filtro $H_{ANS}(\omega)$ se ha considerado una tonalidad fija de la señal de voz a la entrada ($\text{ton} = 0.8$) de tal manera que se evita el cálculo de la planitud espectral SMF (189).

La mejora que ahora se propone es la incorporación de un conformador superdirectivo en la banda de baja frecuencia, diseñado para la cancelación de ruido difuso. La teoría sobre la conformación superdirectiva fue descrita en el punto 2.2.3 de la Parte 1 de esta Tesis. Un conformador superdirectivo de este tipo, ofrece una elevada directividad (y por tanto atenuación) ante direcciones de incidencia no coincidentes con la principal, y sólo lo hace a

aquellas frecuencias en las que dicho ruido difuso tiene un alto grado de coherencia espacial intercanal.

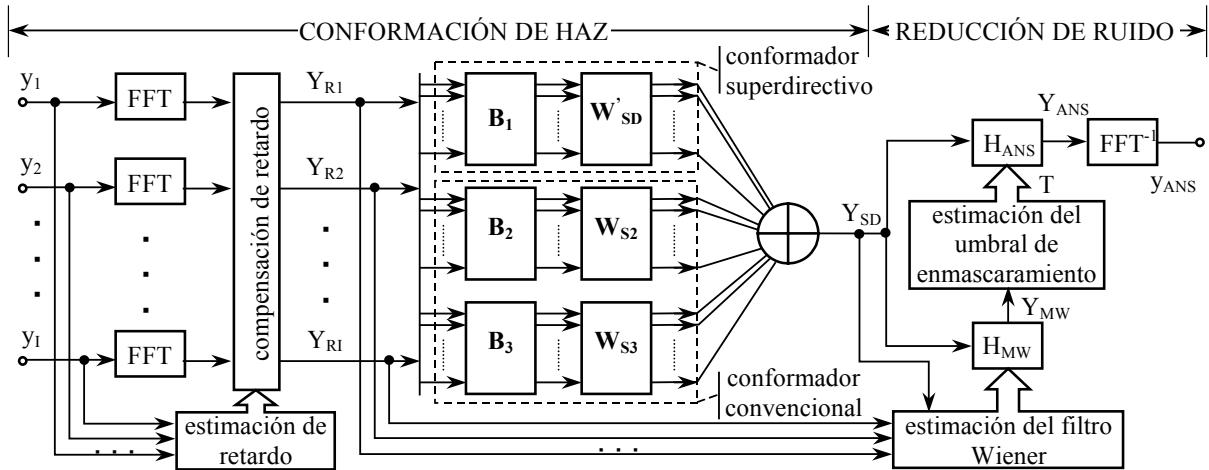


Figura 128. Esquema del procesador SD-ANS-MW.

El aumento de la directividad en baja frecuencia (la máxima alcanzable para ruido difuso) se consigue mediante combinaciones de sumas y restas de micrófonos, que en cierta medida se puede asimilar a combinaciones de pares bidireccionales de receptores. Estas combinaciones se expresan en un vector de ponderaciones $\mathbf{W}_{SD}(\omega)$ de (100) ó (103), ó $\mathbf{W}'_{SD}(\omega)$ de (112) si que se aplica a los espectros de los micrófonos previamente alineados en tiempo. Este vector, que aparece en la Figura 128, se aplica a los $I_{B1} = 7$ micrófonos correspondientes a la banda primera de baja frecuencia, una vez su espectro ha pasado por el filtro B_1 que selecciona la banda de baja frecuencia (Figura 129). Los $I_{B2} = 7$ e $I_{B3} = 7$ micrófonos correspondientes a las bandas de media y alta frecuencia, una vez han sido filtrados por B_2 y B_3 respectivamente, son ponderados y sumados (conformador convencional) mediante los coeficientes de suma (107):

$$\mathbf{W}_{S2} = \frac{1}{7} \underbrace{[1, 1, \dots, 1]}_7^T = \mathbf{W}_{S3} \quad (292)$$

Una vez recompuesta la banda total del espectro mediante un sumador, la salida que se obtiene es $Y_{SD}(\omega)$, correspondiente al conformador superdirectivo combinado.

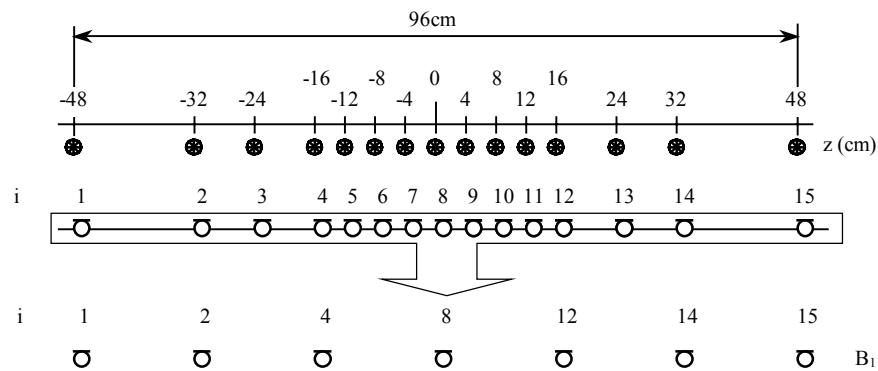


Figura 129. Índices i para el subarray superdirectivo de baja frecuencia.

Los coeficientes $\mathbf{W}'_{SD}(\omega)$ deben ser precalculados e introducidos en el DSP una vez conocida la dirección de apuntamiento (véase la Figura 125). Como el juego de coeficientes es diferente para cada dirección de apuntamiento θ_0 , se han precalculado 90 conjuntos diferentes para otros tantos ángulos de apuntamiento pertenecientes al intervalo $[0^\circ, 180^\circ]$, con una precisión de 2° . En total $256 \times 90 = 23040$ valores complejos, correspondientes a la banda lateral inferior de una FFT de 512pt. De ese modo, cuando se apunta el array, se eligen los coeficientes más apropiados para la posición de la fuente.

En la Figura 130 se representa el vector de coeficientes $\mathbf{W}'_{SD}(\omega)$ (en módulo y fase) para la subbanda de baja frecuencia B_1 del array superdirective apuntado a $\theta_0 = 0^\circ$ (*endfire*), con un parámetro de restricción –véase (103)– de $\mu = -10\text{dB}$ y $\mu = -15\text{dB}$. Téngase en cuenta que según (292), para las bandas B_2 y B_3 , los coeficientes del conformador convencional en dB son $W_{S2i} [\text{dB}] = W_{S3i} [\text{dB}] = -16.9\text{dB}$. Puede observarse cómo el vector módulo $|\mathbf{W}'_{SD}(\omega)|$ en todos los casos tiende hacia este valor de -16.9dB al crecer la frecuencia y la fase $\arg[\mathbf{W}'_{SD}(\omega)]$ tiende a cero. Es decir, la conformación superdirective optimizada para ruido difuso tiende a un conformador convencional en alta frecuencia y por eso es interesante la implementación superdirective sólo en la banda B_1 . Por otra parte, en los dos casos de la Figura 130, el módulo $|\mathbf{W}'_{SD}(\omega)|$ tiende a subir a medida que la frecuencia decrece. Esta subida es menos acusada si μ tiene un valor grande.

Como se sabe y se manifestó en el punto 2.2.3, el parámetro de restricción μ evita que los coeficientes superdirectivos crezcan mucho en baja frecuencia para conseguir una respuesta en frecuencia plana del array en θ_0 . Cuanto mayor sea μ , menor amplificación será necesaria para obtener esta respuesta en frecuencia constante, pero más alejado de la solución óptima estará el conformador así configurado. El valor del parámetro $\mu = -15\text{dB}$ no es peligroso en cuanto a excesiva amplificación de la baja frecuencia ya que sólo produce un aumento de 13dB en 30Hz respecto al conformador convencional –véase $|\mathbf{W}'_{SD}(\omega)|$ en la Figura 130(b)–. En la Figura 131 se representa $|\mathbf{W}'_{SD}(\omega)|$ y $\arg[\mathbf{W}'_{SD}(\omega)]$ para un apuntamiento $\theta_0 = 90^\circ$ (*broadside*). En este caso –Figura 131(b)–, la amplificación máxima proporcionada por los coeficientes con respecto al conformador convencional, es de unos 18dB , a una frecuencia de 90Hz aproximadamente. Este valor no es excesivamente alto y no se proporciona a todos los micrófonos simultáneamente, con lo que se asegura que el ruido no coherente de baja frecuencia, representado sobre todo por el ruido eléctrico interno de los micrófonos, no se amplificará excesivamente.

Hay que destacar que los valores de $\mathbf{W}'_{SD}(\omega)$ mostrados en las Figuras 130 y 131, y también los que se utilizan en el prototipo implementado, sirven para la aproximación de campo lejano, en la práctica para $r_0 > 5\text{m}$. Quiere decir que quizás, cuando la fuente se sitúe a distancias más próximas, la respuesta polar del array de desviará ligeramente de la esperada, en el sentido de que el máximo de captación no estará exactamente la posición de la fuente (r_0, θ_0) y por tanto la cancelación de ruido y reverberación laterales no será óptima.

Siguiendo con las Figuras 130 y 131, puede apreciarse cómo el apuntamiento *endfire* proporciona el mismo peso a las parejas de micrófonos $i = 1, 15$; $i = 2, 14$ e $i = 4, 12$ (véase la Figura 129), pero con fases conjugadas. Para el apuntamiento *broadside* sin embargo las parejas anteriores de micrófonos son ponderadas con el mismo coeficiente superdirective, en módulo y fase, pero en casi toda la baja frecuencia existe una diferencia de fase de 180° entre dos parejas diferentes, de tal manera que la directividad total del array tenga un mínimo de captación alrededor $\theta_0 = 0^\circ$ y $\theta_0 = 180^\circ$.

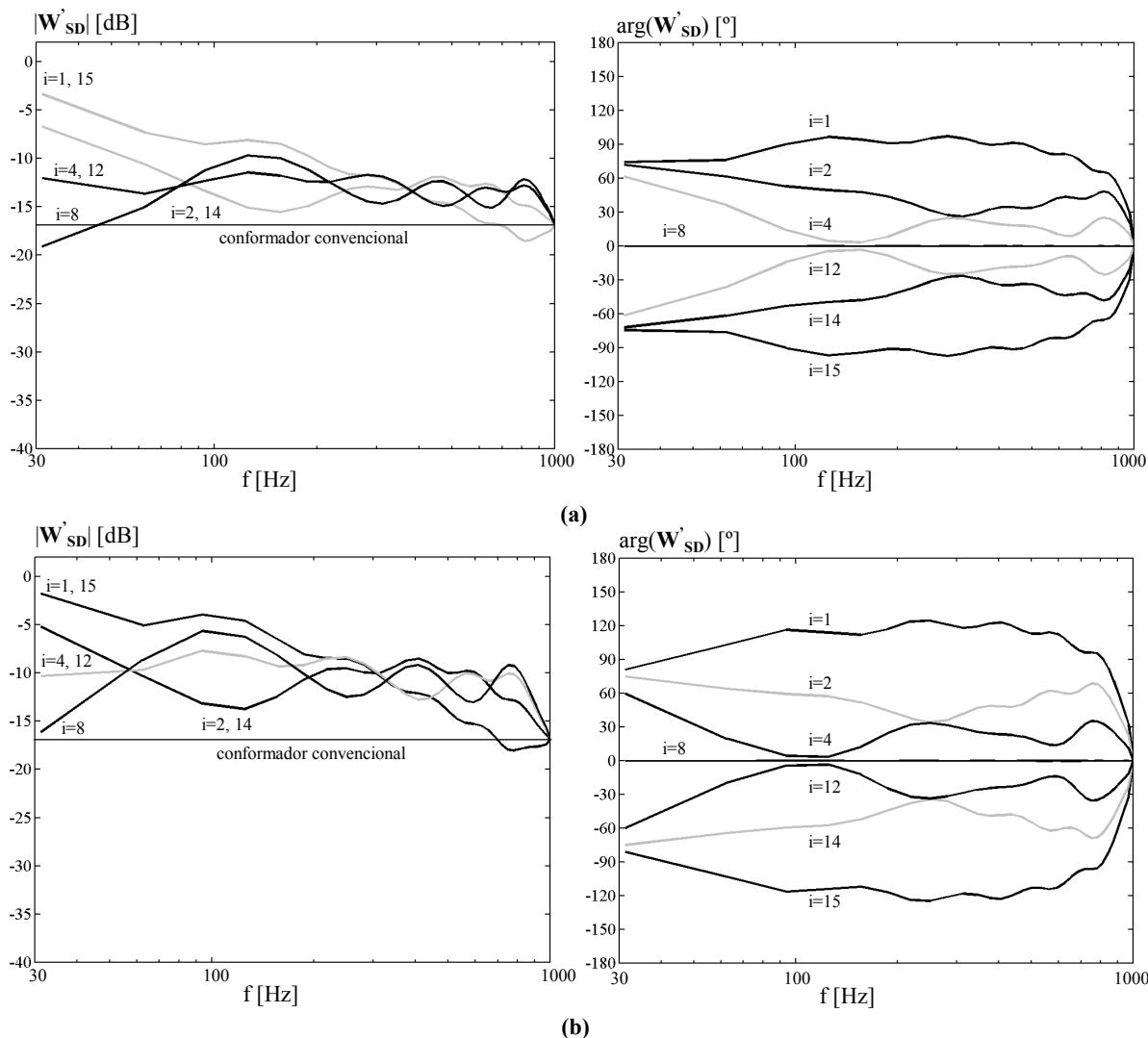


Figura 130. Vector de ponderaciones $\mathbf{W}'_{SD}(\omega)$ (módulo y fase) para la subbanda de baja frecuencia del array superdirektivo propuesto. El índice i de los micrófonos corresponde al subarray de baja frecuencia (véase la Figura 129). En esta figura se considera apuntamiento *endfire* ($\theta_0=0^\circ$). **(a)** $\mu[\text{dB}]=-10\text{dB}$. **(b)** $\mu[\text{dB}]=-15\text{dB}$.

A continuación se representan las curvas polares de directividad $D(\theta)$ de la banda superdirektiva B_1 en tercios de octava, respectivamente para apuntamientos *broadside*, *endfire* y lateral. Las curvas correspondientes a las bandas B_2 y B_3 no se representan, ya que corresponden al conformador convencional y ya han sido mostradas en las Figuras 61, 62 y 63 del punto 6.1 donde se describía al conformador convencional de 15 canales anidados.

En las Figuras 132, 133 y 134 se muestran las directividades del array con los dos valores considerados anteriormente para la constante de restricción μ , es decir $\mu = -10\text{dB}$ ó $\mu[\text{dB}] = -15\text{dB}$.

El menor valor de μ (curvas de la derecha en cada pareja representada) proporciona unos mínimos más profundos en el diagrama polar y por tanto una mayor directividad global y mayor cancelación de ruido y reverberación, al menos en teoría. Las diferencias entre las parejas de curvas van disminuyendo a medida que la frecuencia aumenta, debido a que el conformador se acerca cada vez más a la solución convencional de retardo y suma, en la que el parámetro μ no influye.

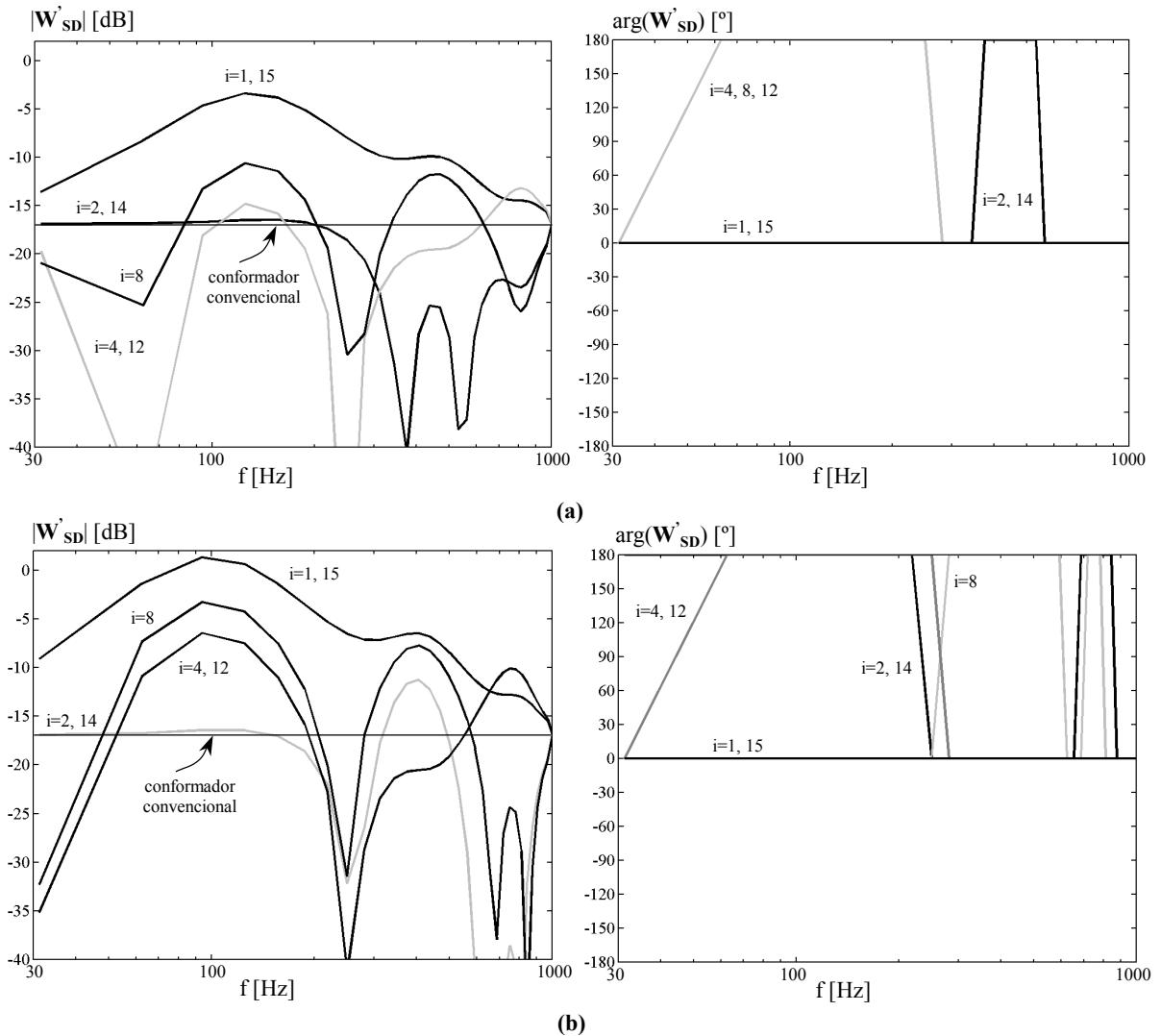


Figura 131. Vector de ponderaciones $\mathbf{W}'_{SD}(\omega)$ (módulo y fase) para la subbanda de baja frecuencia del array superdirectivo propuesto. El índice i de los micrófonos corresponde al subarray de baja frecuencia (véase la Figura 129). En esta figura se considera apuntamiento *broadside* ($\theta_0=90^\circ$). **(a)** $\mu[\text{dB}]=-10\text{dB}$. **(b)** $\mu[\text{dB}]=-15\text{dB}$.

Un fenómeno curioso, que puede ser observado en la Figura 134 para un apuntamiento lateral ($\theta_0 = 45^\circ$), es que para baja frecuencia, los máximos de directividad no corresponden exactamente a la dirección de apuntamiento, es decir a $\theta = 45^\circ$. Este fenómeno se manifiesta de forma más patente para el μ mayor (-10dB). A partir de 400Hz se consigue ya un apuntamiento correcto hacia la dirección de la fuente θ_0 . Este apuntamiento incorrecto se origina porque un parámetro de restricción μ grande en (103) modifica ligeramente la condición de máximo para la posición de apuntamiento en el numerador y desvía el eje de máxima captación del array. Curiosamente éste efecto no aparece para las configuraciones endfire y broadside por la simetría en los coeficientes de ponderación $\mathbf{W}'_{SD}(\omega)$, según se visualiza en las Figuras 130 y 131.

Comparando las curvas de directividad del array superdirectivo con las del conformador convencional, queda patente la ganancia de selectividad espacial en la banda de baja frecuencia, que es, por cierto, donde se suelen situar los mayores problemas de excesivo ruido y reverberación cuando se capta la voz de un locutor en una situación real.

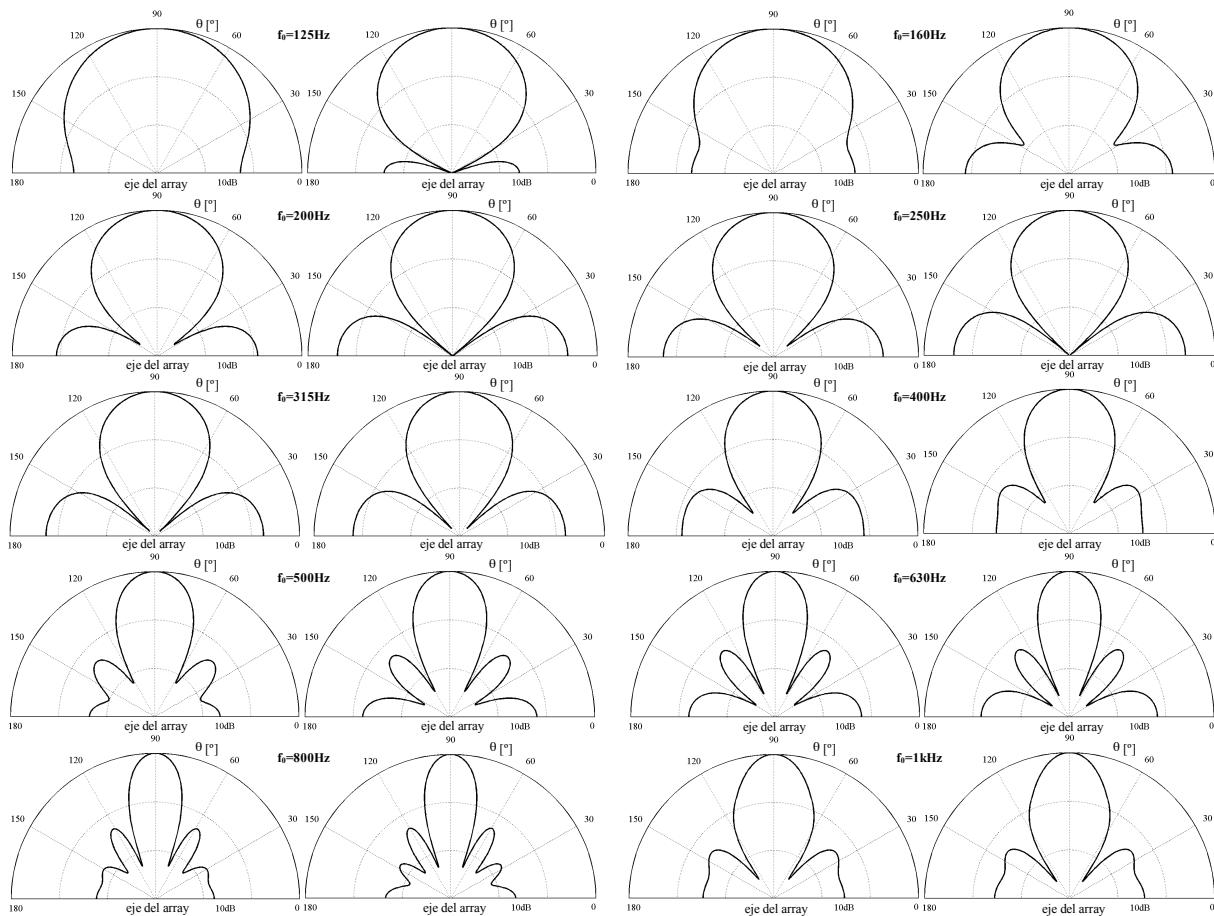


Figura 132. Curvas polares de directividad $D(\theta)$ [dB] para el subarray superdirectivo B_1 en bandas de 1/3 de octava. Apuntamiento *broadside* ($\theta_0=90^\circ$). Se ha usado la aproximación de campo lejano. En cada pareja de curvas, la de la izquierda corresponde a $\mu[\text{dB}]=-10\text{dB}$ y la de la derecha a $\mu[\text{dB}]=-15\text{dB}$.

En la Figura 135 se representan las curvas de directividad de valor promedio en la banda superdirectiva, entre 125Hz y 1kHz. Se manifiesta que, globalmente consideradas estas frecuencias, la selectividad espacial en la banda conflictiva B_1 es similar a las bandas de alta frecuencia del conformador convencional. Para ello compárese la Figura 135 con la Figura 64 correspondiente a las bandas consideradas como de directividad constante para el conformador anidado.

En la Figura 136 se representa el comportamiento global promedio del array anidado propuesto, combinando la banda superdirectiva con las dos bandas anidadas de alta frecuencia. Se ha evitado incluir en el promedio frecuencias superiores a 4kHz, ya que a partir de ahí, el array pierde sus propiedades de directividad constante, el lóbulo principal se estrecha y aparece *aliasing* espacial para ciertas direcciones de apuntamiento. El haz principal puede considerarse suficientemente estrecho en la banda considerada, sobre todo si se considera que están incluidas las bajas frecuencias. Además, de forma ventajosa, la directividad específica de la banda B_1 (Figura 135) que es la más difícil de estrechar, no difiere excesivamente de la de todo el array (Figura 136).

No debe perderse de vista que, debido a la simetría de revolución alrededor del eje del array, para apuntamientos diferentes de *endfire*, el lóbulo principal se duplica, aumentando peligrosamente la captación posterior. De esa manera la configuración *endfire* siempre será la más recomendable si se usan micrófonos omnidireccionales (como es el caso del prototipo implementado), que no atenúen convenientemente la captación trasera.

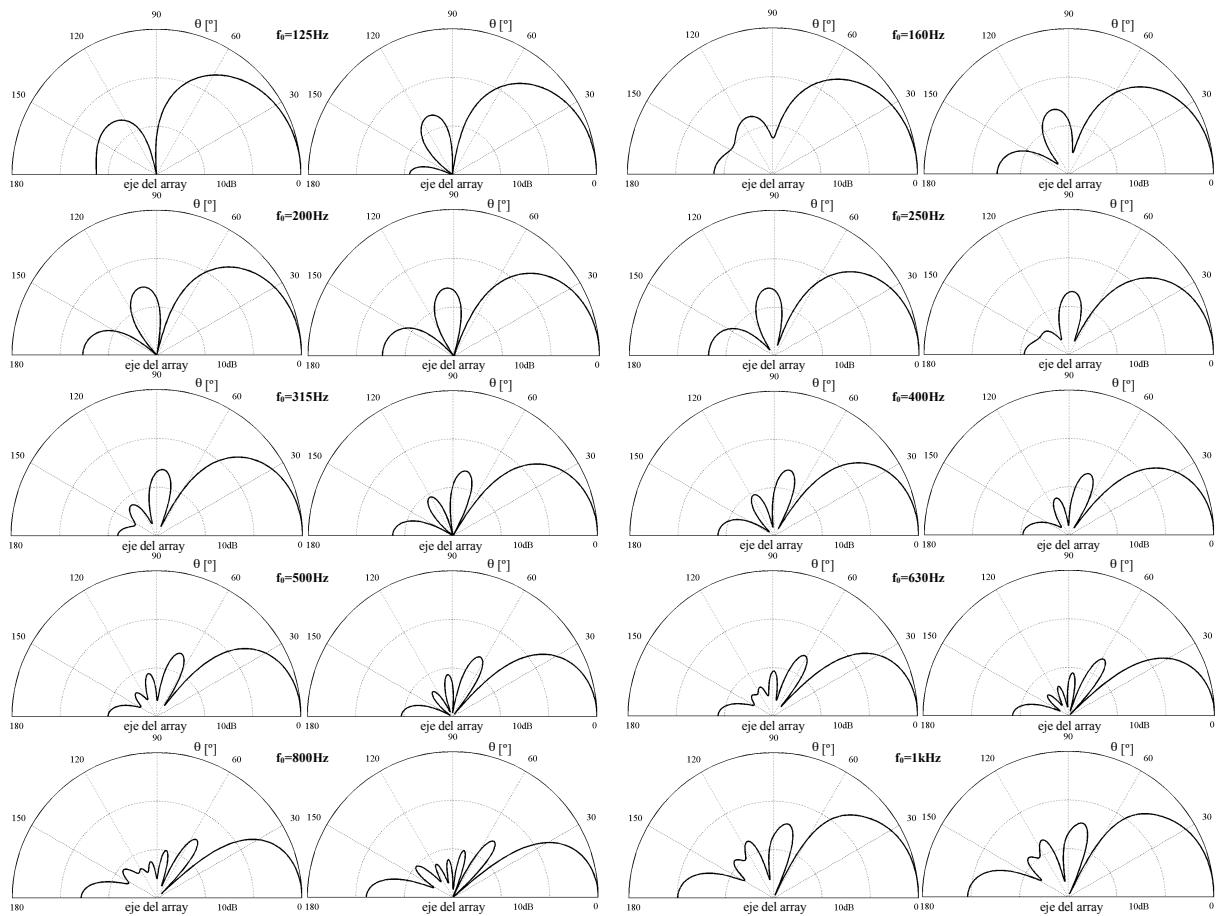


Figura 133. Curvas polares de directividad $D(\theta)$ [dB] para el subarray superdirectivo B_1 en bandas de 1/3 de octava. Apuntamiento *endfire* ($\theta_0=0^\circ$). Se ha usado la aproximación de campo lejano. En cada pareja de curvas, la de la izquierda corresponde a μ [dB]=-10dB y la de la derecha a μ [dB]=-15dB.

