Universidad
Carlos III de Madrid

TESIS DOCTORAL

Contribuciones a la Aplicación de la Factorización de Matrices No Negativas a las Tecnologías del Habla

Autor:

Jimmy D. Ludeña Choez

Director:

Dra. Ascensión Gallardo Antolín

Dpto. de Teoría de la Señal y Comunicaciones

Leganés, Abril 2015

TESIS DOCTORAL

CONTRIBUCIONES A LA APLICACIÓN DE LA FACTORIZACIÓN DE MATRICES NO NEGATIVAS A LAS TECNOLOGÍAS DEL HABLA

Autor:
JIMMY D. LUDEÑA CHOEZ

Director:
Dra. ASCENSIÓN GALLARDO ANTOLÍN

Firma del Tribunal Calificador:

Firma

Presidente:

Vocal:

Secretario:

Calificación:

Leganés, de de

RESUMEN

El funcionamiento de los sistemas de procesamiento y clasificación de audio (incluída la voz) en escenarios reales, depende, en gran medida, de una adecuada representación de la señal de audio, tanto en condiciones limpias como ruidosas. Por este motivo, en esta Tesis abordamos la problemática del diseño de nuevos esquemas de preprocesamiento y extracción de características acústicas con aplicación a dos tareas distintas: reconocimiento automático del habla y clasificación de eventos acústicos. El nexo de unión de los métodos propuestos es la utilización de la técnica denominada factorización de matrices no negativas (NMF, *Non-Negative Matrix Factorization*) que ha demostrado ser una herramienta poderosa para el análisis de la señal de audio.

En primer lugar, en este trabajo de tesis se propone un método de eliminación de ruido en señales de voz basado en NMF, que, a diferencia de otras aproximaciones previas, no asume un conocimiento a priori acerca de la naturaleza del ruido. La técnica es evaluada tanto para mejora de voz como para reconocimiento automático de habla mostrando un mejor funcionamiento que la técnica convencional de sustracción espectral.

En segundo lugar, se proponen tres parametrizaciones novedosas para la tarea de clasificación de eventos acústicos. La primera de ellas es una extensión de los parámetros convencionales mel-cepstrales y consiste en el filtrado paso alto de la señal de audio. El segundo esquema consiste en una mejora de la técnica de integración temporal de características llamada coeficientes de banco de filtros (FC, *Filter bank Coefficients*) en el que NMF se utiliza como método no supervisado para el aprendizaje del banco de filtros FC óptimo. Finalmente, en el último nuevo parametrizador se propone la inclusión de características cepstrales derivadas de los coeficientes de activación o ganancia de NMF, motivada por la robustez al ruido que NMF ofrece. Los experimentos realizados muestran que, en términos generales, estos tres esquemas mejoran el funcionamiento del sistema de clasificación de eventos acústicos con respecto al de referencia tanto en condiciones limpias como ruidosas.

ABSTRACT

In real scenarios, the performance of audio processing and classification systems depends largely on an adequate representation of the signal in both clean and noisy conditions. Therefore, in this Thesis we face the problem of designing new methods to preprocess audio signals and extract acoustic features with the intention of being applied to two different tasks: Automatic Speech Recognition (ASR) and Acoustic Event Classification (AEC). The proposed methods are based on the well-known Non-Negative Matrix Factorization (NMF) technique, which has proven to be a powerful tool for analyzing audio signals.

Firstly, a method for speech denoising is proposed, that unlike other previous approaches it does not assume a prior knowledge about the nature of the kind of noise. The method is evaluated for both, speech enhancement and ASR, showing better performance than one of the state of art techniques known as Spectral Subtraction (SS).

Secondly, we propose three new parameterization schemes for AEC. The first one is an extension of the conventional Mel Frequency Cepstral Coefficients (MFCC) and can be seen as a high-pass filtering of the audio signal. The second scheme is an improvement of the temporal feature integration technique named Filterbank Coefficients (FC), in which the NMF technique is used in an unsupervised manner, allowing to discover an optimal FC filterbank. Finally, the last new parameterization scheme proposes the use of cepstral features derived from the NMF activation coefficients, this is mainly motivated by the robustness shown by NMF in noisy conditions. Experiments have shown that, in general terms, these three feature extraction modules improve the performance of the acoustic event classification systems with respect to the baseline based on MFCC, for both, clean and noisy conditions with different noises at different signal-to-noise ratio (SNR) levels.

AGRADECIMIENTOS

Sin lugar a duda no alcanzarían palabras suficientes para agradecer a todas personas que en algún momento me han acompañado en el desarrollo de este trabajo Tesis. Pasé muchos momentos de impaciencia y desesperación, especialmente cuando los resultados no eran favorables, convirtiendose en un reto personal sacarlos adelante, motivándome en todo momento a continuar en este proyecto.

A mi Director de Tesis, Dra. Ascensión G. Antolín, por su paciencia y apoyo constante desde que empecé en este mundo de la investigación, muchas gracias Profesora Ascensión.

A mis grandes amigos Gonzalo, Efraín y Juan José, sin algún orden en particular, por darme la oportunidad y confiar en mí en este reto de la investigación y recibir recomendaciones siempre acertadas. Por supuesto como no agradecer a mis compañeros de la liga 300, aquellos miércoles o jueves, que me ayudaba a distraerme y relajarme un rato de las labores de la investigación. A mis amigos de la UNSA, Alonso, Guillermo, Ebert,..., cuando nos reuniamos los domingos a jugar futbol y me hacían renegar cuando perdiamos, gracias por su amistad.

Al Dr. German Chávez, Rector de la UCSP. Gracias por confiar en mí y darme la oportunidad en este reto de la investigación. También agradecer a Fredy, Erika y a todos los profesores de la Escuela Profesional de Ingeniería Electrónica y de Telecomunicaciones de la UCSP, gracias a todos ellos por su amistad.

Agradecer a mi novia Claudia, por su paciencia y comprensión, durante estos años del desarrollo de este proyecto y por la gran noticia que me ha dado. Por su puesto como no dejar de agradecer a mi hermano Jean Paul, y en especial a mis padres Juliana y Luis Edilberto, gracias a ellos por su sacrificio, esfuerzo y constante aliento, siempre animándome a terminar este proyecto y hacerme recordar de no dejar las cosas a medias. Este trabajo de Tesis, es el resultado de sus enseñanzas y formación y como retribución, éste trabajo, lo dedico a ustedes.

Gracias a todos.

" *No hay que confundir nunca el conocimiento con la sabiduría.*
El primero nos sirve para ganarnos la vida;
la sabiduría nos ayuda a vivir."

Sorcha Carey,
1943 - ?

Índice general

Índice de figuras

Índice de tablas

Glosario

AEC Acoustic Event Classification

ALS NMF Alternating Least Squares NMF

ASC Audio Spectrum Centroid

ASE Audio Spectral Envelope

ASF Audio Spectrum Flatness

ASS Audio Spectrum Spread

CEA Clasificación de Eventos Acústicos

CIFE Conditional Informative Feature Extraction

CMIM Conditional Mutual Information Maximization

CMN Cepstral Mean Normalization

CondRed Conditional Redundancy

DBNN Deep Belief Neural Networks

DCT Discrete Cosine Transform

DISR Double Input Symmetrical Relevance

EEG Electroencefalograma

ERB Equivalent Rectangular Bandwidth

FBEC Filter Bank Energy Coefficients

FC Filterbank Coefficients

FS Feature Selection

GMM Gaussian Mixture Models

HMM Hidden Markov Models

HTK Hidden Markov Model Toolkit

JMI Joint Mutual Information

KL Kullback-Leibler Divergence

LPC Linear Prediction Coefficients

LPCC Linear Prediction Cepstral Coefficients

LSF Line Spectral Frequency

MFCC Mel Frequency Cepstral Coefficients

MI Mutual Information

MIFS Mutual Information Feature Selection

mRMR Minimum-Redundancy Maximum-Relevance

NMF Non-Negative Matrix Factorization

OND Offline Noise Data

PCA Principal Components Analysis

PESQ Perceptual Evaluation of Speech Quality

PLP Perceptual Linear Prediction

PLPCC Perceptual Linear Prediction Cepstral Coefficients

RAH Reconocimiento Automático del Habla

RBFNN Radial Basis Function Neural Networks

RR Recognition Rate

SAH Sistema Auditivo Humano

SBV Spectral Basis Vectors

SNR Signal Noise Rate

SS Spectral Substraction

STFT Short Time Fourier Transform

SVM Support Vector Machines

VAD Voice Activity Detector

VADND Voice Activity Detector Noise Data

VQ Vectorial Quantization

ZCR Zero Crossing Rate

Capítulo 1

Introducción

En este primer capítulo, se establece el contexto general en que se desarrolla la tesis, presentando algunos conceptos fundamentales de la señal del habla como un medio natural de transmitir información, dando a conocer las dificultades que se presentan en condiciones reales en diversas tareas relacionadas con las tecnologías del habla. Iniciamos el capítulo con una explicación breve acerca de la producción del habla. A continuación se realiza una breve descripción de las técnicas de mejora de la voz para la aplicación al reconocimiento automático del habla (RAH). Tras lo anterior se exponen los objetivos perseguidos en este trabajo, para finalizar el capítulo con la descripción de la estructura del contenido de esta tesis.

1.1. El habla

Las últimas decadas han sido testigo de la aparición de una nueva generación de interfaces hombre-máquina, que combinan varias tecnologías del habla, permitiendo a las personas conversar con las computadoras usando el diálogo para acceder, crear y procesar información. Hoy en día gran cantidad de información está disponible a través de internet y redes sociales y pueden ser utilizada para una gran cantidad de propósitos diversos: educación, toma de decisiones, finanzas, entretenimiento, etc.

Figura 1.1: Modelo fuente-filtro para señal del habla.

Del mismo modo gran cantidad de la población está interesada en acceder a la información cuando están en movimiento, desde cualquier lugar y en su propio idioma. Una solución prometedora es otorgar a las máquinas capacidades como similares a la de los humanos de modo que puedan' "hablar" y "escuchar" de la misma manera que las personas interactúan.

De esta modo, el lenguaje hablado es muy atractivo porque es la forma más natural, eficiente y flexible de comunicación humana, permitiendo la transmisión de la información a través del uso de palabras y expresiones. A través del habla también es posible transmitir información concerniente al locutor como es la identidad, el género, el estado de salud y las emociones.

Esto es posible debido a las variaciones espectro-temporales de la señal de voz. Estas variaciones, que conforman los diferentes sonidos del habla, son producidas por el mecanismo de producción de la voz, en el que participan los pulmones, laringe, cavidad del tracto vocal, cavidad nasal, dientes y los labios. En la figura 1.1, podemos apreciar el proceso de producción de la señal de voz utilizando el denominado modelo fuente-filtro, en el que la señal de la fuente o excitación es modulada de acuerdo a la frecuencia de apertura o cierre de los pliegues glotales, generándose una señal cuasiperiódica que es la entrada al tracto vocal en el que se enfatizan ciertas frecuencias resonantes llamadas formantes. En otras palabras, el habla se produce como resultado de un proceso de modulación de la fuente de energía sonora (señal cuasiperiódica) a través de un filtro con función de transferencia variante con el tiempo, determinado por la forma y tamaño del tracto vocal. De esta manera, el modelo fuente describe la estructura fina o detallada de la señal del habla, mientras que el modelo filtro describe la envolvente del espectro de la voz [Vaseghi, 2007].

El propósito del modelado y parametrización del habla es encontrar una adecuada representación, en función de un conjunto eficiente y compacto de características para tareas tales como: codificación, reconocimiento, síntesis y mejora de la voz. Por ejemplo para codificación de la voz, muchos de los codificadores comerciales están basados en modelos de predicción lineal, mientras que para reconocimiento automático del habla, muchas de las características acústicas comúnmente utilizadas se obtienen a partir de la envolvente espectral, parámetros cepstrales y sus dinámicas en el tiempo [Vaseghi, 2007].

En la presente tesis, nos centramos en el reconocimiento automático del habla que consiste en obtener la transcripción automática de las expresiones orales pronunciadas por un determinado locutor. Aunque, en la actualidad, los sistemas de reconocimiento de voz funcionan bien en tareas controladas y en condiciones limpias (cuando no hay presencia de ruido aditivo u otras distorsiones), uno de los principales retos a los que deben enfrentarse es mejorar su funcionamiento en ambientes adversos (también denominados condiciones ruidosas), en los que su rendimiento se degrada significativamente, debido principalmente a la presencia de ruido de fondo.

En una parte de este trabajo de tesis, nos hemos enfocado en esta problemática que hemos abordado mediante el diseño de una etapa de eliminación de ruido cuyo objetivo es mejorar la calidad de la señal de voz antes de ser reconocida por el sistema de reconocimiento de habla. Para ello, hemos propuesto un método basado en la factorización de matrices no negativas (NMF, *Non-Negative Matrix Factorization*), tal y como explicaremos en esta tesis.

Por otra parte, debemos de tener en cuenta que muchas de las distorsiones que afectan la calidad de la señal de la voz y por ende a los sistemas de RAH, están constituidas por otros tipos de sonidos como risas, toses, timbres telefónicos, etc. llamados de forma genérica, eventos acústicos, por lo que es recomendable el diseño de sistemas de clasificación y detección de estos sonidos, con la finalidad de detectar su presencia y eliminarlos o por lo menos aminorar su impacto sobre la señal de la voz, con lo que se espera incrementar la robustez de los sistemas de RAH.

A continuación comentamos otras aplicaciones de la tarea de clasificación de eventos acústicos (AEC, *Acoustic Event Classification*) con más detalle.

1.2. Los eventos acústicos

La importancia de los sonidos o eventos acústicos, es que permiten a los seres humanos sentir y comprender al mundo físico que los rodea. Por ello la necesidad de desarrollar herramientas basadas en procesamiento de señal que permitan extraer información útil a partir de los datos acústicos del mundo real; además de ser lo suficientemente robustos. En las últimas decadas, el despliegue masivo de teléfonos celulares y micrófonos han permitido mejorar la calidad y bajar el costo de los sistemas de adquisión de datos acústicos, permitiendo incrementar la capacidad de cómputo para el análisis de los eventos acústicos en tiempo real. A pesar de este potencial, el análisis de los eventos acústicos (que no sea el reconocimiento de voz) no ha llegado ser tan estudiado y desplegado.

Hoy en día, las técnicas usadas para el análisis acústico estan desarrolladas para aplicaciones muy específicas (por ejemplo en la detección de submarinos a través del SONAR, detección de llamadas de ballenas de una especie específica; detección de balas, etc.), siendo muy dependientes de la tarea a realizar. Sería por lo tanto, recomendable usar nuevas herramientas del procesamiento de señales con la finalidad de realizar un adecuado análisis acústico extrayendo información útil y representativa de la señal acústica, pudiendo ser además generales en el sentido que puedan ser aplicadas a otras tareas de clasificación de audio.

Los sistemas de monitoreo basado en audio tienen un impacto significativo en aplicaciones tales como de vigilancia, seguridad pública y evaluación de ruido urbano en zonas residenciales. Actualmente, el riesgo de seguridad y robos se incrementa en zonas públicas como bares, lugares de recreación y ocio. Es crucial por lo tanto, detectar estos casos de emergencia de manera oportuna y alertar a la policía para prevenir daños mayores.

Del mismo modo, el incremento de actividades de tráfico, negocios e inclusive recreativas contribuyen a empeorar los efectos del ruido urbano en la salud humana, mediante la exposición de niveles excesivos de ruido. Debido a esto, es necesario implementar sistemas como mapas de ruido y predicción que permitan evaluar los niveles de ruido en una zona urbana y predecir los cambios del ambiente ruidoso. La creación de mapas de ruido, permitirá definir políticas futuras para planificación, construcción, tráfico y transporte dentro de la comunidad.

Muchas de la técnicas desarrolladas se han implementado para la producción y análisis de la voz, por lo que no se han estudiado modelos de producción para todos los eventos acústicos. En esta tesis se van a desarrollar varias parametrizaciones para AEC, algunas de las cuales están basadas en NMF.

1.3. Objetivos

La elección del tema de tesis ha estado motivada por la necesidad de mejorar las técnicas de análisis y procesado de señales de voz y audio en escenarios reales, en los que la presencia de ruido y otros tipos de distorsiones degradan en gran medida el funcionamiento de los sistemas basados en tecnologías del habla y audio.

En este contexto, el objetivo de esta tesis doctoral es profundizar en la aplicación de métodos basados en la factorización de matrices no negativas (NMF, *Non - Negative Matrix Factorization*) al análisis, caracterización y mejora de la calidad de las señales de voz y audio en diversas tareas relacionadas con las tecnologías del habla y audio. En concreto, en este trabajo se estudia la potencialidad de NMF tanto para el análisis espectral y la obtención de nuevas representaciones paramétricas como para la eliminación de ruido en señales de voz y audio.

Como se ha mencionado anteriormente en la presente tesis, se han considerado los siguientes ámbitos de aplicación: reconocimiento automático del habla en condiciones adversas (ruido de fondo) y clasificación de eventos acústicos (pasos, toses, risas, etc.) en entornos de oficina con diversas condiciones de ruido.

1.4. Estructura del documento

La organización del resto de capítulos de esta tesis se realiza de la siguiente manera: en el **capítulo 2**, se dan a conocer los fundamentos del método de factorización de matrices no negativas (NMF) y de los principales métodos de extracción de características tanto para reconocimiento automático del habla como para clasificación de eventos acústicos. En el **capítulo 3**, se presenta un método basado en NMF para la eliminación de ruido para mejora de la señal de voz con aplicación al reconocimiento automático del habla. En el **capítulo 4**, se presenta el análisis de las características espectrales de diversos eventos acústicos usando NMF y la nueva parametrización desarrollada a partir de este análisis para la tarea de clasificación de eventos acústicos, que está basada en el filtrado paso alto de la señal de audio. En el **capítulo 5**, se describe una nueva parametrización para esta última tarea que está basada en la integración temporal de características acústicas en la que se utiliza NMF para adaptar dicha técnica al caso particular de los eventos acústicos. En el **capítulo 6**, se desarrolla un nuevo método de extracción de características para AEC que está basada en la selección automática de bandas espectrales. El **capítulo 7** trata sobre un nuevo procedimiento de parametrización para AEC en el que se utilizan los coeficientes de activación o ganancia obtenidos con NMF en combinación con las características mel-cepstrales. Finalmente en el **capítulo 8**, se resumen las principales contribuciones de la presente tesis doctoral, del mismo modo que se describen algunas líneas futuras de investigación.

Capítulo 2

Estado del arte

En este capítulo se desarrolla la base teórica del método de Factorización de Matrices No Negativas (NMF, *Non - Negative Matrix Factorization*) y se realiza un repaso de las principales técnicas de extracción de características usadas en tareas tales como: el Reconocimiento Automático del Habla (RAH) y la tarea de Clasificación de Eventos Acústicos (CEA).

2.1. Factorización de Matrices No Negativas (NMF)

Un problema común en la mayoría de las aplicaciones basadas en aprendizaje automático es la necesidad de encontrar una representación adecuada de los datos de entrada. En los últimos años el algoritmo de la factorización de matrices no - negativas está siendo utilizado para este propósito con resultados satisfactorios en diferentes ámbitos relacionados con el procesamiento de señales y datos: procesamiento de audio [Virtanen, 2007], [Wilson et al., 2008], aprendizaje de características acústicas [Schuller et al., 2010], imágenes [Sandler and Lindenbaum, 2011], electroencefalogramas (EEG) [Chen et al., 2006], [Damon et al., 2013], bioinformática [Brunet et al., 2004], agrupamiento [Jingu and Haesun, 2008] y minería de texto [Wei et al., 2003],

NMF realiza una representación lineal no supervisada de los datos, siendo los coeficientes de la combinación lineal positivos, a diferencia de otros métodos como son el Análisis de Componentes Principales (PCA, *Principal Components Analysis*) y cuantización vectorial (VQ, *Vectorial Quantization*) que realizan una representación holística de los datos.

Básicamente, con NMF es posible obtener una representación de los datos basada en partes a través del uso de la restricción de no - negatividad, por la que solo se permiten combinaciones aditivas de los diversos componentes básicos (o *partes*). De esta forma, se consigue una mejor interpretabilidad de los resultados de la descomposición. La restricción de la no-negatividad se refuerza por el hecho de que muchos de los datos del mundo real son no-negativos y sus correspondientes componentes fundamentales ocultas tienen un significado físico solo cuando son no-negativos, por ejemplo: las intensidades de los píxeles en una imagen, el espectro de amplitud, el consumo de energía y alimentos, etc. Existe además evidencia psicológica y fisiológica de que nuestro cerebro realiza una representación basada en la partes y ciertas teorías computacionales aplicadas al reconocimiento de objetos confían en tales representaciones [Lee and Seung, 1999].