En la Figura 137 se representan los mapas de directividad $D(\theta, f)$ de todo el array para los tres apuntamientos y los dos valores de μ considerados anteriormente. Estos mapas deben compararse con los de la Figura 60 para el array convencional. Como se ve las dos subbandas superiores no difieren del array convencional por retardo y suma, aunque sí lo hace la subbanda B_1 , que tiene un haz de captación principal manifestamente más estrecho. Por otra parte, los mapas para $\mu = 15$ dB muestran una directividad ligeramente mayor en B_1 , debido al efecto ya expresado del parámetro de restricción.

En la Figura 138 está representado, en función de la frecuencia, el índice de directividad $DI(\theta_0)$ en la dirección de apuntamiento –equivalente a DI_{MAX} de (55)–. Nótese que en este caso se ha evitado igualar DI_{MAX} a $DI(\theta_0)$ ya que, debido a los defectos de apuntamiento reseñados anteriormente para $\theta_0 = 45^\circ$ en baja frecuencia, la captación máxima puede diferir del ángulo θ_0 . Es decir el array no se apunta correctamente a la DOA prevista. La Figura 138 se debe comparar con la Figura 67 para el conformador convencional. Se manifiesta un aumento muy conveniente de $DI(\theta_0)$ en baja frecuencia para el apuntamiento *endfire*, que en 30Hz incluso supera al obtenido en las bandas B_2 y B_3 . También manifiestan esa mejora, aunque en menor medida, los otros dos apuntamientos considerados.

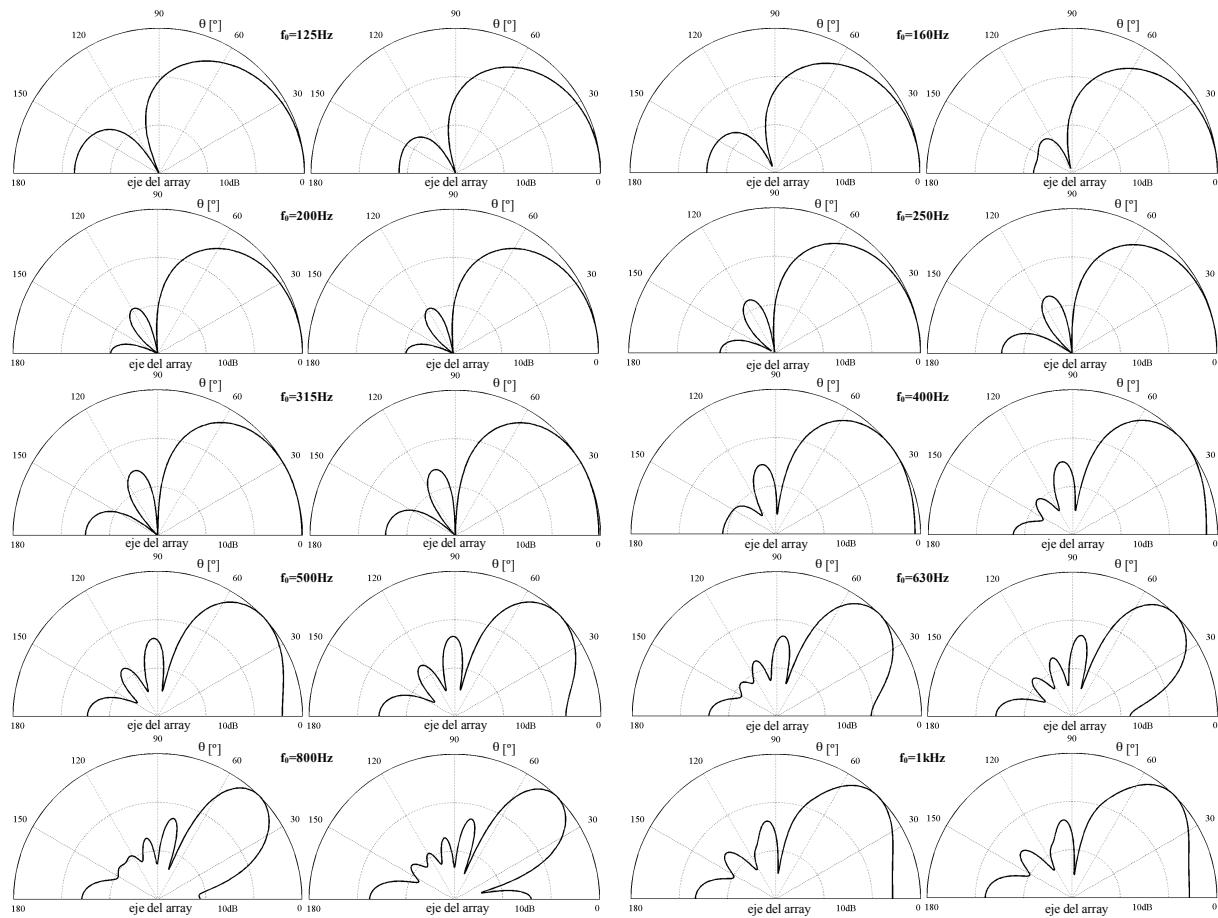


Figura 134. Curvas polares de directividad $D(\theta)$ [dB] para el subarray superdirectivo B_1 en bandas de 1/3 de octava. Apuntamiento lateral ($\theta_0=45^\circ$). Se ha usado la aproximación de campo lejano. En cada pareja de curvas, la de la izquierda corresponde a μ [dB]=-10dB y la de la derecha a μ [dB]=-15dB.

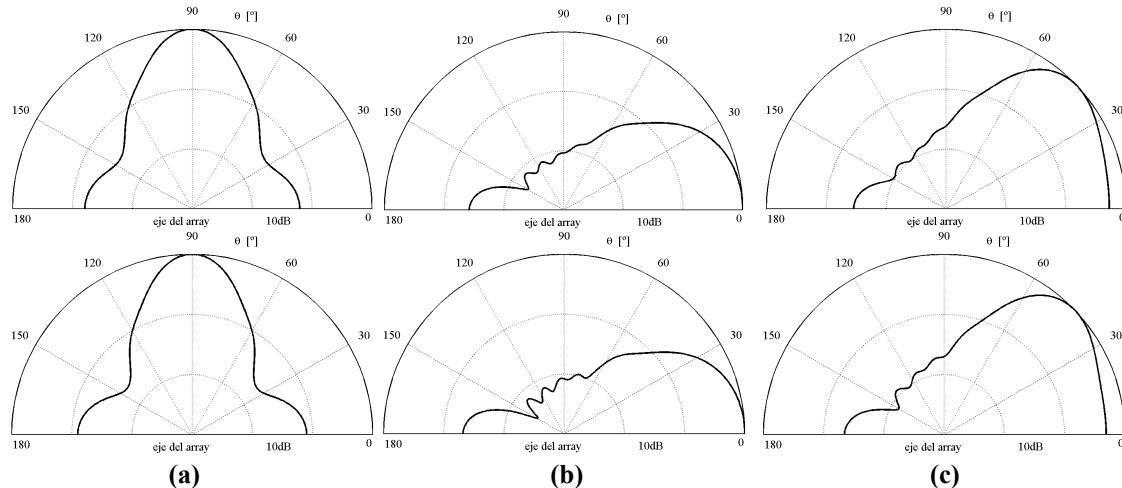


Figura 135. Curvas polares de directividad $D(\theta)$ [dB] de la banda superdirectiva B_1 , considerado el promedio en la banda [125Hz, 1kHz] (compárese con la Figura 64). Aproximación de campo lejano. Arriba se considera μ [dB]=-10dB y abajo μ [dB]=-15dB. (a) Broadside ($\theta_0=90^\circ$). (b) Endfire ($\theta_0=0^\circ$). (c) Lateral ($\theta_0=45^\circ$).

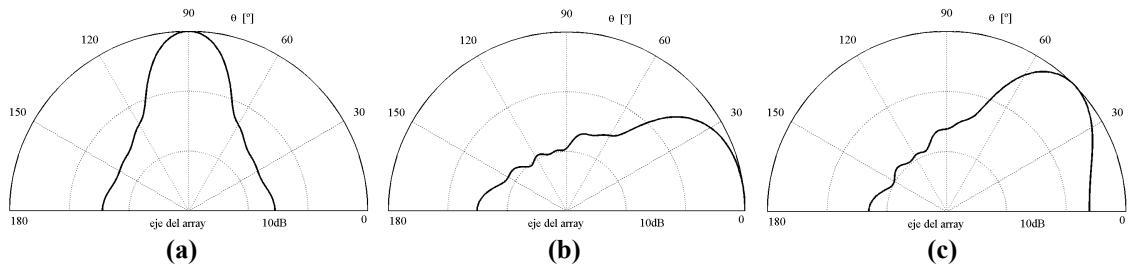


Figura 136. Curvas polares de directividad $D(\theta)$ [dB] del array anidado de 15 canales, considerando el promedio en la banda [125Hz, 4kHz], (compárese con la Figura 64). Aproximación de campo lejano. Sólo se considera μ [dB]= -15dB ya que con μ [dB]= -10dB se obtienen resultados muy parecidos. (a) *Broadside* ($\theta_0=90^\circ$). (b) *Endfire* ($\theta_0=0^\circ$). (c) Lateral ($\theta_0=45^\circ$).

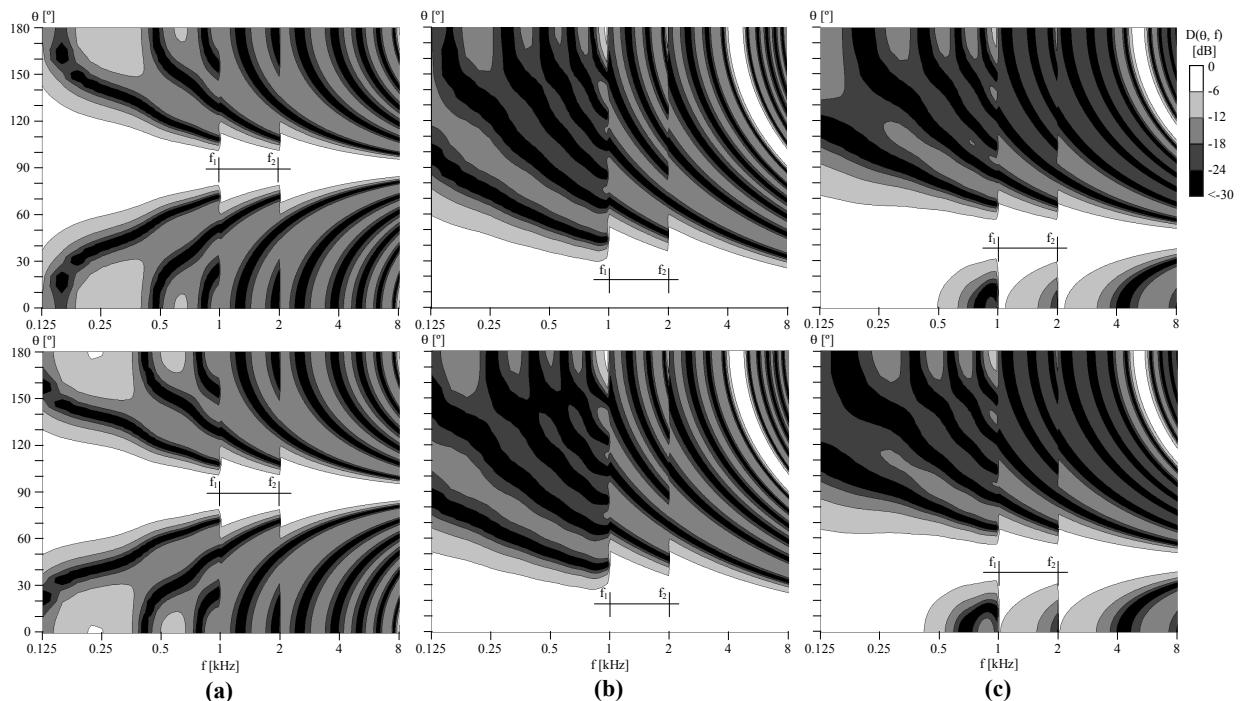


Figura 137. Mapa de directividad $D(\theta, f)$ [dB] del array anidado de 15 micrófonos con banda B₁ superdirectiva, en la aproximación de campo lejano para $f_1=1\text{kHz}$ y $f_2=2\text{kHz}$. Arriba μ [dB]= -10dB y abajo μ [dB]= -15dB. (a) *Broadside* ($\theta_0=90^\circ$). (b) *Endfire* ($\theta_0=0^\circ$). (c) Apuntamiento lateral ($\theta_0=45^\circ$).

Todo el desarrollo considerado en el presente apartado, sirve para obtener dos importantes conclusiones, en cuanto a la implementación final del prototipo.

En primer lugar, el apuntamiento *endfire* es el más conveniente, y se debe partir de esta situación, para hacer correcciones, si se desea (mediante un apuntamiento fijo o basado en *Web Cam*), sobre esta configuración de partida. Es decir, el array se deberá disponer con su eje positivo (determinado por el micrófono $i = 15$, véase la disposición de la Figura 6) dirigido a la escena donde se encuentra el locutor.

En segundo lugar, parece que el valor del parámetro de restricción $\mu = -15\text{dB}$ es el que consigue apuntamientos más selectivos (aunque las diferencias son pequeñas), sin por ello requerir una excesiva amplificación de los micrófonos y por consiguiente de su ruido eléctrico interno. Por tanto será la solución adoptada preferentemente en las pruebas y resultados que se exponen a continuación.

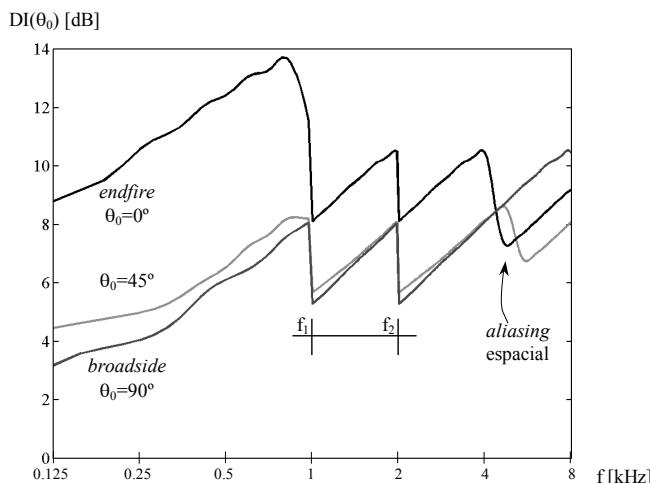


Figura 138. Índice de directividad $DI(\theta_0)$ hacia la DOA, en función de la frecuencia, para el array anidado de 15 micrófonos con banda B_1 superdirectiva (μ [dB]=-15dB). Se representan las configuraciones de apuntamiento tipo *broadside* ($\theta_0=90^\circ$), *endfire* ($\theta_0=0^\circ$) y lateral ($\theta_0=45^\circ$). Compárese con la Figura 67. Se ha usado la aproximación de campo lejano.

9.2 RESULTADOS ELECTROACÚSTICOS DEL CONFORMADOR DE HAZ SUPERDIRECTIVO

Las primeras pruebas a las que va a ser sometido el prototipo de array de 15 canales son de tipo electroacústico. El array va a ser calibrado, y se va a medir su directividad operando en tiempo real en diferentes configuraciones de conformación y en condiciones de prueba estándar, es decir en una cámara anecoica.

El montaje de laboratorio para estas pruebas se muestra en la Figura 139. La cámara anecoica pertenece al Departamento de Ingeniería Audiovisual y Comunicaciones (DIAC) de la U.P.M en la E.U.I.T. Telecomunicación.

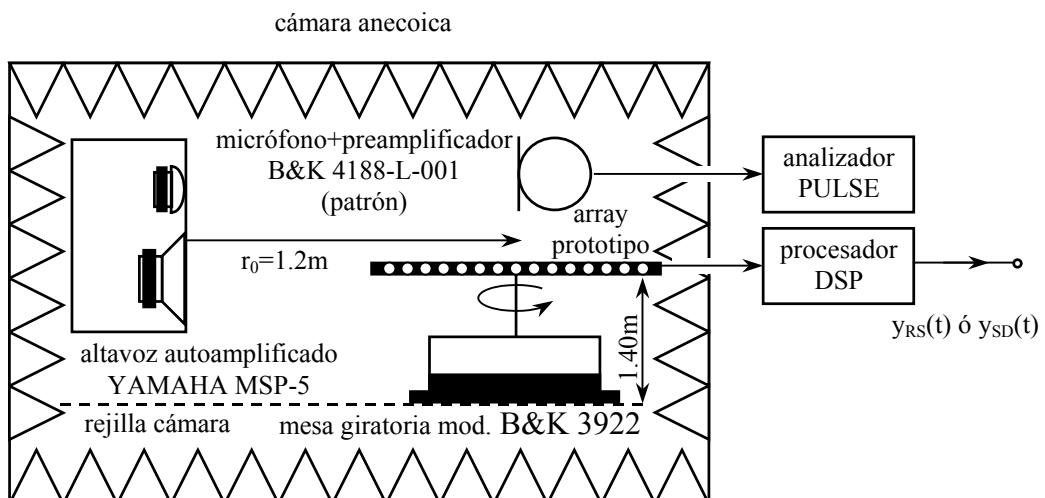


Figura 139. Disposición utilizada para medir el comportamiento electroacústico del array. Se ha utilizado el analizador PULSE™ 3560C de B&K y una mesa giratoria para medir directividad.

Para las pruebas de directividad se monta el array en una mesa giratoria, de tal manera que dé un giro completo alrededor del eje X en disposición de partida *broadside* (véase la Figura 6 para la convención utilizada en el eje coordenado). La señal de salida $y_{RS}(t)$ (conformador convencional, método de retardo y suma) o $y_{SD}(t)$ (conformador superdirectivo en la banda B_1) es grabada en disco duro para ser analizada posteriormente. Como puede verse en la Figura 139, la distancia de la fuente al array se ha fijado en $r_0 = 1.2\text{m}$ para las medidas de directividad, aunque puede ser menor para las medidas de calibración. La distancia de 1.2m permite el giro del array sin chocar con la fuente y a la vez es una distancia suficientemente corta, buscando el campo directo de la fuente que puede no mantenerse en baja frecuencia, a pesar de la cámara anecoica.

La disposición de medida utilizada no es del todo óptima y puede influir negativamente en las pruebas de calibración y de directividad que se detallan más abajo. Los problemas que plantea dicha disposición son los siguientes:

- La cámara anecoica no es ideal. En baja frecuencia (por debajo de 250Hz aproximadamente) aparecen algunos modos propios (ondas estacionarias debidas a las paredes) que se hacen muy patentes para frecuencias inferiores a 100Hz. Esto obliga a acercar bastante el array a la fuente, para que el campo acústico en el que se encuentra dicho array pueda ser considerado como libre. En este sentido, los micrófonos laterales se alejan del campo libre cuando pasan por la posición *broadside*, a medida que la mesa giratoria rota. También se aleja del campo libre el micrófono posterior cuando pasa por la posición *endfire*.
- La fuente de señal es un altavoz de banda ancha, que no es omnidireccional. Este hecho se hace más patente cuando la distancia del array al altavoz es pequeña, y en la posición *broadside*, ya que en este caso la desviación de los micrófonos laterales con respecto al eje del altavoz es mayor, y la señal captada por los mismos se reduce debido a la directividad de la fuente. Según esto último, se debe calibrar al array en baja frecuencia, en la banda de 100 - 250Hz, donde el altavoz está más cercano a la omnidireccionalidad y la respuesta de la cámara anecoica es suficientemente adecuada.

9.2.1 Calibración

Se ha procedido al cálculo de los coeficientes de calibración C , utilizando el método descrito en el apartado 8.3. Es decir, se ha excitado al array con una señal senoidal de 150Hz mediante una fuente situada, en posición *broadside*, a una distancia de $r_0 = 1\text{m}$, algo menor de la reseñada en la Figura 139, para el cálculo de la directividad del array. Con ello se ha obtenido el vector de coeficientes de calibración C que se representan en la Tabla 27 y la Figura 140. Estos coeficientes igualan la salida eléctrica de cada canal del array, recogida por la tarjeta de adquisición, a la frecuencia de 150Hz. Tienen en cuenta la corrección de nivel debida las diferentes distancias de la fuente a cada uno de los micrófonos del array y se calculan por la siguiente expresión:

$$C_i = \frac{y_{RMS8} \cdot r_i}{y_{RMSi} \cdot r_8} \quad (293)$$

con y_{RMS} el valor cuadrático medio de tensión eléctrica, medido en el tiempo de promediado T , y obtenido por el procesador con el subprograma de cálculo de coeficientes de calibración. La distancia r es la correspondiente de cada micrófono del array al altavoz. Los coeficientes C están estrechamente ligados a las posiciones del mando de amplificación de cada canal del

preamplificador MLA7. Si se modifica cualquiera de los mandos de dicho preamplificador habrá que proceder a una nueva calibración.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
SPL'_i-SPL'_8 [dB]	-1.33	-0.53	-0.34	-0.15	-0.06	-0.04	0.02	0.00	0.02	-0.04	-0.16	-0.25	-0.44	-0.73	-1.63
C _i [dB]	-2.87	-1.48	-0.38	1.37	-1.51	-3.30	0.21	0.00	-2.17	1.20	-1.73	-0.84	-5.21	-2.92	-1.53
C'_i [dB]	-1.54	-0.96	-0.04	1.52	-1.46	-3.27	0.19	0.00	-2.19	1.24	-1.58	-0.60	-4.77	-2.20	0.09

Tabla 27. Cálculo de los coeficientes de calibración en cámara anecoica, con un tono puro de 150Hz y a una distancia $r_0=1m$. La diferencia de nivel de presión entre cada canal y el central, $SPL'_i-SPL'_8$, ha sido medida con el micrófono patrón. En esta última medida se han corregido las pérdidas por distancia entre la fuente y cada micrófono del array y por tanto sólo incluye los efectos de la cámara anecoica y de la directividad del altavoz. C_i [dB] es la amplificación de correspondiente a los coeficientes de calibración y es la que se proporciona a cada canal en régimen de funcionamiento normal del procesador prototipo. C'_i [dB] es la amplificación corregida por los efectos de directividad del altavoz y de la cámara anecoica. Normalmente estos efectos no se tienen en cuenta.

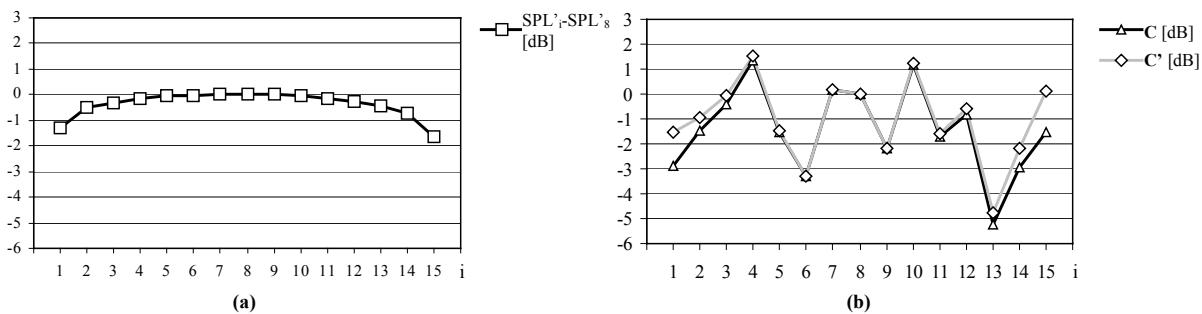


Figura 140. Representación gráfica de la Tabla 27. (a) Diferencia de nivel de presión entre cada canal y el central $SPL'_i - SPL'_8$, descontando las pérdidas por distancia. (b) Vectores C y C' de coeficientes de calibración.

Además, con el micrófono patrón de la Figura 139 se ha medido el nivel de presión sonora SPL ocasionado por el tono de 150Hz, en cada una de las posiciones de los micrófonos del array. En la Tabla 27 y la Figura 140 se representa la diferencia $SPL'_i - SPL'_8$ entre cada canal y el central. Aquí se han corregido también las pérdidas por distancia entre la fuente y cada punto del array mediante la siguiente expresión:

$$SPL'_i - SPL'_8 = SPL_i - SPL_8 + 20 \log \frac{r_i}{r_8} \quad (294)$$

donde SPL_i y SPL_8 son los niveles medidos por el micrófono patrón. Por tanto la diferencia $SPL_i - SPL_8$ sólo incluye los efectos de la cámara anecoica y de la directividad del altavoz.

A la vista de la Figura 140(a), parece que el único efecto encontrado que puede perjudicar al proceso de calibración está producido por la directividad del altavoz a la frecuencia de 150Hz, ya que puede observarse una ligera atenuación del orden de 1dB en la diferencia $SPL_i - SPL_8$, sólo en los micrófonos laterales del array. Si la diferencia de nivel en cada micrófono del array estuviese producida por la falta de anecoicidad de la cámara, el comportamiento de la Figura 140(a) sería más errático. En la Figura 140(b) se representan también el vector de coeficientes de calibración C' corregido por la diferencia $SPL_i - SPL_8$ y calculado mediante:

$$C'_i = C_i 10^{-\frac{SPL'_i - SPL'_8}{20}} \quad (295)$$

El vector C' es el que se debería aplicar en el DSP para realizar una correcta calibración de los canales, sin embargo esta posibilidad no se considera en la práctica, puesto que exigiría una medida del nivel de presión sonora en los micrófonos del array cada vez que se realizase un cálculo de coeficientes C . Si se eligen cuidadosamente las condiciones de calibración, las diferencias entre C y C' no serán muy grandes y los coeficientes C serán admisibles para que puedan ser utilizados sin mucho error en la calibración del sistema propuesto.

Para evitar los defectos de directividad del altavoz en la prueba de calibración, otra posibilidad es apuntar el array hacia la fuente en posición *endfire*, de tal manera que todos los micrófonos vean al altavoz desde su eje. Esta posibilidad no ha sido probada ya que las pequeñas dimensiones de la cámara anecoica en la que se ha calibrado el array hacen que el micrófono posterior quede muy alejado relativamente del altavoz, y muy próximo a la pared posterior, haciéndolo más susceptible de captar la onda estacionaria debida a esa superficie.

9.2.2 Directividad en cámara anecoica

Se ha medido la directividad de array en cámara anecoica, en diferentes configuraciones de conformación, con la disposición de la Figura 139. Para ello se excita al array desde el altavoz con un ruido rosa y se hace rotar a la mesa giratoria obteniéndose la señal procesada sólo por el conformador, $y_{RS}(t)$ para el conformador convencional o $y_{SD}(t)$ para el conformador superdirectivo. Se utiliza el vector de calibración C destacado anteriormente. También se ha obtenido, en el punto del procesador indicado en la Figura 125, la señal multicanal procedente del array –vector $y(t)$ –, una vez aplicada la calibración.

Medida de la velocidad del sonido

La alineación temporal de todos los canales del array se hace mediante la estimación de la diferencia de caminos acústicos entre la fuente y cada micrófono del array. Para pasar de distancia a tiempo hay que conocer la velocidad del sonido. Ciento es que si se propone el valor $c = 343 \text{ m} \cdot \text{s}^{-1}$, que es el que normalmente se supone para las condiciones ambientales normales, la desviación temporal de la alineación de canales no puede ser muy grande. Sin embargo, para ser más precisos se ha medido c utilizando la señal multicanal $y(t)$ y el método PHAT descrito en el punto 3.2.2 de la Parte 1 de esta Tesis. Un resumen de los resultados se representa en la Figura 141 y la Tabla 28.

En la Figura 141 se representa, a medida que la mesa va girando, el retardo estimado entre tres parejas representativas de micrófonos, en función del ángulo de apuntamiento del array θ_0 . Se considera el retardo máximo τ_{\max} y mínimo τ_{\min} y se estima c mediante la siguiente expresión:

$$c = \frac{\tau_{\max} - \tau_{\min}}{2 |z_i - z_j|} \quad (296)$$

siendo z_i y z_j las coordenadas de posición en el array de la pareja de micrófonos implicada en la medida. Con ello se obtienen los resultados de la Tabla 28. El valor $c = 349 \text{ m} \cdot \text{s}^{-1}$ parece el más consistente y es el adoptado para el ajuste de retardos del conformador del DSP.

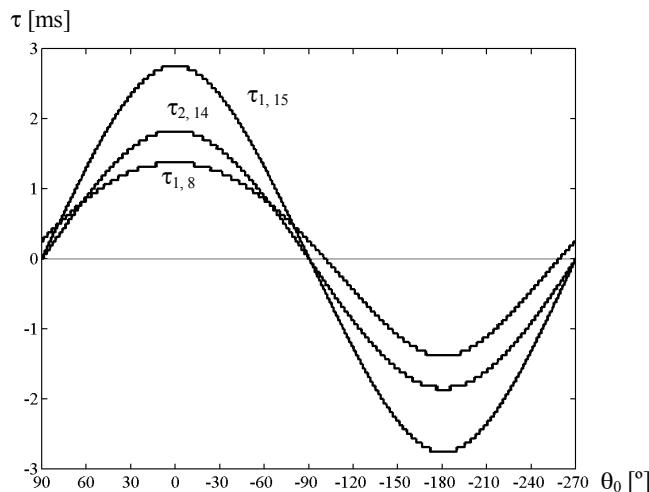


Figura 141. Retardo obtenido, a medida que la mesa giratoria está rotando, entre tres parejas significativas de micrófonos del array, en función de la dirección de la fuente.

	$\tau_{1,15}$ [ms]	$\tau_{2,14}$ [ms]	$\tau_{1,8}$ [ms]
τ_{\max} [ms]	2.7	1.8	1.4
τ_{\min} [ms]	-2.8	-1.9	-1.4

c [m/s]	349	347	349
---------	-----	-----	-----

Tabla 28. Estimación de la velocidad del sonido c mediante los retardos de los pares de micrófonos medidos, según la Figura 141.

Los resultados mostrados en la Tabla 28 y la Figura 141, además de ser útiles para el cálculo de c, corroboran en cierta medida el correcto funcionamiento del array implementado. El valor obtenido en los retardos intermicrofónicos se aproxima con mucha exactitud a lo predicho por la teoría, y el valor de c calculado es totalmente compatible con las condiciones ambientales en las que se hicieron las medidas electroacústicas.

Medidas de la directividad en cámara anecoica

Se ha medido la directividad del prototipo de array con el montaje expuesto Figura 139. Para ello se han considerado dos modelos de conformación. Por una parte el conformador combinado superdirectivo –salida $y_{SD}(t)$ –, con una constante de restricción $\mu = -15\text{dB}$, tal y como se expone en el punto 9.1 de la Tesis. Por otra parte el conformador convencional en tres subbandas –salida $y_{RS}(t)$ – como se describe en el punto 6.1. Téngase en cuenta que, para que los resultados teóricos puedan compararse con las medidas obtenidas en la cámara anecoica, en ambos conformadores se ha ajustado una distancia de apuntamiento de $r_0 = 1.2\text{m}$ en el procesador, que al ser tan pequeña influye en la respuesta del mismo. Por lo tanto, estos resultados difieren levemente de los considerados para la aproximación de campo lejano.

En las Figuras 142, 143 y 144 se representan las curvas de directividad $D(\theta)$ obtenidas para el conformador superdirectivo, promediando los resultados en las bandas normalizadas de 1/3 de octava. En éas figuras se representa igualmente la respuesta teórica esperada, en las mismas condiciones que las reales de medida.

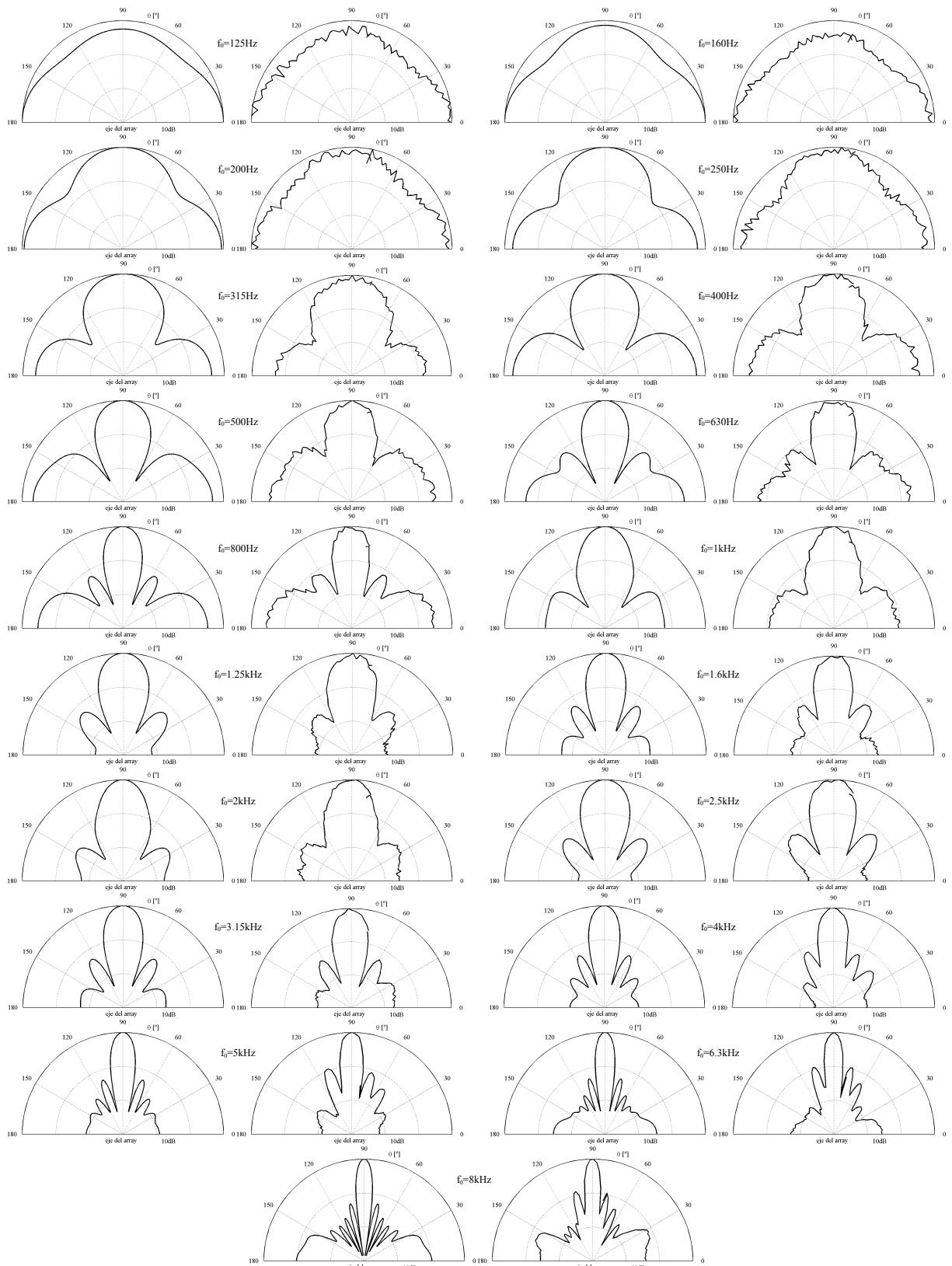


Figura 142. Curvas polares de directividad $D(\theta)$ [dB] en bandas de $1/3$ de octava para el array anidado de 15 micrófonos, con la banda B_1 superdirectiva y usando μ [dB] = -15 dB. Apuntamiento *broadside* ($\theta_0 = 90^\circ$, $r_0 = 1.2\text{m}$). En cada pareja de curvas, la de la izquierda corresponde a la respuesta teórica del array y la de la derecha a la medida en cámara anechoica.

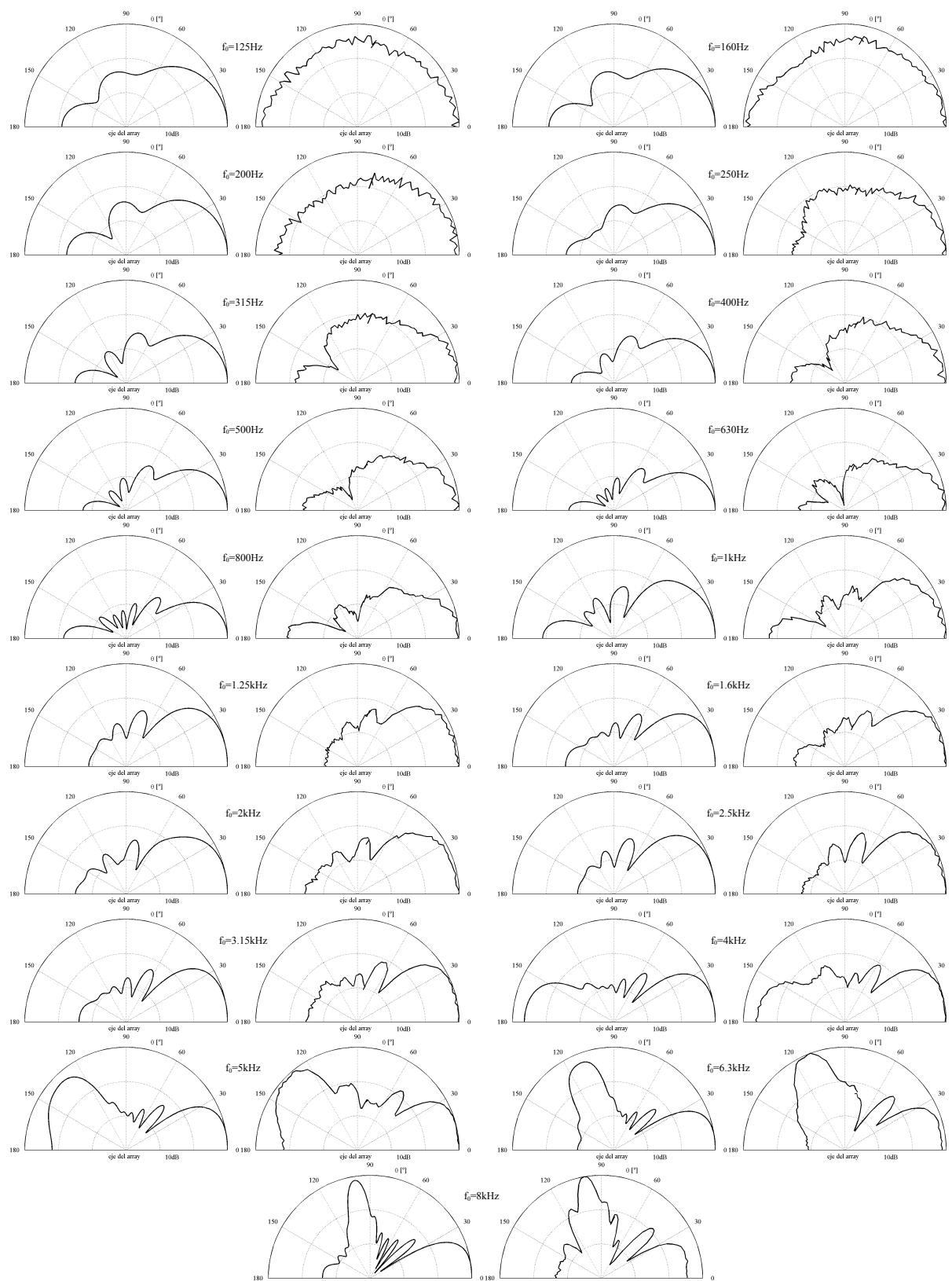


Figura 143. Curvas polares de directividad $D(\theta)$ [dB] en bandas de $1/3$ de octava para el array anidado de 15 micrófonos con la banda B_1 superdirectiva y usando μ [dB]=-15dB. Apuntamiento *endfire* ($\theta_0=0^\circ$, $r_0=1.2m$). En cada pareja de curvas, la de la izquierda corresponde a la respuesta teórica del array y la de la derecha a la medida en cámara anecoica.

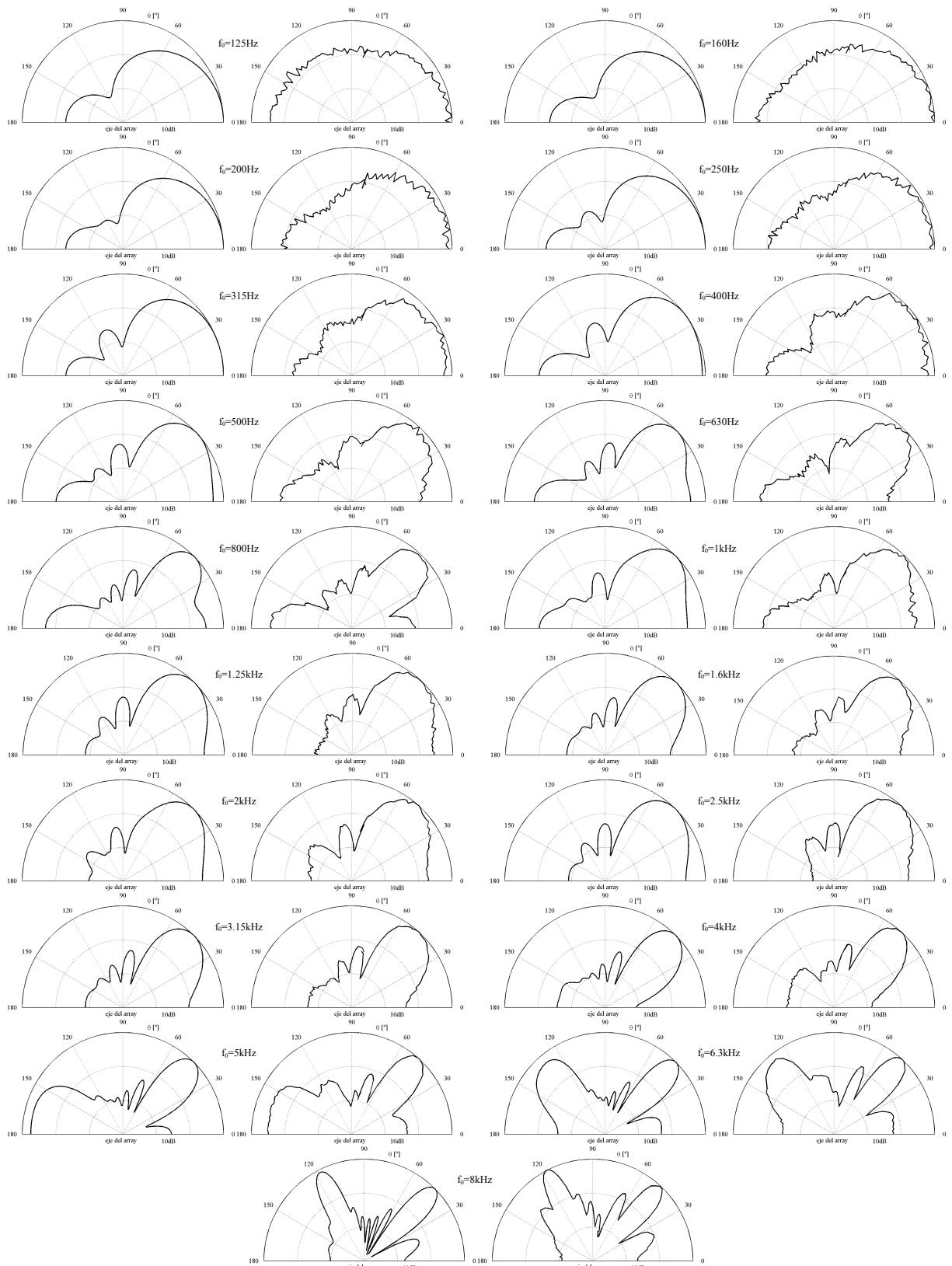


Figura 144. Curvas polares de directividad $D(\theta)$ [dB] en bandas de $1/3$ de octava para el array anidado de 15 micrófonos con la banda B_1 superdirectiva y usando μ [dB]=-15dB. Apuntamiento lateral ($\theta_0=45^\circ$, $r_0=1.2\text{m}$). En cada pareja de curvas, la de la izquierda corresponde a la respuesta teórica del array y la de la derecha a la medida en cámara anechoica.

A la vista de los resultados anteriores, cabe destacar en primer lugar, la excelente concordancia entre la respuesta del procesador implementado con la esperada teóricamente. Esta concordancia es equivalente a que la calibración del array se ha hecho con bastante precisión, ya que la conformación superdirectiva es muy sensible a las diferencias de nivel entre las señales captadas por micrófonos correspondientes, según se explicó en el punto 9.1 de la Tesis. Quizás las mayores discrepancias entre teoría y práctica se den en la parte más baja de la banda B_1 , para la conformación *endfire* (Figura 143) y el tercio de 125Hz para $\theta_0 = 45^\circ$ (Figura 144).

Debe resaltarse también, que en algunas curvas el máximo de captación no se produce en la dirección de apuntamiento θ_0 , y esa circunstancia sucede tanto para la representación teórica como para la práctica. Aparte de que puede influir que el factor μ sea un poco alto (véase el punto 9.1 para $\theta_0 = 45^\circ$), la razón fundamental de este defecto es que los coeficientes superdirectivos $\mathbf{W}'_{SD}(\omega)$ cargados en el DSP están ajustados para la aproximación de campo lejano, que va a ser la situación real de funcionamiento del prototipo. Sin embargo, la problemática ya expuesta que se produce en las medidas en cámara anecoica, sobre la distancia a la que debe colocarse la fuente, exige un r_0 pequeño para las medidas de directividad, lo que no quiere decir que el apuntamiento en campo lejano vaya a ser defectuoso. No se piensa que sea así, ya que si el comportamiento del prototipo es muy bueno en campo cercano (ya que las curvas teóricas y prácticas de directividad son muy parecidas en este caso), por extensión debe serlo también en campo lejano, por lo que habría que remitirse a las respuestas teóricas representadas en el punto 9.1 para la aproximación de campo lejano, que no pueden ser fácilmente determinadas de forma práctica.

En la Figura 145 se representa la respuesta de valor promedio del prototipo de array en la banda superdirectiva B_1 [125Hz, 1kHz] –salida $y_{SD}(t)$ –, para los tres apuntamientos considerados. En la Figura 146 se muestra lo mismo para la banda global [125Hz, 4kHz]. Estas representaciones dan idea, con una apreciación global, de la selectividad espacial del array real implementado. Se descartan las frecuencias por encima de 4kHz por estar fuera de la actuación del array de directividad constante considerado como tal. Otra vez la concordancia entre teoría y práctica vuelve a ser muy buena en todas las medidas.

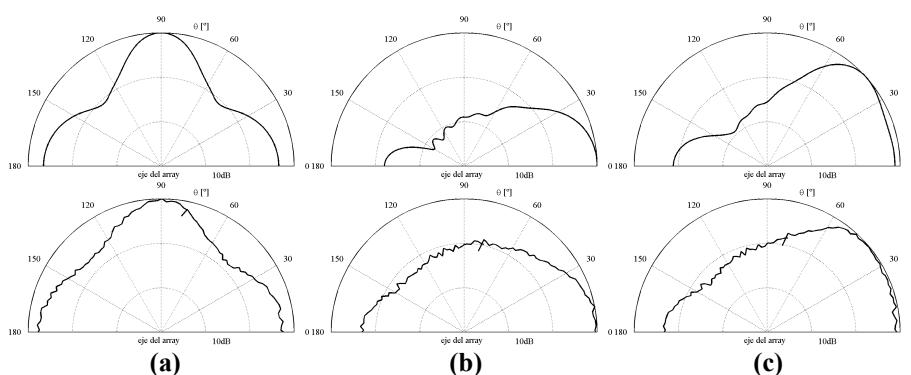


Figura 145. Curvas polares de directividad $D(\theta)$ [dB] de la banda superdirectiva B_1 con μ [dB]=−15dB, considerado el promedio en la banda [125Hz, 1kHz]. Arriba, se considera la respuesta teórica y abajo la respuesta real, procedente de la medida del prototipo de array. **(a)** *Broadside* ($\theta_0=90^\circ$, $r_0=1.2\text{m}$). **(b)** *Endfire* ($\theta_0=0^\circ$, $r_0=1.2\text{m}$). **(c)** *Lateral* ($\theta_0=45^\circ$, $r_0=1.2\text{m}$).

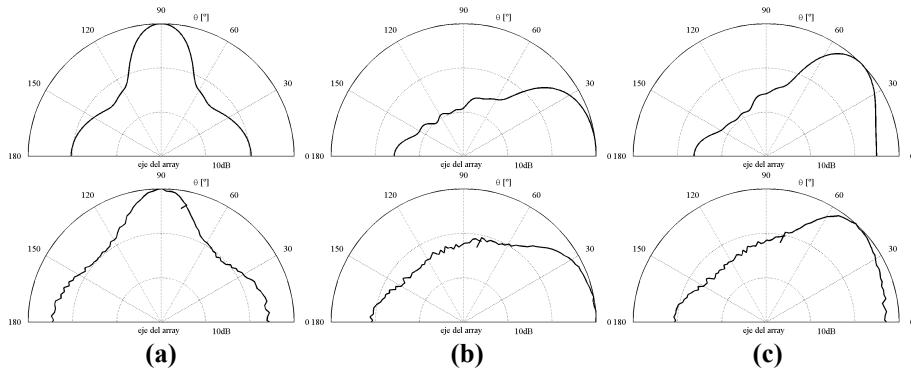


Figura 146. Curvas polares de directividad $D(\theta)$ [dB] de la banda superdirectiva B_1 con μ [dB]=-15dB, considerado el promedio en la banda [125Hz, 4kHz]. Arriba, se considera la respuesta teórica y abajo la respuesta real procedente de la medida del prototipo de array. (a) *Broadside* ($\theta_0=90^\circ$, $r_0=1.2\text{m}$). (b) *Endfire* ($\theta_0=0^\circ$, $r_0=1.2\text{m}$). (c) Lateral ($\theta_0=45^\circ$, $r_0=1.2\text{m}$).

En las Figuras 147, 148 y 149 se visualizan las curvas de directividad $D(\theta)$ en bandas de 1/3 de octava, medidas para el conformador convencional –salida $y_{RS}(t)$ –, y comparadas con la respuesta teórica del array en campo cercano. En estas representaciones sólo se ha considerado la banda B_1 de baja frecuencia, ya que las bandas B_2 y B_3 coinciden con las correspondientes para el conformador superdirective, mostradas en las Figuras 142, 143 y 144.

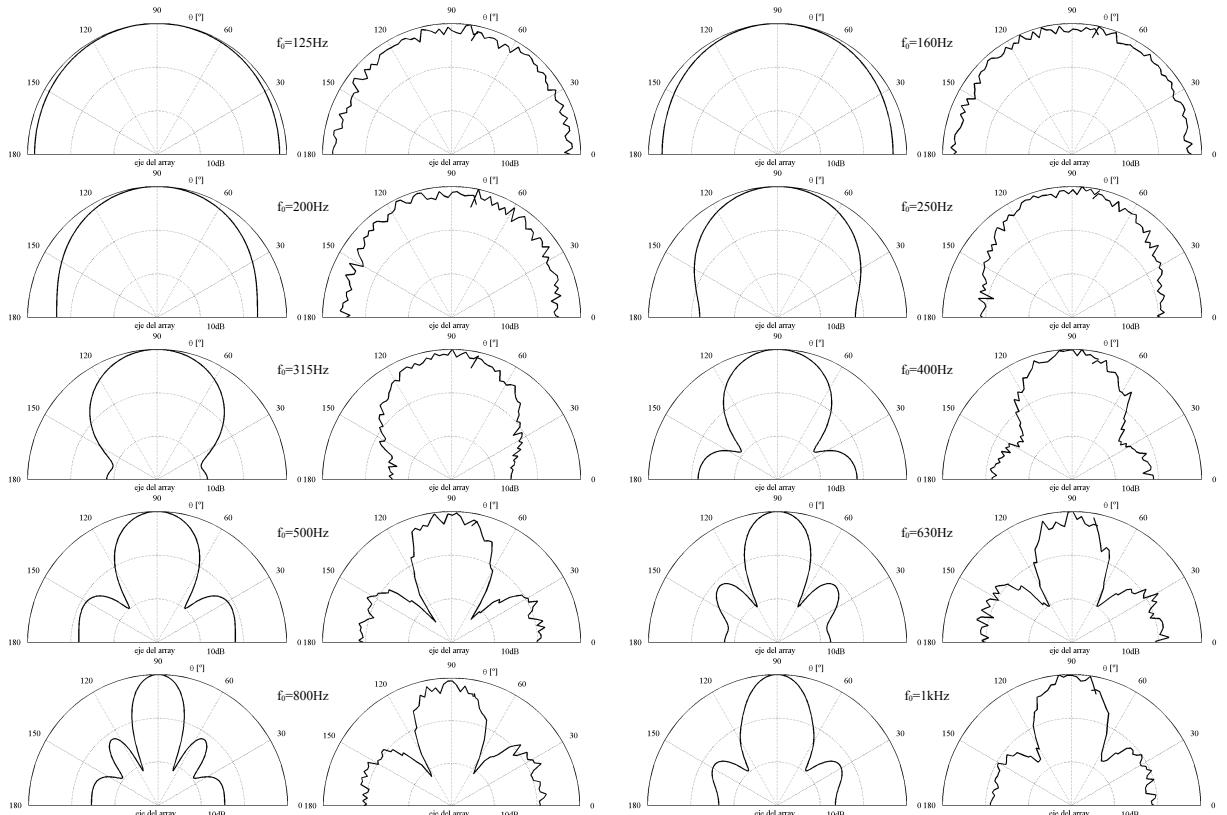


Figura 147. Curvas polares de directividad $D(\theta)$ [dB] en bandas de 1/3 de octava para el array anidado de 15 micrófonos usando el conformador convencional. Sólo se muestra la banda B_1 que es la que difiere de la mostrada en la Figura 142 para conformación superdirective. Apuntamiento *broadside* ($\theta_0=90^\circ$, $r_0=1.2\text{m}$). En cada pareja de curvas, la de la izquierda corresponde a la respuesta teórica del array y la de la derecha a la medida en cámara anechoica.

Como en el caso superdirective, otra vez se manifiesta una concordancia muy alta entre las predicciones teóricas y el resultado de las medidas con el prototipo. Esta concordancia ahora es mayor si cabe, ya que como se sabe, los defectos de calibración influyen menos en el conformador convencional que en el superdirective.

Comparando los resultados obtenidos por el conformador convencional en la banda B₁ con los del conformador superdirective, se muestra la mayor selectividad espacial en baja frecuencia de este último, lo que justifica la propuesta del conformador SD-ANS-MW, ya que con poco gasto computacional extra –los coeficientes $\mathbf{W}'_{SD}(\omega)$ se calculan previamente–, el mayor rechazo de ruido y reverberación laterales producirá previsiblemente una mayor mejora de señal de habla en la banda B₁.

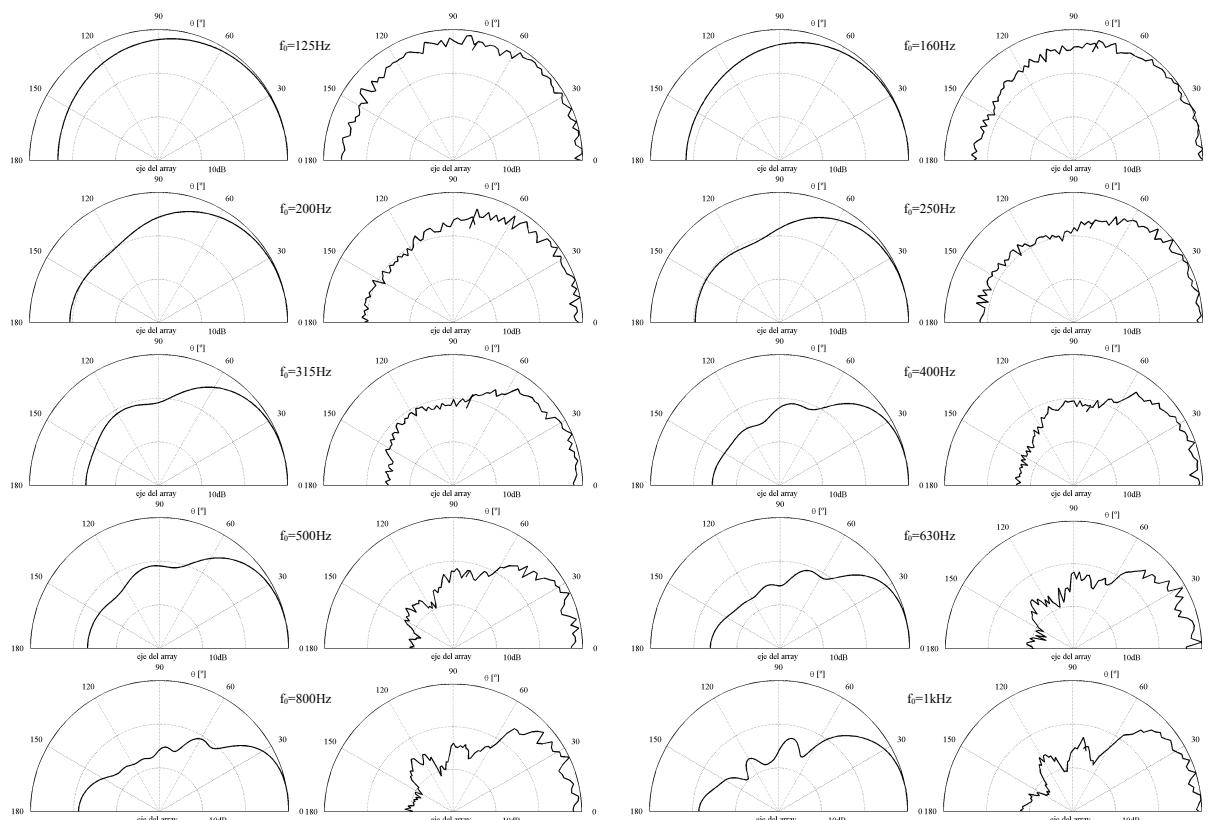


Figura 148. Curvas polares de directividad $D(\theta)$ [dB] en bandas de 1/3 de octava para el array anidado de 15 micrófonos usando el conformador convencional. Sólo se muestra la banda B₁ que es la que difiere de la mostrada en la Figura 143 para conformación superdirective. Apuntamiento *endfire* ($\theta_0=0^\circ$, $r_0=1.2\text{m}$). En cada pareja de curvas, la de la izquierda corresponde a la respuesta teórica del array y la de la derecha a la medida en cámara anecoica.

En la Figura 150 se representa la respuesta de valor promedio del conformador convencional en la banda [125Hz, 4kHz]. Si se compara con la Figura 146, puede corroborarse la mayor directividad global de la aproximación superdirective. No obstante, en esta última representación dicha mejora parece ligera debido a que la banda B₁ contribuye poco a la respuesta global, aunque subjetivamente sí es muy influyente.