Por otra parte, la no - negatividad también induce a una representación dispersa (*sparse*) de los datos, que es fundamental cuando es deseable que dichos datos sean representados con la menor cantidad de componentes relevantes [Cichocki et al., 2009].

2.1.1. Formulación Matemática de NMF

El problema básico que resuelve NMF puede establecerse de la siguiente manera: Dada una matriz $\mathbf{V} \in \mathbb{R}_{+}^{F \times T}$, donde cada columna corresponde a un vector de datos, el algoritmo de la factorización de matrices no negativas lo aproxima como el producto de dos matrices de bajo rango no negativa W y H (figura 2.1), tal que

$$V \approx WH \tag{2.1}$$

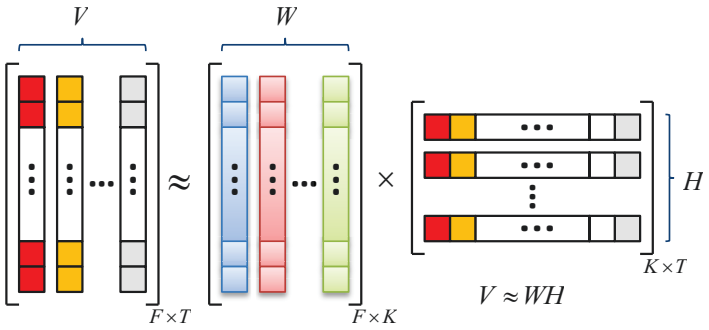


Figura 2.1: Representación Básica NMF.

donde $W \in \mathbb{R}_{+}^{F \times K}$ y $H \in \mathbb{R}_{+}^{K \times T}$ y normalmente $K \leq min\,(F, T)$. De esta manera, cada columna de W puede ser escrita como una combinación lineal de los K vectores base (columnas de W), ponderadas por los coeficientes de activación o ganancia localizados en las correspondientes columnas de H. NMF puede verse como un método de reducción de la dimensionalidad de los vectores de datos desde un espacio F dimensional hasta un espacio K dimensional. Esto es posible si las columnas de W descubren la estructura latente (oculta) en los datos [Lee and Seung, 1999].

La factorización se logra por una minimización iterativa de una determinada función de coste como por ejemplo, la distancia euclídea, dada en la ecuación 2.2 o la divergencia generalizada de Kullback-Leibler (KL), dada en la ecuación 2.3.

$$D_{\mathrm{EU}}\,(V \| WH) = \sum_{ij} \left(V_{ij} - (WH)_{ij}\right)^2 \tag{2.2}$$

$$D_{\mathrm{KL}}\,(V \| WH) = \sum_{ij} \left(V_{ij} log \frac{V_{ij}}{(WH)_{ij}} - (V - WH)_{ij}\right) \tag{2.3}$$

A lo largo del desarrollo de la tesis se ha considerado a la divergencia KL como función de coste dado que en la literatura se ha mostrado sus buenos resultados en tareas relacionadas con el procesamiento de voz, la separación de fuentes de sonido [Virtanen, 2007], mejora de la señal de voz [Wilson et al., 2008] o extracción de características acústicas [Schuller et al., 2010]. La divergencia KL tiene las siguientes

ventajas: presenta mejores resultados perceptuales, buenas propiedades de convergencia, razonable coste computacional [Schuller et al., 2009] y es apropiada cuando los datos presentan un amplio rango dinámico [Bertrand et al., 2008a].

Existen diferentes métodos para la obtención del valor óptimo local para la divergencia KL entre V y (WH). En este trabajo, hemos optado por utilizar uno de los más extendidos: el esquema iterativo con reglas de aprendizaje multiplicativo propuesto en [Lee and Seung, 1999] y establecido en la ecuación 2.4,

$$W \leftarrow W \otimes \frac{\frac{V}{WH}H^T}{1H^T} \qquad H \leftarrow H \otimes \frac{W^T\frac{V}{WH}}{W^T1} \tag{2.4}$$

donde 1 es una matriz de tamaño V, cuyos elementos son todos unos y las multiplicaciones $\otimes$ y divisiones son operaciones componente a componente.

El algoritmo NMF no asume dispersión de los datos (*sparsity*) o independencia estadística mutua entre las columnas de W. Sin embargo, NMF suele proporcionar una descomposición dispersa de los datos [Lee and Seung, 1999], lo que facilita su interpretabilidad. Hay varios caminos para lograr algún control sobre dicha representación dispersa. En esta tesis, hemos seguido la aproximación propuesta en [Cichocki et al., 2006] y [Cichocki et al., 2009] para la función de coste KL, donde NMF se regulariza usando proyecciones no lineales sobre (2.4). Aplicando este procedimiento, las reglas de aprendizaje son las siguientes,

$$W \leftarrow \left[W \otimes \frac{\left[\frac{V}{WH}H^T\right]^{\omega}}{1H^T}\right]^{(1+\alpha_w)} \qquad H \leftarrow \left[H \otimes \frac{\left[W^T\frac{V}{WH}\right]^{\omega}}{W^T1}\right]^{(1+\alpha_h)} \tag{2.5}$$

donde $\alpha_w \geqslant 0$ y $\alpha_h \geqslant 0$ son los parámetros de regularización o factores *sparse* y $\omega \in (0, 2)$ es un parámetro de relajación que también controla el grado de dispersión, además de la velocidad de convergencia del algoritmo. Es importante notar que al considerar los parámetros de regularización, el exponente de las reglas de aprendizaje son mayor que uno, lo que implica que los valores ms pequeños en las matrices nonegativas tienden a ser cada vez más próximos a cero a medida que el número de iteraciones se incrementa [Cichocki et al., 2009] mientras que los valores grandes se acrecientan.

2.1.2. Consideraciones prácticas de NMF

Una de la dificultades que se presenta en NMF es la elección adecuada del valor de K (número de bases o componentes), que por lo general depende de la matriz de datos y de la aplicación concreta. Mientras más grande sea el valor de K mejor es la aproximación de los datos; sin embargo, un valor pequeño de K conduce a un modelo menos complejo. La influencia del número de componentes se mostrará en la experimentación realizada para las tareas de eliminación de ruido en señales de voz (capítulo 3) y clasificación de eventos acústicos (capítulos 4, 5, 6 y 7).

Otro factor a tener en cuenta es la forma de inicializar el algoritmo iterativo a partir del cuál se obtienen las matrices W y H. La solución y convergencia del algoritmo NMF depende de estas condiciones de inicialización, de tal manera que se debe realizar una adecuada selección de las matrices iniciales W y H, pues de lo contrario puede resultar afectada la eficiencia del algoritmo NMF. Una pobre inicialización conducirá a un mínimo local y conducir por lo tanto a una solución incorrecta e irrelevante. El problema llega a ser aún más crítico para problemas NMF de gran escala o cuando se imponen ciertas restricciones sobre las matrices factorizadas. Por otro lado, una buena inicialización para un conjunto de datos no garantiza que sea buena para otros conjuntos de datos distintos. Con respecto a la función de coste, el proceso de inicialización cumple un rol fundamental ya que la minimización de dicha función puede contener múltiples mínimos locales y la minimización alternante intrínsica en las reglas de aprendizaje de NMF no es convexa, incluso aunque la función de coste sea estrictamente convexa con respecto a una de las variables. Por ejemplo, las funciones de coste dadas en las ecuaciones 2.2 y 2.3, son estrictamente convexas con respecto a solo uno de la variables (W y H) pero no a ambas [Cichocki et al., 2009].

Con el objetivo de solventar esta problemática en la medida de lo posible, en esta tesis las matrices factorizadas se inicializaron usando el algoritmo de inicialización múltiple dado en [Cichocki et al., 2009], de tal modo que se generan 10 pares de matrices aleatorias uniformes (W_0 y H_0) y la factorización que produce la distancia

euclídea más pequeña entre V y (W_0H_0) se escoge para inicialización. Estas matrices iniciales se refinaron posteriormente mediante la minimización de la divergencia KL entre V y sus correspondientes matrices factorizadas (W_0H_0) usando el esquema iterativo y reglas de aprendizaje dadas en 2.4 estableciendo como punto de parada del algoritmo el número máximo de iteraciones (en nuestro caso se eligió 200).

Para mayor información sobre NMF se aconsejan las referencias [Cichocki et al., 2009], [Lee and Seung, 1999] y [Cichocki et al., 2006].

2.2. Reconocimiento Automático del Habla (RAH)

Los sistemas de Reconocimiento Automático del Habla (RAH) tienen un amplio rango de aplicaciones desde un simple sistema de reconocimiento de palabras aisladas (por ejemplo, para marcación de nombres en un teléfono celular, servicios al cliente automatizado, control de maquinarias y coches a través del uso de la voz) hasta el reconocimiento de voz continua como auto dictado, transcripción de noticias, etc. En la figura 2.2 se muestra el diagrama esquemático típico de un sistema RAH, donde podemos observar que normalmente la señal de voz es preprocesada con la finalidad de eliminar o atenuar posibles distorsiones como son, fundamentalmente, la presencia de ruido aditivo. Posteriormente, la señal de voz preprocesada pasa por un proceso de un extracción de características acústicas, cuya finalidad es retener la información más relevante generándose vectores de parámetros representativos de la señal de voz. Para obtener la palabra o frase reconocida se comparan dichos vectores de características u observaciones con unos patrones o modelos acústicos mediante un proceso denominado decodificación acústica, que suele llevarse a cabo utilizando el algoritmo de Viterbi. Típicamente, los modelos acústicos son modelos estadísticos que se obtienen durante la fase de entrenamiento del sistema, siendo los Modelos Ocultos de Markov (HMM, *Hidden Markov Models*) los más extensamente utilizados.

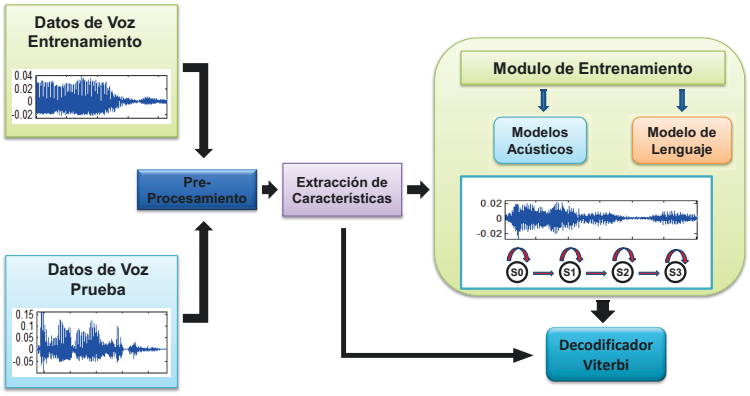


Figura 2.2: Diagrama esquemático de un sistema de RAH.

2.2.1. Extracción de Características Acústicas.

El proceso de extracción de características acústicas consiste en extraer un conjunto de características espectro - temporales que representen de forma adecuada a la señal de voz, capturando información esencial para la identificación de sonidos y palabras. Para el reconocimiento automático del habla, las características comúnmente utilizadas se estiman a partir de la envolvente espectral de la señal de voz, debido a la alta correlación que presenta un fonema formado por el tracto vocal y su espectro. Otra característica utilizada en la tarea del RAH es el cálculo de la energía por trama; sin embargo, es fundamental considerar los cambios de energía local para ayudar a la discriminación entre fonemas con mayor y menor energía [O'Shaughnessy, 2013].

2.2.1.1. Transformada de Fourier a corto plazo (STFT, Short Time Fourier Transform)

El método tradicional para realizar el análisis espectral de una señal es la transformada de Fourier; sin embargo no es adecuada para el análisis de señales no estacionarias como es el caso de la voz, porque solo proporciona información frecuencial (espectral) de la señal, pero no proporciona información sobre el momento en el que la frecuencia está presente. La transformada de Fourier a corto plazo enventanada (STFT) proporciona la información temporal acerca del contenido de frecuencia de

la señal. En nuestro caso vamos a utilizar la transformada discreta de Fourier a corto plazo dada en la ecuación 2.6.

$$X\left(f,t\right)=\sum_{n=0}^{N-1}x\left(n\right)w\left(n-t\right)e^{-j\frac{2\pi fn}{N}} \tag{2.6}$$

$$w\left(n\right)=0{,}54-0{,}46\cos(\tfrac{2\pi n}{N-1}),\quad 0<n<N-1 \tag{2.7}$$

donde, $X\left(f,t\right)$ es la transformada discreta de Fourier a corto plazo de la señal $x(n)$ con $n=0,1,...,N-1$ y $w(n-t)$ es la ventana de análisis usada y desplazada una cantidad de tiempo t. Usualmente la ventana que se utiliza es la ventana Hamming, mostrada en la ecuación 2.7.

Para distinguir entre diferentes sonidos, no es necesario solo tener en cuenta la energía total de la señal sino obtener la mayor cantidad de detalles del espectro de la voz. Con frecuencia se usa la Transformada de Fourier Discreta a corto plazo (STFT), que se aplica a una sección enventanada (*trama*) de la señal, asegurando que la señal enventanada sea aproximadamente estacionaria. Esta operación se repite periódicamente desplazando la ventana, creando de esta forma una versión segmentada de la señal de la voz. La duración de la ventana y su desplazamiento están relacionados con las variaciones temporales de la articulación del tracto vocal y son un compromiso entre la maximización de la precisión y la minimización del coste computacional. Como se ha mencionado anteriormente, habitualmente se utiliza la ventana de Hamming con la finalidad de prevenir el fenómeno de Gibbs producido por las discontinuidades al aplicar la ventana de análisis.

Otra alternativa a la STFT son las ondículas (*wavelets*) [Evangelista, 1993] [Hlawatsch and Boudreaux-Bartels, 1992]. La principal limitación de la STFT es que utiliza una longitud fija de muestras que es determinada por la duración de la ventana de análisis; de esta manera si la longitud de la ventana de análisis es pequeña se empeora la resolución espectral (incrementando la resolución temporal); mientras que se mejora la resolución espectral cuando la longitud de la ventana de análisis es grande. Por otro lado a nivel perceptual, el oído usa una escala de frecuencia no lineal

(Mel) y este hecho no se toma en cuenta cuando la ventana de análisis se mantiene fija. En el caso de las *wavelets*, la ventana de análisis varía a través de operaciones de escalamiento y desplazamiento temporal, siendo la ventana de análisis de corta duración en altas frecuencias (donde la resolución espectral es pobre a nivel perceptual) y de larga duración para bajas frecuencias (donde se requiere una mejor resolución espectral). Debido a esto, las *wavelets* han sido usadas en muchas aplicaciones de procesado de voz. Sin embargo, su no linealidad y difícil interpretación han limitado su uso en la tarea del RAH [O'Shaughnessy, 2013].

2.2.1.2. Coeficientes cepstrales basado en predicción lineal (LPCC, Linear Prediction Cepstral Coefficients)

Los coeficientes cepstrales basado en predicción lineal se obtienen a partir del cálculo de los coeficientes de predicción lineal (LPC, *Linear Prediction Coefficients*). LPC es un método de análisis de la señal de voz que es muy útil para la codificación de la voz a bajas tasas de bits. LPC se basa en la suposición de que la señal de la voz se produce como un sonido vibrante (producido por la glotis y caracterizado por su intensidad y frecuencia) al final de un tubo (formado por el tracto vocal y caracterizado por sus frecuencias resonantes llamados formantes), permitiendo a los LPCC reflejar las diferencias de la estructura biológica del tracto vocal humano. El cálculo de los LPCC es se puede realizar a través de una recursión de los parámetros LPC de la siguiente manera:

$$\begin{cases} c_1 = a_1 \\ c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k} & , 1 < m \leq P \\ c_m = \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k} & , m > P \end{cases} \tag{2.8}$$

donde c_m son los coeficientes cepstrales, a_p son los coeficientes de predicciońn lineal y P es el orden de predicción.

Se ha mostrado que LPC es muy eficiente para los sonidos vocálicos de la señal de la voz, siendo menos eficiente para regiones no vocálicas, transitorias y no esta-

cionarias [Zbancioc and Costin, 2003], [Yujin et al., 2010]. Para una señal en tiempo discreto, $x(n)$, LPC considera que cada muestra puede ser aproximada como la combinación lineal de P muestras precedentes $x(n-p), p = 1...P$. Los factores de la ponderación en esta combinación (a_p) son los llamados coeficientes LP. Usualmente los coeficientes LP (a_p) se calculan usando el algoritmo de Levinson - Durbin [Vaseghi, 2007]. Para el RAH, un aspecto importante del LPC es que modela bien los detalles espectrales alrededor de los picos espectrales (pricipalmente los sonidos vocálicos); sin embargo, una debilidad del análisis LPC es que trata a todas las frecuencias sin tener en cuenta la característica de frecuencia no lineal del oído humano. Otro aspecto importante es que se debe escoger a priori el valor de P del modelo LPC (con frecuencia, $P = 10$ para voz telefónica a 8000 muestras / segundo). En principio, el valor de P es proporcional al ancho de banda de la voz, de tal modo que si se escoge un valor de P muy grande puede conducir a modelar picos espectrales que corresponden a los armónicos y no a los formantes y si el valor de P es pequeño se suaviza el espectro ocultando a los formantes. Cuando el análisis LPC se utiliza en codificación de voz, tales desviaciones no son tan importantes porque los oyentes pueden tolerar pequeñas distorsiones en el procesos de resíntesis de la voz. Para RAH, estas desviaciones pueden ocasionar tasas de reconocimiento bajas.

En muchas aplicaciones, la representación LP básica se transforma en un conjunto de coeficientes de reflexión, que son más eficientes para la transmisión y tienen el beneficio de corresponder al modelo del tracto vocal; sin embargo, tal modelo no es el único ya que muchas formas del tracto vocal pueden conducir al mismo espectro de la señal de voz. Otra aproximación más eficiente para transmisión son las líneas de frecuencia espectrales (LSF, *Line Spectral Frequency*). Los LSFs representan a los polos de la representación LP dentro del círculo unitario en el plano z, que permite una representación más sencilla de los formantes de la voz. Ambas aproximaciones (coeficientes de reflexión y LSFs) permiten una interpretación más fácil que la STFT en términos de resonancias del tracto vocal; no obstante, ninguna de ellas es eficiente para propósitos del RAH [O'Shaughnessy, 2013].

2.2.1.3. Coeficientes cepstrales en escala de frecuencia Mel (MFCC, Mel Frequency Cepstral Coefficients)

Estudios psicofísicos han mostrado que la percepción humana del contenido en frecuencia del sonido para señales de voz no sigue una escala lineal. Los coeficientes cepstrales en escala de frecuencia mel (MFCC), convierten el espectro lineal en un espectro no - lineal (Mel), tratando de imitar el comportamiento de la membrana basilar del oído interno que determina los anchos de banda críticos del oído humano en función de la frecuencia. A día de hoy, es el método de extracción de características más utilizado para RAH.

Para su extracción, primero se determina el espectro de magnitud $V = \mid X(f,t) \mid$ de la señal de voz segmentada usando la STFT con N-puntos. Es usual aplicar un filtro de preénfasis (filtro paso alto) como una etapa de preprocesamiento a la señal de voz antes de realizar su transformación al dominio espectral. Este filtro se aplica con la finalidad de producir una relación constante a través de toda la banda de frecuencia compensando los efectos producidos por la fuente glotal y radiación de los labios. El espectro de amplitud es entonces multiplicado por un conjunto de filtros pasobanda triangulares a diferentes escalas de frecuencia llamado banco de filtros auditivo inspirado por la selectividad en frecuencia que realiza la membrana basilar (cóclea). Este banco de filtros trata de simular el fenómeno de percepción auditiva por el que para un tono con amplitud y frecuencia fija, se reduce la sensibilidad del oído a otros tonos de frecuencia similar, llegando a ser inaudibles si caen dentro de la banda crítica [Mcloughlin, 2009]. A este producto se aplica el logaritmo que cumple la función de convertir una multiplicación espectral en una suma, lo que permite separar la envolvente espectral del tracto vocal del conjunto de armónicos generados por la fuente de la señal de voz. Dicha separación facilita la tarea del RAH. Finalmente se aplica la tranformada discreta del coseno (DCT, *Discrete Cosine Transform*). Este procesamiento está justificado en parte por el hecho de que la DCT ortogonaliza las log-energías en banda, dando lugar a un conjunto de características acústicas no correladas que son aproximadamente gaussianas. El uso de la escala logarítmica

considera aspectos relevantes de la percepción y producción de la voz, lo que hace posible que la precisión del RAH mejore con respecto al uso del análisis LP, en el que no es fácil integrar dichos aspectos [O'Shaughnessy, 2013]. El banco de filtros auditivo se puede construir usando diferentes escalas de frecuencia (Mel, Bark y ERB), de tal manera que los coeficientes cepstrales se obtienen de acuerdo a la ecuación 2.9.

$$c_{t,m} = \alpha(m) \sum_{q=0}^{Q-1} log\left(\phi(q,t)\right) cos\left[m\left(\frac{2q+1}{2}\right)\frac{\pi}{Q}\right] \tag{2.9}$$

$$\phi(q,t) = \varphi(q,f) \mid X(f,t) \mid \tag{2.10}$$

$$\alpha(m) = \begin{cases} \sqrt{\frac{1}{Q}} & m = 0 \\ \sqrt{\frac{2}{Q}} & m \neq 0 \end{cases} \tag{2.11}$$

donde $c_{t,m}$ es el m-simo coeficiente cepstral correspondiente a la trama t-sima, con $m = 1, 2, ..., M$ siendo M el número deseado de coeficientes cepstrales y $\varphi(q,f)$ es el q-simo filtro pasobanda del banco de filtros auditivo con $q = 1, 2, ..., Q$ siendo Q el número de filtros pasobanda que forman el banco de filtros auditivo.