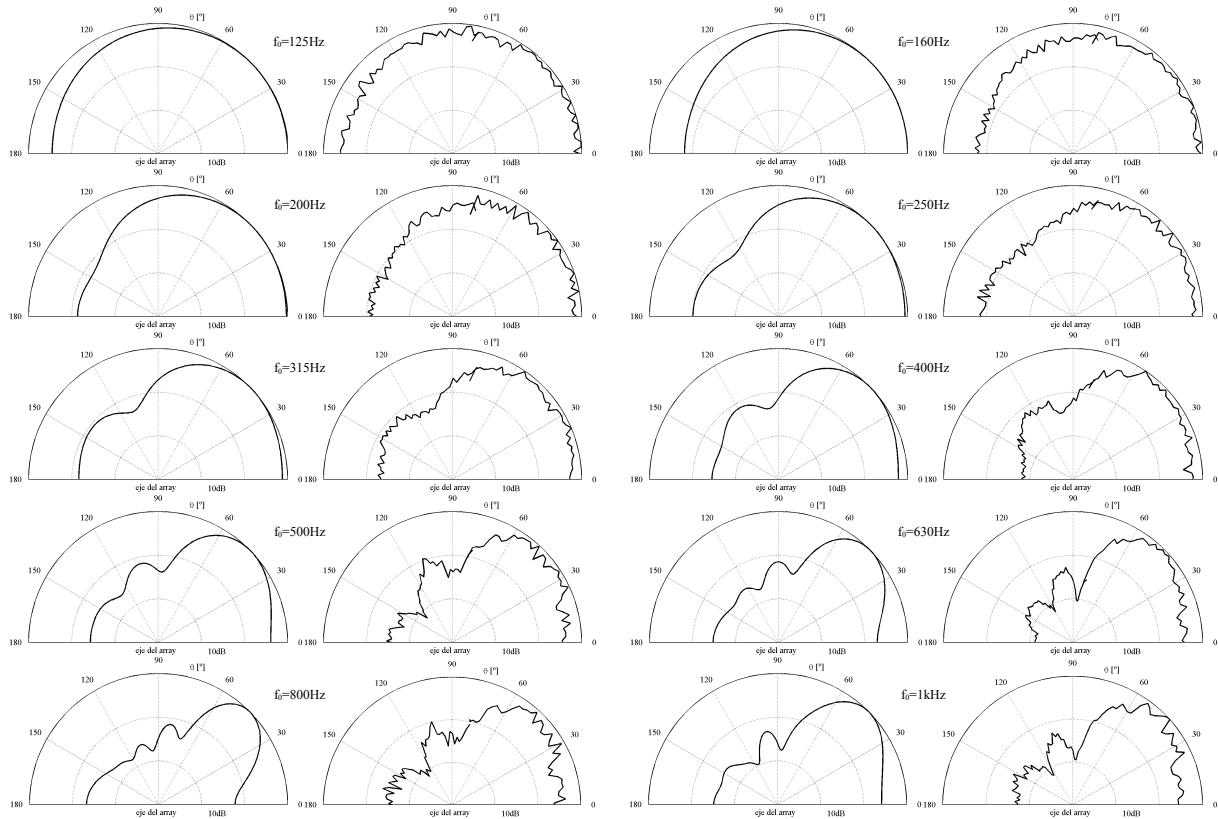


Figura 149. Curvas polares de directividad $D(\theta)$ [dB] en bandas de $1/3$ de octava para el array anidado de 15 micrófonos usando el conformador convencional. Sólo se muestra la banda B_1 que es la que difiere de la mostrada en la Figura 144 para conformación superdirectiva. Apuntamiento lateral ($\theta_0=45^\circ$, $r_0=1.2m$). En cada pareja de curvas, la de la izquierda corresponde a la respuesta teórica del array y la de la derecha a la medida en cámara anecoica.

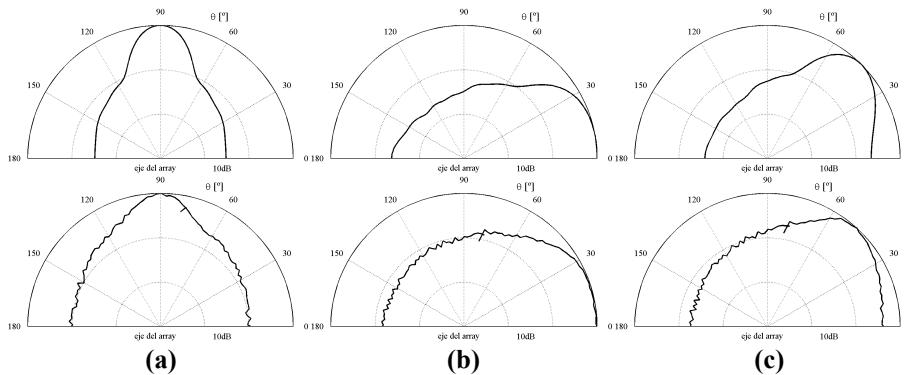


Figura 150. Curvas polares de directividad $D(\theta)$ [dB] para el conformador convencional considerándose el promedio en la banda [125Hz, 4kHz] (compárese con la Figura 146). Arriba, se considera la respuesta teórica y abajo la respuesta real procedente de la medida del prototipo de array. **(a)** Broadside ($\theta_0=90^\circ$, $r_0=1.2m$). **(b)** Endfire ($\theta_0=0^\circ$, $r_0=1.2m$). **(c)** Lateral ($\theta_0=45^\circ$, $r_0=1.2m$).

A continuación se representan los mapas de directividad $D(\theta, f)$ para el conformador superdirectivo (Figura 151) y el conformador convencional (Figura 152), en los que puede apreciarse con mayor detalle el comportamiento de la directividad con la frecuencia. Estos mapas representan la respuesta del array relativa a la captación máxima del mismo, producida en un determinado ángulo θ y una determinada frecuencia f . Por lo tanto la respuesta en

frecuencia relativa del array, según un ángulo concreto de desviación respecto a su eje, se puede extraer tomando una horizontal sobre cualquiera de los mapas.

El mapa de directividad $D(\theta, f)$ en la Figura 151 (conformador superdirective) correspondiente a las bandas B_2 y B_3 es ligeramente distinto al de la Figura 152, para el conformador convencional de retardo y suma, aunque en ambos casos, para esas frecuencias, el método de conformación haya sido el mismo. Los mapas no son realmente distintos, lo que varía es la asignación de colores, debido a que el conformador superdirective, considerada la DOA, no tiene respuesta en frecuencia $D(\theta_0, f)$ plana, alcanzándose un máximo de captación en baja frecuencia (banda B_1), debido a los errores de apuntamiento anteriormente comentados. Este incremento en la respuesta en frecuencia produce una pérdida de nivel relativo en las bandas B_2 y B_3 del mapa de la Figura 151. Por eso, en el conformador superdirective las bandas B_2 y B_3 son más oscuras (menor nivel) que en el conformador convencional, aunque mantengan la misma forma, como se manifiesta en las Figuras 151 y 152.

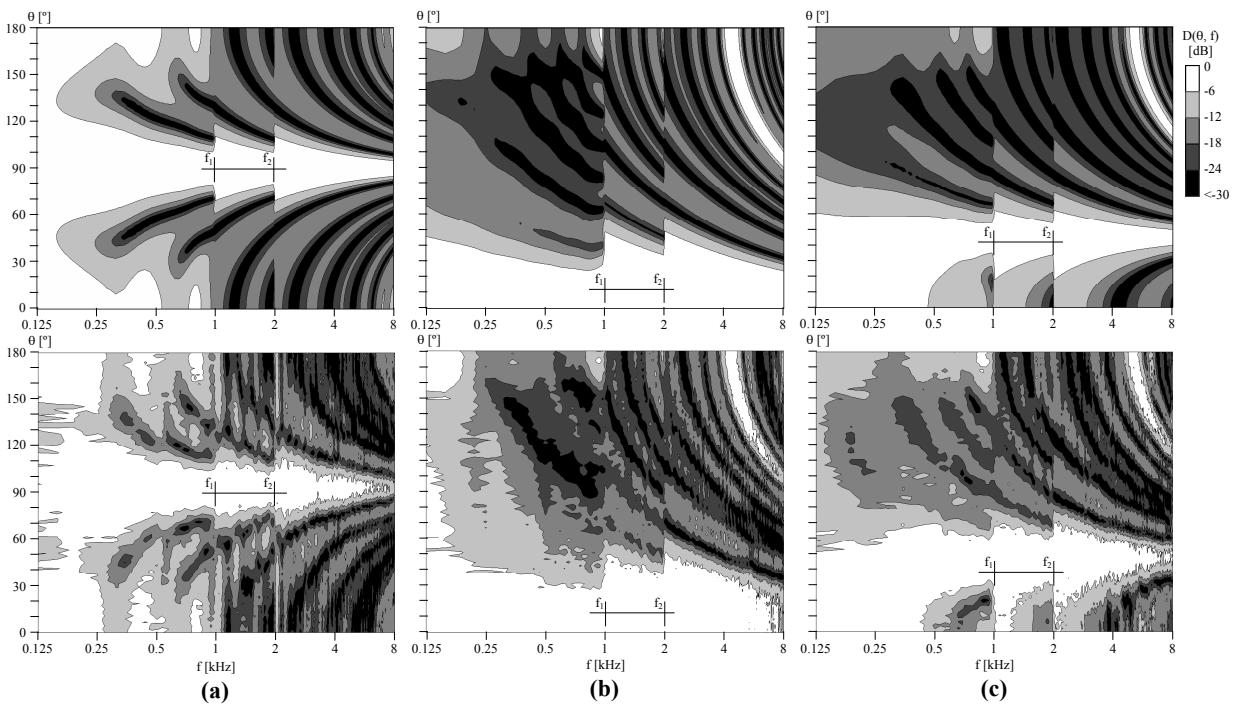


Figura 151. Mapa de directividad $D(\theta, f)$ [dB] del array anidado de 15 micrófonos con banda B_1 superdirective y μ [dB]=-15dB ($f_1=1\text{kHz}$ y $f_2=2\text{kHz}$). Arriba directividad teórica y abajo directividad medida con el prototipo de array. **(a)** *Broadsire* ($\theta_0=90^\circ$, $r_0=1.2\text{m}$). **(b)** *Endfire* ($\theta_0=0^\circ$, $r_0=1.2\text{m}$). **(c)** Apuntamiento lateral ($\theta_0=45^\circ$, $r_0=1.2\text{m}$).

En las Figuras 153 y 154 se representa respectivamente, el índice de directividad $DI(\theta_0)$ en la dirección de apuntamiento, para los casos de conformación superdirective y conformación convencional. Como se ha venido diciendo, la única zona de frecuencias donde el conformador superdirective se aleja significativamente del comportamiento teórico, es por debajo de unos 200Hz, en la banda B_1 , y en mayor medida con el apuntamiento *endfire*. Aun así, en esta zona de frecuencias el conformador superdirective apuntado en *endfire* es de 1 a 4dB más selectivo que el conformador convencional, comparando la Figura 153 con la Figura 154, lo cual justifica la utilización del primero.

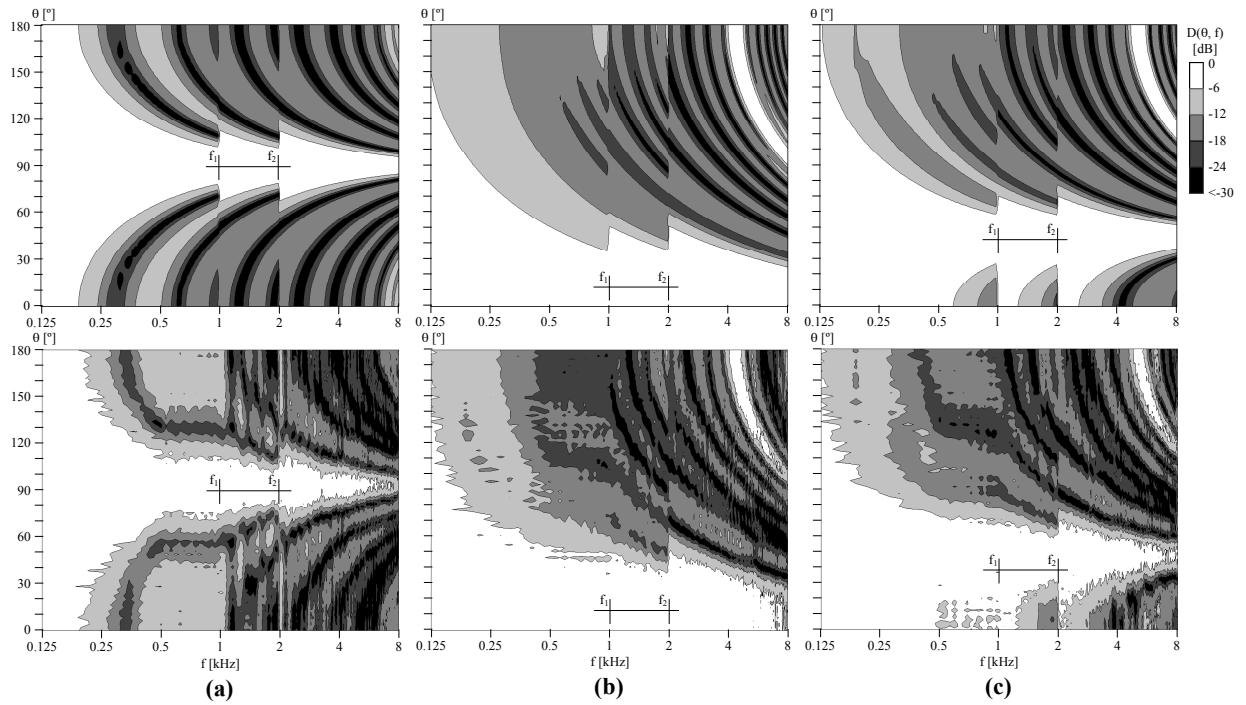


Figura 152. Mapa de directividad $D(\theta, f)$ [dB] del array anidado de 15 micrófonos usando conformación convencional en las tres bandas, B_1 , B_2 y B_3 . ($f_1=1\text{kHz}$ y $f_2=2\text{kHz}$). Arriba directividad teórica y abajo directividad medida con el prototipo de array. **(a)** *Broadside* ($\theta_0=90^\circ$, $r_0=1.2\text{m}$). **(b)** *Endfire* ($\theta_0=0^\circ$, $r_0=1.2\text{m}$). **(c)** Apuntamiento lateral ($\theta_0=45^\circ$, $r_0=1.2\text{m}$).

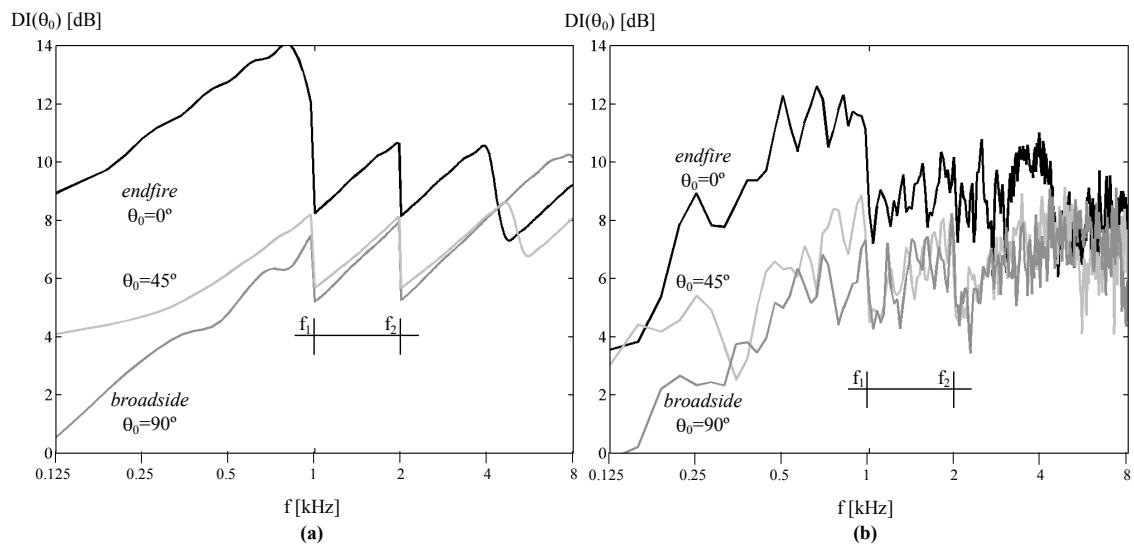


Figura 153. Índice de directividad $DI(\theta_0)$ en función de la frecuencia para el array anidado de 15 micrófonos con banda B_1 superdirectiva (μ [dB]=-15dB) en las configuraciones de apuntamiento tipo *broadside* ($\theta_0=90^\circ$, $r_0=1.2$), *endfire* ($\theta_0=0^\circ$, $r_0=1.2$) y lateral ($\theta_0=45^\circ$, $r_0=1.2$). **(a)** Aproximación teórica. **(b)** Medida sobre el prototipo de array obtenida mediante la transformada FFT promediada cada 2° de incremento del ángulo θ .

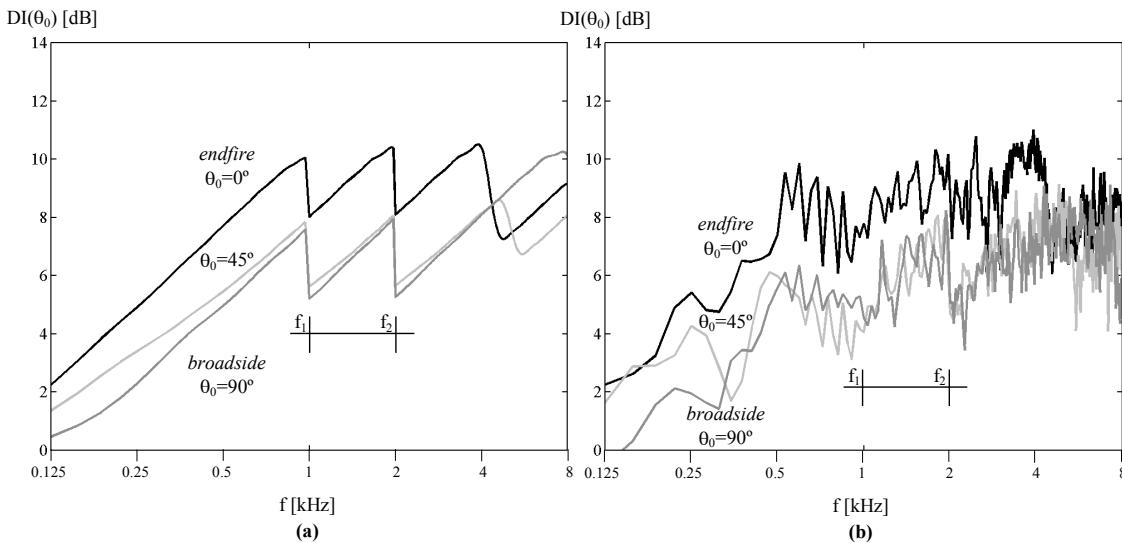


Figura 154. Índice de directividad $DI(\theta_0)$ en función de la frecuencia para el array anidado de 15 micrófonos usando conformación convencional en las tres bandas B_1 , B_2 y B_3 , en las configuraciones de apuntamiento tipo *broadside* ($\theta_0=90^\circ$, $r_0=1.2$), *endfire* ($\theta_0=0^\circ$, $r_0=1.2$) y lateral ($\theta_0=45^\circ$, $r_0=1.2$). **(a)** Aproximación teórica. **(b)** Medida sobre el prototipo de array obtenida mediante la transformada FFT promediada cada 2° de incremento del ángulo θ .

9.3 RESULTADOS DE MEJORA DE VOZ CON EL PROCESADOR SD-ANS-MW SOBRE UNA BASE DE DATOS REAL

Una vez que se han hecho las medidas oportunas de tipo electroacústico que confirman el funcionamiento correcto del procesador propuesto desde el punto de vista de la conformación de haz, es necesario probar los elementos de reducción de ruido de dicho procesador, en un escenario real pero con condiciones controladas de ruido y reverberación. Para ello ha sido generada por el autor una base de datos multicanal – $y(t)$ – en condiciones acústicas adversas. Es la que aquí se llama base de datos UPM. Junto con la base de datos se graba simultáneamente la señal procesada $y_{ANS}(t)$ y la señal de referencia $x_0(t)$, en condiciones que se describen después. De esa manera se obtienen los resultados de mejora a la salida del array, grabados *in situ*, y además se dispone de la señal multicanal $y(t)$ que ha producido esa señal mejorada. Esta señal multicanal servirá para los ajustes del array SD-ANS-MW en un procesador clónico del de tiempo real, generado mediante *software* en un ordenador de tipo PC y que, por contra, no funciona en tiempo real. Con la salida de habla producida por el procesador clónico, se han obtenido apreciaciones subjetivas del grado de mejora y medidas objetivas del estilo a las realizadas en las pruebas preliminares del capítulo 7. En concreto se ha evaluado la mejora producida por el procesador SD-ANS-MW con estimadores de calidad de tipo SNR: GNMR, GSNRA y GAI (véase el punto 5.1) y medidas con el índice E-RASTI propuesto por el autor en esta Tesis (según se describe en el punto 7.2).

9.3.1 Base de datos multicanal UPM

La base de datos multicanal UPM ha sido generada en el escenario representado en la Figura 155. Para ello se ha emitido una señal de habla pregrabada limpia, perteneciente a la base de datos AHUMADA en español [Ortega-García 00], desde un altavoz de alta calidad (altavoz de referencia). Se han seleccionado cinco locuciones, tres de voz masculina y dos de

voz femenina. Simultáneamente se ha emitido por unos altavoces de baja calidad (de tipo PC) un ruido de banda ancha [20Hz, 8kHz]. Este ruido corresponde a un automóvil de la marca Volvo (Figura 156) y ha sido seleccionado de la base de datos SpEAR [Wan 03].

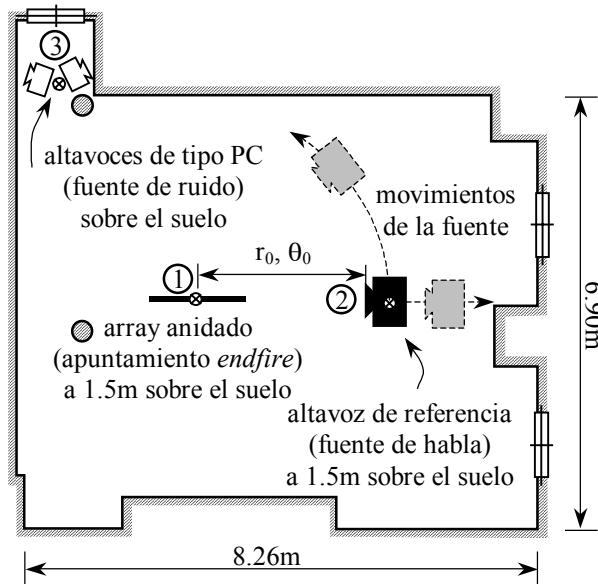


Figura 155. Geometría de la sala donde se ha generado la base de datos multicanal UPM y configuración de los experimentos. La altura de la sala es de 2.4m.

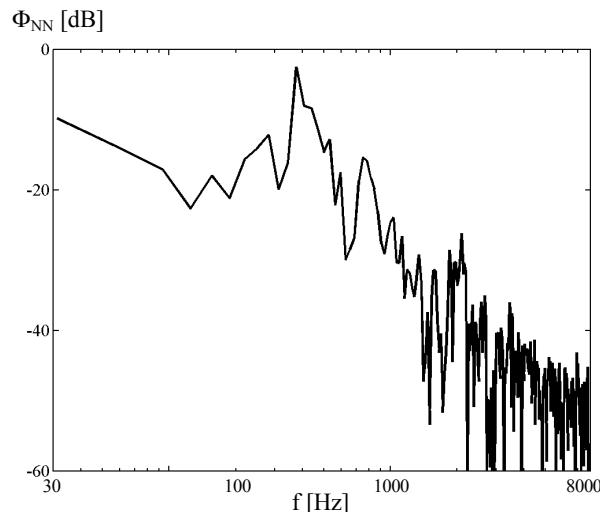


Figura 156. Espectro promedio de potencia del ruido introducido en la base de datos UPM, medido en el punto 1 de la Figura 155.

Los altavoces que generan el ruido se han situado enfocados a la pared, en una zona de la sala que no tiene visión directa con el array. De esa manera se potencia la componente difusa del ruido y de la reverberación asociada a dicho ruido (alta incoherencia espacial de dicho ruido en el array), con lo que la situación se parece más a la realidad.

El altavoz de referencia se enfoca hacia el centro del array, a diferentes distancias r_0 y ángulos θ_0 respecto del array según la Figura 155, de tal manera que el array se mantenga fijo y lo que varíe sea la posición de la fuente.

Se ha medido el tiempo de reverberación T_{60} de la sala en bandas de tercio de octava, obteniéndose los resultados que se muestran en la Figura 157. Los valores de esa figura muestran la reverberación captada con un micrófono patrón en el centro del array, y con la señal de excitación situada, bien en el punto donde se sitúa la fuente en apuntamiento *endfire* (punto 2, $r_0 = 2\text{m}$ y $\theta_0 = 0^\circ$) o bien donde se sitúa el ruido (punto 3 de la sala). Puede apreciarse que el tiempo de reverberación obtenido en el array es bastante similar en ambas condiciones de medida, lógicamente un poco mayor con la señal desde punto 3 por estar ésta más cerca de la pared y excitar en mayor medida los modos propios de la sala. En cualquier caso, la reverberación existente es bastante alta ($T_{60} > 1\text{s}$ en baja frecuencia) para una habitación tan pequeña, de tal manera que las condiciones acústicas en las que trabaja el array son bastante exigentes.

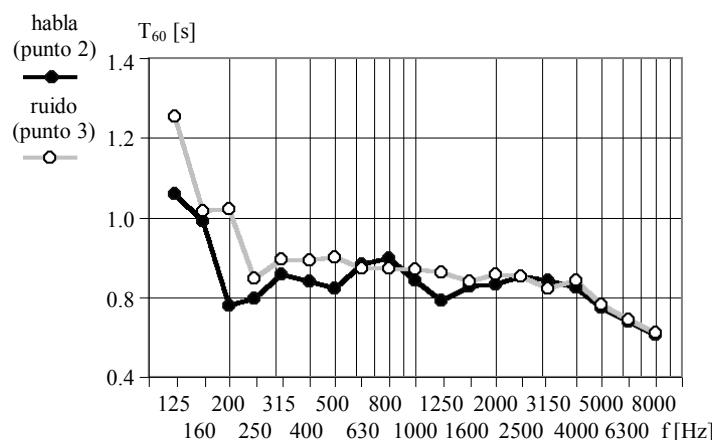


Figura 157. Tiempo de reverberación (T_{60}) medido en el punto de la sala de pruebas donde se sitúa el array (punto 1 de la Figura 155) mediante un analizador SYMPHONIE™ de 0.1dB. Las dos curvas corresponden a una excitación situada en punto 2 ($r_0=2\text{m}$ y $\theta_0=0^\circ$) donde se pone el altavoz de referencia (fuente de habla) y en el punto 3 donde se coloca la fuente de ruido.

En la Tabla 29 se detallan las condiciones de SNR en las que se han grabado los archivos de audio de la base de datos.

identificador de subcorpus	calidad relativa en función de SNR_s	Nº fragmentos generados	θ_0 [°]	r_0 [m]
B	buena	16	0	2
			90	4
	regular	16	0	2
			90	4
R	regular	16	0	2
			90	4
	mala	16	0	2
			90	4

Tabla 29. Condiciones de ruido y posición (según la Figura 155) con las que se ha generado la base de datos multicanal UPM.

Se han considerado tres calidades relativas en función de relación señal a ruido a posteriori SNR_8 introducida en el canal 8. La relación SNR_8 ha sido calculada con (241) que da una estimación de SNR midiendo el valor RMS de la señal en las tramas de voz y del ruido en las tramas de ausencia de voz (usando una segmentación manual de la actividad de voz). Estas tres calidades han sido catalogadas como “buena”, subcorpus **B**; “regular”, subcorpus **R** y “mala”, subcorpus **M**. En total 48 locuciones multicanal de 15 canales cada una –señal $y(t)$ antes del procesador–, más el canal de referencia $x_0(t)$ y la señal procesada $y_{\text{ANS}}(t)$ mediante el método SD-ANS-MW.

La señal limpia o de referencia $x_0(t)$ ha sido grabada con el micrófono nº 16 del array, a 10 cm del centro del altavoz que simula la fuente de voz. Debido a las limitaciones del funcionamiento en tiempo real del procesador implementado, ha sido imposible grabar simultáneamente los 15 canales con la señal limpia $x_0(t)$ y la salida $y_{\text{ANS}}(t)$, ya que éste no dispone de recursos suficientes para almacenar los 15 canales preprocesados $y(t)$ y a la vez la señal $y_{\text{ANS}}(t)$, puesto que en ese momento el procesador está trabajando con el máximo esfuerzo computacional. Además, como el micrófono de referencia para obtener $x_0(t)$ es omnidireccional (del mismo modelo que los del array), tiene excesiva tendencia a captar el ruido y la reverberación ajenos a la señal de voz. Por estas dos razones, la señal de referencia se grabó en un periodo de tiempo diferente al de generación de la señal multicanal sucia, en condiciones de ausencia de ruido añadido. Esto no ha sido inconveniente, puesto que al utilizar fragmentos de voz pregrabados, en ambas medidas no simultáneas se emite exactamente la misma señal. Posteriormente se ha requerido una alineación temporal de $x_0(t)$ con $y_8(t)$ e $y_{\text{ANS}}(t)$ usando el método PHAT. No reviste demasiada dificultad realizar la alineación temporal con una precisión de una muestra, incluso con el subcorpus **M** de alto ruido según la Tabla 29.

9.3.2 Experimentos y resultados

En la Tabla 30 se resumen los parámetros más importantes utilizados por el procesador prototípico en las pruebas sobre la base de datos real UPM.

SD-ANS-MW																				
enventanado			SD	ANS		MW												Coherencia		
						LBF														
						VAD polo simple y dos lados						frecuencias límite								
						señal			ruido			ataque			caída					
L[pt]	N[pt]	S _t [pt] /S[%]	μ [dB]	η	ton	λ _{Sa}	t _{ΔSa} [ms]	λ _{Sc}	t _{ΔSc} [ms]	λ _{Na}	t _{ΔNa} [s]	λ _{Nc}	t _{ΔNc} [ms]	f _{R1} [kHz]	f _{R2} [kHz]	λ _{coh}	t _{Δcoh} [ms]	α		
512	512	171 /67	-15	1 ó 10	0.8	0.343	10	0.586	20	0.997	4	0.343	10	0.5	3	0.8	143	50		

Tabla 30. Parámetros seleccionados para el procesador SD-ANS-MW en las pruebas sobre la base de datos real UPM.

Como se puede apreciar, se ha utilizado un enventanado de 512 puntos con un solapamiento de 2/3, sin *zero padding*. El conformador superdirectivo para la banda B₁ utiliza un parámetro de restricción $\mu[\text{dB}] = -15\text{dB}$ o de forma equivalente $\mu = 0.0316$. Por otra parte se ha considerado un parámetro de sobresupresión $\eta = 1$ ó 10, dependiendo de si se desea forzar o no la supresión auditiva ANS. La tonalidad de la señal de voz de entrada al array se fija al valor ton = 0.8. Para el filtro H_{MW}(ω) con el que se realiza la detección de umbrales auditivos, se ha utilizado un detector de actividad de voz VAD, con una estimación recursiva

de doble lado y polo único (véase el 4.1.4 de esta Tesis) con unos tiempos de ataque y caída para la estimación de señal de $t_{\Delta S_a} = 10\text{ms}$ y $t_{\Delta S_c} = 20\text{ms}$ respectivamente e igualmente para la estimación de ruido de $t_{\Delta N_a} = 4\text{s}$ y $t_{\Delta N_c} = 10\text{ms}$. El margen de frecuencias con el que el VAD toma la decisión es el comprendido entre $f_{R1} = 500\text{Hz}$ y $f_{R2} = 3\text{kHz}$. Los parámetros de coherencia son similares a los utilizados en los procesadores anteriormente propuestos para las pruebas preliminares del capítulo 7.

En estas condiciones se han realizado diferentes pruebas sobre la base de datos UPM de la Tabla 29 con dos condiciones de supresión ANS, determinadas por el valor de sobresupresión $\eta = 1$ ó 10 . Es decir, en total se han obtenido $48 \times 2 = 96$ resultados representados por la señal $y_{ANS}(t)$. También se han logrado en cada caso las señales correspondientes a las anteriores, tanto de referencia o limpias $x_0(t)$, como de entrada al array o sucias $y_8(t)$ (canal central del array). Con todas estas pruebas se han realizado los experimentos que se detallan a continuación, con sus resultados correspondientes.

Operación en tiempo real

El sistema final trabaja en el dominio de la frecuencia y como se expone en el apartado 8.3.1, el bloque FFT es el que más recursos del sistema consume. En las pruebas iniciales de eficiencia computacional, se evaluaba la tasa t_F de FFT's de 512pt –según (291)– que es capaz de realizar el procesador por segundo, sin recortar la señal analógica a la salida. Esta tasa se aumenta haciendo más pequeño el parámetro S_1 de desplazamiento interventana. En estas pruebas iniciales se midió una capacidad máxima de $t_F \approx 4000$ FFT's de 512pt por segundo, que marca el límite de funcionamiento en tiempo real del prototipo. Al parámetro de desplazamiento interventana elegido en el procesador probado (Tabla 30), $S_1 = 171\text{pt}$ ($S = 67\%$), le corresponde una tasa $t_F = 1497$ FFT's por segundo, incluyendo la reconversión temporal del canal de salida. Esto ofrece una cantidad estimada de recursos libres en el procesador del 63%. Esta disponibilidad estimada inicialmente es muy optimista ya que en el proceso de implementación del procesador final ha habido que incluir elementos adicionales de procesado y comunicación con el *Host PC*, que han reducido la cantidad de recursos libres del procesador.

Un análisis detallado de la operación en tiempo real más exigente, con el procesador SD-ANS-MW, para el subprograma de “operación” (véase la Figura 127) y utilizando un apuntamiento mediante Web Cam, ha ofrecido un diagrama de tiempos como el reflejado en la Figura 158.

El tiempo disponible para realizar todo el procesado necesario corresponde a los 10.7ms de desplazamiento intertrama. Dentro de ese lapso, la conformación de haz (incluye las transformadas FFT de análisis) consume un 36% del tiempo disponible, y el postfiltrado, con todas las estimaciones recursivas consideradas, necesita un 24% de ese tiempo, quedando por tanto disponible un 40% de tiempo remanente. Sin embargo, en ese tiempo no siempre el procesador está desocupado, puesto que cada 0.5s , se produce una comunicación del *HOST PC* con el DSP, para transferirle los datos de apuntamiento. Por tanto, en la práctica se ha determinado en un 15% aproximadamente el tiempo que queda libre en el procesador.

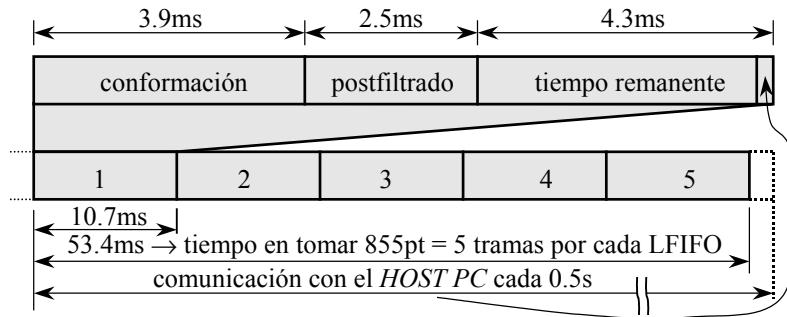


Figura 158. Diagrama temporal de operación en tiempo real para el procesador SD-ANS-MW. Cada 53.4ms, equivalentes a 855 muestras temporales (longitud LFIFO del ADC) se procesan cinco tramas de 512pt ($n_v=5$). Entre la adquisición de dos tramas consecutivas de 512pt se tarda un tiempo de 10.7ms (171pt) correspondiente a S_1 . Cada 0.5s el *HOST PC* se comunica con el DSP y le transfiere, mediante el teclado o un clic de ratón, la posición de la fuente (r_0, θ_0).

Estimación de umbrales auditivos $T(b)$ ó $T(\omega)$

Aunque en las pruebas preliminares sobre la supresión auditiva, descritas en el punto 7.3, se había verificado de modo informal la idoneidad del filtrado $H_{MW}(\omega)$ para la estimación de umbrales auditivos $T(b)$, hasta el momento no se ha hecho un estudio detallado de si la señal $y_{MW}(t)$ produce unos umbrales auditivos que se aproximen a los obtenidos de la señal original $x_0(t)$, que es el modelo que se trata de alcanzar. En este apartado se comparan los umbrales $T(b)$ obtenidos por el procesador prototipo mediante el método MW, con los umbrales $T_0(b)$ originales o de referencia obtenidos a partir de $x_0(t)$.

Se han obtenido con el procesador clónico al prototipo, 48 juegos de umbrales auditivos $T(b, k)$. Cada juego contiene, para cada una de las tramas k de señal con actividad de habla, los umbrales auditivos $T(b, k)$ de las 22 bandas críticas consideradas. En este experimento, la detección VAD se ha hecho de modo manual, observando la señal original $x_0(t)$. Estos 48 juegos se confrontan con los 48 juegos correspondientes a los umbrales originales $T_0(b)$. Se determina la relación $T(b, k)/T_0(b, k)$ para cada uno de los juegos y cada una de las tramas de voz implicadas.

En la Figura 159 se representa un ejemplo de estas comparaciones para una trama de voz particular de la base de datos UPM. En ella se muestran los umbrales de referencia y los estimados sobre la señal $y_{MW}(t)$.

En la Figura 160 se representa una muestra de tratamiento estadístico para uno de los 48 fragmentos de voz considerados. En ella puede verse, considerado un mismo fragmento de voz, la dispersión de $T(b, k)/T_0(b, k)$ relativa a las tramas k con actividad de habla. Como se aprecia, la naturaleza espectral del ruido añadido (Figura 156) influye en el cálculo de umbrales. Efectivamente, en baja frecuencia se tiende a sobreestimar el valor de $T(b, k)$ debido a que es mayor la presencia de ruido, que no es capaz de eliminar completamente el filtro de Wiener multicanal. La mayoría de la dispersión está en el entorno de ± 2 dB. Hay que considerar, que la relación SNR del fragmento elegido para el estudio estadístico de la Figura 160 es muy malo (véase la Figura 162). En alta frecuencia, a partir de 1kHz aproximadamente, la estimación de $T(b)$ es mucho mejor y muy poco variable con la frecuencia.

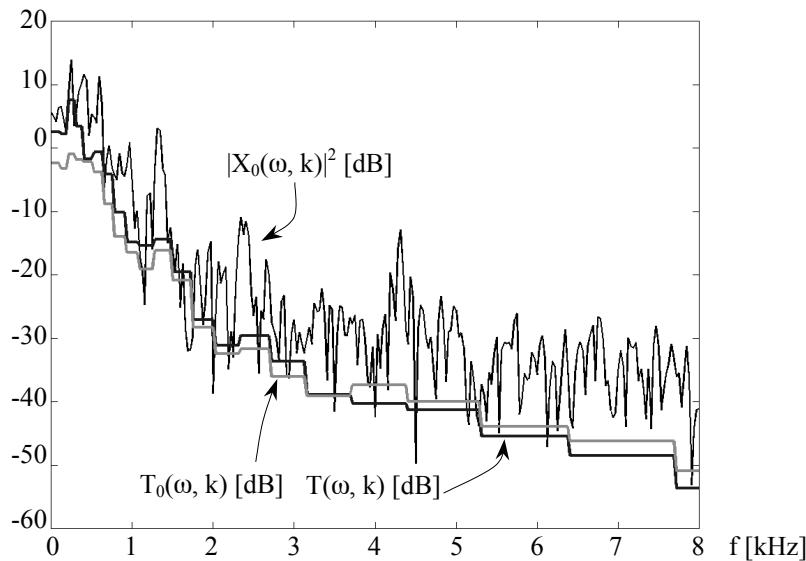


Figura 159. Ejemplo de cálculo de umbrales $T(\omega, k)$ para una trama k de voz en el instante $t=3s$ correspondiente a la Figura 162 de la base de datos UPM. $|X_0(\omega, k)|^2$ [dB] es la potencia de la señal de referencia, y $T(\omega, k)$ y $T_0(\omega, k)$ son respectivamente los umbrales estimados –a partir de $y_{MW}(t)$ – y de referencia –a partir de $x_0(t)$ –.

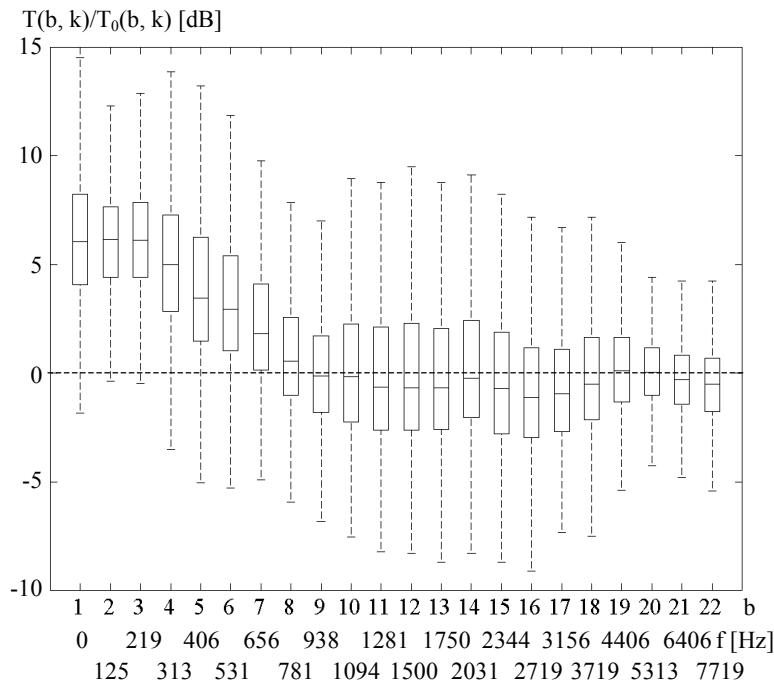


Figura 160. Diferencia en dB, para cada banda crítica b , entre el umbral $T(b, k)$ obtenido de la señal sucia y el umbral $T_0(b, k)$ obtenido de la señal de referencia, ambos correspondientes al ejemplo de la Figura 162. Las cajas representan la dispersión intertrama de los resultados, considerando sólo las tramas de actividad de voz. Están limitadas por los cuartiles superior e inferior, y se marca la mediana en su interior. En el eje de frecuencias se indica el límite inferior de cada banda crítica considerada.

Se ha realizado un estudio estadístico del mismo tipo, extendido a toda la base de datos multicanal UPM. La clasificación inicial de la base de datos (Tabla 29) en tres subcorpus (**M**, **R** y **B**) se hizo a priori, dosificando de una forma aproximada la cantidad de ruido aportada por los altavoces de tipo PC situados en el punto 3 de la Figura 155. Con posterioridad, se ha preferido clasificar la base de datos de forma más rigurosa, en función de la relación señal a ruido a posteriori (241) observada en el canal 8 del array. Para ello se han seleccionado tres tramos de SNR_8 :

- 1.- $\text{SNR}_8 < 5\text{dB}$
- 2.- $5\text{dB} < \text{SNR}_8 < 10\text{dB}$
- 3.- $10\text{dB} < \text{SNR}_8$

Se ha podido comprobar que estos tres tramos corresponden aproximadamente a la clasificación por el identificador inicial de subcorpus: **M**, **R** y **B**. Con esto, el estudio estadístico de la estimación de umbrales en cada uno de los 48 fragmentos de voz, se ha clasificado según la división anterior en tamos de SNR_8 .

En la Tabla 31 y la Figura 161 se resumen los resultados más significativos obtenidos. Hay que decir que la relación de umbrales $T(b, k)/T_0(b, k)$ ha sido corregida por igual en cada fragmento de voz para que el valor medio pase por $T(b, k)/T_0(b, k) = 1$ (ó 0dB). Esta operación es equivalente a igualar el valor RMS de $x_0(t)$ con el valor RMS de $y_{MW}(t)$ y se hace necesaria porque, a la hora de grabar la base de datos, no se ha conservado la información de amplitud relativa entre la señal de referencia $x_0(t)$ y la señal multicanal $y(t)$ y hay que hacer el ajuste de niveles posteriormente, si se quieren comparar los umbrales extraídos de dos señales diferentes. Por eso todas las curvas promedio de la Figura 161 están centradas en 0dB.

Los resultados que se muestran en la Tabla 31 y la Figura 161 representan el promedio y la desviación estándar intertrama de la diferencia entre umbrales $T(b, k)/T_0(b, k)$.

b		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
$T(b, k)/T_0(b, k)$ [dB]																							
$\text{SNR}_8 < 5\text{dB}$	media	4.4	4.4	3.7	2.5	2.3	2.0	1.2	0.1	-0.4	-0.9	-1.2	-1.4	-1.6	-1.5	-1.4	-1.2	-1.3	-1.3	-1.3	-1.8	-2.5	-2.9
	desviación estándar	4.3	3.7	3.6	3.9	4.3	3.9	3.5	3.5	3.4	3.6	3.6	3.7	3.7	3.7	3.5	3.3	3.2	3.2	3.1	3.2	3.4	3.9
$T(b, k)/T_0(b, k)$ [dB]																							
$5\text{dB} < \text{SNR}_8 < 10\text{dB}$	media	4.3	4.4	3.7	2.7	2.6	2.0	1.0	0.2	-0.1	-0.8	-1.3	-1.4	-1.4	-1.1	-1.0	-1.3	-1.6	-1.6	-1.5	-2.0	-2.7	-3.2
	desviación estándar	4.4	4.1	4.1	3.9	4.2	3.9	3.7	3.4	3.5	3.5	3.7	3.8	3.8	3.6	3.5	3.3	3.2	3.2	3.1	3.2	3.5	3.9
$T(b, k)/T_0(b, k)$ [dB]																							
$\text{SNR}_8 > 10\text{dB}$	media	3.5	4.2	4.3	3.3	2.8	2.2	1.2	0.0	-0.4	-0.7	-0.8	-1.0	-1.1	-0.8	-0.6	-1.0	-1.6	-1.8	-2.0	-2.7	-3.3	-3.7
	desviación estándar	4.2	4.0	3.8	3.5	3.9	3.8	3.6	3.3	3.3	3.4	3.6	3.7	3.6	3.4	3.2	3.1	3.0	3.2	3.3	3.6	4.1	4.6

Tabla 31. Diferencias de umbrales $T(b, k)/T_0(b, k)$ [dB] obtenidas utilizando la base de datos UPM, clasificadas en tres tramos de SNR_8 . Se muestra, en función del índice b de banda crítica, la media intertrama (en decibelios), considerando todas las tramas con actividad de voz de los 48 fragmentos analizados de la base de datos e igualmente la desviación estándar obtenida sobre el conjunto de valores $T(b, k)/T_0(b, k)$, ambas expresadas en decibelios.

Llaman la atención dos hechos significativos. En primer lugar que no existen diferencias apreciables en los resultados cuando se comparan los tres tramos de calidad seleccionados en función de SNR_8 . En segundo lugar que las desviaciones promedio del valor 0dB no son muy grandes, lo que avala que el cálculo de umbrales auditivos se está haciendo correctamente, incluso en condiciones de ruido muy alto. Las mayores desviaciones del valor

0dB (+4dB aproximadamente) corresponden a la baja frecuencia, donde la presencia de ruido es muy alta, siendo las estimaciones correspondientes a frecuencias medias bastante más correctas. Las desviaciones estándar observadas alrededor del valor promedio están comprendidas entre 3 y 4dB en la mayoría de los casos.

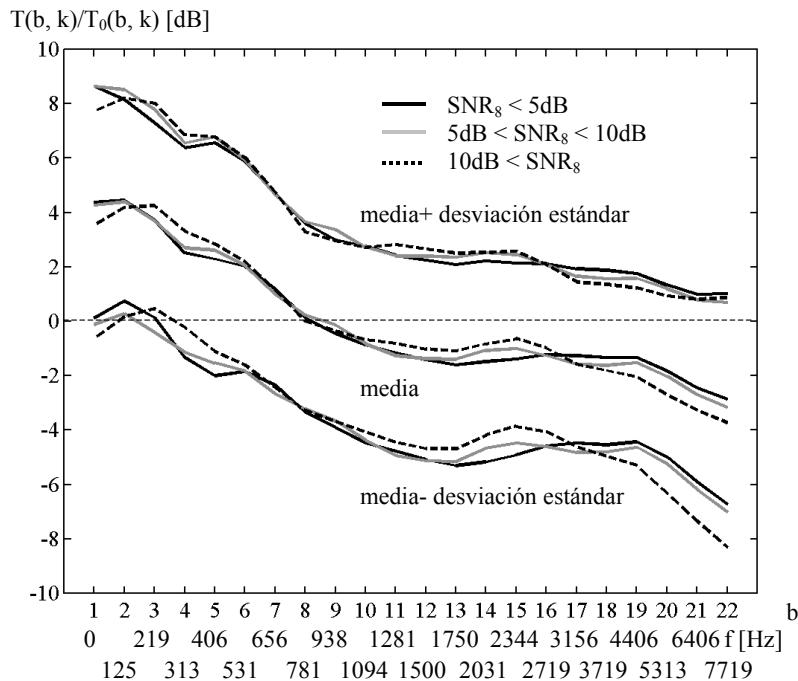


Figura 161. Representación gráfica de los resultados de la Tabla 31. Para cada tramo de SNR_8 considerado, se representa el valor medio intertrama de $T(b, k)/T_0(b, k)$ [dB] en función del índice b de banda crítica y los correspondientes desplazamientos hacia arriba o hacia abajo de la curva de valor medio, en una cantidad igual a la desviación estándar observada. El eje de frecuencias indica el límite inferior de cada banda crítica considerada.

Debe tenerse en cuenta que, en un gran porcentaje de las tramas de voz consideradas para cada fragmento, la relación señal a ruido local existente es de muchos decibelios negativos, por la poca amplitud de $x_0(t)$ en esa trama. Eso provoca que se calculen incorrectamente los umbrales. Aun así, las desviaciones de los valores promedio son relativamente pequeñas, y aunque la desviación estándar no lo indica, se ha verificado que existe prácticamente la misma probabilidad de desviación positiva de $T(b, k)/T_0(b, k)$ con respecto a la media que de desviación negativa (hecho que sí puede observarse en Figura 160 para uno de los elementos de la base de datos, puesto que las cajas de desviación son prácticamente simétricas alrededor de la mediana).

Evaluación del procesador SD-ANS-MW sobre la base de datos UPM

Se han realizado pruebas de mejora de señal de voz multicanal utilizando el array microfónico prototipo sobre la base de datos UPM. En la Tabla 30 se resumen los parámetros utilizados por el procesador SD-ANS-MW. A continuación se exponen y estudian los resultados de mejora de voz obtenidos, utilizando evaluadores objetivos de tipo SNR y con el método E-RASTI propuesto en la Tesis (véase el capítulo 5).

En la Figura 162 se visualiza una muestra de actuación del procesador sobre uno de los elementos de la base de datos. En ella se representan tanto las señales temporales como los espectrogramas correspondientes.

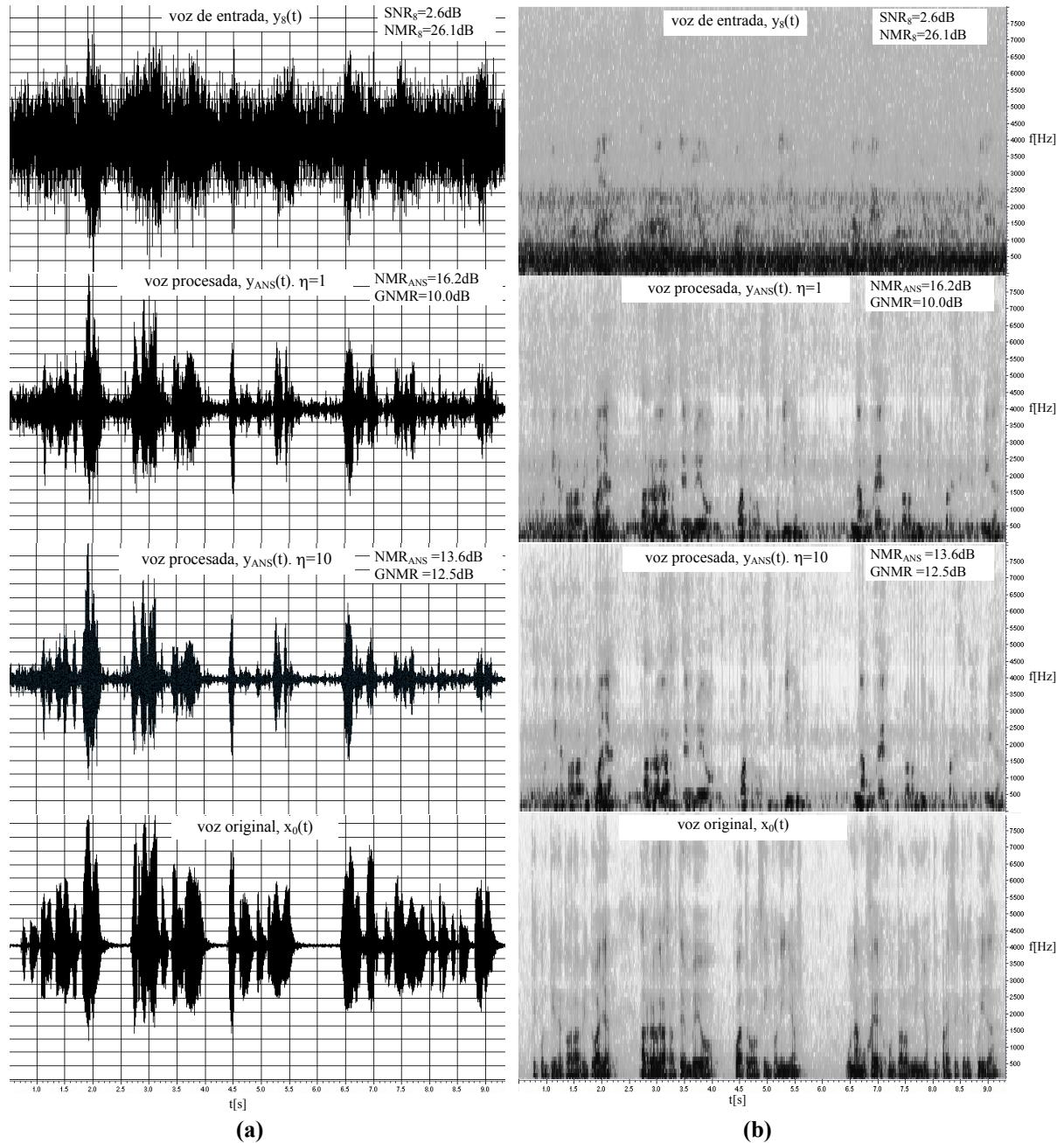


Figura 162. Señal $y_8(t)$ (antes de procesar) e $y_{ANS}(t)$ (después de procesar) para $\eta=1$ y $\eta=10$ junto con la señal de referencia $x_0(t)$ u original. Las señales mostradas pertenecen a la base de datos UPM con la fuente a 4m del centro del array en posición $endfire$ ($\theta_0=0^\circ$). **(a)** Señales temporales. **(b)** Espectrogramas.

En la parte superior de la Figura 162(a) se muestra la señal contaminada $y_8(t)$, antes del procesador. A continuación dos salidas procesadas $y_{ANS}(t)$, asociadas a los dos valores considerados del parámetro de sobresupresión ($\eta = 1$ y $\eta = 10$). Recuérdese que la propuesta original de procesador auditivo ANS, según aparece en [Tsoukalas 97] adopta $\eta = 1$. En la parte inferior se representa la señal original $x_0(t)$ correspondiente. Puede verse que la cantidad de ruido eliminado en $y_{ANS}(t)$ es menor para $\eta = 1$ que para $\eta = 10$. El parámetro de

sobresupresión $\eta = 1$ produce una voz más natural pero con mayor presencia de ruido de fondo. En cualquiera de los casos cotejados se ha podido apreciar subjetivamente la distorsión por ruido musical, debida al procesado mediante análisis STFT, tanto mayor cuanto menor es la calidad de partida de la señal $y_8(t)$. No obstante, como quedó patente en las pruebas preliminares del capítulo 7, atendiendo al ruido musical remanente, el procesado auditivo es mejor que otros tipos de postfiltrado usados tradicionalmente sobre señal de habla.

Hay que considerar que, en las pruebas realizadas con el procesador prototipo sobre la base de datos UPM, se ha utilizado el detector VAD basado en recursión de polo simple y dos lados, que a pesar de funcionar bien, tiene un comportamiento no comparable al utilizado en las pruebas del capítulo 7, que era un VAD manual como ya se advirtió, basado en la detección de las tramas de actividad de voz usando la señal original $x_0(t)$. En ese sentido, puede entenderse que la apreciación global subjetiva de los resultados del prototipo sea peor que la que se obtuvo en las pruebas preliminares. Aunque también en esta apreciación influye significativamente que el ruido aditivo de la base de datos UPM pueda ser más difícil de eliminar, por sus características espectrales o nivel, que el correspondiente a CMU y a simCMU-2.

La gran discrepancia existente entre los valores de SNR_8 y NMR_8 de la Figura 162 (ambos son en cierta medida relación señal a ruido –cambiando el signo para NMR ya que ésta es una relación “ruido a señal”–), y que se mantiene en todos los resultados, se debe a que la evaluación del ruido $N(\omega)$ mediante (223) supone que la reverberación (que es muy grande) es considerada también como ruido. Por eso los valores de NMR_8 son siempre muy malos (muy altos), incluso en señales con relativamente poco ruido en una apreciación subjetiva. Sin embargo, esta forma de evaluar el ruido informa con más detalle de la presencia de reverberación y ruido, que es incapaz de detectar el método tradicional de evaluación a posteriori (241) con el que se calcula SNR_8 .

A continuación se muestran los resultados obtenidos en forma de gráficas y tablas. En todos ellos sólo se han considerado las tramas de actividad de voz, que se han seleccionado manualmente mediante la señal de referencia $x_0(t)$. Esta detección manual de las tramas de voz sólo se hace en la estimación de los resultados, y no durante el funcionamiento del procesador SD-ANS-MW, donde se hace una detección VAD automática.