Cuando la escala de frecuencia utilizada es Mel (la más habitual), entonces se obtienen los coeficientes cepstrales en escala de frecuencia Mel (MFCC) tal y como se muestra en la figura 2.3. Una breve descripción de la diferentes escalas de frecuencia se muestran en la subsección 2.2.2.

2.2.1.4. Coeficientes cepstrales de predicción lineal perceptual (PLPCC, Perceptual Linear Prediction Cepstral Coefficients)

Los PLPCC se basan en la obtención de las características espectrales de predicción lineal perceptual (PLP, *Perceptual Linear Prediction*). PLP es un método de análisis que incorpora una escala de frecuencia no lineal al igual que los MFCC y otras propiedades conocidas de la psicofísica de la audición (curva de igual sonoridad y la ley de la intensidad de potencia). La parametrización PLPCC se muestra en la

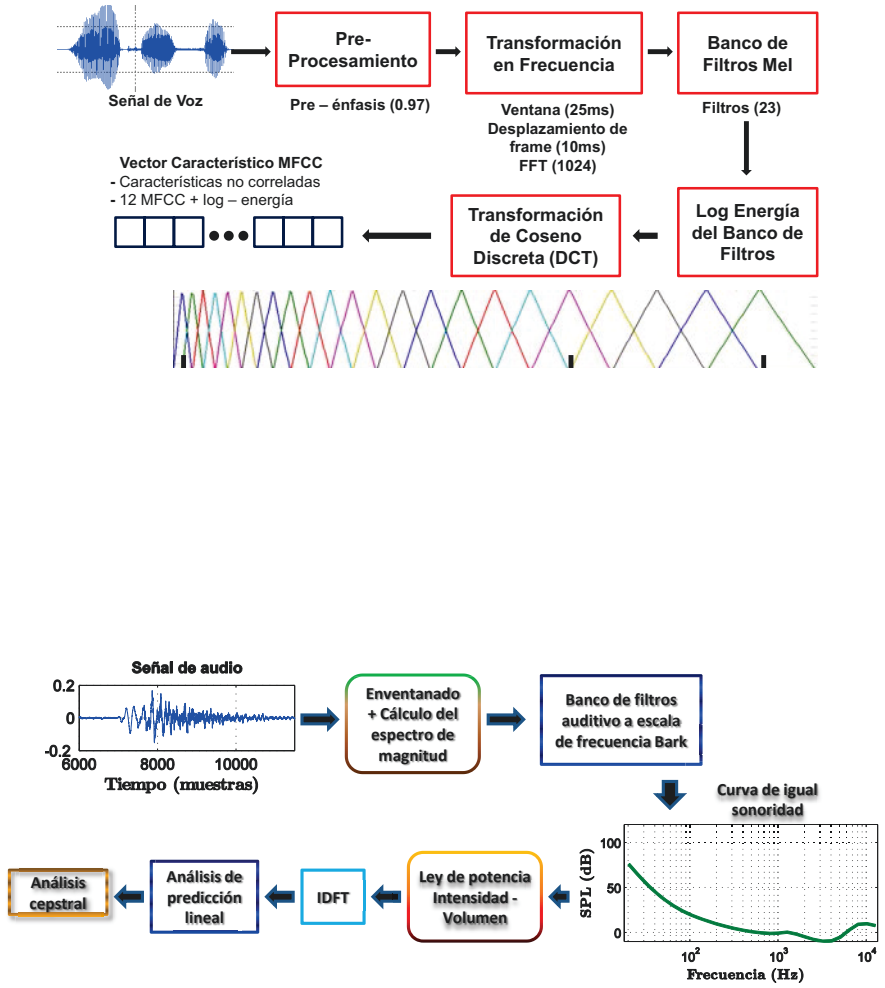


Figura 2.4: Diagrama de bloques de extracción de los coeficientes PLPCCs.

figura 2.4, donde podemos observar que una vez que el espectro se obtiene después de aplicar la transformada de Fourier, se transforma a una escala Bark y se enfatiza por medio del uso de una función que aproxima la sensibilidad del oído humano a diferentes frecuencias (curva de igual sonoridad). La salida se comprime para aproximarse a la relación no lineal entre la intensidad del sonido y su sonoridad percibida. Posteriormente se calculan los coeficientes de predicción lineal a través del análisis LPC para finalmente transformarlos a coeficientes cepstrales [Vachhani and Patil, 2013].

2.2.2. Escalas de frecuencia

En el diseño del banco de filtros auditivo para la extracción de los parámetros cepstrales, en este trabajo hemos considerado tres escalas de frecuencia logarítmica (Mel, Bark y ERB). Estas escalas logarítmicas son usualmente escogidas en tareas relacionadas al procesamiento de audio y voz, debido a que está ampliamente aceptado que el Sistema Auditivo Humano (SAH) realiza una compresión logarítmica en el rango auditivo, de modo que los intervalos de alta frecuencia están representados con menos detalle que los rangos de baja frecuencia. Esta observación se ha derivado de una serie de experimentos psicoacústicos realizados para determinar las denominadas bandas críticas (ancho de banda de frecuencia alrededor de un frecuencia central cuyas componentes afectan el nivel de sonido y la percepción del *pitch* de la frecuencia central) [Zwicker and Terhardt, 1980].

La escala Mel es una de las escalas de frecuencia de transformación logarítmica más conocida y ha sido desarrollada por Stevens, Volkman y Newmann en 1937 [Stevens and Newman, 1937],

$$F_m(f) = 2595 \log\left(1 + \frac{f}{0{,}7}\right) \tag{2.12}$$

con m en Mel y f en KHz. Esta transformación en frecuencia es la base para el procedimiento de extracción de los MFCC que utiliza un banco de filtros formado de filtros triangulares solapados que están uniformemente distribuidos según esta escala.

La escala Bark fue desarrollada por Zwicker. En la ecuación 2.13 se muestra la fórmula de transformación a partir de la escala de frecuencia lineal a la escala Bark [Zwicker and Terhardt, 1980],

$$F_z(f) = 13 \arctan(0{,}76f) + 3{,}5 \arctan\left(\frac{f}{7{,}5}\right)^2 \tag{2.13}$$

Con z en Bark y f en KHz.

La escala ERB es una escala logarítmica basada en el ancho de banda rectangular equivalente (ERB, *Equivalent Rectangular Bandwidth*), que es una medi-

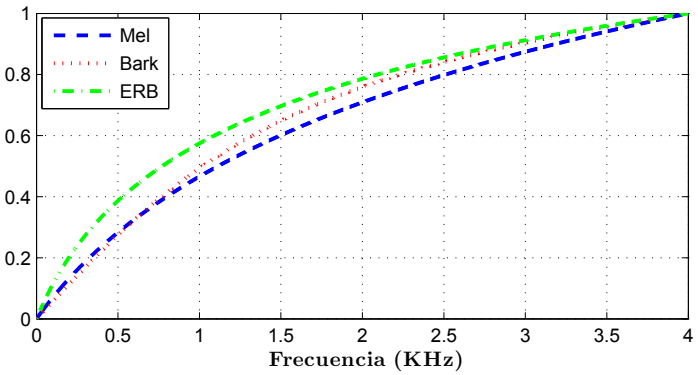


Figura 2.5: Escalas de frecuencia para el banco de filtros auditivo.

da más ajustada de la banda crítica. Está definida mediante la siguiente ecuación [Moore and Glasberg, 1983],

$$F_{ERB}(f) = 11{,}17 \ln\left(\frac{f + 0{,}312}{f + 14{,}675}\right) + 43{,}0 \tag{2.14}$$

con f en KHz. En la figura 2.5 se muestran las diferentes escalas de frecuencia (Mel, Bark y ERB) normalizadas, donde se puede observar que las escalas Mel y Bark son aproximadamente lineales a bajas frecuencias (hasta aprox. $1kHz$) y logarítmica en el resto.

2.2.3. Reconocimiento de voz usando Modelos Ocultos de Markov (HMM, Hidden Markov Models)

Es recomendable en muchos casos usar una etapa de pre - procesamiento para la seal de voz, principalmente cuando la señal de la voz se degrada por la presencia de ruido de fondo. Existen varias técnicas para la mejora de la señal de la voz, entre ellas tenemos técnicas basadas en substraccin espectral [Berouti et al., 1979], filtro de wiener [Scalart and Vieira, 1996]. En este trabajo de tesis se realiza el proceso de eliminación de ruido basado en el algoritmo NMF, creando un modelo acústico para la señal de la voz y del ruido.

La metodología básica para el RAH ha empleado la arquitectura de los modelos ocultos de Markov para la generación de los patrones o modelos acústicos, usando como entradas la salida del proceso de extracción de características, principalmente aquellas basadas en los parámetros MFCC (típicamente 13 coeficientes estáticos que modelan la posición del tracto vocal, más sus primera, delta, y segunda derivadas, delta - delta, que indirectamente caracterizan la velocidad y aceleración del tracto vocal).

Los modelos ocultos de Markov se usan para realizar el modelamiento estadístico de procesos de señal no - estacionario tales como las señales de voz, secuencia de imágenes, etc. La suposición fundamental del HMM es que la señal de la voz se encuentra bien caracterizada como un proceso aleatorio parámetrico y que estos parámetros pueden estimarse de una manera bien definida. Se basan en los procesos de Markov (proceso cuyo estado o valor en algún tiempo t depende de su estado o valor en el tiempo anterior $t-1$ y es independiente de la historia del proceso antes de $t-1$). HMM modela las variaciones en el tiempo de los estadísticos de un proceso aleatorio con una cadena Markoviana de subprocesos estacionarios dependiente de estado.

HMM se caracteriza por tres matrices A, B y π, donde:

- A - representa la matriz de probabilidad de transiciones de estado $(N \times N)$
- B - representa la matriz de distribución de probabilidad del símbolo de observación $(N \times M)$
- π - representa la matriz de distribución de estados inicial $(N \times 1)$

  N es el número de estados del modelo y M es el número de símbolos de observación distintos por estado. Usualmente se usa una notación compacta $\lambda = \{A, B, \pi\}$.

El proceso del reconocimiento de voz incluye los siguientes pasos [Rabiner, 1989]:

- Vectores de observaciones: son generados a partir del análisis espectral y temporal (ver subsección 2.2.1) de la señal de la voz. Estos vectores de observaciones se usan para entrenar los HMMs los cuales caracterizan los sonidos del habla considerados.

- Unidades acústicas: se debe realizar una adecuada selección de la unidades acústicas. Esta selección suele depender del tamaño del vocabulario, por ejemplo, para reconocimiento de voz con vocabulario pequeño es razonable y práctico utilizar la palabra como unidad acústica. Para el caso de vocabularios grandes es recomendable usar unidades de sub - palabra (fonemas, difonemas y trifonemas) como unidades acústicas. Cada una de estas unidades es caracterizada por un HMM cuyos parámetros son estimados a partir del conjunto de entrenamiento de los datos de voz.

- Módulo de comparación: El módulo de comparación determina las verosimilitudes entre todas las secuencias de unidades acústicas permitidas en el sistema con la entrada de voz desconocida, determinando el mejor *score* sujeto a restricciones sintácticas y léxicas. Por ejemplo, supongamos que el diccionario de un sistema de RAH consta de L palabras distintas. Durante la fase de entrenamiento del sistema, para cada palabra l en el vocabulario se construye un modelo HMM λ_l, cuyos parámetros $\{A, B, \pi\}$ se estiman optimizando la verosimilitud del conjunto de observaciones de entrenamiento para la l-sima palabra utilizando habitualmente el algoritmo de Baum-Welch [Rabiner, 1989]. Para cada palabra desconocida que va a ser reconocida, se deben calcular las verosimilitudes entre la secuencia de observaciones (características acústicas de la palabra a reconocer) de entrada al sistema $O = (o_1, o_2, ..., o_T)$, con todos los posibles modelos, $P(O/\lambda_l)$ con $1 \leq l \leq L$, seleccionando aquella palabra con la verosimilitud más alta de la siguiente manera:

$$l^* = \underset{1 \leq l \leq L}{argmax} \left[P(O/\lambda_l) \right] \tag{2.15}$$

Para este proceso suele utilizarse el algoritmo de Viterbi [Rabiner, 1989].

- Decodificación léxica: En esta etapa, se establecen restricciones sobre el sistema de comparación, de modo que se debe considerar solo aquellas unidades acústicas que se encuentren en el diccionario (o léxico) de la tarea. Este procedimiento implica que el vocabulario usado para el reconocimiento se debe especificar en términos de las unidades acústicas escogidas. Cuando las unidades acústicas son palabras, este paso de decodificación léxica se elimina, simplificando de esta forma la estructura del reconocedor.

- Análisis sintáctico: Al igual que en la etapa anterior, se establecen restricciones donde las palabras deben seguir una secuencia apropiada determinada por la gramática de la tarea, que puede ser representada por un modelo de lenguaje llamado n-grama, donde se toma en cuenta las probabilidades de secuencias de n palabras. Los modelos de lenguaje bigramas y trigramas son los más utilizados. Para tareas de control y comandos, únicamente se requiere una palabra del conjunto finito de palabras para ser reconocida y por lo tanto no es necesario la gramática. A este tipo de tareas se le denomina tarea de reconocimiento de voz de palabras aisladas. Para otras aplicaciones como, por ejemplo secuencia de dígitos, se requiere una gramática muy simple (solo un dígito puede continuar después de algún otro dígito). Hay, sin embargo, tareas donde la gramática es un factor dominante y aunque adiciona una restricción más al proceso de reconocimiento, produce una mejora en el rendimiento del sistema global.

- Análisis semántico: Al igual que en las dos etapas anteriores, en este proceso se imponen restricciones semánticas para la tarea del reconocimiento.

  Desde el punto de vista práctico, hay un factor adicional que debe tenerse en cuenta en la implementación de un reconocedor de voz y es el problema de separar el silencio (o ruido de fondo) de la señal de voz propiamente dicha. Hay al menos dos caminos razonables para lograr esta tarea:

- a) La detección explícita de la presencia de voz a través de técnicas que discriminan el silencio de la señal de voz a partir del cálculo de la energía, cruces por cero, duración de los segmentos de voz y silencio y otros parámetros acústicos adecuados.
- b) Inclusión de un modelo de silencio dentro del repertorio de los patrones acústicos del sistema. Es habitual considerar dos modelos distintos de silencio: silencio largo, correspondiente a los silencios iniciales y finales de cada elocución y silencio corto, correspondiente a las pausas entre palabras.

2.2.4. Reconocimiento automático del habla en condiciones ruidosas

Como se ha comentado en el capítulo de introducción, uno de los principales retos a los que deben enfrentarse los sistemas de reconocimiento automático del habla en la actualidad es el de mejorar su funcionamiento en ambientes adversos (también denominados condiciones ruidosas), en los que su rendimiento se degrada significativamente, debido principalmente a la presencia de ruido de fondo.

En el reconocimiento automático del habla robusto al ruido, habitualmente se realiza la suposición de que el desajuste acústico entre los datos de entrenamiento y los datos de prueba (condiciones reales de funcionamiento), que es el causante de la degradación en la tasa de reconocimiento del sistema, puede modelarse como una transformación que puede ser determinada mediante diversos métodos y que puede aplicarse tanto a los parámetros acústicos de la voz a reconocer como a los modelos o patrones acústicos del sistema.

A partir de esta suposición, se han propuesto diversas estrategias para abordar esta problemática, tales como [O'Shaughnessy, 2013]:

- Mejora o realce de la señal de voz. Consiste en "limpiar" o eliminar el ruido de la voz ruidosa usando técnicas de procesado de señal en un proceso previo al

reconocedor de voz en sí.

- Parametrizaciones robustas: Consisten en encontrar características acústicas que representen la voz y que sean robustas a la presencia de ruido y distorsiones.
- Compensación de características: Consiste en transformar las características acústicas extraídas de la voz degradada en parámetros limpios.
- Compensación de modelos acústicos: Consiste en adaptar los modelos o patrones acústicos, que generalmente han sido construidos a partir de habla limpia, a las condiciones ruidosas del entorno.

2.2.4.1. Mejora de la señal de voz

En las últimas decadas, se han sugerido una gran diversidad de esquemas para mejorar la calidad de la voz degradada, principalmente, en escenarios ruidosos, como calles, coches, metro, trenes o lugares públicos como salas de exposiciones. Su principal propósito es atenuar las distorsiones tanto como sea posible, preservando la señal de la voz. Dicha reducción de ruido puede beneficiar a un rango amplio de aplicaciones tales como teléfonos móviles, teléfonos a manos libres, teleconferencia, audífonos, servicios de voz automáticos basados en reconocimiento y síntesis de voz, forense y grabaciones antiguas.

En la literatura se han propuesto diversos métodos de reducción de ruido, entre los que destaca la técnica de substracción espectral y el filtrado de Wiener, que atenúan la señal de entrada en aquellos rangos de frecuencia donde el valor de la relación señal a ruido (SNR, *Signal Noise Rate*) es baja. El método de substracción espectral (SS, *Spectral Substraction*) [Scalart and Vieira, 1996] consiste en la resta del espectro de ruido a partir del espectro de la señal ruidosa. El filtrado de Wiener [Berouti et al., 1979] es similar al método de sustracción espectral, con la diferencia que en vez de calcular el espectro, diseña un filtro para eliminar el ruido en regiones de baja SNR. En ambos casos, el ruido se estima durante las porciones de la señal de entrada con amplitud relativamente baja, bajo la suposición de que tales porciones

son menos probables de contener voz. Ambos métodos mejoran la calidad de la voz, aunque tienen la desventaja de producir un ruido residual desagradable llamado ruido musical.

En los últimos años el método de la factorización de matrices no - negativas se ha usado en distintas áreas del procesamiento de la señal, especialmente en procesado de audio [Virtanen, 2007], [Wilson et al., 2008], logrando buenos resultados, especialmente, en condiciones ruidosas y mostrando, por tanto, su robustez frente al ruido [Schuller et al., 2010]. El éxito de NMF radica en el hecho de su capacidad de encontrar componentes relevantes ocultos de la señal de voz y su inmunidad al ruido ya que enfatiza aquellas componentes relevantes características de la señal y reduce la influencia de las componentes ruidosas.

2.3. Clasificación de Eventos Acústicos (CEA).

En los últimos años, el problema de la clasificación y detección de eventos acústicos que no correspondan a voz, han atraído la atención de numerosos investigadores. Aunque la voz es el evento acústico más informativo, otras clases de sonidos (tales como risas, toses, tipeado utilizando un teclado, etc.) pueden proporcionar pistas relevantes acerca de la presencia y actividad humana en ciertos escenarios (por ejemplo en una oficina). Esta información podría ser usada en diferentes aplicaciones, principalmente en aquellas con interfaces "amigables" tales como habitaciones "inteligentes" [Temko and Nadeu, 2006], aplicaciones en automóviles [Muller et al., 2008], robots trabajando en diversos ambientes [Chu et al., 2006] o en sistemas de vigilancia [Clavel et al., 2005]. Adicionalmente, los sistemas de clasificación y detección de eventos acústicos pueden utilizarse como una etapa de preprocesamiento para sistemas de reconocimiento automático del habla de tal manera que esta clase de sonidos puedan ser eliminados antes del proceso del reconocimiento de voz en sí, incrementando la robustez del sistema completo.