En la Figura 163 y la Tabla 32 se manifiestan los resultados en cuanto a ganancia GNMR para los dos procesadores con $\eta = 1$ y $\eta = 10$. La elección del factor η depende de la calidad de la señal a limpiar. Cuando la calidad es muy mala (mucho ruido y/o reverberación), es preferible aumentar η , porque, aunque el resultado es menos natural, la claridad de la señal resultante es mayor. Para señales de calidad media/buena es preferible utilizar $\eta = 1$, porque proporciona menos distorsión a la señal de salida del procesador y la mejora de voz es similar que con $\eta = 10$.

Analizando la Figura 163 se verifica cómo un aumento del factor η de sobresupresión sólo produce beneficios, en cuanto a la mejora de la relación NMR, cuando la calidad de la señal sucia $y_8(t)$ es muy mala ($\text{SNR}_8 < 10\text{dB}$).

En la Tabla 32 se han evaluado en detalle los resultados para discernir alguna dependencia de la mejora conseguida con el ángulo de apuntamiento θ_0 y la distancia r_0 . Para ello se han realizado promedios agrupando las locuciones evaluadas con la misma distancia y ángulo de apuntamiento. Se puede apreciar que un aumento de la distancia r_0 empeora el comportamiento del procesador, lo que es bastante previsible, porque la reverberación captada aumenta con la distancia. También se puede apreciar una ligerísima mejora en la señal

procesada cuando se utiliza el conformador con apuntamiento *endfire* ($\theta_0 = 0^\circ$) en lugar del apuntamiento *broadside* ($\theta_0 = 90^\circ$).

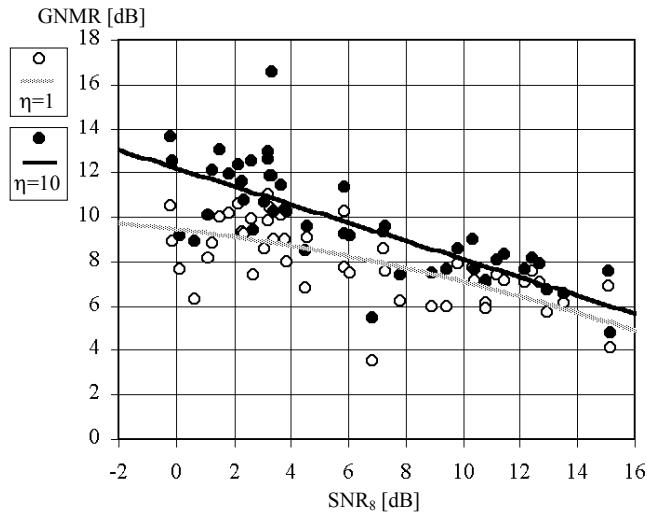


Figura 163. GNMR en función de SNR_8 para dos valores del factor de sobresustracción $\eta=1$ y $\eta=10$. Los datos se han ajustado con un polinomio de orden 2.

		GNMR[dB]			
η	$\theta_0 [^\circ]$	$r_0 [m]$	$\text{SNR}_8 < 5\text{dB}$	$5\text{dB} < \text{SNR}_8 < 10\text{dB}$	$\text{SNR}_8 > 10\text{dB}$
1	0	2	9.97	9.09	7.21
		4	8.91	5.68	6.36
	90	2	9.19	8.18	6.84
		4	8.94	6.55	6.03
10	0	2	12.47	9.95	7.97
		4	11.33	7.55	7.57
	90	2	11.17	9.29	7.36
		4	10.80	8.02	6.79
promedio total			10.29	7.82	6.85

Tabla 32. Resultados en GNMR para los tres tramos considerados de relación señal a ruido SNR_8 (véase la Figura 163). Los resultados de GNMR en cada casilla muestran los promedios de todos los resultados que cumplen los valores de β , r_0 y θ_0 de esta tabla.

Los resultados en cuanto a GSNR_A (Figura 164 y Tabla 33) y GAI (Figura 165 y Tabla 34) corroboran las conclusiones anteriores, aunque con claridad inferior, ya que hay una menor diferencia en los resultados con $\eta = 1$ y $\eta = 10$, que no se ha manifestado en las escuchas informales de la señal procesada.

En la Figura 166 y la Tabla 35 se muestran los resultados obtenidos utilizando el índice E-RASTI como método de evaluación objetiva de la calidad de habla. Se evalúa la salida procesada $y_{\text{ANS}}(t)$ para los dos valores $\eta = 1$ y $\eta = 10$ y también la señal de referencia $x_0(t)$, en todo los casos haciendo una comparación con la señal sucia $y_8(t)$ correspondiente al canal central del array.

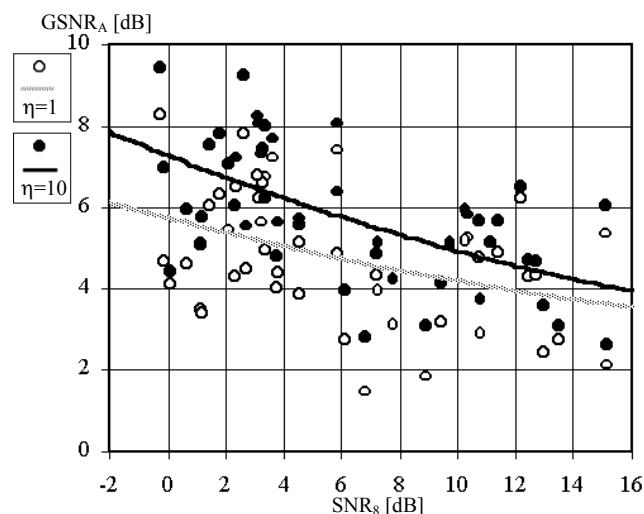


Figura 164. GSNR_A en función de SNR₈, como en la Figura 163.

GSNR _A [dB]				
η	θ_0 [°]	r_0 [m]	GSNR _A [dB]	
1	0	2	6.21	
		4	6.37	
	90	2	4.77	
		4	4.50	
10	0	2	7.23	
		4	7.67	
	90	2	6.08	
		4	6.04	
promedio total		6.11	4.29	
			4.49	

Tabla 33. Como la Tabla 32 pero mostrando resultados en GSNR_A (véase la Figura 164).

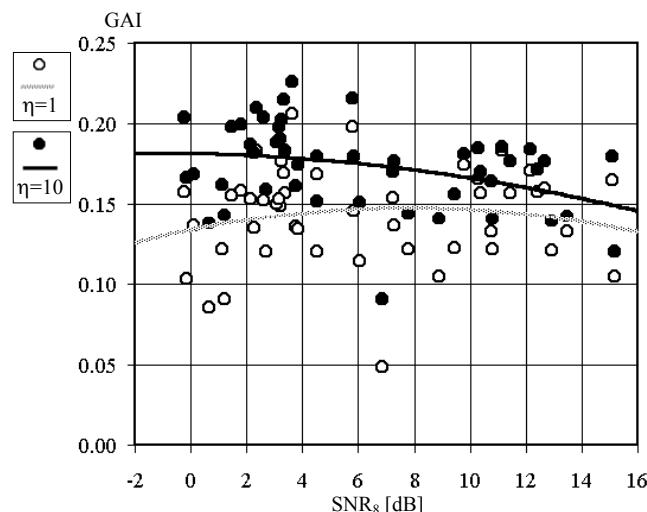


Figura 165. GAI en función de SNR₈, como la Figura 163.

GAI				
η	θ_0 [°]	r_0 [m]		
			SNR ₈ <5dB	
1	0	2	0.17	
		4	0.15	
	90	2	0.14	
		4	0.13	
10	0	2	0.20	
		4	0.19	
	90	2	0.18	
		4	0.17	
promedio total		0.16	0.15	
			0.15	

Tabla 34. Como la Tabla 32 pero mostrando resultados en GAI (véase la Figura 165).

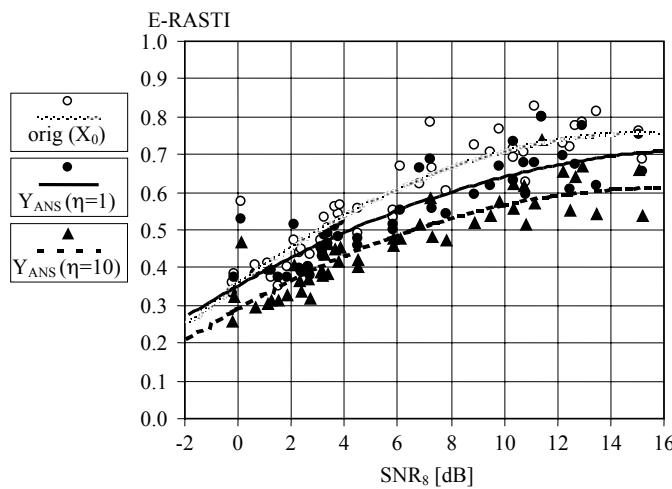


Figura 166. Índice E-RASTI en función de SNR₈, como en la Figura 163. Las tres curvas representan respectivamente el índice E-RASTI considerando como señal de entrada la señal $x_0(t)$ de referencia u original, la señal de salida $y_{ANS}(t)$ con el parámetro de sobresupresión $\eta=1$ y la señal de salida $y_{ANS}(t)$ con el parámetro de sobresupresión $\eta=10$. Para los tres juegos de resultados la señal de salida en el evaluador E-RASTI ha sido $y_8(t)$.

El primer hecho significativo observado en la evaluación de E-RASTI, es que la calidad de las señales procesadas parece ser erróneamente mayor (menor E-RASTI en la Figura 166) que la de la señal original $x_0(t)$. Este defecto del índice E-RASTI ya fue manifestado y explicado en las pruebas preliminares de los apartados 7.2 y 7.3. Simplemente pone de manifiesto que el procesador aumenta mucho la modulación de la señal procesada, normalmente de forma forzada (y por tanto introduciendo distorsión), de tal manera que la modulación resultante puede ser incluso mayor que en la señal original, aunque no por ello la calidad de la señal procesada sea mayor. Como ya se explicó, el índice E-RASTI sólo sirve para comparar con precisión señales del mismo tipo, contaminadas o procesadas de forma similar, en este caso las salidas $y_{ANS}(t)$ para $\eta = 1$ y $\eta = 10$, ya que la comparación de dos señales muy diferentes $-x_0(t)$ e $y_{ANS}(t)$ lo son – puede dar ese tipo de resultados discordantes.

E-RASTI. Señal original $x_0(t)$					
$\theta_0 [^{\circ}]$	$r_0 [m]$	$SNR_8 < 5dB$	$5dB < SNR_8 < 10dB$	$SNR_8 > 10dB$	
0	2	0.48	0.65	0.80	
	4	0.44	0.66	0.72	
90	2	0.48	0.67	0.77	
	4	0.45	0.66	0.71	
promedio total		0.46	0.66	0.75	

(a)

E-RASTI. Señal procesada $y_{ANS}(t)$ ($\eta=1$)					
$\theta_0 [^{\circ}]$	$r_0 [m]$	$SNR_8 < 5dB$	$5dB < SNR_8 < 10dB$	$SNR_8 > 10dB$	
0	2	0.45	0.59	0.67	
	4	0.42	0.61	0.69	
90	2	0.44	0.60	0.61	
	4	0.44	0.56	0.70	
promedio total		0.44	0.59	0.67	

(b)

E-RASTI. Señal procesada $y_{ANS}(t)$ ($\eta=10$)					
$\theta_0 [^{\circ}]$	$r_0 [m]$	$SNR_8 < 5dB$	$5dB < SNR_8 < 10dB$	$SNR_8 > 10dB$	
0	2	0.40	0.53	0.60	
	4	0.35	0.51	0.62	
90	2	0.39	0.52	0.55	
	4	0.36	0.49	0.60	
promedio total		0.38	0.51	0.59	

(c)

Tabla 35. Resultados de índice E-RASTI para el procesador prototipo SD-ANS-MW (véase la Figura 166). (a) Señal original $x_0(t)$ comparada con la señal sucia $y_8(t)$. (b) Señal procesada $y_{ANS}(t)$ ($\eta=1$) comparada con la señal sucia $y_8(t)$. (c) Señal procesada $y_{ANS}(t)$ ($\eta=10$) comparada con la señal sucia $y_8(t)$.

Hecha esta aclaración, puede interpretarse según la Figura 166 que el grado de mejora aumenta (el índice E-RASTI disminuye) a medida que SNR_8 disminuye y que la mejora es mayor para mayor supresión de ruido ($\eta = 10$). Esto es lógico, corresponde a la apreciación subjetiva y ya ha sido observado en las demás pruebas con GNMR, GSNR_A y GAI.

Sin embargo, lo que no se corresponde con los resultados observados de los estimadores anteriores es que la diferencia de E-RASTI entre los casos $\eta = 1$ y $\eta = 10$, aumenta con SNR_8 , es decir, parece que el grado de mejora es mayor cuando la señal de partida tiene buena calidad (SNR_8 alta). Esta discrepancia se basa en lo siguiente. Los estimadores fundamentados en la relación señal a ruido (GNMR, GSNR_A y GAI) evalúan la cantidad de ruido remanente en la señal procesada. Desde ese punto de vista hay menos diferencia (en dB) en el ruido remanente producido por tasas de sobresustracción distintas cuando la calidad de la señal de partida es alta. El valor $\eta = 10$ equivale a una multiplicación por 10 del producto $\delta(\omega) \cdot \eta$ en el filtro H_{ANS} (286), e induce un mayor aumento de la supresión de ruido cuando SNR_8 es baja. Efectivamente, si se observa con detenimiento la Figura 34(b), que da la respuesta del filtro H_{ANS} en función de SNR_{post} (que aquí es SNR_8), se verifica que la diferencia de ganancia entre la curva H_{ANS} con $\delta = 1$ ó $\delta = 10$ (equivalente aquí a usar $\eta = 10$) es mucho mayor para SNR_{post} pequeña que para SNR_{post} alta. El estimador E-RASTI está basado en la profundidad de modulación, es decir en la diferencia entre máximos y mínimos de $y^2(t)$. Esta profundidad de modulación (o índice de modulación m) aumentará bastante

cuando las tramas de ausencia de voz se atenúen fuertemente por el filtro H_{ANS} (286) y las tramas de presencia de voz se dejen pasar tal cual ($H_{ANS} \approx 1$). Parece ser que una mayor supresión ($\eta = 10$) aplicada a una señal de SNR_8 alta induce una mayor modulación, como se explica a continuación.

En la Figura 167 se ilustra el efecto del procesador sobre dos elementos de la misma base de datos que, aunque comparten la misma referencia $x_0(t)$ (es decir han sido generados utilizando la misma locución), se han contaminado inicialmente con diferentes cantidades de ruido y por tanto tienen diferentes SNR_8 asociadas. Puede comprobarse cómo, yendo de izquierda a derecha en la Figura 167, el incremento del factor de supresión η produce un mayor aumento de la modulación, es decir la diferencia entre máximos y mínimos aumenta, cuando SNR_8 es alta. En este caso, SNR_8 está cerca del recodo del filtro H_{ANS} mostrado en la Figura 34(b) (por ejemplo, cerca del punto de dicha gráfica correspondiente a SNR_{post} [dB] = 9dB para $\delta = 10$ y $\epsilon = 1$), de tal manera que, debido a que el análisis se hace trama a trama, las tramas con actividad de voz no se ven modificadas y las tramas sin actividad de voz se ven muy atenuadas. Cuando el factor de sobresupresión pasa de $\eta = 1$ a $\eta = 10$ las tramas de voz siguen atravesando al procesador sin apenas modificación (la SNR local es alta), pero las tramas de ruido son aminoradas de forma más intensa, ya que el filtro H_{ANS} tiene una atenuación mayor al pasar de $\eta = 1$ a $\eta = 10$ para SNR_{post} bajas. De esta manera se produce una mayor modulación y por tanto un mayor valor del índice E-RASTI.

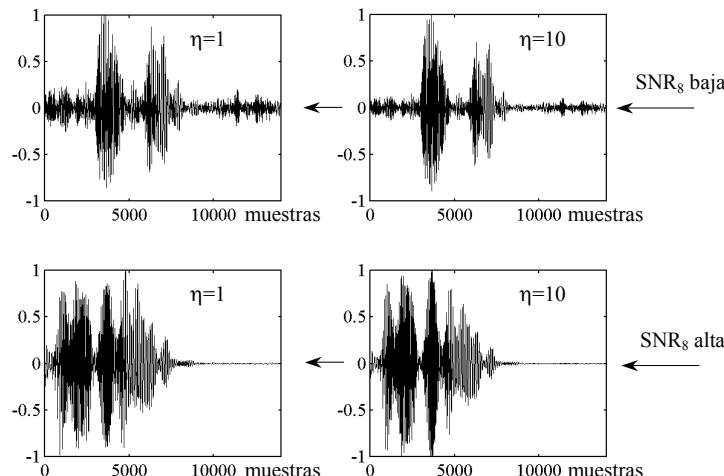


Figura 167. Comparación de cuatro fragmentos de señal procesada $y_{ANS}(t)$ pertenecientes a dos elementos de la base de datos. Arriba: señal de salida del procesador correspondiente a una entrada con baja SNR_8 , a la izquierda con $\eta=1$ y a la derecha con $\eta=10$. Abajo: ídem de lo anterior pero con una señal de entrada al procesador de alta SNR_8 . Cuando SNR_8 es baja (pareja superior) la diferencia de modulación es menor que cuando SNR_8 es alta (pareja inferior).

De lo explicado anteriormente se puede inferir que los supresores de ruido basados en postfiltrado que usan una transformada instantánea como la STFT, pueden engañar en cierta medida al evaluador E-RASTI. Este tipo de filtro tiende a sobrefiltrar en las ausencias de voz, y puede inducir un aumento de modulación que no está asociado necesariamente a un aumento de la calidad de la señal a la salida del procesador. De hecho, esto muchas veces es equivalente a un aumento de la distorsión de tipo musical a la salida, según se expuso en el punto 4.1.2.

Una vez hechas estas consideraciones, se puede finalizar el análisis de los resultados de la Tabla 35 sobre el índice E-RASTI. Parece que la ligera tendencia mostrada con los estimadores basados en SNR ,según la cual el apuntamiento *endfire* era ligeramente superior al apuntamiento *broadside*, no se muestra aquí. Tampoco se muestra ninguna relación clara del E-RASTI con la distancia.

A pesar de los defectos comentados, como ya se trató más específicamente en el apartado 7.2, los resultados del E-RASTI son equiparables a los obtenidos con otros parámetros objetivos más convencionales, especialmente la ganancia GNMR (compárese la Figura 163 con la 166), con la ventaja de que el índice propuesto no necesita estimar el ruido mediante la sustracción (223), sino que simplemente analiza las diferencias de modulación observadas entre la entrada y la salida del array microfónico, y por ello no se requiere una perfecta alineación temporal de los fragmentos de voz a comparar. Además puede considerarse como especialmente adecuado para medir derreverberación, como se demostró también en el apartado 7.2 de la Tesis.

Para concluir con el análisis global de los resultados expuestos en este apartado, se puede decir lo siguiente. En primer lugar, que el procesador SD-ANS-MW consigue unas resultados más que aceptables si se usa una apreciación subjetiva informal. Téngase en cuenta que el ruido añadido era de banda ancha y aleatorio, con lo que aumentan las dificultades en cuanto a su eliminación. También, que todos los estimadores objetivos utilizados (GNMR, GSNR_A, GAI y E-RASTI) muestran resultados parecidos a grosso modo, aunque son un poco diferentes analizados en detalle. Es decir todos ellos muestran la mejora (más que patente) de la señal procesada, y que esta mejora es mayor cuando la calidad de la misma es pequeña. Los estimadores basados en SNR no detectan la distorsión por ruido musical mientras que el E-RASTI detecta un aumento de modulación cuando se produce un sobrefiltrado de las tramas sin voz.

EPÍLOGO

10 CONCLUSIONES Y LÍNEAS FUTURAS DE TRABAJO

Los objetivos marcados al iniciar este trabajo de Tesis eran muy ambiciosos. Aparte de los más puramente científicos, en el sentido de realizar propuestas novedosas en el ámbito del procesado en array de señales acústicas, la pretensión de implementar un prototipo funcional que trabajase en tiempo real era una meta difícil situada al final de una ruta tortuosa. No obstante, las contribuciones científicas que ha aportado y que puede aportar en un futuro el prototipo implementado son de gran interés, puesto que el procesado multicanal de habla en particular, y más ampliamente de señales acústicas, es un tema de estudio con gran proyección desde el punto de vista de la ingeniería y de la investigación aplicada.

Lo cierto es que podemos considerar en este punto que los objetivos previstos inicialmente, básicamente se han cumplido. Dentro del grupo de investigación que acoge a este trabajo se ha cimentado una buena base sobre procesado acústico multicanal, y se han generado herramientas para continuar por la misma línea de investigación. Por supuesto que el trabajo realizado es mejorable, tanto desde el punto de vista de los algoritmos propuestos como el de la implementación del prototipo. Son muchas las alternativas que en el momento presente ofrece la familia científica para el procesado multicanal y muchos los avances tecnológicos por venir que ayudarán, sin duda, en este ámbito del tratamiento de señales acústicas. Aquí se ha dado una visión parcial del problema de la captación multicanal y mejora de habla, con soluciones muy enfocadas algunas veces a la implementación práctica de un prototipo prevista inicialmente. Como tal, el trabajo desarrollado en esta Tesis puede ser continuado, complementado y mejorado desde muchos otros puntos de vista.

A continuación se hace una breve síntesis de las conclusiones más importantes que se pueden ofrecer tras finalizar este trabajo de investigación y desarrollo, la mayoría de las cuales ya han sido reseñadas dentro de cada uno de los grandes bloques considerados anteriormente en esta Tesis.

10.1 CONCLUSIONES

10.1.1 Propuestas de mejora de señal de voz mediante postfiltrado multicanal

La mejora multicanal de señal de habla mediante postfiltrado está muy relacionada con el concepto de conformación de haz. Como se ha demostrado, un filtrado óptimo, es equivalente desde algunos puntos de vista a una conformación óptima.

Para hacer una conformación óptima es fundamental que el array propuesto tenga una buena selectividad espacial, especialmente en baja frecuencia, por debajo de 1kHz, ya que es aquí donde se concentran normalmente los problemas de ruido y reverberación asociados a la señal vocal. Después de cotejar algunas alternativas, finalmente, en el capítulo 9 de la Tesis, se ha propuesto trabajar con un array anidado en tres subbandas compuesto por 15

micrófonos, implementando una conformación superdirective en la banda de baja frecuencia y una conformación convencional en las otras dos bandas superiores. Sobre esta base, se han propuesto combinaciones novedosas de métodos de postfiltrado monocanal ya conocidos, adaptados para el trabajo multicanal y para la operación en tiempo real.

El método de Wiener modificado por coherencia (MW), que ya se propuso en otra Tesis dentro del grupo ATVS [González-Rodríguez 99-a], ha sido experimentado con más extensión y detalle, obteniéndose buenos resultados con evaluaciones tanto objetivas como subjetivas. Se ha tenido que realizar con cuidado el ajuste de los parámetros del procesador, en especial el de las constantes λ de actualización de las estimaciones recursivas, puesto que éstas influyen apreciablemente en la cantidad de perturbación eliminada y de la distorsión remanente en la señal mejorada.

El método de derreverberación basado en descomposición cepstral multicanal (MPAP) ha sido evaluado tanto de forma individual como colectivamente, junto con otros esquemas complementarios de postfiltrado, probando los efectos que produce en la señal procesada y los valores más adecuados de los parámetros de enventanado, para una evaluación usando tramas de duración finita. Se ha desecharido finalmente este procesador para la implementación del prototipo en tiempo real, por su complejidad computacional, que no es compensada suficientemente por la ligera mejora que ofrece en cuanto a reverberación, cuando se compara al procesador MPAP con los métodos de conformación de haz más frecuentes.

Finalmente, como contribución más original en este apartado, se ha propuesto un método de filtrado auditivo multicanal (ANS) que utiliza como base las propuestas originales de D. E. Tsoukalas *et al* [Tsoukalas 97] para supresión monocanal de ruido y las aplica al array anidado de 15 micrófonos. El método ANS multicanal se combina con el filtrado Wiener para originar el que aquí se ha llamado método ANS-MW. De esta manera el filtro de Wiener, anteriormente probado, produce una primera mejora en la señal de voz, suficiente para estimar adecuadamente los umbrales auditivos, que son necesarios para implementar el método ANS. Con unas evaluaciones objetivas cada vez más elaboradas, se ha demostrado la superioridad del supresor auditivo con respecto a procesador en array basado en filtrado de Wiener anteriormente propuesto. La razón es que la señal procesada a la salida del array es subjetivamente más natural y tiene menor distorsión de tipo ruido musical, lo que ha sido verificado también mediante medidas objetivas.

Una cuestión importante en el campo del postfiltrado de la señal de habla es si éste se hace en el dominio del tiempo mediante filtrado digital y reconstrucción perfecta o en el dominio de la frecuencia mediante la transformada STFT, unido a un proceso de enventanado. Aunque el procesado en el dominio de la frecuencia tiene una mayor complejidad computacional debido a que normalmente se usa con una elevada precisión en frecuencia, su especial adaptación a las propuestas planteadas en la Tesis ha hecho que éste sea el elegido como método de trabajo.

Otra cuestión relacionada con la forma de procesado ha sido la selección de los parámetros adecuados para el enventanado, si se desea una reconstrucción temporal óptima de la señal de salida. Este asunto está ligado de forma estrecha con la capacidad computacional del procesador usado para la implementación del prototipo. Se han probado diversas alternativas, demostrándose que los solapamientos del 50% y del 67% con ventana Hanning son óptimos cuando la alineación temporal de los canales del array se hace en el dominio del tiempo. Si se desea una mayor precisión temporal de la corrección de retardos, hay que utilizar una alineación en el dominio de la frecuencia. En este caso, las dos tasas de solapamiento anteriores son subóptimas, ya que producen cierta distorsión en la reconstrucción temporal, que puede ser especialmente audible en el caso del valor de 67%.

Otro asunto que ha consumido bastantes esfuerzos desde el punto de vista de la implementación práctica, ha sido la detección de actividad de voz (VAD). Aunque la detección VAD no se planteó inicialmente como tema científico a desarrollar con profundidad dentro de la Tesis, la idea de la implementación de un prototipo autónomo hizo necesaria la articulación de un método sencillo de detección que pudiese ser incorporado al procesador DSP, sin consumir excesivos recursos del sistema. Si bien en el transcurso de la implementación del prototipo fueron probados (mediante experimentos que no se exponen en esta Tesis) detectores VAD convencionales, basados en estimaciones de la energía y los cruces por cero de la señal vocal, finalmente se ha optado por incorporar un VAD recursivo por análisis de potencia, que ha dado muy buenos resultados en los experimentos finales, incluso con muestras de voz de muy mala calidad.

10.1.2 Evaluación objetiva de los resultados a la salida del procesador

Aunque las escuchas subjetivas de la señal procesada siempre han de estar presentes, es necesario considerar una evaluación objetiva de la misma para analizar en detalle los resultados y los efectos de los parámetros del sistema y comparar pormenorizadamente los diferentes procesadores probados.

Se han usado tres tipos diferentes de estimadores objetivos de la mejora producida sobre la señal de habla. Éstos son, la ganancia en valor de SNR, la distancia en parámetros LP y la mejora de inteligibilidad basada en la evaluación del índice RASTI. De entre los dos primeros grupos de evaluadores, el método que más paralelismo tiene con la apreciación subjetiva de la mejora de habla es la ganancia GNMR. Los métodos que utilizan distancias LAR y cepstrales se han mostrado poco útiles en esta tarea, ya que parecen medir más la distorsión del procesador que la reducción global de ruido y la mejora de la calidad en el habla de salida. Tampoco el método basado en la ganancia en índice de articulación GAI ha evaluado con suficiente precisión los resultados obtenidos.

Por otra parte, en esta Tesis se propone el método E-RASTI para considerar de forma más clara la reverberación presente en el habla, especialmente cuando no se dispone de una muestra de voz limpia perfectamente alineada en tiempo con la señal a evaluar. Este método es una derivación del método RASTI convencional, utilizado en la evaluación acústica de recintos sonoros y de sonorizaciones electroacústicas. En la propuesta aquí presentada se aplica directamente sobre tramas de voz procesada, de tal manera que se mide la ganancia de modulación de una señal de entrada ya tratada por el array, sobre una señal de salida sucia o antes de procesar. La gran ventaja del método, como se apuntaba antes, es que las señales a comparar no tienen que estar perfectamente alineadas en tiempo, y que no se necesita la señal original, si no es como test de verificación adicional.

En general, los resultados del E-RASTI han sido buenos, obteniéndose una mejora de habla bastante similar a la proporcionada por la ganancia GNMR (que también evalúa, a su modo, la derreverberación) y acorde a la impresión subjetiva ante la escucha de la señal de salida del procesador. El principal defecto mostrado por el índice E-RASTI es que puede interpretar la sobresustracción de ruido como mejora de señal, debido a que la distorsión ocasionada se traduce en una modulación artificial de la señal de habla. En ese sentido, el E-RASTI debe ser utilizado con precaución cuando se analizan procesadores basados en el postfiltrado por tramas temporales de señal, como los propuestos en esta Tesis, en los que normalmente la distorsión de salida se traduce en sobremodulación de la señal tratada. El índice E-RASTI sin embargo, es más adecuado para cuantificar la derreverberación

ocasionada con otro tipo de métodos de respuesta más estacionaria, como el procesador MPAP, que no conlleva el problema de la sobresustracción.