Con la finalidad de distinguir entre las diferentes clases acústicas, se

ha experimentado con distintos esquemas de clasificación, entre los que destacan los modelos de mezclas de gaussianas (GMM, *Gaussian Mixture Models*) [Temko and Nadeu, 2006], modelos ocultos de Markov (HMM) [Cotton and Ellis, 2011], máquinas de vectores soporte (SVM, *Support Vector Machines*) [Temko and Nadeu, 2006] [Mejia Navarrete et al., 2011], redes neuronales con funciones de base radial (RBFNN, *Radial Basis Function Neural Networks*) [Dhanalakshmi et al., 2008] and redes neuronales profundos (DBNN, *Deep Belief Neural Networks*) [Kons and Toledo-Ronen, 2013]. La alta correlación entre el rendimiento de diferentes clasificadores sugiere que el principal problema no es la técnica de clasificación usada, sino que el diseño de un proceso adecuado de extracción de características para AEC es fundamental [Kons and Toledo-Ronen, 2013].

Por lo tanto, el diseño de un adecuado proceso de extracción de características para AEC es una tarea importante. Se han propuesto varios esquemas en la literatura, algunos de ellos basados sobre las características a corto plazo, tales como los coeficientes cepstrales en escala de frecuencia Mel (MFCC) [Temko and Nadeu, 2006] [Zieger, 2008] [Zhuang et al., 2010] [Kwangyoun and Hanseok, 2011], log-energías en banda [Zhuang et al., 2010], predicción lineal perceptual (PLP) [Portelo et al., 2009], log-energía, flujo espectral, entropía fundamental y tasa de cruces por cero [Temko and Nadeu, 2006]. Muchas de estas características son a corto plazo o a nivel de trama en el sentido de que se calculan trama a trama (típicamente, el período de trama usado para análisis de voz / audio es aproximadamente de $10 - 20ms$).

No obstante, otras aproximaciones están basadas en la aplicación de diferentes técnicas de integración temporal sobre estos parámetros a corto plazo [Meng et al., 2007], de modo que se extraen características a escalas de tiempo más largas, combinando de alguna manera la información de los coeficientes a corto plazo contenidos en segmentos (conjunto de tramas consecutivas) o ventanas temporales de duración más larga [Mejia Navarrete et al., 2011], [Zhang and Schuller, 2012]. Dichas caractéristicas se denominan a largo plazo o segmentales,

En la siguiente subsección se describen los parámetros acústicos más comúnmente

utilizados para clasificación y detección de eventos acústicos.

2.3.1. Extracción de características acústicas para CEA

El principal objetivo del proceso de extracción de características es encontrar una transformación de la señal que logre capturar la información más importante y representativa de la misma, reduciendo su dimensionalidad. A continuación se describen distintas aproximaciones para la extracción de características acústicas para AEC. Como se ha mencionado antes, se pueden clasificar en dos grandos grupos: parámetros a corto plazo o a nivel de trama y parámetros a largo plazo o segmentales.

2.3.1.1. Características a corto plazo

En los primeros trabajos sobre parametrización de eventos acústicos de propuso el uso de características acústicas similares a las utilizadas para reconocimiento automático de habla o de locutor, como, por ejemplo, los parámetros mel-cepstrales, log-energías en banda, etc. Más recientemente se ha experimento con parámetros derivados de los descriptores de audio MPEG-7 [Kim et al., 2005], como el flujo espectral, centroide espectral, etc.

- **A. Coeficientes de log-energías en banda (FBEC, *Filter Bank Energy Coeficcients*):**

  El cálculo de las características FBEC es similar al cálculo de los MFCC (ver subsección 2.2.1.3), con la diferencia de que no se aplica la transformada discreta del coseno como paso final. Al igual que los MFCCs, primero, la señal de audio se divide en tramas (con una longitud típica de $20ms - 30ms$) con solapamiento de, usualmente, $10ms$ a $15ms$, usando una ventana de análisis de Hamming. Estas tramas se transforman al dominio de la frecuencia usando la transformada discreta de Fourier. A continuación, el espectro de magnitud se pasa a través de un banco de filtros triangulares a escala de frecuencias Mel definida en la ecuación 2.11. Por último se calcula el logaritmo de la energía de

la salida de cada filtro, obteniéndose los coeficientes de log-energías en banda (FBEC).

Las caractersíticas FBEC, permiten realizar un análisis espectro - temporal de la señal de audio; sin embargo, presenta el problema de la resolución, ya que el uso de una ventana de análisis pequeña implica una mejor resolución en el tiempo, a costa de una pobre resolución en la frecuencia; mientras que ventanas más grandes proporcinan una mejor resolución en la frecuencia y una pobre resolución en el tiempo. Por lo tanto, la selección del tamaño de la ventana es un parámetro crítico.

- **B. Coeficientes cepstrales en escala de frecuencia Mel (MFCC, *Mel Frequency Cepstral Coefficients*):**

  Es una de las características más usadas para el procesamiento de audio derivadas del análisis espectral y motivadas perceptualmente. La obtención y cálculo de los MFCCs se encuentra descrito en la subsubsección 2.2.1.3. Tal y como se expondrá con mayor profundidad en el capítulo 4, este tipo de parámetros acústicos puede no ser necesariamente el más apropiado para la tarea de clasificación y segmentación de eventos acústicos puesto que se diseñaron teniendo en cuenta la estructura espectral de la señal de voz, que es bastante diferente de la de los eventos acústicos.

- **C. Tasa de cruces por cero (ZCR, *Zero Crossing Rate*):**

  Es una característica temporal de audio que se define como el número de veces que la amplitud de la señal de audio cruza el valor de cero. En el dominio del tiempo los cruces por cero dan una medida del ruido de la señal. En la ecuación 2.16 se muestra el cálculo de la tasa de cruce por cero para la t-sima trama $x\,(n)$.

$$Z_t = \frac{1}{2}\sum_{n=0}^{N-1} \mid sign\,(x\,(n)) - sign\,(x\,(n-1)) \mid \tag{2.16}$$

Donde N es el número total de muestras de $x\,(n)$ y la función signo $sign\,(x\,(n))$ está definida en la ecuación 2.17.

$$sign\,(x\,(n)) = \begin{cases} +1, & x\,(n) \geq 0 \\ -1, & x\,(n) < 0 \end{cases} \tag{2.17}$$

- **D. Flujo espectral (F):**

  El flujo espectral es un indicativo de cuan rápido cambia el espectro de potencia de una señal, es decir, es una medida de la variación del espectro de potencia entre tramas sucesivas y se define mediante la ecuación 2.18. El flujo espectral puede usarse para determinar el timbre de una señal de audio.

$$F_t = \frac{1}{N} \sum_{k=0}^{Nf-1} \left[X_t\,(k) - X_{t-1}\,(k)\right]^2 \tag{2.18}$$

  Donde, $X_t\,(k)$ y $X_{t-1}\,(k)$ son los espectros de magnitud para la trama t y $t-1$, respectivamente; siendo Nf el número total de bins de frecuencia.

- **E. *Rolloff* espectral (F_{ro}):**

  Es una medida que define la frecuencia F_{ro} debajo de la cuál reside el 85 % del espectro de magnitud acumulado y está definido por la ecuación 2.19. Donde, $X_t\,(k)$ es el espectro de magnitud para la trama t. Este parámetro refleja la simetría del espectro.

$$\sum_{k=0}^{F_{ro}} \mid X_t\,(k) \mid = 0{,}85 \sum_{k=0}^{Nf-1} \mid X_t\,(k) \mid \tag{2.19}$$

- **F. Envolvente espectral de audio (ASE, *Audio Spectral Envelope*):**

  Este parámetro representa una descripción compacta del espectrograma y se obtiene como la suma de la energía del espectro de potencia dentro de una serie de bandas de frecuencia b, las cuales están logarítmicamente distribuidas entre dos límites de frecuencia (lo_b y hi_b) y se encuentra definido en la ecuación 2.20.

$$ASE_t(b) = \sum_{k=lo_b}^{hi_b} P(k) \tag{2.20}$$

donde $P(k)$ son los coeficientes del espectro de potencia, lo_b (hi_b) son, respectivamente, el límite inferior y superior de la banda b. ASE presenta la desventaja de ser dependiente del nivel amplitud de la señal, por lo que si se extrae a partir del mismo sonido pero con diferentes factores de amplificación, los vectores de características resultantes diferirán significativamente.

- **G. Centroide espectral de audio (ASC, *Audio Spectrum Centroid*):**

  Esta característica describe el centro de gravedad del espectro. Se usa para describir el timbre de una señal de audio. También indica si las bajas o altas frecuencias son dominantes en el espectro de potencia y puede ser considerado como una aproximación de la intensidad (*sharpness*) perceptual de la señal.

- **H. Espectro disperso de audio (ASS, *Audio Spectrum Spread*):**

  Esta característica describe cómo se distribuye el espectro de potencia de la señal de audio alrededor de su centro. Un valor bajo indica que el espectro se encuentra concentrado alrededor de su centro; mientras que un valor alto refleja una distribución de la potencia a través de un amplio rango de frecuencias.

- **I. Planicidad del espectro de audio (ASF, *Audio Spectrum Flatness*):**

  Este parámetro se usó originalmente para calcular el umbral de enmascaramiento de ruido en codificación de voz. Esta característica permite caracterizar el espectro de audio y permite cuantificar el grado en que un sonido se parece a un ruido o un tono.

2.3.1.2. Integración temporal de características

La integración temporal de características consiste en el proceso de combinar un conjunto de parámetros extraídos a corto plazo (nivel de trama) y contenidos en una

ventana de longitud dada en un único vector de características (nivel de segmento), de modo que se capture la información de la evolución temporal de los parámetros dentro de dicha ventana [Meng et al., 2007]. Usualmente se utilizan como características a corto plazo las mencionadas en la subsección 2.2.1 (principalmente, FBEC o MFCC), puesto que se ha observado que la información sobre su comportamiento dinámico puede ser relevante para distinguir unos eventos acústicos de otros. La integración temporal también permite una reducción de los datos, de forma que la información relevante puede ser resumida eficientemente en poco espacio [Meng et al., 2007].

La integración temporal de características puede expresarse como una secuencia de T características a corto plazo de dimensión D_x, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ que se divide en k segmentos, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_K\}$ de la siguiente manera,

$$\mathbf{y}_k = f\left(\mathbf{x}_{k \cdot H_s}, \mathbf{x}_{k \cdot H_s + 1}, ..., \mathbf{x}_{k \cdot H_s + L_s - 1}\right) \tag{2.21}$$

donde L_s es el tamaño del segmento, H_s es el desplazamiento del segmento, ambos definidos en tramas, y f es la función que combina los parámetros a nivel de trama en parámetros a nivel de segmento. En este trabajo de tesis se han considerado dos formas distintas de combinación: estadísticos de las características a corto plazo y coeficientes de banco de filtros (FC, *Filterbank Coefficients*).

- **A. Estadísticos de las características a corto plazo:**

  Esta aproximación consiste en aplicar la integración temporal sobre un conjunto de parámetros a corto plazo dentro una ventana deslizante con una duración de varios segundos y calcular sus estadísticos (media, desviación estándar, simetría, etc.) sobre dicha ventana [Meng et al., 2007], [Mejia Navarrete et al., 2011], [Zhang and Zhou, 2004].

$$\mathbf{z}_k = \left[\begin{array}{ccc} mean\left(\mathbf{y}_k\right) & std\left(\mathbf{y}_k\right) & skewness\left(\mathbf{y}_k\right) \end{array} \right] \tag{2.22}$$

  donde $mean\left(\mathbf{y}_k\right)$, $std\left(\mathbf{y}_k\right)$ y $skewness\left(\mathbf{y}_k\right)$ son la media, desviación estándar y simetría calculados sobre los parámetros contenidos en el k-simo segmento.

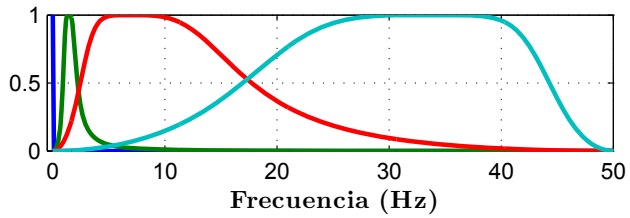


Figura 2.6: Banco de filtros predefinido **U** usado para parametrización FC.

- **B. Coeficientes de banco de filtros (FC):**

  Inicialmente, la aproximación FC fue aplicada en tareas de clasificación de géneros musicales y audio en general [Meng et al., 2007], [McKinney and Breebaart, 2003]. FC ayuda a capturar la estructura dinámica de las características a corto plazo [Meng et al., 2007], calculando su espectro de modulación a través del uso de un banco de filtros (**U**) inspirado en el sistema auditivo humano. Habitualmente, dicho banco consta de cuatro filtros correspondiente a las siguientes bandas de frecuencia [McKinney and Breebaart, 2003]:

  - Filtro 1: 0 Hz (filtro DC)
  - Filtro 2: 1 - 2 Hz (energía de modulación)
  - Filtro 3: 3 - 15 Hz (energía de modulaciín)
  - Filtro 4: 20 - 43 Hz (aspereza *roughness* perceptual)

  En esta aproximación se calcula el periodograma de cada dimensión de las características a corto plazo contenidas en el k - simo segmento $\mathbf{y}_k$ y la información contenida en dicho periodograma se resume mediante la obtención de la potencia en las diferentes bandas de frecuencia del banco de filtros predefinido (**U**),

$$\mathbf{z}_k = \mathbf{P}_k \mathbf{U} \tag{2.23}$$

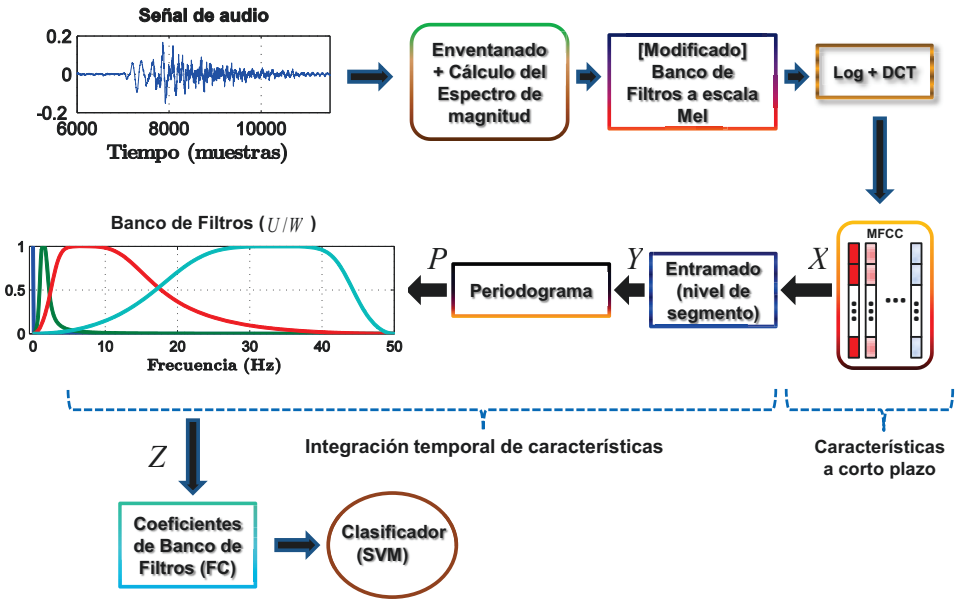


Figura 2.7: Diagrama de bloque del proceso de extracción de características.

donde $\mathbf{P}_k$ representa los periodogramas de la secuencia de coeficientes a corto plazo perteneciente al k - simo segmento, $\mathbf{U}$ es la magnitud de la respuesta en frecuencia del banco de filtros predefinido y $\mathbf{z}_k$ es el vector de características final, que son la entrada al clasificador correspondiente. En la figura 2.7, se muestra el diagrama de bloques para esta parametrización.

Capítulo 3

Eliminación de ruido con NMF para aplicación en la mejora de voz y el reconocimiento automático de habla

Los sistemas de reconocimiento automático de habla entrenados en condiciones controladas (es decir, usando voz limpia) degradan significativamente su funcionamiento en condiciones reales, especialmente por la influencia del ambiente acústico, como, por ejemplo, la presencia de ruido aditivo que afecta significativamente la calidad de la voz. El ruido en la señal de la voz es un problema común en muchas aplicaciones como son el uso de reconocedores de voz en comunicaciones telefónicas fijas y móviles, etc. De hecho, uno de los principales problemas a los que tiene que enfrentarse los sistemas de RAH es el reconocimiento en entornos ruidosos dado que la capacidad de comprensión de la máquinas aún está lejos de parecerse al de los seres humanos en estas condiciones. Por este motivo y con objeto de mejorar las prestaciones de los sistemas de RAH, en ocasiones, se utiliza una etapa de preprocesamiento para mejora de la señal de voz mediante la eliminación de ruido.

En este capítulo concentramos nuestros esfuerzos en la mejora de la señal de voz para el reconocimiento automático del habla. En la literatura, se han propuesto

varios métodos para reducir la influencia del ruido. Entre ellos, destacan la técnica del filtro de Wiener [Scalart and Vieira, 1996] y el método convencional de sustracción espectral (SS, *Spectral Subtraction*), [Berouti et al., 1979] que consiste en la resta de una estima del espectro del ruido del espectro de la señal de voz ruidosa. Ambos métodos producen una señal más inteligible; sin embargo, tienen la desventaja de producir ruido residual molesto para el oyente (y para el reconocedor de voz) llamado ruido musical.

Recientemente, la factorización de matrices no-negativas se ha utilizado satisfactoriamente en diferentes áreas relacionadas con el procesamiento de voz, incluyendo eliminación de ruido en señales de voz [Wilson et al., 2008], separación de sonidos [Virtanen, 2007], separación de locutores [Schmidt and Olsson, 2006], y extracción de características [Schuller et al., 2010]. NMF proporciona una forma de descomponer una señal en una combinación convexa de bloques de construcción no negativos (llamados también vectores base) mediante la minimización de una función de coste dada. Funciones de coste típicas son la distancia euclídea y la divergencia de Kullback - Leibler (KL). En la sección 2.1 del capítulo 2 puede encontrase una descripción de los fundamentos matemáticos principales de NMF y la solución del proceso de factorización.

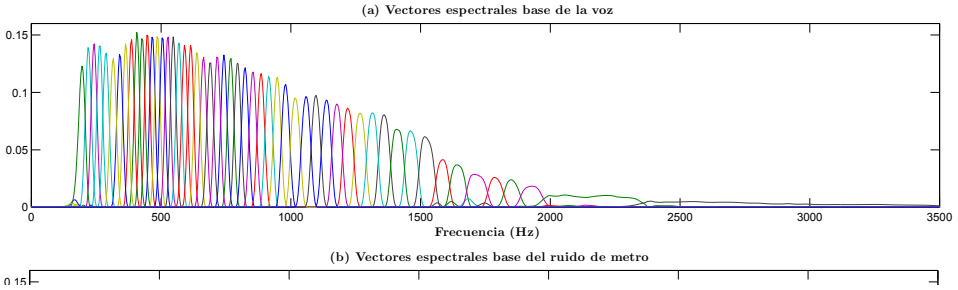
En los trabajos previamente mencionados, se ha mostrado que NMF es capaz de separar fuentes de sonido cuando sus bloques de construcción son suficientemente distintos como es el caso de la voz y ruido. En este capítulo proponemos usar un método basado en NMF para la eliminación de ruido en señales de voz, que está basado en el desarrollado en [Wilson et al., 2008] para la tarea de la mejora de la voz. La técnica en [Wilson et al., 2008] se sustenta en el desarrollo de un modelo a priori de la voz y ruido, y por lo tanto, asume un conocimiento a priori del tipo de ruido que contamina la voz. En contraste, nuestro método no usa informacin explícita acerca del ruido, ya que el modelo de ruido se estima a partir de los segmentos de silencio/ruido de las elocuciones a reconocer, obtenidos con la ayuda de un detector de actividad de vocal (VAD, *Voice Activity Detector*). Mientras que en [Wilson et al., 2008] sólo

se presentan resultados para mejora de voz, en este capítulo se muestra también el funcionamiento del método en un sistema de reconocimiento automático del habla.

Por otro lado, varios estudios recientes indican que puede resultar beneficioso realizar un control explícito del grado de dispersión (*sparsity*) en las descomposiciones NMF para el caso de separación de sonidos y locutores. En este sentido, el método para separación de locutor propuesto en [Schmidt and Olsson, 2006] introduce un término de penalización en el algoritmo NMF con distancia euclídea que permite controlar dicho grado de dispersión de la solución. Sin embargo, en procesamiento de voz, se han reportado mejores resultados usando NMF con divergencia KL [Schuller et al., 2010], [Virtanen, 2007]. Por esta razón, nuestro método basado en NMF para eliminar ruido en señales de voz además combina el uso de la divergencia KL con restricciones de dispersión siguiendo el procedimiento general descrito en [Cichocki et al., 2006].

3.1. Eliminación de ruido en señales de voz usando NMF

Los métodos basados en factorización de matrices no negativas permiten la eliminación (al menos, parcial) de ruido en señales de voz bajo la hipótesis de que las señales de voz ruidosas son una mezcla aditiva de dos fuentes suficientemente distintas: voz y ruido. NMF se aplica al espectro de magnitud de la señal de voz ruidosa, $|V_{\text{mix}}|$, de tal forma que puede ser expresado como la combinación lineal de varias componentes diferentes, aquellos que solo representan el espectro de magnitud de la voz (W_{speech}) y aquellos que solo representen al espectro de magnitud del ruido (W_{noise}). Estas componentes se denominan vectores espectrales base (SBV, *Spectral Basis Vectors*) y pueden interpretarse como los bloques constructivos de la voz y el ruido. En la figura 3.1, se muestran los vectores espectrales base para la voz y el ruido de metro, donde claramente observamos que son distintas. Además podemos observar que la distribución de los vectores espectrales base de la voz (en nuestro



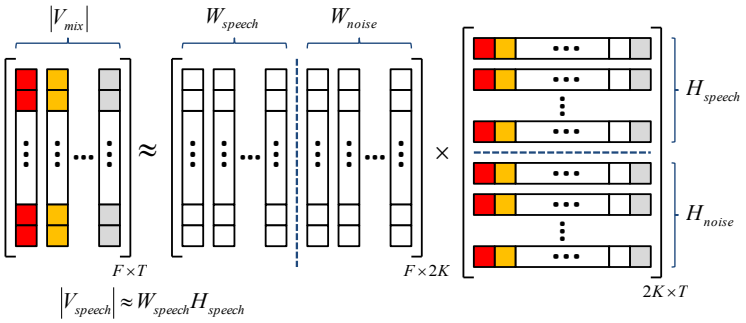


Figura 3.2: Representación NMF de señales de voz ruidosas.

caso 50 SBV) se asemeja al banco de filtros auditivo a escala de frecuencia Mel, concentrando mayor cantidad de vectores espectrales (filtros) en la región de baja de frecuencia que en la de alta frecuencia.

La representación NMF de una señal de voz ruidosa se muestra en la figura 3.2, en la que los SBVs de la voz (W_{speech}) y sus correspondientes coeficientes de activación (H_{speech}) pueden usarse para reconstruir la señal de voz limpia ($|V_{\text{speech}}| \approx W_{\text{speech}} H_{\text{speech}}$), mientras que los SBVs del ruido (W_{noise}) y sus correspondientes coeficientes de activación (H_{noise}) pueden usarse para reconstruir la señal de ruido ($|V_{\text{noise}}| \approx W_{\text{noise}} H_{\text{noise}}$).

El proceso de mejora de la señal de voz consiste de dos etapas: entrenamiento y eliminación de ruido propiamente dicho, tal y como se detalla a continuación.

3.1.1. Etapa de entrenamiento.

En esta etapa se determinan los vectores espectrales base que representan a las señales de voz y ruido, mediante la aplicación de NMF sobre los datos de entrenamiento de voz limpia (sin ruido de fondo) y ruido. Para ello, primero, se calcula el espectro de magnitud de la voz limpia ($|V_{\text{speech}}|$) y del ruido ($|V_{\text{noise}}|$). A continuación, se minimiza la divergencia de Kullback-Leibler entre el espectro de magnitud y sus correspondientes matrices factorizadas ($W_{\text{speech}}H_{\text{speech}}$) y ($W_{\text{noise}}H_{\text{noise}}$) usando las reglas de aprendizaje dadas en la ecuación 2.4 de la sección 2.1 del capítulo 2. De esta forma, en esta etapa de entrenamiento se han obtenido los vectores espectrales base de la voz y el ruido, que están contenidos en las matrices W_{speech} y W_{noise}, respectivamente, y que se utilizan en la siguiente etapa como modelos de voz y ruido. En la figura 3.3 se muestra este proceso de determinación de los modelos de voz y ruido usando NMF. Es importante destacar que dado que NMF es un algoritmo iterativo, es importante realizar un adecuado proceso de inicialización de las matrices factorizadas. En la subsección 2.1.2 del capítulo 2 y en la sección experimental de este capítulo se encuentran más detalles sobre la inicialización del algoritmo para esta tarea.

En la práctica, para construir el modelo de voz, se asume que se encuentran disponibles suficientes datos de voz limpia. No obstante, para el modelo de ruido, se han explorado dos alternativas diferentes:

- *Offline Noise Data (OND).* En esta aproximación se supone un conocimiento a priori del tipo de ruido como en [Wilson et al., 2008]. Por lo tanto, para cada tipo de ruido considerado, se entrena su modelo de ruido correspondiente usando todos los datos de ruido disponibles en la base de datos. Esta aproximación proporciona un límite superior al rendimiento del método de eliminación de

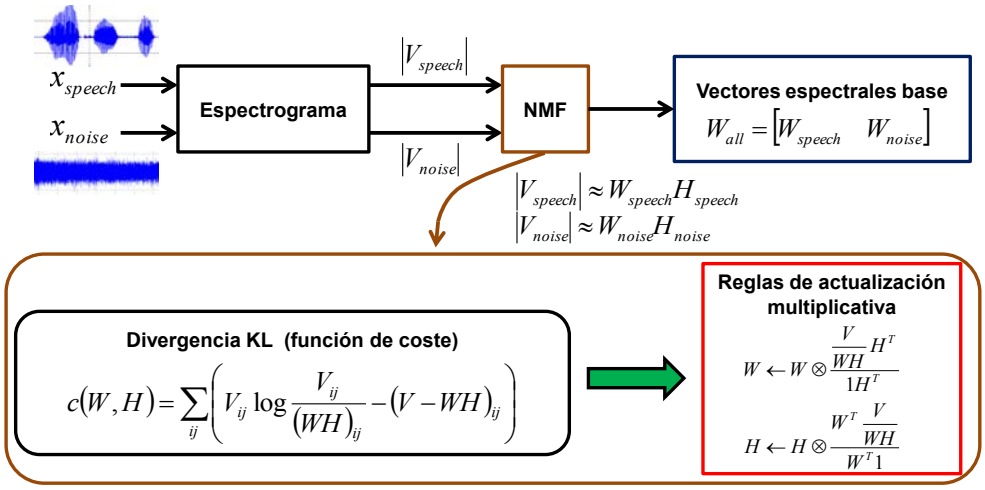


Figura 3.3: Diagrama de bloque para la obtención de los modelos de voz y ruido usando NMF.

ruido basado en NMF.

- *Voice Activity Detector Noise Data (VADND).* En esta aproximación, se usa un detector de actividad de voz (VAD) con la finalidad de determinar explícitamente las zonas de las elocuciones de voz ruidosas que contienen solo ruido. El modelo de ruido se construye a partir de estos segmentos, por lo que es necesario entrenar un modelo de ruido para cada elocución utilizando únicamente los datos de ruido contenidos en dicha elocución. Esta alternativa es computacionalmente más costosa, pero evita la necesidad de tener un conocimiento a priori del tipo de ruido, lo cual no siempre es posible.

3.1.2. Etapa de eliminación de ruido.

En la etapa de eliminación de ruido en sí (*denoising*), se supone que W_{speech} y W_{noise} son vectores espectrales base adecuados para describir la voz y ruido. Bajo esta suposición estos vectores no necesitan ser reentrenados, por lo que mantienen fijos y se concatenan para formar un único conjunto de SBVs denominado W_{all}, que pueden considerarse como modelo de la voz ruidosa al contener componentes de

voz y ruido. Dado el espectro de magnitud de la señal de voz ruidosa ($|V_{\text{mix}}|$), se calcula su factorización $|V_{\text{mix}}| \approx W_{\text{all}} H_{\text{all}}$ minimizando la divergencia KL entre $|V_{\text{mix}}|$ y $(W_{\text{all}} H_{\text{all}})$, actualizando únicamente la matriz de activaciones H_{all}.

Una de las novedades del método propuesto en este capítulo es la utlilización en el algoritmo NMF de la distancia KL junto con una serie de factores que controlan el grado de dispersión (*sparseness*) de las matrices factorizadas, puesto que presuponemos que una representación dispersa de los datos puede beneficiar el proceso de eliminación del ruido de la señal de voz. Tal y como se comenta en el capítulo 2, para este propósito hemos seguido la aproximación propuesta en [Cichocki et al., 2006] y [Cichocki et al., 2009] para la función de coste KL y las reglas de aprendizaje modificadas indicadas en la ecuación (2.5) de dicho capítulo. Puesto que la matriz W_{all} permanece fija en esta etapa, las reglas de actualización sólo se aplican sobre la matriz H_{all} con los parámetros (ω y α_h) adecuados (ver sección 3.2).

Una vez recalculada la matriz H_{all}, el espectro de magnitud de la voz regenerada se estima como $|V_{\text{speech}}| \approx W_{\text{speech}} H_{\text{speech}}$, siendo H_{speech} las filas de H_{all} correspondientes a los coeficientes de W_{speech}. Finalmente, el espectrograma de la voz regenerada se recupera usando la fase del espectro de la señal de voz ruidosa original y la correspondiente señal de voz en el dominio del tiempo se obtiene mediante la aplicación del método de *overlap-add* convencional. El proceso completo de la eliminación de ruido en señales de voz se muestra en el diagrama de bloques de la figura 3.4.

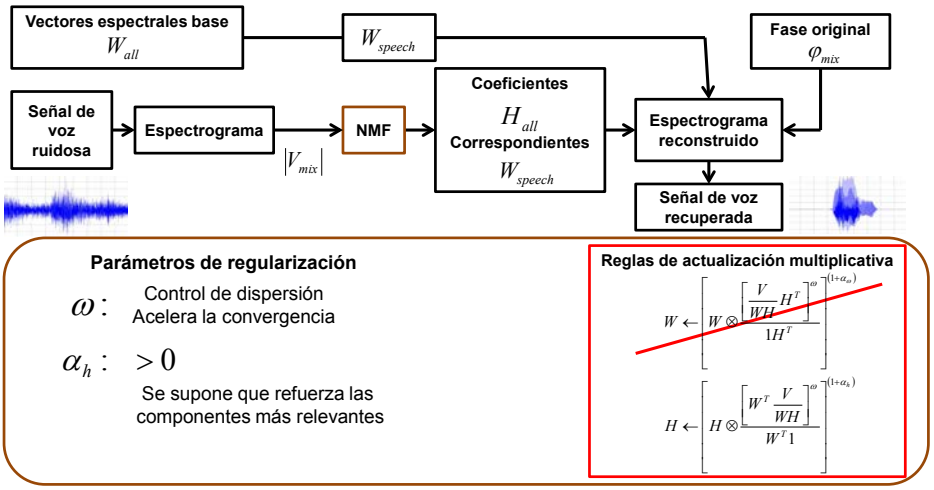


Figura 3.4: Diagrama de bloque del proceso de eliminación de ruido en señales de voz usando NMF.

3.2. Aplicación a la mejora de la señal de voz.

En esta sección se describe la experimentación realizada par evaluar los métodos propuestos basado en NMF (*OND* y *VADND*) en la tarea de eliminación de ruido en señales de voz con aplicación a la mejora de voz.

3.2.1. Base de datos y protocolo experimental.

La base de datos usada para los experimentos de mejora de la señal de voz es la AURORA-2 [Hirsch and Pearce, 2000], que está creada a partir de la base de datos TIDIgits y contiene las grabaciones de 52 hombres y 52 mujeres adultos norteamericanos pronunciando secuencias de dígitos en inglés. Originalmente la base de datos fue grabada en condiciones limpias a una frecuencia de muestreo es $8KHz$ y subsecuentemente contaminada con varios tipos de ruido a diferentes relaciones señal a ruido (SNR, *Signal-to-Noise Ratio*). El principio y fin de las elocuciones fue determinado automáticamente utilizando el detector de actividad vocal del estándar de codificación G.729.

Para entrenar los vectores espectrales base de la voz se usaron 420 archivos de voz limpia pertenecientes al conjunto de entrenamiento de la base de datos AURORA-2. En el método *OND*, los modelos de ruido se entrenaron usando las correspondientes ficheros de ruido incluidos en la base de datos. En la aproximación *VADND*, el modelo de ruido para cada elocución fue entrenado usando las tramas de solo ruido del principio de cada elocución determinadas por el detector de actividad vocal del del estándar de codificación G.729. Con la finalidad de realizar el estudio de la subsección 3.2.2 se usaron 6006 archivos de voz del conjunto de prueba denominado TEST A de la base de datos AURORA-2, que corresponde a diferentes versiones ruidosas de 1001 archivos contaminados con ruido de coche a SNRs desde $-5dB$ hasta $20dB$ con pasos de $5dB$. Finalmente, los experimentos finales (subsección 3.2.3) se realizaron utilizando 24024 archivos del conjunto de prueba TEST A que contienen voz contaminada con ruidos de metro, voces, coche y sala de exposiciones a las SNRs mencionadas anteriormente.

Para evaluar el funcionamiento de los métodos propuestos, se usó la medida de evaluación perceptual de la calidad de la voz (PESQ, *Perceptual Evaluation of Speech Quality*), recomendada por la ITU-T para valorar la calidad de la voz. PESQ es capaz de predecir la calidad subjetiva con buena correlación con la percepción humana en un amplio rango de condiciones (ruido, filtrado, distorsiones debidas a la codificación, etc.) [Lee and Seung, 2002] y usa una escala de 5 puntos indicando como 1 el peor valor y 5 el mejor valor. Los valores PESQ se calcularon usando el código disponible en [Hu and Loizou, 2011] y considerando la señal de voz limpia como referencia. Los resultados se presentan en términos de la siguiente medida de eficiencia relativa,

$$Ef_{\text{rel}} = \frac{PESQ_{\text{denoised}} - PESQ_{\text{noisy}}}{PESQ_{\text{noisy}}} \times 100\,\% \tag{3.1}$$

donde $PESQ_{\text{noisy}}$ y $PESQ_{\text{denoised}}$ son los valores $PESQ$ antes y después de la aplicación del proceso de mejora de la voz, respectivamente. Un incremento en es-

te valor implica una mejora de la calidad y una disminución significa que la voz procesada ha sufrido una degradación con respecto a la voz ruidosa correspondiente.

3.2.2. Estudio de la influencia de los parámetros NMF.

En este apartado se muestran los resultados obtenidos en un conjunto de experimentos realizados con la finalidad de estudiar el impacto de varios parámetros del algoritmo NMF en la calidad de la voz mejorada. Los parámetros considerados fueron la longitud de la ventana de análisis y el desplazamiento de trama usado para el cálculo de los espectrogramas, el número de vectores espectrales base y los valores de los parámetros de regularización, ω y α_h.

En todos los casos, NMF fue inicializado usando el esquema de inicialización descrito en la subsección 2.1.2 del capítulo 2. En particular, en cada experimento se ejecutó 10 veces el algoritmo NMF de mínimos cuadrados alternantes (ALS NMF, *Alternating Least Squares NMF*) [Cichocki et al., 2009], de tal manera que la factorización que produjo la distancia euclídea menor entre la matriz original V y (WH) se escogió para inicializar el algoritmo. Posteriormente, estas matrices iniciales se refinaron mediante la minimización de la divergencia KL con restricciones de dispersión utilizando la regla de aprendizaje de la ecuación 2.5. Antes de estudiar el efecto de los parámetros antes mencionados, se realizaron una serie de experimentos para observar el funcionamiento de la distancia euclídea y la divergecia KL como funciones de coste, concluyendo que la distancia euclídea daba lugar a peores resultados en términos del PESQ que la distancia KL, corroborando las conclusiones obtenidas en estudios previos [Virtanen, 2007], [Wilson et al., 2008], [Schuller et al., 2010].

A continuación se resumen los principales experimentos y resultados obtenidos:

- Con respecto al número de SBVs, se probaron diversos valores desde 10 hasta 80 en pasos de 10 con una longitud de ventana de $20ms$ y desplazamiento de trama de 2,5ms. Los resultados mostraron que la calidad de la voz procesada se degradaba cuando se usaba un número pequeño de SBVs (debajo de 30),

mientras que los mejores valores de PESQ se obtuvieron en el rango de 40 a 80 SBVs. Estos resultados indican que para una adecuada representación de la señal de voz en NMF, parece necesario considerar más de 30 SBVs. Estos resultados se muestran en la figura 3.5(a).

- El desplazamiento de la trama, mostrado en la figura 3.5(b), se estudió en el rango desde $1ms$ hasta $10ms$ con una longitud de ventana de $20ms$ y 50 SBVs. En este caso, la calidad de la voz mejoraba cuando el desplazamientos de trama disminuía. Los mejores valores de PESQ se encontrados en el rango entre $1ms$ y $5ms$.

- La longitud de ventana se varió desde $10ms$ hasta $45ms$ con pasos de $5ms$ con un desplazamiento de trama de $2{,}5ms$ y 50 SBVs, como se muestra en la figura 3.5(c). A partir de este conjunto de experimentos, se observó que los valores de PESQ aumentaban conforme la longitud de la ventana se incrementaba, obteniendo los mejores resultados en el rango entre $25ms$ hasta $45ms$.

- Con respecto a los factores que controlan la dispersión de la matriz de activaciones, se realizaron varios experimentos variando α_h desde 0 hasta 1,2 y ω desde 1 hasta 2,5 utilizando el procedimiento de búsqueda por rejilla *grid search*. Los resultados para la aproximación *OND* se muestran en la figura 3.6, indicando con una celda de color rojo una mejora en la calidad y con color azul una degradación de la calidad con respecto a la señal ruidosa. Para el método VADND se observaron tendencias similares.

  Como se puede observar, los valores de PESQ empeoran cuando no se utiliza ninguna regularización (este caso corresponde a $\alpha_h = 0$ y $\omega = 1$ mostrado en la figura 3.6) o cuando se incrementan de forma conjunta ambos parámetros (por ejemplo $\alpha_h = 1{,}2$ y $\omega = 2{,}5$). Sin embargo, cuando solo alguno de los valores de estos factores se incrementan, la calidad de la voz mejora, encontrando el mejor rendimiento para la combinación de α_h y ω mostrados con celdas de color rojo en la figura 3.6 (por ejemplo cuando $\alpha_h = 1$ y $\omega = 1$).

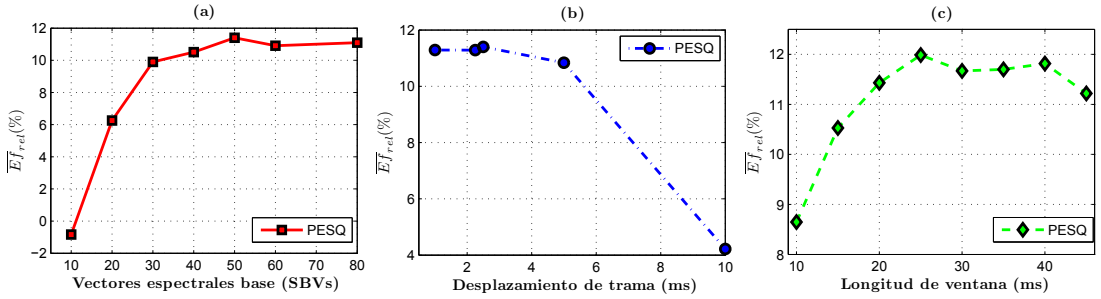


Figura 3.5: Influencia de varios parámetros de NMF en el proceso de eliminación de ruido. a) Número de vectores espectrales base (SBVs), b) Desplazamiento de trama y c) Longitud de la ventana de análisis.

3.2.3. Resultados experimentales.

En esta subsección se compara el rendimiento de las dos aproximaciones de eliminación de ruido en señales de voz basado en NMF (*OND* y *VADND*) con el método convencional de substracción espectral (SS) utilizando la medida del PESQ relativo. De acuerdo a los resultados obtenidos en la subsección previa, para los métodos basados en NMF se usó una longitud de ventana de $40ms$, un desplazamiento de trama de $2{,}5ms$, 50 SBVs, $\omega = 1$ y $\alpha_h = 1$. Para realizar una adecuada comparación, para SS se utilizaron los mismos valores para la longitud de ventana y desplazamiento de trama.

La figura 3.7 muestra la medida relativa del PESQ con respecto a la señal ruidosa para los cuatro tipos de ruido considerados a varias SNRs. Para los ruidos de metro y voces, los dos métodos basados en NMF superan a SS para valores de SNR bajos y medios ($-5dB$ - $10dB$). Para el ruido de sala de exposiciones, el rendimiento del método *OND* es superior a SS en valores de SNR bajos y medios ($-5dB$ - $15dB$) mientras que *VADND* supera a SS para SNR medias ($5dB$ - $15dB$).