10.1.3 Localización de fuente

La localización de la fuente sonora mediante arrays microfónicos es una materia de gran interés en la actualidad y que por sí sola constituye un tema de estudio suficientemente amplio. Un array con 15 elementos, como la propuesta ofrecida aquí, tiene bastantes posibilidades a la hora de detectar eficientemente el número y la posición de las diferentes fuentes de voz presentes en el escenario de operación. Aunque no se muestra en la Tesis, se han hecho experimentos de localización de fuente mediante los métodos convencionales mostrados en los puntos 3.1 y 3.2, basados respectivamente en la maximización de potencia de salida y en la estimación de retardos mediante métodos de correlación cruzada entre los elementos del array. Los resultados obtenidos no se pueden considerar como buenos en la mayoría de los casos, debido sobre todo a las características no estacionarias de la señal vocal y a la poca robustez de los métodos anteriores ante las malas condiciones acústicas planteadas en esta Tesis. Por todo ello se ha desecharido finalmente incorporar ningún detector automático de fuente en el procesador final en tiempo real. Aunque en el capítulo 3 se ha hecho una breve revisión del estado del arte sobre la localización de fuente basada en métodos de subespacios, su complejidad computacional ha hecho desestimar inicialmente estos métodos para la aplicación presentada en la Tesis, aunque se prevé estudiar su viabilidad en desarrollos futuros.

10.1.4 Directividad

Una vez implementado el prototipo final, se realizaron múltiples pruebas de conformación de haz utilizando el método convencional de retardo y suma en las tres subbandas del array, con excelentes resultados en cuanto a cumplimiento de las curvas teóricas de directividad, incluso en la subbanda B_1 de baja frecuencia.

Lo primero en desechar fue el apuntamiento en la dirección *broadside* ($\theta_0 = 0^\circ$), ya que la simetría del array en $\theta = 0^\circ$ expande peligrosamente el lóbulo de captación principal alrededor de dicho eje, de tal manera que se capta mucha reverberación, procedente del suelo y el techo del recinto de pruebas. Se optó finalmente por la implementación de un array superdirective de tipo Griffiths-Jim no-adaptativo, para intentar aumentar la selectividad en la banda B_1 . Con esto, se consigue mayor selectividad en baja frecuencia con respecto al conformador convencional, pero menor paralelismo con las curvas de directividad teóricas por las razones de desalineación intermicrofónica en módulo y fase explicadas a lo largo de la Tesis. En este último sentido, tienen gran importancia, como se ha demostrado, las pruebas de calibración del array. Una precisa calibración en una cámara anecoica consigue una mayor directividad en la banda de baja frecuencia, lo que se traduce finalmente en una mayor eliminación de ruido y reverberación, que ha sido apreciada tanto desde el punto de vista subjetivo como en evaluaciones objetivas. Los últimos resultados obtenidos por el conformador son excelentes en las bandas B_2 y B_3 (el conformador convencional de retardo y suma es mucho menos sensible a despareamientos intercanal en el array) y más que aceptables en la banda B_1 , con el conformador superdirective, que consigue aumentar la directividad en la problemática banda de baja frecuencia.

También se han probado diferentes valores del coeficiente superdirective de restricción μ , optándose finalmente por el valor correspondiente a -15dB.

10.1.5 Implementación en DSP de un prototipo de array microfónico

La realización de un prototipo real de un array microfónico de múltiples canales tiene bastantes complicaciones prácticas. Por una parte hay que ser cuidadoso en la elección y emplazamiento del tipo de micrófonos del array. Aunque inicialmente se barajaron diferentes posibilidades finalmente se optó por usar micrófonos omnidireccionales, que aunque son menos selectivos espacialmente, su respuesta electroacústica es mucho más controlable, y la versatilidad en cuanto a apuntamiento del array final es mucho mayor. No hay que perder de vista que las pruebas preliminares del capítulo 7 se hicieron con un array (base de datos real CMU) compuesto por micrófonos directivos en configuración *broadside*, que pierde sus buenas características directivas cuando el apuntamiento empieza a ser lateral.

Para la realización del prototipo implementado se han empleado 16 micrófonos omnidireccionales con números de serie consecutivos lo que favorece el que los micrófonos tengan respuestas parecidas. Como se ha explicado, desde un punto de vista electroacústico los micrófonos de presión (omnidireccionales) presentan menor posibilidad de intervariabilidad que los micrófonos de gradiente (familia cardioide). Se ha medido en una cámara anecoica la respuesta de cada micrófono en su emplazamiento final en el array obteniéndose desviaciones de respuesta poco significativas para la aplicación propuesta. Asimismo se buscó el emplazamiento óptimo de los micrófonos en el array, en una barra cilíndrica hueca que distorsionase lo menos posible el campo acústico captado. Aunque esta posible distorsión es inapreciable en baja frecuencia, sí se nota en media y alta frecuencia (como se comprobó), pero esta posibilidad se ha evitado.

Otro asunto muy relevante, atendiendo a la implementación práctica, ha sido todo lo relativo al diseño del procesador en tiempo real sobre una tarjeta con un DSP. La perspectiva de diseño algorítmico cambia mucho si éste se tiene que implementar de forma práctica. Las propuestas de procesadores relativamente sencillos, que no ofrecen dificultades cuando operan en una simulación *software*, normalmente necesitan ser rediseñadas para la operación en tiempo real. Se ha encontrado especial dificultad en la optimización algorítmica del cálculo FFT de 15 canales simultáneamente, con un solapamiento intertrama relativamente alto que exigía una gran velocidad de procesado. Con la tecnología disponible para esta Tesis, el resultado ha sido que se tiene que sacrificar un poco de relación señal a ruido (por la limitación en la precisión numérica requerida, que ha sido finalmente de 13bits) si se quiere operar en el dominio de la frecuencia con suficiente rapidez. No obstante, todas las muestras de habla procesadas con el prototipo de array han ofrecido un nivel de calidad similar al de las simulaciones *software* realizadas con un procesador clónico del de tiempo real, con una relación señal a ruido teórica mucho mayor, inherente a una precisión numérica superior.

A la vista de lo expuesto anteriormente, las principales aportaciones conseguidas en este trabajo de Tesis consisten en el perfeccionamiento de las técnicas convencionales de mejora multicanal de voz, y de su evaluación objetiva, y en la implementación de un prototipo en tiempo real, para ser empleado con voz en muy malas condiciones acústicas, y que pueda ser usado en futuras investigaciones con otros procesadores diferentes a los ya propuestos.

En la Tabla 36 se resumen las principales aportaciones ofrecidas por los trabajos desarrollados en esta Tesis.

principales aportaciones	principales facultades / logros conseguidos
Procesador ANS-MW	<ul style="list-style-type: none"> - Implementación multicanal de los métodos auditivos de supresión de ruido utilizando una limpieza previa de la señal de voz mediante filtrado Wiener - Consigue unos resultados de limpieza con menor percepción de distorsión que los métodos convencionales de sustracción espectral
Método E-RASTI de evaluación objetiva de habla	<ul style="list-style-type: none"> - Método de evaluación objetiva de la calidad de habla especialmente adecuado para detectar derreverberación - No necesita una alineación temporal perfecta de las señales a comparar - Se han contrastado los resultados de valoración en diferentes situaciones de ruido y reverberación, comparándolos con otros métodos convencionales de evaluación
Prototipo de array en tiempo real SD-ANS-MW	<ul style="list-style-type: none"> - Procesado simultáneo de 15 canales en tiempo real y en el dominio de la frecuencia, con una frecuencia de muestreo de 16kHz - Posibilidad de calibración automática de los micrófonos del array con la que se consiguen unas curvas de directividad muy parecidas a las predichas teóricamente - Posibilidad de apuntamiento manual o mediante cámara Web a la fuente de voz - Detección automática de la actividad de habla - Comunicación en tiempo real con el <i>HOST PC</i> lo que permite obtener bases de datos multicanal

Tabla 36. Principales aportaciones conseguidas en el trabajo de Tesis junto a los principales logros obtenidos y facultades de las propuestas ofrecidas.

10.2 LÍNEAS FUTURAS DE TRABAJO

Son varios los frentes abiertos con el trabajo de investigación y desarrollo efectuado, que pueden ser completados y continuados en un futuro.

En primer lugar, a muy corto plazo, se plantean posibles mejoras puntuales sobre los elementos *software* ya implementados en el prototipo.

Aunque desde el principio se planteó la utilización de un procesador en el dominio de la frecuencia mediante la transformada FFT (lo que como se ha visto es muy costoso computacionalmente), es posible reducir el tiempo de procesado mediante un filtrado temporal de reconstrucción perfecta, eso sí, utilizando un número de bandas de análisis apropiado para los métodos de mejora de habla propuestos, pero suficientemente pequeño para reducir el tiempo de operación. Este ahorro computacional permitirá quizás la incorporación de nuevas mejoras en el procesador, por ejemplo las referidas a la localización automática de fuente.

Otra de las cuestiones puntuales, que ha quedado pendiente, es la de la alineación temporal de la señal multicanal. Como se ha referido en el capítulo 7, las pruebas preliminares se hicieron con una alineación temporal en el dominio del tiempo, atrasando o adelantando las señales de salida de cada micrófono tanto como fuese necesario. Para aumentar la precisión temporal, se implementó una interpolación de x 4 muestras. Como esto no es factible en el procesador en tiempo real, se optó por realizar en éste la alineación en el dominio de la frecuencia. Aunque este método introduce mayor distorsión que el anterior (y es más costoso computacionalmente ya que necesita las FFT's), es mucho más preciso desde el punto de vista

de la alineación temporal, y produce por tanto curvas de directividad como las predicciones teóricas. Consecuentemente una tarea próxima será la de probar en el procesador en tiempo real la alineación temporal de canales en el dominio del tiempo, con una precisión de $1/f_s$, para estudiar qué tal se comporta en cuanto a directividad y a resultados de mejora de voz.

Se plantea también, como tarea a corto plazo, la necesidad de probar el array implementado en condiciones más reales y menos controladas que las correspondientes a las pruebas realizadas en el capítulo 9. Es decir, en ambientes reales con alto ruido y/o reverberación.

Otra de las labores previstas de forma inmediata es la de probar el prototipo en aplicaciones de identificación de locutor. La aplicación de los arrays microfónicos para este uso es un campo de investigación ya abierto, y en este caso puede complementar las líneas de investigación sobre identificación biométrica ya trazadas en el ATVS.

A medio plazo será necesario abrir una línea de trabajo relativa a la localización/identificación de fuentes de habla mediante arrays microfónicos. Como se ha referido a lo largo de la Tesis, la localización de fuente es un asunto especialmente complejo y no se ha tratado con profundidad en las propuestas realizadas. Los siguientes esfuerzos en este sentido se enfocarán a la implementación de un array que sea capaz de hacer un seguimiento automático de la fuente principal y la discriminación entre las demás fuentes consideradas como de ruido. En este sentido, tienen especial interés para el autor, los métodos basados en subespacios combinados con ML.

También quedan pendientes muchas tareas relativas a la implementación práctica, sobre todo referidas a la portabilidad del sistema. En este sentido, el principal problema del prototipo implementado lo constituye la interfaz analógica-analógica, integrada por el conjunto cableado + preamplificador microfónico. Se prevé construir en un futuro un prototipo inalámbrico, de tal manera que el array sea relativamente portable, y que se comunique con una estación fija receptora y de procesado.

La elección de la plataforma *hardware* de procesado sobre la que deben hacerse las futuras implementaciones de otros prototipos de arrays microfónicos es otra cuestión que plantea dudas. En el transcurso del trabajo desarrollado en la Tesis, la tecnología de procesadores digitales de señal ha progresado de forma importante (existen ya chips DSP de hasta 4GIPS). Sin embargo, si se desean utilizar plataformas suficientemente universales y compatibles, teniendo en cuenta el gran avance que están teniendo los procesadores de los ordenadores domésticos, hay que plantearse la posibilidad de hacer la futura implementación de un prototipo de array usando el procesador de un PC, evitando la idea de un DSP, que suele ser relativamente costoso y su evolución es más lenta que los microprocesadores de uso general.

BIBLIOGRAFÍA

[Acousticmagic] Acousticmagic, available in <http://www.acousticmagic.com>

[Abe 84] Abe, Y., Miyaji, N. and Iwahara M., “Practical application and digital control of the microphone array”, in *Preprints of the 76th AES Convention*, no.2116 (E-1), New York (USA), October 1984

[Affes 96] Affes, S., Gazor, S. and Grenier, Y., “An algorithm for multisource beamforming and multitarget tracking”, in *IEEE Trans. on Signal Processing*, vol.44, no.6, pp.1512-1522, June 1996

[Affes 97] Affes, S. and Grenier, Y., “A signal subspace tracking algorithm for microphone array processing of speech”, in *IEEE Trans. on Speech and Audio Processing*, vol.5, no.5, pp.425-437, September 1997

[Ahnoff 00] Ahnoff, Mattias, “Extended precision complex radix-2 FFT/IFFT implemented on TMS320C62x”, Application Report, SPRA696, Texas Instruments, September 2000

[Akbari 95] Akbari, A., Le-Bouquin, R. and Faucon, G., “Optimizing speech enhancement by exploiting masking properties of the human ear”, in *Proc. ICASSP 1995*, pp.800-803, 1995

[Allen 77-a] Allen, J.B., Berkley and Blauert, J. “Multimicrophone signal processing technique to remove room reverberation from speech signals”, in *J. Acoust. Soc. Am.*, no.62, 912-915, 1977

[Allen 77-b] Allen, J.B. and Berkley D.A., “Image method for efficiently simulating small-room acoustics”, in *J. Acoust. Soc. Am.*, vol.65, no.4, pp.943-950, 1977

[Alonso 03] Alonso Valdesueiro, Javier, *Software de comunicaciones y pruebas de un sistema de procesado de voz en array basado en el DSP TMS320C6701*, Proyecto Fin de Carrera, E.U.I.T. Telecomunicación, Universidad Politécnica de Madrid, Madrid, Julio 2003

[Andrea] Andrea Electronics Corporation, available in <http://www.andreaelectronics.com>

[Angus 93] Angus J.A.S., Kershaw, S. and Lewis, A., “An adaptive beam-steering microphone array implemented on the Motorola DSP56000 digital signal processor”, in *Preprints of the 95th AES Convention*, no. 3761 (A3-AM-1), New York (USA), October 1993

[Alvarado 90] Alvarado, V.M. and Silverman, H.F. “Experimental results showing the effects of optimal spacing between elements of a linear microphone array”, in *Proc. ICASSP 1990*, pp.837-840, Albuquerque (USA), 1990

- [Araki 01] Araki, S., Makino, S., Nishikawa, T. and Saruwatari, H., “Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech”, in *Proc. ICASSP 2001*, pp.2737-2740, Salt Lake City (USA), 2001
- [Asano 00] Asano, F., Satoru, H., Yamada, T. and Nakamura, S., “Speech enhancement based on the subspace method”, in *IEEE Trans. on Speech and Audio Processing*, vol.8, no.5, pp.497-507, 2000
- [Barabell 83] Barabell, A. J., “Improving the resolution performance of eigenstructure based direction-finding algorithms”, in *Proc. ICASSP 1983*, pp. 336-339, Boston, April 1983
- [Bartlett 90] Bartlett, B. and Billingsley, M. “An improved stereo microphone array using boundary technology: theoretical aspects”, in *J. Audio Eng. Soc.*, vol.38, no.7/8, pp.543-552, July/August 1990
- [Bédard 94] Bédard, S., Champagne, B. and Stéphanne, A., “Effects of room reverberation on time-delay estimation performance”, in *Proc. ICASSP 1994*, pp.261-264, Adelaida (Australia), Abril 1994
- [Beltrán 99] Beltrán, F.A., Beltrán, J.R., Holzem, N. and Gogu A., “Matlab implementation of reverberation algorithms”, in *2nd COST-G6 Workshop on Digital Audio Effects, (DAFx99)*, Trondheim (Norway), 1999
- [Bendjama 00] Bendjama, A. and Bourennane, S., “Blind focusing wide-band array processing”, in *Proc. EUSIPCO 2000*, pp.781-784, Tampere (Finland), 2000
- [Ben-Jebara 00] Ben-Jebara, S., Venaza-Benyahia, A. and Ben-Khelifa A., “Reduction of musical noise generated by spectral subtraction by combining wavelet packed transform and Wiener filtering”, in *Proc. EUSIPCO 2000*, pp.749-752, Tampere (Finland), 2000
- [Berouti 79] Berouti, M., Schwartz, B. and Makhoul, J., “Enhancement of speech corrupted by acoustic noise”, in *Proc. ICASSP 1979*, pp.215-228, 1979
- [Bienvenu 80] Bienvenu, G., and Kopp, L., “Adaptivity to background noise spatial coherence for high resolution passive methods”, in *Proc. ICASSP 1980*, pp.307-310, 1980
- [Billingsley 90] Billingsley, M. and Bartlett, B. “Practical field recording applications: an improved stereo microphone array using boundary technology”, in *J. Audio Eng. Soc.*, vol.38, no.7/8, pp.553-565, July/August 1990
- [Bohme 86] Bohme, J.F., “Estimation of spectral parameters of correlated signals in wavefields”, *Signal Processing*, no.10, pp.329-337, 1986
- [Boll 79] Boll, S.F., “Suppression of acoustic noise in speech using spectral subtraction”, in *IEEE Trans. on Speech and Audio Processing*, vol.ASSP-27, no.2, pp.113-120, April 1979
- [Brandstein 95] Brandstein, M., Adcock, J. and Silverman, H., “A practical time-delay estimator for localizing speech sources with a microphone array”, in *Computer, Speech, and Language*, no.9, pp.153-169, 1995
- [Brandstein 97-a] Brandstein, M. and Silverman, H.F., “A robust method for speech signal time-delay estimation in reverberant rooms”, in *Proc. ICASSP 1997*, pp.375-378, Munich (Germany), 1997

- [Brandstein 97-b] Brandstein, M. and Silverman, H.F., "A practical methodology for speech source localization with microphone arrays", in *Computer, Speech, and Language*, vol.11, pp.91-126, November 1997
- [Brandstein 98] Brandstein, M., "On the use of splicit speech modeling in microphone array applications", in *Proc. ICASSP 1998*, pp.3613-16, Seattle (USA), 1998
- [Brandstein 99] Brandstein, M., "An event-based method for microphone array speech enhancement", in *Proc. ICASSP 1999*, pp.953-56, Phoenix (USA), March 1999
- [Brandstein 00] Brandstein, M.S. and Ward, D.B., "Cell-Based beamforming (CE-BABE) for speech acquisition with microphone arrays", in *IEEE Trans. on Speech and Audio Processing*, vol. 8, no.6, pp.738-743, November 2000
- [Brandstein 01] Brandstein, M. and Ward, D., *Microphone arrays*, Springer Verlag, Berlin, 2001
- [Bresler 86] Bresler, Y. and Macovski, A., "Exact maximum likelihood parameter estimation of superimposed exponential signals in noise" in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.34, pp. 1081-1089, October 1986
- [Buckley 86] Buckley, K.M. "Broad-band beamforming and the generalized sidelobe canceller", in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.34, no.11, pp.1322-1323, October 1986
- [Buckley 90] Buckley, K.M. and Xu, Xiao-Liang, "Spatial-spectrum estimation in a location sector", in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.38, no.11, pp.1842-1852, November 1990
- [Burg 64] Burg, J.P., Three-dimensional filtering with an array of seismometers", in *Geophysics*, vol.39, no.5, pp.693-713, October 1964
- [BWS 99-a] Blue Wave Systems, "PCI/C6600 Application board. Technical reference manual", September 1999
- [BWS 99-b] Blue Wave Systems, *PMC/16IO2, PCI Mezzanine card. User manual*, March 1999
- [BWS 00-a] Blue Wave Systems, *PCI/C6600 DSP Utility library. Function reference manual*, September 2000
- [BWS 00-b] Blue Wave Systems, *Host and MPC860 C GenrHL. User guide for ComStruct boards*, September 2000
- [Cadzow 90] Cadzow, J.A., "Multiple source location-the signal subspace approach", in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.38, no.7, pp.1110-1125, July 1990
- [Cao 95] Cao, Y., Sridharan, S. and Moody M. "An auto-tracking auto-beamforming microphone array for sound recording" in *Preprints of the 5th AES Australian Regional Convention*, no. 4037, Sydney (Australia), April 1995
- [Capon 69] Capon, J., "High-resolution frequency-wave number spectrum analysis," in *Proc. IEEE*, vol.57, no.8, pp.2408-1418, August 1969
- [Carter 73] Carter, G.C., Knapp, C.H. and Nuttall, A.H, "Estimation of the magnitude-squared coherence function via overlapped fast Fourier transform processing", in *IEEE Trans. on Audio and Electroacoustics*, vol.AU-21, no.4, pp.337-344, 1973

- [Carter 87] Carter, G.C. "Coherence and time delay estimation", in *Proc. IEEE*, vol.75, no.2, pp. 236-255, February 1987
- [Chamorro 99] Chamorro-Calvo, María-Teresa, *Implementación y evaluación de sistemas de procesado en array para señal de voz*, Proyecto Fin de Carrera, E.U.I.T. Telecomunicación, Universidad Politécnica de Madrid, Madrid, Junio de 1999
- [Champagne 96] Champagne, B., Bedard, S. and Stephenne, A., "Performance of time-delay estimation in the presence of room reverberation", in *IEEE Trans. on Speech and Audio Processing*, vol.4, no.2, pp.148-152, March 1996
- [Chen 98] Chen (ed.), "Highlights of statistical signal and array processing", in *IEEE Signal Processing Magazine*, pp.21-64, September 1998
- [Cheng 00] Cheng, Yao-Ting, "Autoscaling radix-4 FFT for TMS320C6000", Application Report, SPRA654, Texas Instruments, March 2000
- [Claeson 92] Claeson, I. and Nordholm, S., "A spatial filtering approach to robust adaptive beamforming", in *IEEE Trans. on Antennas Propagat.*, pp.1093-1096, September 1992
- [CMU] Carnegie Mellon University multichannel database, available in <http://fife.speech.cs.cmu.edu/databases/micarray>
- [Cole 97] Cole, D., Moody, M. and Sridharan S., "Position-independent enhancement of reverberant speech", in *J. Audio Eng. Soc.*, vol.45, no.3, March 1997
- [Cook 55] Cook, R.K., Waterhouse, R.V., Berendt, R.D., Edelman, S. and Thompson Jr, M.C., "Measurement of correlation coefficients in reverberant sound fields", in *J. Acoust. Soc. Am.*, vol.27, pp.1072-1077, 1955
- [Cox 87] Cox, H., Zeskind, R.M. and Owen, M.M., "Robust Adaptive Beamforming", in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.35, pp. 1365-1375, October 1987
- [Crochiere 83] Crochiere, R.E., and Rabiner, L.R., *Multirate digital signal processing*, Englewood Cliffs, Prentice Hall, NJ, 1983
- [Cron 62] Cron, B.F. and Sherman, C.H., "Spatial-correlation functions for various noise models", in *J. Acoust. Soc. Am.*, vol.34, pp.1732-1736, 1962
- [Deller 01] Deller, John R., Hansen, John H.L. and Proakis, John G., *Discrete-Time processing of speech signals*, Wiley-IEEE Press, March 2001
- [Denisowski 01] Denisowski, Paul, "How does it sound?", in *IEEE Spectrum*, pp.60-64, February 2001
- [Di Claudio 00] Di Claudio, E., Parisi, R. and Orlandi, G., "Multi-source localization in reverberant environments", in *Proc. EUSIPCO 2000*, pp.1429-1432, Tampere (Finland), 2000
- [Doles 88] Doles III, J.H., and Benedict F.D., "Broad-band array design using the asymptotic theory of unequally spaced arrays", in *IEEE Trans. on Antennas Propagat.*, vol.36, no.1, pp.27-33, January 1988
- [Drews 95] Drews, M. "Time delay estimation for microphone array speech enhancement systems", in *Proc. EUROSPEECH 1995*, vol. 3, pp.2013-2016, Madrid 1995

- [Durbin 60] Durbin, J., "The fitting of time series models", in *Rev. Int. Stat. Inst.*, vol.28, pp.233-244, 1960
- [Elko 00] Elko, G.W., "Superdirective microphone arrays", in *Acoustic signal processing for telecommunications*, Gay, S.L. and Benesty, J., eds., Kluwer Academic Press, 2000
- [Ephraim 84] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol. ASSP-32, December 1984
- [Ephraim 85] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol. ASSP-33, no.2, pp.443-445, April 1985
- [Ephraim 95] Ephraim, Y. and Van Trees, H.L., "A signal subspace approach for speech enhancement", in *IEEE Trans. on Speech and Audio Processing*, vol.3, no.4, pp.251-266, July 1995
- [Etter 94] Etter, W. and Moschytz, G.S., "Noise reduction by noise-adaptive spectral magnitude expansion", in *J. Audio Eng. Soc.*, vol.42, May 1994
- [Farrier 90] Farrier, D.R. and Prosper, L.R., "A signal subspace beamformer", in *Proc. ICASSP 1990*, vol.5, pp.2815-2818, 1990
- [Faucon 89] Faucon, G., Tazi Mezalek, S. and Le Bouquin R., "Study and comparison of three structures for enhancement of noisy speech", in *Proc. ICASSP 1989*, vol.1, pp.385-3888, New York, 1989
- [Fernández 99] Fernández, J., Lleida, E. and Masgrau, E. "Microphone array design for robust speech acquisition and recognition", in *Proc. EUROSPEECH 1999*, vol. 5, pp.2363-236, Budapest (Hungary), 1999
- [Fischer 96] Fischer, S. and Simmer, K.U., "Beamforming microphone arrays for speech acquisition in noisy environments", in *Speech Communication (Special Issue on Acoustic Echo and Noise Control)*, vol.20, no.3-4, pp. 215-227, December 1996
- [Fischer 97] Fischer, S. and Kammeyer, K.D., "Broadband beamforming with adaptive postfiltering for speech acquisition in noisy environments", in *Proc. ICASSP 1997*, vol.1, pp. 359-362, Munich (Germany), April 1997
- [Flanagan 85-a] Flanagan, J.L., "Beamwidth and useable bandwidth of delay-steered microphone arrays", in *AT&T Technical Journal*, vol.64, no.4, pp.983-995, April 1985
- [Flanagan 85-b] Flanagan, J. L., "Use of acoustic filtering to control the beamwidth of steered microphone arrays", in *J. Acoust. Soc. Am.*, vol.78, no.2, pp.423-428, August 1985
- [Flanagan 85-c] Flanagan, J.L., Johnston, J.D., Zahn, R. and Elko G.W., "Computer-steered microphone arrays for sound transduction in large room", in *J. Acoust. Soc. Am.*, vol.78, pp.1508-1518, 1985
- [Flanagan 91] Flanagan, J.L. Berkley, D.A., Elko, G. W., West, J.E. and Sondhi, M.M., "Autodirective microphone systems", in *Acustica*, vol.73, no.1, pp.58-91, 1991
- [Flanagan 93] Flanagan, J.L., Surendran, A.C., and Jan, E.E., "Spatially selective sound capture for speech and audio processing", in *Speech Communication*, no.13, pp.207-222, 1993

- [Flanagan 96] Flanagan, L. and Jan, E.E., "Sound capture form spatial volumes: matched-filter processing of microphone arrays having randomly-distributed sensors", in *Proc. ICASSP 1996*, pp.917-920, Atlanta (USA), 1996
- [Furui 91] Furui, S. and Sondhi, M. M. (ed.), "Quality assessment of codec speech", Chapter 12, in *Advances in speech signal processing*, pp.357-385, Marcel Dekker, New York, 1991
- [Gabrea 00] Gabrea, M. and O'Shaughnessy, D. "Speech signal recovery in white noise using an adaptive kalman filter", in *Proc. EUSIPCO 2000*, pp.159-162, Tampere (Finland), 2000
- [Gade 87-a] Gade, S. and Herlufsen, H., "Use of weighting functions in FFT/DFT analysis (Part I)" in *Technical Review*, Brüel & Kjaer, no.3, 1987
- [Gade 87-b] Gade, S. and Herlufsen, H., "Use of weighting functions in FFT/DFT analysis (Part II)" in *Technical Review*, Brüel & Kjaer, no.4, 1987
- [Gay 00] Gay, S.L. and Benesty, J. (ed.), "Microphone Arrays", Chapter IV, in *Acoustic signal processing for telecommunication*, pp.181-282, Kluwer Academic Publishers, Massachusetts, 2000
- [Gilbert 55] Gilbert, E.N. and Morgan, S.P., "Optimum design of directive antenna arrays subject to random variations", in *Bell Syst. Tech. J.*, pp.637-663, May 1955
- [Giuliani 95] Giuliani, D., Matassoni, M., Omologo, M. and Svaizer, P., "Robust continuous speech recognition using a microphone array", in *Proc. EUROSPEECH 1995*, pp. 2021-2024, Madrid (Spain), 1995
- [Giuliani 96] Giuliani, D., Omologo, M. and Svaizer, P., "Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation", in *Proc. ICSLP 1996*, pp.1329-1332, Philadelphia (USA), 1996
- [Gómez 00] Gómez, P., Álvarez, A., Martínez, R., Nieto, V. and Rodellar, V., "Speech enhancement through binaural negative filtering", in *Proc. EUSIPCO 2000*, pp.187-190, Tampere (Finland), 2000
- [González-Rodríguez 97] González-Rodríguez, J. et al., "Robust speaker recognition through acoustic array processing and spectral normalization", in *Proc. ICASSP 1997*, pp.1103-1106, Munich (Germany), 1997
- [González-Rodríguez 99-a] González-Rodríguez, J., *Influencia y compensación del entorno acústico en sistemas de reconocimiento automático de locutores*, Tesis doctoral, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, Madrid, 1999
- [González-Rodríguez 99-b] González-Rodríguez, J., Sánchez-Bote, J.L. and Ortega-García, J., "Mejora de voz con arrays microfónicos mediante descomposición en componentes de fase mínima y paso todo", in *Libro de actas de TecniAcustica 1999*, Ávila (Spain), 1999
- [González-Rodríguez 00] González-Rodríguez J., Sánchez-Bote J.L. and Ortega-García, J., "Speech dereverberation and noise reduction with a combined microphone array approach", in *Proc. ICASSP 2000*, pp.1037-1040, Istambul (Turkey), 2000
- [Goodman 68] Goodman, J. W., *Introduction to Fourier optics*, McGraw-Hill, San Francisco, 1968