Para el caso del ruido de coche, *OND* es mejor que SS en SNR baja ($-5dB$ - $5dB$); sin embargo SS supera a *OND* para SNR por encima de $15dB$. Para este ruido, *VADND* produce resultados ligeramente peores que SS para SNRs baja y

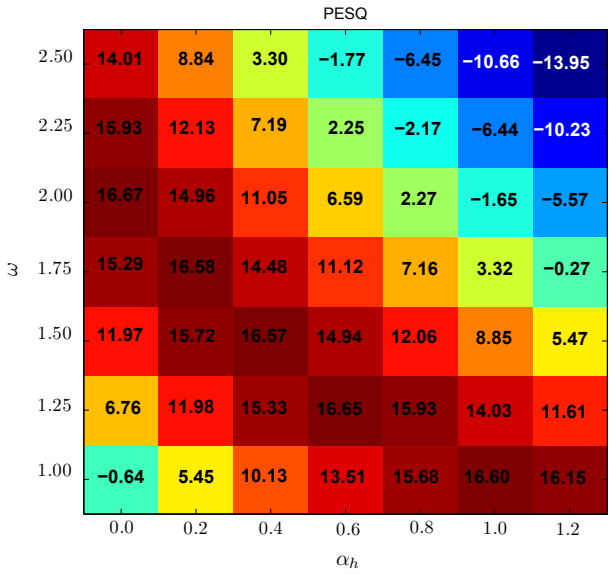


Figura 3.6: Eficiencia relativa del PESQ en función de los parámetros de regularización α_h y ω.

media ($-5dB$ - $10dB$), mostrando mayores degradaciones con respecto a SS para SNRs por encima de los $15dB$.

En general, los resultados muestran que *OND* y *VADND* son más adecuados que SS para rangos de SNR bajos y medios. Sin embargo, para la SNR más alta ($20dB$), el método SS produce mejores resultados que las técnicas basadas en NMF, no siendo tan notoria esta mejora para el caso del ruido de sala de exposiciones.

En la tabla 3.1 se muestra la eficiencia relativa del PESQ promediado sobre los cuatro tipos de ruido y las SNRs consideradas junto con sus respectivos intervalos de confianza al 95 % para los diferentes métodos de eliminación de ruido en señales de voz. Para los ruidos de metro y voces las técnicas basadas en NMF (*OND* y *VADND*) superan a SS, siendo esta mejora estadísticamente significativa; sin embargo, para el ruido de coche, las técnicas *OND* y SS muestran resultados similares y significativamente mejores que la técnica *VADND*. Para el ruido de sala de exposiciones, las técnicas *VADND* y SS muestran en promedio resultados similares, pero son superados por la técnica *OND*.

Tabla 3.1: Eficiencia relativa PESQ [%] promediado sobre los cuatro tipos de ruido.

Ruido	**OND**	**VADND**	**SS**
Metro	$19{,}51 \pm 1{,}00$	$15{,}61 \pm 0{,}92$	$11{,}75 \pm 0{,}81$
Voces	$9{,}94 \pm 0{,}76$	$8{,}04 \pm 0{,}69$	$6{,}57 \pm 0{,}63$
Coche	$16{,}45 \pm 0{,}94$	$13{,}67 \pm 0{,}87$	$16{,}54 \pm 0{,}94$
Sala de exposiciones	$16{,}69 \pm 0{,}94$	$11{,}22 \pm 0{,}80$	$11{,}11 \pm 0{,}80$

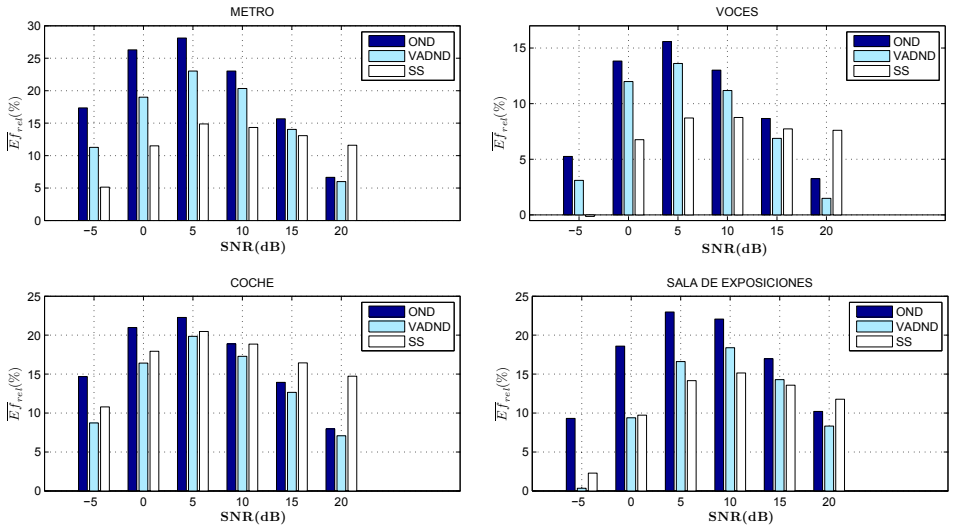


Figura 3.7: Medida relativa PESQ para las técnicas SS, *OND* y *VADND*.

Con respecto a la comparación entre *OND* y *VADND*, en la tabla 3.1 se puede observar que la calidad de la señal de voz procesada es mejor con *OND* en todos los casos. Este resultado es esperable porque *OND* usa mas información que *VADND* en el proceso de eliminación de ruido. De hecho, necesita conocer el tipo de ruido (no sólo el valor de SNR) presente en las elocuciones ruidosas. Sin embargo, *VADND* es capaz de efectivamente eliminar el ruido en la señal de voz usando solo la información contenida en los segmentos de silencio/ruido de cada elocución, especialmente en los casos de ruido de metro y voces.

3.3. Aplicación al reconocimiento automático de habla.

En esta sección se presenta la evaluación de las técnicas propuestas en la tarea del reconocimiento automático de habla ruidosa. En este caso, en primer lugar las señales ruidosas son procesadas en la etapa de eliminación de ruido usando las técnicas basadas en NMF (*OND* y *VADND*) descritas en la sección 3.1 y a continuación. las señales regeneradas se alimentan a un sistema de reconocimiento automático de habla basado en modelos ocultos de Markov.

3.3.1. Base de datos y protocolo experimental

Los experimentos se realizaron sobre la base de datos AURORA-2 [Hirsch and Pearce, 2000], la misma que se utilizó para la tarea de mejora de la señal de voz descrita en la sección 3.2.1. El reconocedor está basado en modelos ocultos de Markov y fue implementado usando el paquete software HTK (Hidden Markov Model Toolkit) [Hirsch and Pearce, 2000] con la configuración incluida en el protocolo experimental estándar de la base de datos. Los modelos acústicos se obtuvieron a partir del conjunto de entrenamiento de la base de datos formado por datos limpios (sin ruido aditivo), mientras que los archivos de prueba correspondían al conjunto completo TEST A de la base de datos. Los resultados se muestran en términos de la tasa de reconocimiento (*RR*, *Recognition Rate*).

Para las extracción de las características acústicas se utilizaron diferentes escalas de frecuencia (Mel, Bark y ERB). En todos los casos, se calcularon doce coeficientes cepstrales cada $10ms$ usando una ventana de análisis de Hamning de $25ms$ y 23 bandas espectrales a las diferentes escalas antes mencionadas.

La log-energía de cada trama y sus correspondientes coeficientes de primera derivada y aceleración también fueron calculados y concatenados a los parámetros estáticos. De esta forma, los vectores de característicos constaban de 39 componentes. Finalmente, se eliminó la media de las componentes de los paraámetros acústicos

mediante la aplicación de la técnica de normalización de la media cepstral (CMN, *Cepstral Mean Normalization*).

3.3.2. Resultados experimentales

La tabla 3.2 muestra las tasas de reconocimiento promediadas sobre todos los SNRs para cada tipo de ruido usando las tres escalas de frecuencia consideradas (Mel, Bark y ERB) incluyendo los intervalos de confianza al 95 %, para los dos métodos basado en NMF, SS y el sistema base (sin eliminación de ruido). Para cada una de la técnicas SS, *OND* y *VADND* se usaron los mismos parámetros de configuración que en el proceso de mejora de la señal de voz, excepto para la longitud de la ventana de análisis que fue fijada a $25ms$ y los parámetros de regularización que, tras una experimentación preliminar, se establecieron en $\alpha_h = 0{,}2$ y $\omega = 1{,}25$. Se puede observar que *OND* y *VADND* superan a la substracción espectral para todos los ruidos, excepto para el de coche en el que los resultados son similares. Esta tendencia se repite en las diferentes escalas; sin embargo, las mejores tasas se obtienen con la escala de frecuencia ERB, siendo estos resultados estadísticamente significativos para una confianza del 95 %. En el caso del ruido de voces, la mejora relativa sobre el sistema base es menor debido a que, en este caso, los vectores base del ruido y la voz son muy similares.

La figura 3.8 muestra las tasas de reconocimiento obtenidas por tipo de ruido y SNR para los dos métodos basados en NMF, sustracción espectral y el sistema base (sin proceso de eliminación de ruido). Por brevedad, sólo se muestran los resultados para el caso de la escala de frecuencia ERB, ya que fue la que produjo mejores prestaciones en media.

Como se puede observar, para los ruidos de metro, voces y sala de exposiciones, las dos técnicas basadas en NMF logran mejores resultados que SS y el sistema base para SNRs bajas y medias (desde $-5dB$ hasta $10dB$). Para SNRs altas, todos los algoritmos presentan un comportamiento similar excepto para SS y el ruido de voces. En este caso la tasa de reconocimiento obtenida por SS en menor que con las

Tabla 3.2: Tasa de reconocimiento promedio [%] para los cuatro tipos de ruido usando diferentes escalas de frecuencia.

Escala de Frecuencia	Ruido	OND	VADND	SS	Base
ERB	**Metro**	$77{,}10 \pm 1{,}06$	$76{,}59 \pm 1{,}07$	$73{,}89 \pm 1{,}11$	$65{,}28 \pm 1{,}20$
	Voces	$70{,}16 \pm 1{,}16$	$69{,}69 \pm 1{,}16$	$65{,}35 \pm 1{,}20$	$66{,}83 \pm 1{,}19$
	Coche	$75{,}12 \pm 1{,}09$	$74{,}81 \pm 1{,}10$	$75{,}72 \pm 1{,}08$	$63{,}86 \pm 1{,}22$
	Sala de exposiciones	$71{,}62 \pm 1{,}14$	$70{,}52 \pm 1{,}15$	$68{,}83 \pm 1{,}17$	$62{,}23 \pm 1{,}23$
Mel	**Metro**	$73{,}53 \pm 1{,}12$	$73{,}24 \pm 1{,}12$	$71{,}44 \pm 1{,}14$	$61{,}96 \pm 1{,}23$
	Voces	$68{,}08 \pm 1{,}18$	$67{,}78 \pm 1{,}18$	$63{,}66 \pm 1{,}22$	$65{,}38 \pm 1{,}20$
	Coche	$72{,}09 \pm 1{,}13$	$71{,}74 \pm 1{,}14$	$74{,}19 \pm 1{,}11$	$61{,}64 \pm 1{,}23$
	Sala de exposiciones	$68{,}40 \pm 1{,}18$	$67{,}24 \pm 1{,}19$	$65{,}62 \pm 1{,}20$	$59{,}10 \pm 1{,}24$
Bark	**Metro**	$73{,}89 \pm 1{,}11$	$74{,}95 \pm 1{,}10$	$72{,}51 \pm 1{,}13$	$64{,}62 \pm 1{,}21$
	Voces	$65{,}35 \pm 1{,}20$	$69{,}85 \pm 1{,}16$	$64{,}29 \pm 1{,}21$	$67{,}58 \pm 1{,}18$
	Coche	$75{,}72 \pm 1{,}08$	$73{,}73 \pm 1{,}11$	$74{,}93 \pm 1{,}10$	$63{,}63 \pm 1{,}22$
	Sala de exposiciones	$68{,}83 \pm 1{,}17$	$69{,}27 \pm 1{,}17$	$65{,}52 \pm 1{,}20$	$61{,}48 \pm 1{,}23$

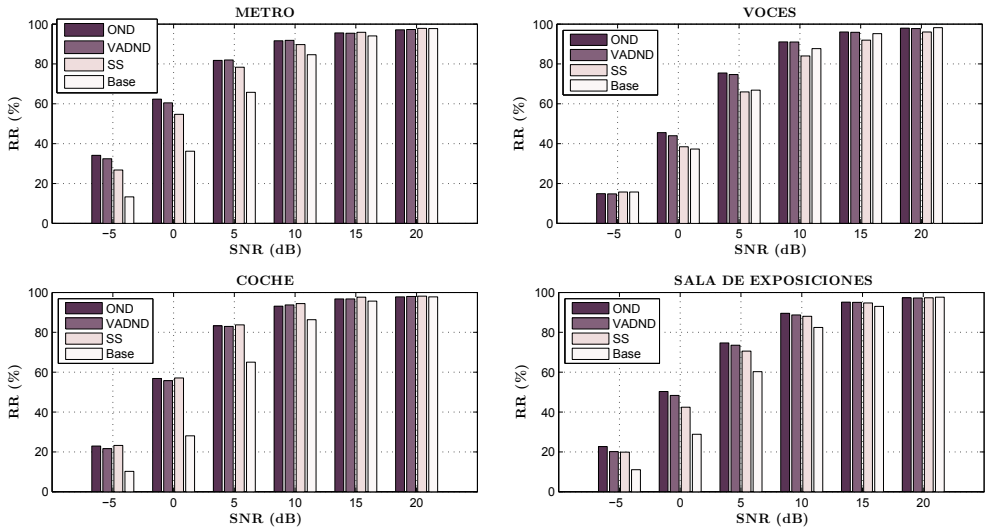


Figura 3.8: Tasas de reconocimiento [%] para el sistema base y las técnicas SS, *OND* y *VADND*.

otras técnicas (incluyendo el sistema base), probablemente debido a las distorsiones introducidas por SS en el proceso de eliminación de ruido. Para el ruido de coche, se alcanzan resultados similares con todas los métodos. Por otro lado, comparando las dos técnicas basadas en NMF para todos los ruidos, *OND* supera ligeramente a *VADND* en la mayoría de los casos, siendo esta diferencia de rendimiento menos notable que en la tarea de mejora de la señal de voz.

3.3.3. Influencia del número de SBVs del ruido en el RAH.

En esta subsección se estudia la evaluación de la tasa de reconocimiento del sistema al variar el número de vectores espectrales base del ruido. En la figura 3.9 se muestran los resultados de la tasa de reconocimiento promediada sobre todas las SNRs ($-5dB$ - $20dB$), en función del número de vectores base del ruido desde 1 hasta 50 usando las dos variantes de eliminación de ruido basado en NMF (*OND* y *VADND*) descritas en la sección 3.1.1 y considerando la escala de frecuencia ERB. De acuerdo a estos resultados se puede observar que para el caso de los ruidos de metro y sala de exposiciones y la técnica *OND*, la tasa de reconocimiento promedio mejora cuando el número de SBVs se incrementa. Sin embargo, con la técnica *VADND*, esta mejora solo se aprecia para el ruido de metro. En el caso del ruido de voces se puede observar una ligera disminución de la tasa de reconocimiento promedio con *VADND*, debido posiblemente a errores del detector de voz que pueden ocasionar que para algunas elocuciones, el modelo de ruido generado no sea adecuado. Para el ruido de coche, no se aprecia una mejora en la tasa de reconocimiento promedio cuando se varía el número de vectores base del modelo de ruido. Esto puede ser debido a que es un tipo de ruido concentrado en ciertas bandas de frecuencia, por lo que puede representarse adecuadamente con un número reducido de vectores base. Es importante mencionar que la mejora en la tasa de reconocimiento promedio se produce cuando el número de vectores base es mayor que 30, lo que es congruente con los resultados obtenidos en la subsección 3.2.2 para la tarea de mejora de voz. Para la escala Mel se observan tendencias similares.

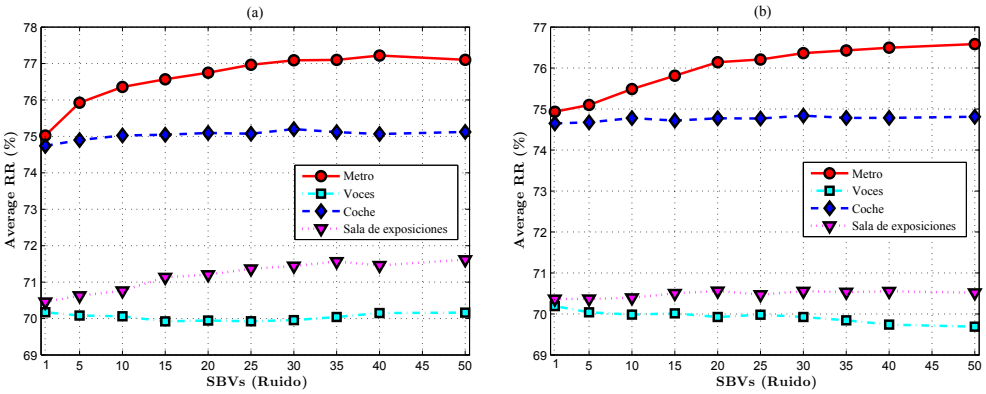


Figura 3.9: Tasas de reconocimiento promedio[%] variando en número de SBVs del ruido usando la escala de frecuencia ERB, para las técnicas basadas en NMF (a) *OND* y (b) *VADND*.

3.4. Conclusiones

En este capítulo se ha mostrado un método basado en NMF para eliminación de ruido de señales de voz que combina el uso de la divergencia de Kullback - Leibler con restriccines de dispersión sobre la matriz de activaciones y no necesita un conocimiento a priori acerca de la naturaleza del ruido. Además, se ha realizado un estudio exhaustivo sobre la influencia de diferentes parámetros de NMF (longitud de ventana, desplazamiento de trama, número de vectores espectrales base y parámetros de regularización) sobre la calidad de la voz mejorada. Hemos comparado el método propuesto con el método de sustracción espectral convencional para las tareas de mejora de la voz y reconocimiento automático del habla bajo diferentes condiciones ruidosas, obteniendo mejoras significativas especialmente en SNRs bajos y medios. El método propuesto es más efectivo con algunos tipos de ruido (ruidos donde sus SBVs son muy distintos al de la voz) que con otros. A pesar de que su funcionamiento es peor que la técnica OND en la tarea de mejora de la voz; sin embargo presenta resultados similares en la tarea de reconocimiento automático del habla.

Capítulo 4

Parametrización basada en filtrado paso alto para clasificación de eventos acústicos

Como se ha comentado en la sección 2.3 del capítulo 2 de la presente tesis, en los últimos años el tema de la clasificación y detección de eventos acústicos, tanto los producidos con el tracto vocal humano (toses, risas, etc.) como otros tipos de sonidos (pasos, tecleo, timbre de teléfono, etc.), ha dado lugar a numerosos trabajos de investigación. En varios de estos estudios, esta tarea se aborda como un problema típico de aprendizaje de patrones en el que los parámetros acústicos de entrada más utilizados son los coeficientes MFCC convencionales y se experimenta con diversos tipos de clasificadores, tales como GMMs [Temko and Nadeu, 2006], HMMs [Cotton and Ellis, 2011], SVMs [Temko and Nadeu, 2006], [Mejia Navarrete et al., 2011] y redes neuronales [Dhanalakshmi et al., 2008], [Kons and Toledo-Ronen, 2013]. Sin embargo, la alta correlación entre el funcionamiento de estos diferentes clasificadores sugiere que el principal problema no es la técnica de clasificación usada, sino el tipo de características acústicas utilizadas [Kons and Toledo-Ronen, 2013].

En este sentido, se han propuesto diversos esquemas de parametrización de eventos acústicos en la literatura, en muchos casos similares a los utilizados para reconocimiento automático del habla o locutor, como los ya mencionados MFCC [Temko and Nadeu, 2006], [Zieger, 2008], [Zhuang et al., 2010], [Kwangyoun and Hanseok, 2011] u otros tales como, log-energías en banda [Zhuang et al., 2010], parámetros de predicción lineal perceptual (PLP) [Portelo et al., 2009], log-energía, tasa de cruces por cero [Temko and Nadeu, 2006], etc. Sin embargo, tal y como se puntualizó en [Zhuang et al., 2010], este tipo de características acústicas convencionales no necesariamente son las más apropiadas para la tarea de clasificación y detección de eventos acústicos, puesto que han sido diseñadas de acuerdo a las propiedades espectrales de la voz que, en general, difieren de la estructura espectral de los eventos acústicos. Algunos autores han tratado de solucionar este problema mediante la utilización de métodos de selección de características para la construcción de una parametrización más adecuada para AEC [Zhuang et al., 2010].

En este capítulo abordamos este problema desde una perspectiva distinta. En primer lugar, estudiamos las características espectrales de diferentes eventos acústicos en comparación con la estructura espectral de la voz. Para ello, utilizamos la técnica de factorizacion en matrices no - negativas sobre el espectro de magnitud de las señales de audio, dado que proporciona una representación compacta y más fácilmente interpretable desde el punto de vista visual del contenido espectral de dichas señales. En segundo lugar, a partir de las conclusiones extraídas de este estudio, proponemos un nuevo esquema de parametrización, que es una extensión de los coeficientes mel-cepstrales basada en el filtrado paso alto de la señal de audio. Finalmente, hemos evaluado la técnica propuesta en condiciones limpias y ruidosas, logrando en ambos escenarios mejoras significativas con respecto al sistema de referencia basado en los MFCC convencionales.

4.1. Análisis espectral de los eventos acústicos basado en NMF.

En esta sección se muestra el estudio realizado sobre las características espectrales de diversos eventos acústicos en comparación con la estructura espectral de las señales de voz. Los eventos acústicos considerados para este análisis pertenecen a 12 clases acústicas distintas que corresponden a 12 tipos de sonidos que típicamente aparecen en ambientes de oficina o salas inteligentes (*smart rooms*): aplausos, toses, movimiento de sillas, tocar la puerta, abrir/cerrar la puerta, teclear usando un teclado, risas, arrugar papel, timbre telefónico, pasos, tintineo de cuchara/taza y tintineo de llaves.

La señal de voz ha sido objeto de una gran cantidad de estudios, por lo que su estructura espectral es bien conocida. En particular, la señal de voz tiene un ancho de banda en torno a $8KHz$. Su espectro se caracteriza por la presencia de energía alta en bajas frecuencias conformando la típica estructura formántica en las zonas correspondientes a sonidos sonoros y en altas frecuencias en sonidos sordos (como, por ejemplo, las fricativas sordas). Sin embargo, en general, los sonidos de naturaleza distinta a la voz no muestran esta estructura espectral. De hecho, en muchos casos, su contenido espectral relevante está localizado en otras bandas de frecuencia y carecen de regiones con estructura formántica, tal y como se muestra en el estudio empírico realizado en esta sección.

Como ejemplo, en la figura 4.1 se representan los espectrogramas de dos instancias distintas del mismo evento acústico, *Timbre telefónico*. Aunque es posible extraer conclusiones acerca de la naturaleza espectral de este evento acústico por medio de la inspección visual de estos espectrogramas, su alta variabilidad, debida en parte a las características frecuenciales intrínsecas del evento acústico y a la presencia de ruido (micrófono, ruido ambiental, etc.), nos motiva a usar un método automático para realizar esta tarea. En este caso, hemos optado por la factorización en matrices no negativas (NMF), dado que nos permite obtener una representación basada en

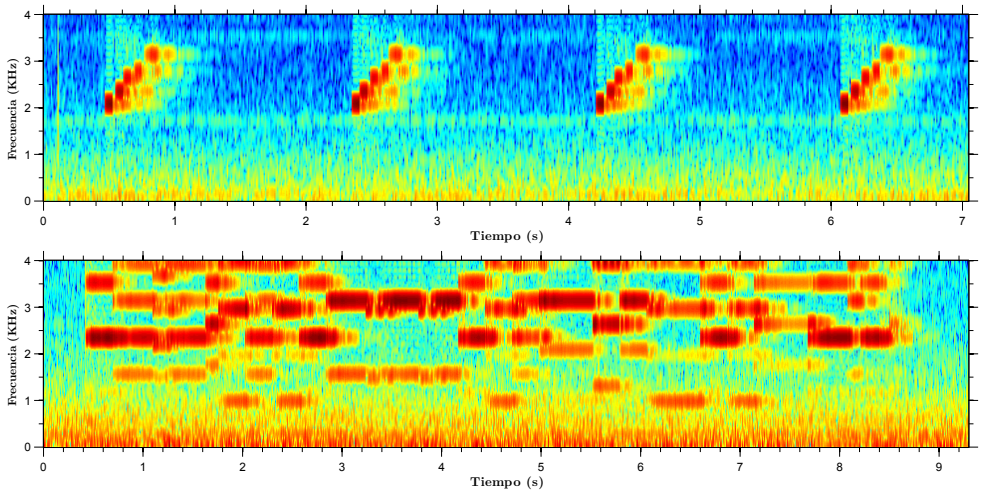


Figura 4.1: Espectrogramas de dos ejemplos diferentes del evento acústico *Timbre telefónico*.

partes más compacta del espectro de magnitud de los eventos acústicos.

Aunque en la sección 2.1 del capítulo 2 puede encontrarse una descripción más extensa sobre NMF, por claridad, en esta sección volvemos a mencionar sus aspectos más relevantes.

Dada una matriz no - negativa $V_e \in \mathbb{R}_+^{F\times T}$, donde cada columna es un vector de datos (en este caso, V_e contiene los espectros de magnitud a corto plazo de un conjunto dado de señales de audio), NMF lo aproxima como el producto de dos matrices no - negativas W_e y H_e, tal que

$$V_e \approx W_e H_e \tag{4.1}$$

donde $W_e \in \mathbb{R}_+^{F\times K}$ y $H_e \in \mathbb{R}_+^{K\times T}$ y F, T y K representan el número de *bins* de frecuencia, tramas y componentes base, respectivamente. La matriz W_e contiene las componentes espectrales base que pueden interpretarse como los bloques básicos a partir de los cuales puede construirse el espectro de magnitud de cualquier señal de audio y H_e contiene los coeficientes de activación o ganancia correspondientes a dichas componentes básicas. De esta forma, cada columna de V_e puede escribirse

como la combinación lineal de los K bloques de construcción ponderados por los coeficiente de activación localizados en la correspondiente columna de H_e. En esta sección estamos interesados en analizar la matriz W_e ya que contienen los vectores espectrales base (SBVs) que encapsulan la estructura frecuencial oculta de los datos contenidos en V_e [Smaragdis, 2004].

Para la obtención de los vectores espectrales base de cada uno de los eventos acústicos considerados, se aplica NMF sobre la correspondiente matriz V_e compuesta por el espectro de magnitud a corto plazo de un subconjunto de archivos de audio de entrenamiento perteneciente a la clase acústica a analizar. En nuestro caso particular, el espectro de magnitud se calculó usando ventanas de análisis de Hamming de $20ms$ con un desplazamiento de trama de $10ms$. En total, se usaron 364,214 tramas de magnitud espectral correspondientes a aproximadamente $60min$ de audio. Las matrices NMF (W_e y H_e) fueron inicializadas usando un algoritmo de inicialización multi - inicio [Cichocki et al., 2009], de tal manera que se generaron 10 pares de matrices aleatorias uniformes (W_e y H_e) y se escogió para inicialización la factorización que produjo la distancia euclídea menor entre V_e y (W_eH_e). A continuación, estas matrices se entrenaron mediante la minimización de la divergencia KL (ecuación 2.3) entre el espectro de magnitud V_e y sus correspondientes matrices factorizadas (W_eH_e) usando el esquema iterativo y las reglas de aprendizaje (ecuación 2.4) propuestas en [Lee and Seung, 1999], siendo el criterio de parada del algoritmo un número máximo de iteraciones (en nuestro caso, 200).

El número de vectores base K se estableció teniendo en cuenta que se debía alcanzar un compromiso entre la precisión de reconstrucción del espectro de magnitud (es decir, el error de aproximación promedio entre V_e y (W_eH_e) calculado sobre todos los eventos acústicos) y una buena visualización de los SBVs (ver apéndice A para más detalle sobre los errores de aproximación obtenidos). Finalmente se utilizaron $K = 23$, valor que corresponde al caso en el cual la variación relativa del error de aproximación promedio entre dos números sucesivos de SBVs es menor del 2 %. Es importante mencionar que cuando el número de vectores base se incrementa,

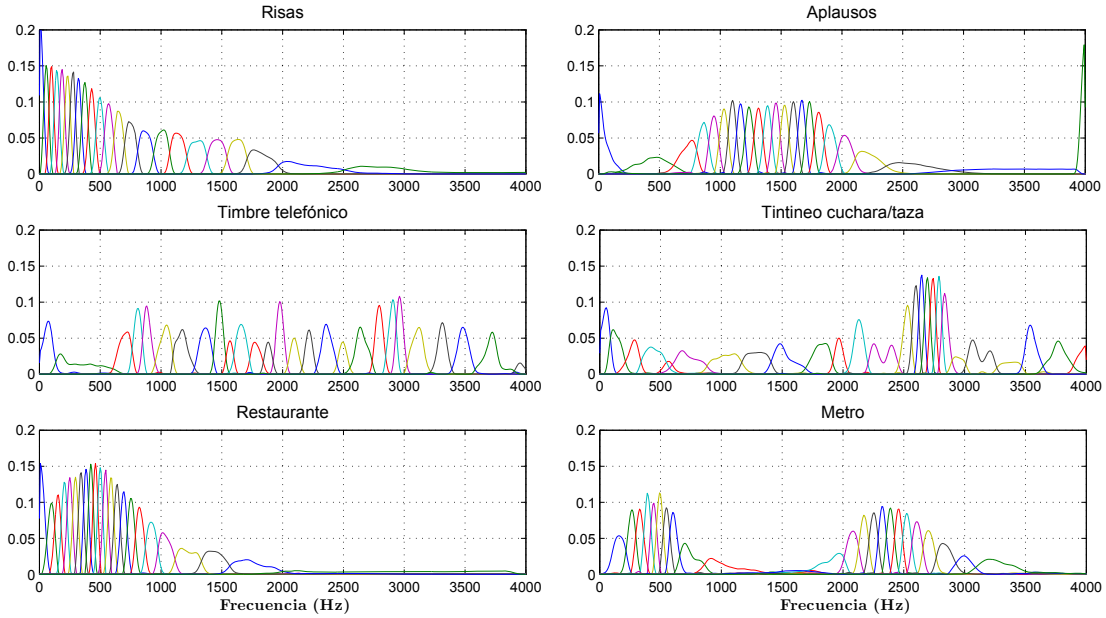


Figura 4.2: Vectores espectrales base (SBVs) para diferentes eventos acústicos y tipos de ruido.

NMF tiende a situar más bandas estrechas (es decir, de ancho de banda pequeño) en las áreas del espectro con energía alta proporcionando mayor resolución en estas regiones y produciendo, por tanto, una reducción del error de reconstrucción [Bertrand et al., 2008b]. Sin embargo, para el propósito de este análisis un valor mayor de K no ofrece una mayor información relevante y, por contra, produce una peor visualización de los SBVs.

En la figura 4.2 se representan los 23 SBVs de cuatro sonidos diferentes que no corresponden a voz (*Risas*, *Aplausos*, *Timbre telefónico* y *Tintineo cuchara/taza*) y dos diferentes clases de ruidos (*Restaurante* y *Metro*). A partir de la figura 4.2 se pueden extraer las siguientes observaciones:

- El contenido espectral de los eventos acústicos es muy diferente entre unos y otros, presentando en general, componentes relevantes en frecuencias medias y altas. Como se ha mencionado previamente, las componentes espectrales de la voz están concentradas en bajas frecuencias, por lo que, de acuerdo a esto,

es posible inferir que las parametrizaciones diseñadas para señales de voz (por ejemplo, los MFCC convencional) pueden no ser lo suficientemente adecuadas para representar otros sonidos diferentes a la voz.

- En todos los casos, las componentes de baja frecuencia se encuentran presentes en mayor o menor grado, de modo que esta parte del espectro parece no ser muy discriminativa cuando se comparan diferentes tipos de eventos acústicos.

- Comparando los SBVs de los sonidos que no corresponden a la voz, se pueden observar grandes diferencias en la parte del espectro medio y alto, sugiriendo que estas bandas de frecuencia son las más adecuadas (o al menos no pueden ser ignoradas) que la parte baja del espectro para distinguir entre diferentes eventos acústicos.

- Distintos tipos de ruidos ambientales presentan diferentes características espectrales. Por ejemplo, en el caso del ruido de *Restaurante*, mucho de su contenido frecuencial está localizado en la banda de frecuencia por debajo de $1KHz$, mientras que los SBVs del ruido de *Metro* se encuentran distribuidos en dos regiones diferentes del espectro: una banda de baja frecuencia por debajo de $750Hz$ y una banda de frecuencia media-alta entre 2 y $3KHz$. El análisis de otras clases de ruidos (*Aeropuerto*, *Voces*, *Tren* y *Sala de exposiciones*) presentan observaciones similares. De esta forma, la distorsión producida sobre las señales de los eventos acústicos debido a la presencia de ruido aditivo variará considerablemente dependiendo de la naturaleza del ruido. Como consecuencia de este hecho, algunos tipos de ruidos serán presumiblemente más dañinos que otros, produciendo degradaciones más notables en el rendimiento del sistema de AEC.

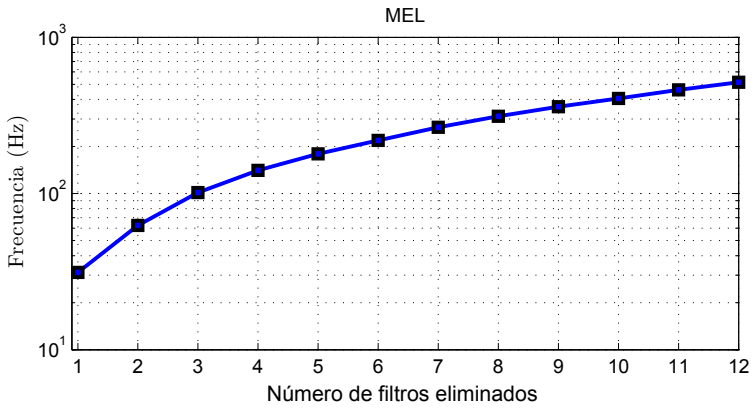


Figura 4.3: Frecuencia superior del banda de paso vs. número de filtros eliminados para la escala Mel.

4.2. Extracción de características para CEA a partir del filtrado paso alto de la señal de audio.

Las conclusiones extraídas de la observación de los SBVs de los diferentes eventos acústicos mostradas en la sección 4.1 han sido la motivación para realizar una extensión de los MFCC convencionales con el objeto de diseñar un módulo de extracción de características acústicas más adecuado para la clasificación de eventos acústicos. Como es bien conocido, MFCC es el procedimiento de parametrización más popular en tareas relacionadas con reconocimiento automático de habla, reconocimiento de locutor y clasificación de audio. La idea básica detrás del nuevo esquema propuesto en este trabajo consiste en considerar explícitamente la especial relevancia de ciertas bandas de frecuencia del espectro dentro del procedimiento de extracción de características mediante de la modificación del banco de filtro auditivo en escala de frecuencia Mel utilizado en el proceso de parametrización.

Una de las principales conclusiones obtenidas a partir del estudio realizado en la sección 4.1 es que las frecuencias medias y altas son especialmente útiles para discriminar entre diferentes eventos acústicos. Por esta razón, estas bandas de frecuencias deberían ser enfatizadas de alguna manera dentro del proceso de parametrización.

Esto se puede conseguir mediante el filtrado paso alto de las tramas de la señal de audio (usando un filtro adecuado) antes de la aplicación del banco de filtros auditivo y el cálculo de los parámetros cepstrales. Sin embargo en este trabajo, adoptamos un método más directo que consiste en modificar el banco de filtros auditivo por medio de la eliminación explícita de ciertos número de filtros situadas en la región de baja frecuencia del espectro. En la figura 4.3 se puede observar la frecuencia superior de la banda de paso en función del número de filtros eliminados en el banco de filtros auditivo para la escala de frecuencia Mel.

En la práctica, este procedimiento consiste en fijar a un valor muy pequeño las energías correspondientes a las salidas de los filtros paso bajo que van a ser eliminados. Este umbral debe ser diferente de cero con la finalidad de evitar problemas numéricos con el logaritmo, siendo en nuestro caso particular, igual a 2^{-52} (el valor del nivel de redondeo *eps* en el lenguaje de programación *Matlab*).

Una vez que el filtrado paso alto es llevado a cabo siguiendo el procedimiento previamente descrito, se procede a calcular las energías del banco de filtros restante, y a continuación se aplica el logaritmo y la DCT sobre ellas como en los MFCC convencionales, dando lugar a un conjunto de coeficientes cepstrales. Finalmente, se aplica una técnica de integración temporal de características que consiste en dividir la secuencia de los coeficientes cepstrales en segmentos de longitud dada y calcular los estadísticos de dichos parámetros (en este caso, la media, desviación estándar y simetría) sobre cada segmento (para más detalles, ver la subsección 2.3.1.2 del capítulo 2). Estas características segmentales son la entrada al clasificador de eventos acústicos, que en este trabajo está basado en máquinas de vectores soporte, tal y como se detallará en la sección 4.3. El proceso completo se muestra en la figura 4.4.

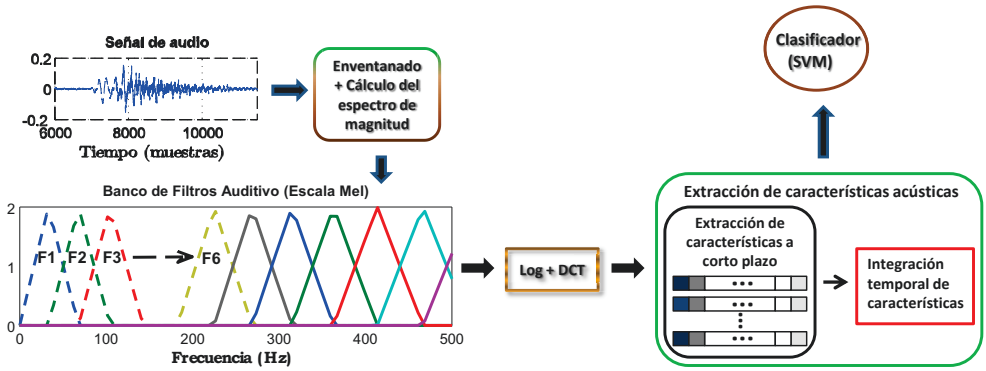


Figura 4.4: Diagrama de bloque del módulo de extracción de caractererísticas acústicas propuesto.

4.3. Experimentos

4.3.1. Base de datos y protocolo experimental.

La base de datos para los experimentos consiste de un total de 2114 instancias de eventos acústicos pertenecientes a 12 clases acústicas diferentes : *aplausos*, *toses*, *movimiento de silla*, *tocar la puerta*, *abrir/cerrar la puerta*, *teclear usando un teclado*, *risas*, *arrugar papel*, *timbre telefónico*, *pasos*, *tintineo de cuchara/taza* y *tintineo de llaves*. La composición de la base de datos completa es similar a la que utilizada en [Zhuang et al., 2010] y se muestra en la tabla 4.1. Los archivos de audio se obtuvieron de fuentes diversas: *websites*, la base de datos FBK-Irst ([S0296, 2009]) y la base de datos UPC-TALP ([S0268, 2008]). Todos los ficheros fueron convertidos al mismo formato y frecuencia de muestreo ($8KHz$). El número total de segmentos de $2s$ de longitud (que corresponde con el tamaño de ventana usado para el cálculo de las características segmentales) en la base de datos completa es de 7775. La figura 4.5 muestra el histograma del número de segmentos por evento acústico en la base de datos. El promedio de segmentos por fichero es de 3,75 segmentos.

Debido a que la base de datos es demasiado pequeña para lograr resultados estadísticamente significativos, se ha utilizado un esquema de validación cruzada de 6

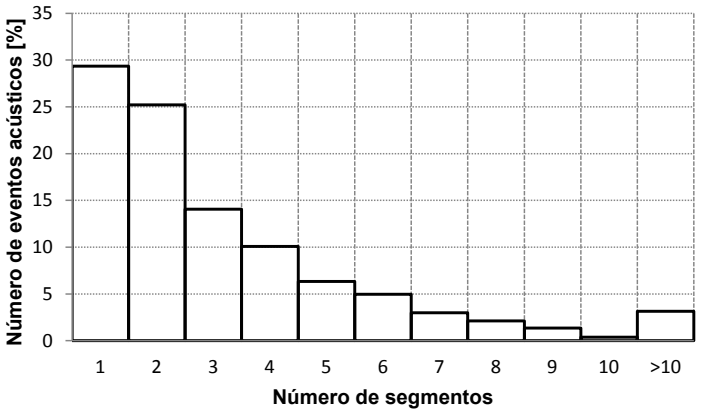


Figura 4.5: Histograma del número de segmentos por evento acústico para la base de datos usada en la experimentación.

grupos (*6-fold*) para extenderla artificialmente. Específicamente la base de datos se ha dividido en seis grupos disjuntos balanceados. Se realizaron 6 subexperimentos, de modo que un grupo diferente se mantuvo para test en cada subexperimento, mientras que el resto se usó para entrenamiento. Finalmente, los resultados que se presentan corresponden al promedio de las tasas obtenidas en los 6 subexperimentos.