- [Grenier 97-a] Grenier, Y. and Affes, S., "Microphone array Response to Speaker Movements" (invited paper), in *Proc. ICASSP 1997*, pp. 247-250, Munich (Germany), April 1997
- [Grenier 97-b] Grenier, Y., "Theoretical and practical aspects of microphone array design", in *Proc. IWAENC 1997*, London (UK), 1997
- [Griebel 01] Griebel, S.M. and Brandstein, M.S., "Microphone array speech dereverberation using coarse channel modeling", in *Proc. ICASSP 2001*, pp.201-204, Salt Lake City (USA), 2001
- [Griffiths 82] Griffiths, Lloyd J. and Jim, Charles W., "An alternative approach to linearly constrained adaptive beamforming", in *IEEE Trans. on Antennas Propagat.*, vol.30, pp.27-34, January 1982
- [Guérin 00] Guérin, A., "A two-sensor voice activity detection and speech enhancement based on coherence with additional enhancement of low frequencies using pitch information", in *Proc. EUSIPCO 2000*, pp.179-182, Tampere (Finland), 2000
- [Gustafsson 98] Gustafsson, S., Jax, P., and Vary, P., "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics", in *Proc. ICASSP 1998*, pp.397-400, New York (USA), 1998
- [Haas 81] Haas, W.H. and Lindquist, C.S., "A synthesis of frequency domain filters for time delay estimation", in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.ASSP-29, no.3, pp.540-548, June 1981
- [Handa 01] Handa, M., Nagai, T. and Kurematsu, A., "Frequency domain multichannel speech separation and its applications", in *Proc. ICASSP 2001*, pp.163-166, Salt Lake City (USA), 2001
- [Harris 78] Harris, F.J., "On the use of windows for harmonic analysis with the discrete Fourier transform", in *Proc. IEEE*, vol.66, pp.51-83, 1978
- [Haykin 75] Haykin, S. and Kesler, J., "Relation between the radiation pattern of an array and the two-dimensional discrete Fourier transform", in *IEEE Trans. on Antennas Propagat.*, vol.AP-23, no.3, pp.419-420, May 1975
- [Haykin 85] Haykin, S. (ed.), *Array signal processing*, Englewood Cliffs-Prentice-Hall, New Jersey, 1985
- [Herre 92] Herre, J., Eberlein, E., Scott, H. and Brandenburg, K., "Advanced audio measurement system using psychoacoustic properties", in *Preprints of the 92nd AES Convention*, no. 3321, Vienna (Austria), March 1992
- [Hoshuyama 99] Hoshuyama, O., Sugiyama, A. and Hirano, A. "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters", in *IEEE Trans. on Signal Processing*, vol.47, no.10, pp.2677-2684, October 1999
- [Houtgast 85] Houtgast, T. and Steeneken, H. J. M., "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria", in *J. Acoust. Soc. Am.*, vol.77, no.3, pp.1069-1077, 1985

- [Hussain 97] Hussain, A., Campbell, D.R. and Moir, T.J., “A multi-microphone sub-band adaptive speech enhancement system employing diverse sub-band processing”, in *ESCA-NATO ETRW Robust Speech Recognition for Unknown Communication Channels*, pp. 123-126, Pont a Mousson (France), April 1997
- [Hwang 96] Hwang, J. J. and Rao, K. R., *Techniques and standards for image, video, and audio coding*, Prentice-Hall, New Jersey, 1996
- [IEC/CD 1672] IEC/CD 1672, *Electroacoustics-sound level meters*, November, 1996
- [Ikram 01] Ikram, M.Z., and Morgan, D.R., “A multiresolution approach to blind separation of speech signals in a reverberant environment”, in *Proc. ICASSP 2001*, pp.2757-2760, Salt Lake City (USA), 2001
- [Ihle 00] Ihle, M, Kroshchel, K. and Riedlinger, R., “A novel noise suppression algorithm using a very small microphone array”, in *Preprints of the 109th AES Convention*, no.5252 Los Angeles (USA), September 2000
- [ISO/IEC 13818-3] ISO/IEC, *Generic coding of moving pictures and associated audio information*, International Standard ISO/IEC 13818-3, pp.84-107, 1998
- [Jabloun 01] Jabloun, F. and Champagne, B., “A multi-microphone signal subspace approach for speech enhancement”, in *Proc. ICASSP 2001*, pp.205-208, Salt Lake City (USA), 2001
- [Jaffer 88] Jaffer, A.G., “Maximum likelihood direction finding of stochastic sources: a separable solution”, in *Proc. ICASSP 1988*, vol.5, pp.2893-2896, 1988
- [Jan 95] Jan, E., Svaizer, P. and Flanagan, J.L., “Matched-filter processing of microphone array for spatial volume selectivity” in *Proc. ISCAS 1995*, vol.2, pp.1460-1463, 1995
- [Jan 96] Jan, E. and Flanagan, J., “Sound capture from spatial volumes: matched-filter processing of microphone arrays having randomly-distributed sensors” in *Proc. ICASSP 1996*, vol.2, pp.917-920, 1996
- [Johnson 89] Johnson, D. H., “Trends in array signal processing”, in *Proc. Sixth IEEE Multidimensional Signal Processing Workshop*, p.61, New York, (USA), 1989
- [Johnson 93] Johnson, D. H., and Dudgeon, D. E., *Array signal processing*, Prentice Hall-Englewood Cliffs, N. J., 1993
- [Johnston 88] Johnston, J. D., “Transform coding of audio signals using perceptual noise criteria”, in *IEEE Journal on Selected Areas in Comm.*, vol.6, no.2, pp.314-323, 1988
- [Junqua 91] Junqua, J.C., Reaves, B. and Mak, B., “A study of endpoint detection algorithms in adverse conditions. Incidence on a DTW and HMM recognize”, in *Proc. EUROSPEECH 1991*, pp.1371-1374, 1991
- [Kahrs 98] Kahrs, M. (Editor) and Brandenburg, K., “Reverberation algorithms”, Chapter 3, in *Applications of digital signal processing to audio and acoustics*, pp.85-131, Kluwer Academic Publishers, Boston, 1998
- [Kamiyanagida 01] Kamiyanagida, H., Saruwatari, H., Takeda, K. and Itakura, F., “Direction of arrival estimation based on nonlinear microphone array”, in *Proc. ICASSP 2001*, pp. 3033-3036, Salt Lake City (USA), 2001

- [Kaneda 86] Kaneda, Y. and Ohga, J., "Adaptive microphone-array system for noise reduction", in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.ASSP-34, pp.1391-1400, 1986
- [Katkovnik 00] Katkovnik, V., "Minimax robust M-beamforming for moving sources and impulse noise environment", in *Proc. EUSIPCO 2000*, pp.1401-1404, Tampere (Finland), 2000
- [Kaveh 90] Kaveh, M. and Bassias, A., "Threshold extension based on a new paradigm for MUSIC-type estimation", in *Proc. ICASSP 1990*, vol.5, pp.2535-2538, 1990
- [Kellerman 91] Kellerman, W. "A self-steering digital microphone array", in *Proc. ICASSP 1991*, vol.5, pp.3581-3584, 1991
- [Khalil 94] Khalil, F., Jullien, J.P. y Gilloire, A. "Microphone array for sound pickup in teleconference systems", in *J. Audio Eng. Soc.*, vol.42, no.9, pp.691-700, Septiembre 1994
- [Kim 00] Kim, N.S. and Chang J.H., "Spectral enhancement based on global soft decision", in *IEEE Signal Processing Letters*, vol.7, no.5, May 2000
- [Klabbers 01] Klabbers, E. and Veldhuis, R., "Reducing audible spectral discontinuities", in *IEEE Trans. on Speech and Audio Processing*, vol.9, no.1, pp.39-51, January 2001
- [Klepko 97] Klepko, J., "5-Channel microphone array with binaural head for multichannel reproduction", in *Preprints of the 103rd AES Convention*, no.4541 (F-4), New York, (USA), September 1997
- [Knapp 76] Knapp, C.H. and Carter, G.C., "The generalized correlation method for estimation of time delay", in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.ASSP-24, pp.320-327, August 1976
- [Krim 96] Krim, H. and Viberg, M., "Two decades of array signal processing research: the parametric approach", in *IEEE Signal Processing Magazine*, vol.13, no.4, pp.67-94, July 1996
- [Kryter 62] Kryter, K., "Methods for the calculation and use of the articulation index", in *J. Acoust. Soc. Am.*, vol.34, p.1689, 1962
- [Kuttruff 91] Kuttruff, H., *Room Acoustics*, 3rd ed., Elsevier Science Publishers, New York, 1991
- [Lacoss 71] Lacoss, R.T., "Data adaptive spectral analysis methods", in *Geophysics*, vol.36, no.4, pp.661-675, 1971
- [Le Bouquin 97] Le Bouquin, R. et al., "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator", in *IEEE Trans. on Speech and Audio Processing*, vol.5, no.5, pp.484-487, September 1997
- [Levinson 47] Levinson, N., "The Wiener RMS error criterion in filter design and prediction", in *Journal of Mathematical Physics*, no.25, pp. 261-278, 1947
- [Liu 96] Liu, Q.G., Champagne, B. and Kabal, P., "A microphone array processing technique for speech enhancement in a reverberant space", in *Speech Communication*, vol.18, pp.317-334, 1996

- [Lockwood 92] Lockwood, P. and Boudy, J., "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars ", in *Speech Communication*, vol.11, pp.215-228, 1992
- [Mahieux 96] Mahieux, Y., Le Tourneur, G. and Saliou, A. "A microphone array for multimedia workstations", in *J. Audio Eng. Soc.*, vol.44. no.5, pp. 365-317, May 1996
- [Mahmoudi 98] Mahmoudi, D. and Drygajlo A., "Combined Wiener and coherence filtering in wavelet domain for microphone array speech enhancement", in *Proc. ICASSP 1998*, pp.385-388, Seattle (USA), 1998
- [Marro 98] Marro, C., Mahieux, Y. and Simmer, U., "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering", in *Trans. on Speech and Audio Processing*, vol.6, no.3, pp.240-259, May 1998
- [Martin 00] Martin, A., Karray, L and Gilloire, A., "High order statistics for robust speech/non-speech detection", in *Proc. EUSIPCO 2000*, pp.469-472, Tampere (Finland), 2000
- [Masgrau 99] Masgrau, E., Aguilar, L. and Lleida, E., "Performance comparison of several adaptive schemes for microphone array beamforming", in *Proc. EUROSPEECH 1999*, vol. 6, pp.2615-2618, Budapest (Hungary), 1999
- [McAulay 80] McAulay, R.J. and Malpass, M.L., "Speech enhancement using a soft-decision noise suppression filter", in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.ASSP-28, April 1980
- [Mittal 00] Mittal U. and Phamdo N., "Signal/Noise KLT based approach for enhancing speech degraded by colored noise", in *IEEE Trans. on Speech and Audio Processing*, vol.8, no.2, pp.159-167, March 2000
- [Munier 87] Munier, J. and Delisle, G.Y., "Spatial analysis in passive listening using adaptive techniques", in *Proc. IEEE*, vol.75, pp.1458-1471, 1987
- [Naidu 01] Naidu, Prabhakar S., *Sensor array signal processing*, Boca Raton, CRC Press, 2001
- [Neely 79] Neely, S.T. and Allen, J.B., "Invertibility of a room impulse response", in *J. Acoust. Soc. Am.*, vol.66, no.1, pp.165-169, 1979
- [Nordholm 92] Nordholm, S., Claesson, I., and Eriksson, P., "The broad-band Wiener solution for Griffith-Jim beamformers", in *IEEE Trans. on Signal Processing*, vol.40, pp.474-478, February 1992
- [Nordholm 99] Nordholm, S., Claesson, I., and Dahl, M., "Adaptive microphone array employing calibration signals: an analytical evaluation", in *IEEE Trans. on Speech and Audio Processing*, vol.7, no.3, pp.241-252, May 1999
- [Omologo 93] Omologo, M. and Svaizer, P., "Talker localization and speech enhancement in a noisy environment using a microphone array based acquisition system". in *Proc. EUROSPEECH 93*, Berlin (Germany), 1993
- [Omologo 96] Omologo, M. and Svaizer, P., "Acoustic source location in noisy and reverberant environment, using CSP analysis", in *Proc. ICASSP 1996*, pp.3099-3102, 1996

- [Omologo 97-a] Omologo M., Matassoni, M., Svaizer, P. and Giuliani, D. "Microphone array based speech recognition with different talker-array positions", in *Proc. ICASSP 1997*, pp.227-230, Munich (Germany), April 1997
- [Omologo 97-b] Omologo, M. and Svaizer, P., "Use of the cross-power spectrum phase in acoustic event localization", in *IEEE Trans. on Speech and Audio Processing*, vol. 5, no.3, pp.288-292, September 1997
- [Ortega-García 00] Ortega-García, J., González-Rodríguez, J. and Marrero-Aguiar, V., "An approach to forensic speaker verification using 'AHUMADA' large speech corpus in Spanish", in *Speech Communication*, Elsevier Science, vol. 31, pp. 255-264, June 2000
- [Ottersten 02] Ottersten, B., Viberg, M. and Kailath, T., "Analysis of subspace fitting and ML techniques for parameter estimation from sensor array data", in *IEEE Trans. on Signal Processing*, vol.40, no. 3, pp.590-600, March 1992
- [Parra 00] Parra, L. and Spence, C., "Convulsive blind separation of non-stationary sources", in *IEEE Trans. on Speech and Audio Processing*, vol.8, no.3, pp.320-327, May 2000
- [Paulraj 86] Paulraj, A., Roy, R., and Kailath, T., "A subspace rotation approach to signal parameter estimation", in *Proc. IEEE*, pp 1044-1045, July 1986
- [Peutz 71] Peutz, V.M.A. "Articulation loss of consonants as a criterion for speech transmission in a room", in *J. Audio Eng. Soc.*, vol.19, p.915, 1971
- [Pohlmann 00] Pohlmann, Ken C., *Principles of digital audio*, pp. 303-361, McGraw-Hill, 2000
- [Poland 99] Poland, Syd, *TMS320C67xx. Divide and square root floating-point functions*, Application Report, SPRA516, Texas Instruments, February 1999
- [Quackenbush 88] Quackenbush, S.R., Barnwell bIII, T.P. and Clemens, M.A., *Objective measures of speech quality*, Englewood Cliffs-Prentice Hall, N.J., 1988
- [Rabiner 77] Rabiner, L.R. and Sambur, M.R., "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem", in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.25, no.4, pp.338-343, August 1977
- [Radlovic 00] Radlovic, B. D. and Kennedy, R. A., "Nonminimum-Phase equalization and its subjective importance in room acoustics", in *IEEE Trans. on Speech and Audio Processing*, vol.8, no.6, pp.728-737, November 2000
- [Reyayee 01] Reyayee, A. and Gazor S., "An adaptive KLT approach for speech enhancement", in *IEEE Trans. on Speech and Audio Processing*, vol.9, no.2, pp.87-94, February 2001
- [Roy 86] Roy, R., Paulraj, A. and Kailath, T., "Direction of arrival estimation by subspace rotation method-ESPRIT", in *Proc. ICASSP 1986*, pp.2495-2498, Tokyo (Japan), 1986
- [Ryan 00] Ryan, G., and Goubran, R.A., "Array optimization applied in the near field of a microphone array", in *IEEE Trans. on Speech and Audio Processing*. vol.8, no.2, pp.173-176, March 2000
- [Sachar 01] Sachar, J., Silverman, H. and Patterson III W. R., "Large vs small aperture microphone arrays: performance over a large focal area", in *Proc. ICASSP 2001*, vol. 5, pp.3049-3052, Salt Lake City (USA), 2001

- [Sánchez-Bote 99] Sánchez-Bote, J.L., *Sistemas de refuerzo sonoro y megafonía*, Dpto. Publicaciones de la EUITT, Universidad Politécnica de Madrid, Madrid, 1999
- [Sánchez-Bote 00] Sánchez-Bote, J.L., González-Rodríguez, J. and Ortega-García, J., “A new approach to dereverberation and noise reduction with microphone arrays”, in *Proc. EUSIPCO 2000*, pp.183-186, Tampere (Finland), 2000
- [Sánchez-Bote 01-a] Sánchez-Bote, J.L., González-Rodríguez, J. and Simón-Zorita, D., “A new auditory based microphone array and objective evaluation using E-RASTI”, in *Proc. EUROSPEECH 2001*, pp.2591-2594, Aalborg (Denmark), 2001
- [Sánchez-Bote 01-b] Sánchez-Bote, J.L., González-Rodríguez, J., Simón-Zorita, D. and Ortega-García, J. “Supresión de ruido audible utilizando arrays microfónicos” (in Spanish), in *Libro de actas URSI 2001*, pp.493-494, Madrid (Spain), 2001
- [Sánchez-Bote 02-a] Sánchez-Bote, J.L., *Micrófonos*, Dpto. Publicaciones de la EUITT, Universidad Politécnica de Madrid, Madrid, 2002
- [Sánchez Bote 02-b] Sánchez-Bote J.L., González-Rodríguez J. and Ortega-García J. “array de micrófonos en tiempo real basado en sustracción perceptual y evaluado con E-RASTI”, in *II Jornadas en Tecnologías del Habla, 2002 (Available in CD)*, Granada (Spain), 2002
- [Sánchez-Bote 03-a] Sánchez-Bote J.L., González-Rodríguez J. and Ortega-García J., “A real-time auditory-based microphone array assessed with E-RASTI evaluation proposal”, in *Proc. ICASSP 2003*, pp.481-484 , Honk Kong (China), 2003
- [Sánchez-Bote 03-b] Sánchez-Bote J.L., González-Rodríguez J. and Ortega-García J., “Audible noise suppression with a real-time broadband superdirective microphone array”, submitted for revision to *IEEE Trans. on Acoust. Speech and Signal Processing*, July 2003
- [Saruwatari 00] Saruwatari, H., Takeda, K. and Itakura, F., “Speech enhancement based on noise adaptive nonlinear microphone array”, in *Proc. EUSIPCO 2000*, pp.175-178, Tampere (Finland), 2000
- [Saruwatari 01] Saruwatari, H., Kurita, S. and Takeda, K., “Blind source separation combining frequency-domain ICA and beamforming”, in *Proc. ICASSP 2001*, pp.2733-2736, Salt Lake City (USA), 2001
- [Schmidt 86] Schmidt, R. O., “Multiple emitter location and signal parameter estimation”, in *IEEE Trans. on Antennas Propagat.*, vol.AP-34, no.3, pp.276-280, March 1986
- [Schroeder 64] Schroeder, M. R., “Improvement of acoustic-feedback stability by frequency shifting”, in *J. Acoust. Soc. Am.*, vol.36, no.9, pp.1718-1724, September 1964
- [Shamsoddini 01] Shamsoddini, A. and Denbigh, P. N., “A sound segregation algorithm for reverberant conditions”, in *Speech Communication*, vol.33, n.0.3, pp.179-196, 2001
- [Silverman 87] Silverman, H. “Some analysis of microphone array for speech data acquisition”, in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.ASSP-35, no.12, pp.1699-1712, 1987
- [Silverman 92] Silverman, H.F. and Kirtman, S.E. “A two-stage algorithm for determining talker location from linear microphone-array data”, in *Computer, Speech, and Language*, vol.2, no.6, pp.129-152, 1992

- [Sinha 93] Sinha, D. and Tewfik, A.H., "Low bit rate transparent audio compression using adapted wavelets", in *IEEE Trans. on Signal Processing*, vol.41, pp.3463-3479, December 1993
- [Stavropoulos 00] Stavropoulos, K. V. and Manikas, A., "Array calibration in the presence of unknown sensor characteristics and mutual coupling", in *Proc. EUSIPCO 2000*, pp.1417-1420, Tampere (Finland), 2000
- [Steeneken 80] Steeneken, H.J.M. and Houtgast, T., "A physical method for measuring speech-transmission quality", in *J. Acoust. Soc. Am.*, no.67, pp.318-326, 1980
- [Steeneken 85] Steeneken, H. J. M. and Houtgas, T., "Rasti: a tool for evaluating auditoria", in *Technical Review*, Brüel & Kjaer, no.3, 1985
- [Stoica 90-a] Stoica, P. and Sharman, K.C., "A novel eigenanalysis method for direction estimation", in *IEEE Proceedings on Radar and Signal Processing*, vol. 137, pp.19-26, no.1, February 1990
- [Stoica 90-b] Stoica, P. and Sharman, K.C., "Maximum likelihood methods for direction-of-arrival estimation", in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.38, no.7, pp. 1132-1143, July 1990
- [Stoica 90-c] Stoica, P. and Nehorai, A., "Performance study of conditional and unconditional direction-of-arrival estimation", in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.38, no.10, pp.1783-1795, October 1990
- [Stoica 91] Stoica, P. and Soderstrom, T., "Statistical analysis of MUSIC and subspace rotation estimates of sinusoidal frequencies", in *IEEE Trans. on Signal Processing*, vol.39, no. 8, pp.1836-1847, August 1991
- [Strobel 00] Strobel, N. and Rabenstein, R., "Robust speaker localization using a microphone array", in *Proc. EUSIPCO 2000*, pp.1409-1412, Tampere (Finland), 2000
- [Strobel 01] Strobel, N. Spors, S. and Rabenstein, R., "Joint audio-video object localization and tracking", in *IEEE Signal Processing Magazine*, January 2001
- [Su 83] Su, G. and Morf, M., "The signal subspace approach for multiple wide-band emmitter location", in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.31, no.12, pp.1502-1522, December 1983
- [Sullivan 93] Sullivan, T. M. and Stern, R. M., "Multi-microphone correlation-based processing for robust speech recognition", in *Proc. ICASSP 1993*, Minneapolis (USA), April 1993
- [Sullivan 96] Sullivan, T. M., *Multi-microphone correlation-based processing for robust speech recognition*, Ph. D. Thesis Dissertation, Carnegie Mellon University, Pittsburg, PA, 1996
- [Tanaka 01] Tanaka, H. and Kobayashi, T., "Estimating positions of multiple adjacent speakers based on music spectra correlation using a microphone array", in *Proc. ICASSP 2001*, pp.3045-3048, Salt Lake City (USA), 2001
- [TI 99] Texas Instruments, *TMS320C6000. Technical brief*, SPRU197D, February 1999
- [TI 00] Texas Instruments, *TMS320C6000 Code Composer Studio manuals, v.120*, 2000

- [TI 02-a] Texas Instruments, *TMS320C62x DSP library programmer's reference*, SPRU402A, Texas Instruments, April 2002
- [TI 02-b] Texas Instruments, *TMS320C67x fastRTS Library programmer's reference*, SPRU100, Texas Instruments, March 2002
- [Tilp 00] Tilp, J., "Single-channel noise reduction with pitch-adaptive post-filtering", in *Proc. EUSIPCO 2000*, pp.171-174, Tampere (Finland), 2000
- [Tohyama 95] Tohyama, M., (Ed.), Suzuki, H. (Ed.) and Ando, Y., (Ed.), *The Nature and Technology of Acoustic Space*, pp.220-253, Academic Press, August 1995
- [Tsoukalas 97] Tsoukalas, D. E., Mourjopoulos, J.N., Kokkinakis, G., "Speech enhancement based on audible noise suppression", in *IEEE Trans. on Speech and Audio Processing*, vol.5, no.6, pp.497-514, 1997
- [Tucker 92] Tucker, R., "Voice activity detection using a periodicity measure", in *IEEE Proceedings in Communications, Speech and Vision*, vol.139, no.4, August, 1992
- [Van Been 88] Van Been, B. D., and Buckley, K.M., "Beamforming: a versatile approach to spatial filtering", in *IEEE ASSP Magazine*, vol. 5, no.2, pp.4-24, April 1988
- [Van Compernolle 90-a] Van Compernolle, D., "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings", in *Proc. ICASSP 1990*, vol.2, pp.833-836, Albuquerque (USA), 1990
- [Van Compernolle 90-b] Van Compernolle, D., Ma, W., Xie, F. and Van Diest, M., "Speech recognition in noisy environments with the aid of microphone arrays", in *Speech Communication*, no.9, pp.433-442, 1990
- [Vaseghi 96] Vaseghi, S.V., "Wiener filters", Chapter 5, pp.140-163 and "Spectral Subtraction", Chapter 9, pp.242-260, in *Advanced signal processing and digital noise reduction*, Wiley-Teubner, 1996
- [Vermaak 01] Vermaak, J. and Blake, A. "Nonlinear filtering for speaker tracking in noisy and reverberant environments", in *Proc. ICASSP 2001*, pp.3021-3024, Salt Lake City (USA), 2001
- [Viberg 91-a] Viberg, M. and Ottersten, B., "Sensor array processing based on subspace fitting", in *IEEE Trans. on Signal Processing*, vol.39, no.5, pp.1110-1121, May 1991
- [Viberg 91-b] Viberg, M. and Ottersten, B. and Kailath T., "Detection and estimation in sensor arrays using weighted subspace fitting", in *IEEE Trans. on Signal Processing*, vol.39, no.11, pp.2436-2449, November 1991
- [Viberg 94] Viberg, M. and Swindlehurst, A.L., "A bayesian approach to auto-calibration for parametric array signal processing", in *IEEE Trans. on Signal Processing*, vol.42, no.12, pp.3495-3507, December 1994
- [Virag 95] Virag, N., "Speech enhancement based on masking properties of the auditory system", in *Proc. ICASSP 1995*, pp. 796-799, vol.1, New York (USA), 1995
- [Virag 99] Virag, N., "Single channel speech enhancement based on masking properties of the human auditory system", in *IEEE Trans. on Speech and Audio Processing*, vol.7, no.2, pp.126-137, 1999
- [VocaLinks] VocaLinks Inc., available in <http://www.vocalinks.com>

- [Wan 03] Wan, E., Nelson, A., and Peterson, Rick, *Speech Enhancement Assessment Resource (SpEAR) Database*, available in http://cslu.ece.ogi.edu/nsel/data/SpEAR_database.html. Beta Release v1.0. CSLU, Oregon Graduate Institute of Science and Technology
- [Wang 85] Wang, H. and Kaveh, M, “Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources”, in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.33, no.4, pp.823-831, August 1985
- [Ward 95] Ward, D.B., Kennedy, R.A. y Williamson, R.C., “Theory and design of broadband sensor arrays with frequency invariant far-field beam patterns”, in *J. Acoust. Soc. Am.*, vol.97, no.2, pp.1023-1034, February 1995
- [Ward 96] Ward, D.B., Kennedy, R.A. y Williamson, R.C., “FIR filter design for frequency-invariant beamformers”, in *IEEE Signal Processing Letters*, vol.3, no. 3, pp. 69-71, March 1996
- [Wax 83] Wax, M. and Kailath, T., “Optimum localization of multiple sources by passive arrays”, in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol. ASSP-31, pp.1210-1217, October 1983
- [Wax 92] Wax, M., “Detection and localization of multiple sources in noise with unknown covariance”, in *IEEE Trans. on Signal Processing*, vol.40, no.1, pp.245-249, January 1992
- [Wiener 49] Wiener, N., *Extrapolation, interpolation, and smoothing of stationary time series*, John Wiley & Sons, Inc., New York, 1949
- [Wiggins 65] Wiggins, R.A. and Robinson, E.A., “Recursive solution to the multichannel filtering problem”, in *J. Geophys. Res.*, no.70, pp.1885-1891, 1965
- [Williams 99] Williams, M. and Guillaume L. D., “Microphone array analysis for multichannel sound recording”, in *Preprints of the 107th AES Convention*, no.4997 (A-5), New York (USA), September 1999
- [Xu 93] Xu, Xiao-Liang and Buckley, K., “An Analysis of beam-space source localization”, in *IEEE Trans. on Signal Processing*, p.501, January 1993
- [Yegnanarayana 98] Yegnanarayana, B. and Murthy P. S., “Enhancement of reverberant speech using LP residual”, in *Proc. ICASSP 1998*, vol.1, pp.405-408, New York (USA), 1998
- [Yegnanarayana 00] Yegnanarayana, B. and Murthy P. S., “Enhancement of reverberant speech using LP residual signal”, in *IEEE Trans. on Speech and Audio Processing*, vol.8, no.3, pp.267-281, May 2000
- [Yong 00] Yong, L.K. and Jung S., “Time-domain approach using multiple Kalman filters and EM algorithm to speech enhancement with nonstationary noise”, in *IEEE Trans. on Speech and Audio Processing*, vol.8, no.3, pp.282-291, May 2000
- [Zelinski 88] Zelinski, R., “A microphone array with adaptive post-filtering for noise reduction in reverberant rooms”, in *Proc. ICASSP 1988*, pp.2578-2581, 1988
- [Ziskind 88] Ziskind, I. and Wax, M., “Maximum likelihood localization of multiple sources by alternating projection”, in *IEEE Trans. on Acoust. Speech and Signal Processing*, vol.36, no.10, pp.1553-1560, October 1988

[Zwicker 99] Zwicker, E. and Fastl H., *Psychoacoustics: facts and models*, Springer Verlag; 2nd edition, Berlin, May 1999