Para los experimentos en condiciones ruidosas, las grabaciones de audio original fueron contaminadas con seis diferentes tipos de ruido (*aeropuerto*, *voces*, *restaurante*, *tren*, *sala de exposiciones* y *metro*) obtenidos de la base de datos AURORA 2 ([Hirsch and Pearce, 2000]) a SNRs desde $0dB$ hasta $20dB$ con pasos de $5dB$. Con la finalidad de calcular la cantidad de ruido que hay que añadir a las grabaciones limpias, la potencia del audio y del ruido se calcularon siguiendo el procedimiento descrito en [Steeneken, 1991], el cual tiene en cuenta las características no - estacionarias de las señales de audio.

El sistema de clasificación de eventos acústicos está basado en un clasificador de tipo SVM con configuración uno contra uno y kernel RBF [Mejia Navarrete et al., 2011]. La entrada a dicho clasificador fueron las características segmentales anteriormente descritas, previamente normalizadas. El sistema se desarrolló usando el software LIBSVM [Chih-Chung and Chih-Jen, 2011]. En lo con-

Tabla 4.1: Base de datos usada en los experimentos.

Clase	Tipo de evento	No. de ocurrencias
1	Aplausos [ap]	155
2	Toses [co]	199
3	Movimiento sillas [cm]	115
4	Tocar puerta [kn]	174
5	Abrir/cerrar puerta [ds]	251
6	Teclear [kt]	158
7	Risas [la]	224
8	Arrugar papel [pw]	264
9	Timbre telefónico [pr]	182
10	Pasos [st]	153
11	Tintineo cuchara/taza [cl]	108
12	Tintineo llaves [kj]	131
Total		2,114

cerniente al entrenamiento de las SVM, para cada uno de los subexperimentos, se realizó una validación cruzada de 5 grupos (*5-fold*) para la determinación de los valores óptimos de los parámetros del kernel RBF usando datos limpios (es decir, estos parámetros no fueron optimizados para condiciones ruidosas). En la etapa de prueba, como el clasificador fue alimentado con características segmentales, las decisiones de clasificación se realizaron a nivel de segmento. Con la finalidad de obtener una decisión a nivel de evento acústico, las salidas del clasificador correspondientes a los diferentes segmentos que conformaban un archivo de audio se integraron usando un esquema de mayoría de votos *majority voting*, de tal forma que la etiqueta más frecuente fue la finalmente asignada a la grabación completa [Geiger et al., 2013].

4.3.2. Experimentos en condiciones limpias.

Este conjunto de experimentos se llevó acabo con la finalidad de estudiar el rendimiento del esquema propuesto en condiciones limpias (cuando ningún ruido se suma a los archivos de audio originales). Para los experimentos de referencia, se

Tabla 4.2: Tasa de clasificación promedio [%] (segmento) en condiciones limpias.

Param.	**Número de filtros eliminados**												
	Base	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
CC	75.10	77.47	77.66	77.58	77.63	78.16	76.95	78.11	76.87	76.12	77.23	77.23	76.10
CC+ΔCC	77.57	79.43	79.45	79.22	79.36	79.07	79.20	79.55	79.41	78.47	77.81	78.77	78.55

extrajeron 12 coeficientes cepstrales ($C1$ al $C12$) cada $10ms$ usando una ventana de análisis de Hamming de $20ms$ de longitud y un banco de filtros auditivo en escala de frecuencia Mel compuesto de 40 bandas espectrales. También se calcularon y añadieron a los coeficientes cepstrales la log-energía de cada trama (en lugar del coeficiente cepstral de orden cero) y sus primeras derivadas (cuando se indica). Los vectores de características finales consistieron en los estadíticos de estos parámetros cesptrales (media, desviación estándar y la simetría) calculados sobre segmentos de $2s$ de longitud con un solape de $1s$.

Las tablas 4.2 y 4.3 muestran, respectivamente, los resultados obtenidos en términos de la tasa de clasificación promedio a nivel de segmento (porcentaje de segmentos correctamente clasificados) y a nivel de evento acústico (porcentaje de eventos correctamente clasificados), variando el número de bandas de baja frecuencia eliminadas en el banco de filtro auditivo. También se incluyen los resultados para el sistema de referencia (cuando no se elimina ninguna banda de frecuencia) . Ambas tablas contienen las tasas de clasificación para dos conjuntos de parámetros acústicos diferentes, CC (coeficientes cepstrales + log-energía) y CC + ΔCC (coeficientes cepstrales + log-energía + sus primeras derivadas).

Como se puede observar para la parametrización CC, los resultados obtenidos con el filtrado paso alto de la señal del evento acústico supera al del experimento base, siendo la mejora más notable cuando el número de filtros eliminados varía desde 3 hasta 7. A partir de la figura 4.3 se puede observar que estos rangos de filtros eliminados corresponden aproximadamente a una banda de paso que se extiende desde $0Hz$ hasta los $100 - 275Hz$. En particular, el mejor rendimiento se obtiene

Tabla 4.3: Tasa de clasificación promedio [%] (evento) en condiciones limpias.

Param.	**Número de filtros eliminados**												
	Base	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
CC	81.07	82.28	82.04	82.42	82.42	81.89	81.31	83.20	81.27	80.78	80.69	81.75	79.72
CC+ ΔCC	81.41	82.62	83.39	83.58	83.49	83.15	82.38	82.71	82.81	80.06	81.12	81.55	81.22

cuando no se consideran los primeros siete filtros de baja frecuencia en el cálculo de los coeficientes cepstrales. En este caso, la diferencia en rendimiento a nivel de segmento con respecto al experimento base es estadísticamente significativo con un 95 % de nivel de confianza. La reducción del error relativo con respecto al experimento base es de alrededor del 12,1 % a nivel de segmento y de alrededor del 11,2 % a nivel de evento.

Para la parametrización CC + ΔCC se pueden extraer similares observaciones. En este caso los mejores resultados se obtienen cuando las bajas frecuencias (por debajo de $100 - 275Hz$) no se consideran en el proceso de extracción de características. Cuando se compara con CC para el caso de los 7 primeros filtros eliminados, se puede observar que CC + ΔCC logra una mejora de aproximadamente 1,4 % absoluto a nivel de segmento y una disminución alrededor de 0,5 % absoluto a nivel de evento acústico sobre CC. Sin embargo, estas diferencias no son estadísticamente significativas.

Adicionalmente, se han realizado experimentos con otras escalas de frecuencia (en particular, ERB y Bark), observando, como es esperado, un comportamiento similar al de la escala de frecuencia Mel con respecto a la eliminación de bandas de baja frecuencia. Sin embargo, la escala Mel produce resultados ligeramente mejores que ERB y Bark. Se pueden encontrar más detalles acerca de estos experimentos en [Ludena-Choez and Gallardo-Antolin, 2013] y en el apéndice B de esta tesis.

Con la finalidad de realizar un análisis más detallado del rendimiento del sistema de clasificación de eventos acústicos, hemos analizado las matrices de confusión producidas por el experimento base y el esquema propuesto. Como ejemplo, las figuras 4.6(a) y (b) muestra las matrices de confusión a nivel de segmento para los

parámetros CC + ΔCC del experimento base y la versión modificada de esta parametrización con los 7 primeros eliminados. En ambas tablas, las columnas corresponden a la clase correcta, las filas son la clase hipotetizada y los valores dentro de ellas se han calculado como el promedio de los resultados de los 6 subexperimentos. Como se puede observar, en el sistema base las clases que menos se confunden (con una tasa de clasificación mayor del 80 %) son *aplausos*, *teclear usando un teclado*, *risas*, *arrugar papel* y *timbre telefónico*, mientras que las clases que más se confunden son *toses*, *movimiento de silla*, *tocar la puerta* y *tintineo cuchara/taza*. En particular, el 23 % de los segmentos de la clase *toses* se clasifican como la clase *risas* y 12 % de los de la clase *movimiento de silla* y los de la clase *tocar la puerta* se asignan a la clase *pasos*. Es de destacar que el número de confusiones entre sonidos producidos por el tracto vocal humano que no corresponden a la voz (por ejemplo las clases *toses* y *risas*) es elevado. Este hecho ya ha sido observado previamente en otros trabajos de la literatura ([Temko and Nadeu, 2006]). Con la parametrización propuesta, la tasa de reconocimiento de todas las clases acústicas se incrementa con la excepción de la clases *toses* y *tintineo de llaves*. Los eventos acústicos que se clasifican mejor son los mismos que en el experimento base, mientras que hay solo dos eventos con una tasa de clasificación menor que el 70 % (las clases *toses* y *movimiento de silla*). La principal mejora en la parametrización propuesta se debe a que las clases *tocar la puerta* y *tintineo cuchara/taza* reducen significativamente su número de confusiones en comparación con el experimento de referencia.

4.3.3. Experimentos en condiciones ruidosas.

Con la finalidad de estudiar el impacto de ambientes ruidosos sobre el funcionamiento del sistema de clasificación de eventos acústicos, se llevaron a cabo una serie de experimentos usando seis tipos de ruido diferentes (*aeropuerto*, *voces*, *restaurante*, *tren*, *sala de exposiciones* y *metro*) a SNRs desde $0dB$ hasta $20dB$ con pasos de $5dB$. Por brevedad, en esta subsección solo se muestran los resultados para el experimento base y para el parametrizador propuesto en el caso de parámetros CC + ΔCC.

(a)

	1	2	3	4	5	6	7	8	9	10	11	12
1	90,40	0,51	0,00	0,25	0,53	0,12	0,44	0,54	0,00	0,15	0,00	1,07
2	0,00	67,95	3,09	3,56	1,60	2,04	7,06	0,54	0,41	1,04	6,67	0,00
3	0,62	1,79	65,73	2,54	2,67	0,54	1,43	0,76	0,68	5,06	3,23	0,71
4	0,00	0,51	2,25	68,19	3,73	1,26	0,11	0,11	0,27	5,95	0,22	0,71
5	0,00	0,51	0,56	5,85	75,20	0,72	0,33	0,76	1,08	1,49	1,51	0,89
6	0,00	3,08	4,49	1,27	4,27	80,54	2,76	7,34	2,71	4,02	6,24	2,84
7	8,05	23,33	4,78	4,07	4,80	0,66	84,23	2,37	4,74	1,64	3,44	2,31
8	0,62	1,03	2,25	0,51	1,33	11,71	1,54	80,58	3,11	4,32	1,29	12,08
9	0,31	0,26	1,40	0,00	1,33	0,42	1,32	1,83	84,57	0,45	3,66	1,95
10	0,00	0,00	12,36	12,21	3,20	0,96	0,22	1,73	0,14	74,26	5,81	2,13
11	0,00	1,03	0,56	0,00	0,80	0,42	0,11	0,76	1,22	1,04	64,95	2,13
12	0,00	0,00	2,53	1,53	0,53	0,60	0,44	2,70	1,08	0,60	3,01	73,18

(b)

	1	2	3	4	5	6	7	8	9	10	11	12
1	93,50	0,26	0,00	1,02	0,80	0,00	0,33	0,54	0,54	0,15	0,43	0,89
2	0,31	65,64	2,25	3,56	1,60	0,78	5,73	0,76	0,27	0,60	3,66	0,53
3	0,62	1,54	68,26	1,27	2,13	0,18	1,65	0,86	0,14	2,83	1,94	0,71
4	0,00	2,05	3,93	74,55	4,80	0,42	0,66	0,11	0,00	5,95	0,22	0,00
5	0,00	2,82	1,97	5,85	76,00	0,78	0,44	0,86	0,95	1,64	1,72	1,24
6	0,00	2,56	4,21	0,51	2,13	83,00	2,21	6,69	3,11	2,98	5,38	4,97
7	4,64	22,05	3,93	4,07	3,20	0,90	85,01	0,76	2,30	1,64	5,81	1,60
8	0,31	1,28	1,97	0,25	1,87	9,61	1,76	81,55	2,30	4,61	2,15	9,59
9	0,31	1,03	1,40	0,25	2,93	0,60	0,66	2,16	86,87	0,89	4,52	1,60
10	0,31	0,26	8,99	7,38	2,93	1,68	0,44	1,73	0,14	77,38	1,08	4,97
11	0,00	0,00	0,28	0,76	0,80	1,08	0,66	0,32	2,57	0,60	70,75	1,78
12	0,00	0,51	2,81	0,51	0,80	0,96	0,44	3,67	0,81	0,74	2,37	72,11

Figura 4.6: Matrices de confusión [%] a nivel de segmento para la parametrización CC+ ΔCC: (a) Experimento base; (b) Parametrización propuesta con los 7 primeros filtros de baja frecuencia eliminados.

En la figura 4.7 se representa el promedio de la reducción del error relativo para cada tipo de ruido con respecto al experimento base (condiciones ruidosas sin filtrado paso alto de la señal de audio) calculado a través de las SNRs consideradas en función del número de filtros de baja frecuencia eliminados a nivel de segmento y de evento acústico. También se indica la media de la reducción del error relativo sobre todos los tipos de ruido y SNRs. Con la finalidad de observar con mayor detalle el comportamiento del sistema de clasificación de eventos acústicos con respecto a todos los tipos de ruido y SNRs, en la tabla 4.4 se muestran las tasas de clasificación a nivel de segmento para el experimento base y para el módulo de extracción de características propuesto para varias SNRs seleccionadas (20, 10 y $0dB$) y para los seis tipos de ruido evaluados y el rango del número de filtros eliminados desde 7 hasta 12.

Aunque todos los tipos de ruido producen una disminución dramática de la tasa

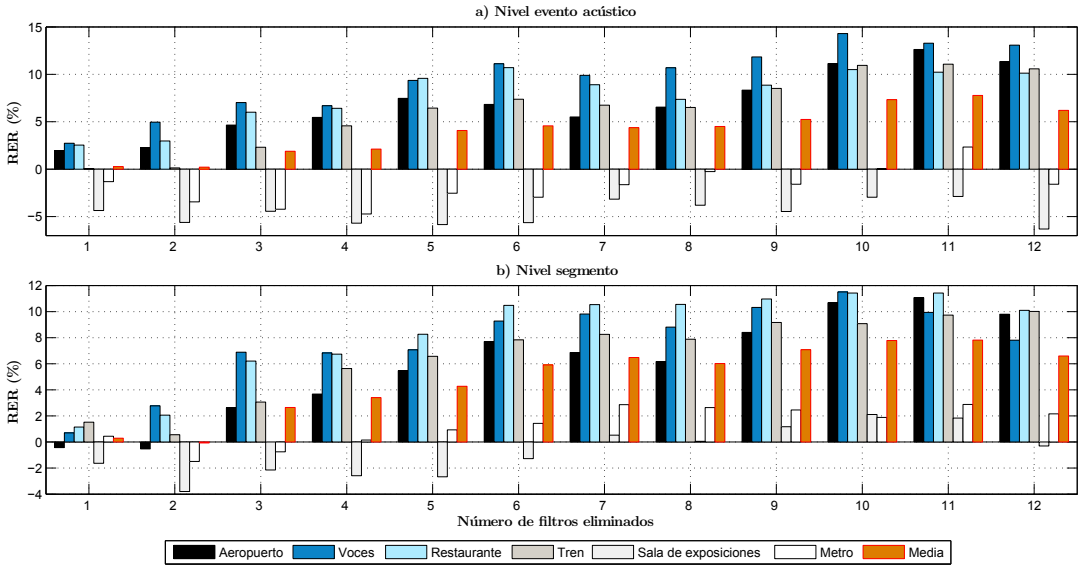


Figura 4.7: Reducción del error relativo [%] con respecto al experimento base para la parametrización CC+ΔCC y la escala Mel: (a) a nivel de evento; (b) a nivel de segmento.

de clasificación, los resultados en la tabla 4.4 sugieren que cada tipo de ruido presenta diferente efecto sobre el funcionamiento del sistema, siendo algunos tipos de ruido (*aeropuerto*, *voces*, *restaurante* y *tren*) menos dañinos que otros (*sala de exposiciones* y *metro*). Este hecho puede explicarse por el análisis de las características espectrales de cada tipo de ruido. En la figura 4.2 se representan los SBVs de los tipos de ruido *restaurante* y *metro*. En el primer caso, la mayor parte de su contenido espectral se encuentra concentrado en bajas frecuencias y, por esta razón, esta clase de ruido afecta en menor medida a las frecuencias más relevantes de los eventos acústicos. Sin embargo, en el segundo caso, parte de los SBVs se distribuye sobre las frecuencias media-altas, y por lo tanto, este tipo de ruido puede ser capaz de enmascarar considerablemente la estructura espectral fundamental de los eventos acústicos. Notése que, por un lado, los SBVs de los tipos de ruido *aeropuerto*, *voces* y *tren* están concentrados en el mismo rango de frecuencias del ruido *restaurante* y por otro lado, los SBVs del ruido de *sala de exposiciones* y *metro* presentan características similares.

A partir de los resultados mostrados en la figura 4.7 se puede observar que con respecto al funcionamiento del esquema propuesto, para los tipos de ruido *aeropuerto*, *voces*, *restaurante* y *tren*, las tasas de clasificación a nivel de segmento mejoran considerablemente cuando el número de filtros eliminados se incrementa, especialmente para SNRs bajas y medias (ver las filas correspondientes etiquetadas como $0dB$ y $10dB$ en la tabla 4.4). En este caso los valores óptimos se obtienen cuando las frecuencias por debajo de $400 - 500Hz$ no son consideradas en el cálculo de los parámetros acústicos, lo que corresponde a la eliminación de los $10 - 11$ primeros filtros de baja frecuencia. Se pueden extraer observaciones similares analizando los resultados a nivel de evento. Para los tipos de ruido *sala de exposiciones* y *metro*, los resultados a nivel de segmento sufren una ligera variación con respecto al número de filtros eliminados, logrando mejoras pequeñas con respecto al experimento base comparado con los otros tipos de ruido. A nivel de evento acústico, las variaciones son más grandes, conduciendo en muchos casos a una disminución en la tasa de clasificación para estos dos tipos de ruido.

Sin embargo, en promedio, con el parametrizador propuesto cuando se eliminan 11 filtros, se obtienen reducciones del error relativo con respecto a la experimento base (ver figura 4.7) de aproximadamente 7,81 % a nivel de segmento y 7,78 % a nivel de evento.

Además se realizaron otro conjunto de experimentos utilizando otras escalas de frecuencia (Bark y ERB) y con la parametrización CC. En todos los casos, los resultados siguen tendencias similares en comparación a la escala Mel y los parámetros CC + ΔCC.

Tabla 4.4: Tasa de clasificación promedio [%] (segmento) para la parametrización CC + ΔCC y diferentes tipos de ruido y SNRs.

Ruido	**SNR (dB)**	**Número de filtros eliminados**						
		Base	**7**	**8**	**9**	**10**	**11**	**12**
Aeropuerto	**20**	66.51	69.47	67.82	68.33	68.82	68.52	68.17
	10	49.92	53.45	52.94	54.29	55.63	56.08	55.28
	0	29.01	33.60	34.59	35.57	37.48	38.20	36.97
Voces	**20**	67.09	68.77	68.45	68.89	68.94	67.94	67.33
	10	52.27	56.45	56.69	56.99	57.44	56.85	56.08
	0	27.59	36.92	35.74	37.16	39.12	37.69	35.28
Restaurante	**20**	67.43	69.40	68.89	68.89	69.22	68.80	68.62
	10	53.09	57.14	56.97	57.34	57.26	57.32	56.69
	0	25.65	37.35	37.91	38.22	38.22	38.65	36.72
Tren	**20**	71.18	72.92	72.82	72.80	72.74	72.27	72.68
	10	58.69	61.72	61.67	62.91	62.90	63.44	63.27
	0	40.46	45.81	45.88	46.40	46.70	47.32	46.83
Sala de exposiciones	**20**	58.00	57.68	57.13	58.01	58.35	57.76	56.49
	10	42.66	42.98	42.41	43.46	44.02	43.65	42.50
	0	22.00	23.45	24.02	23.83	24.66	24.99	23.61
Metro	**20**	56.90	56.38	55.97	56.23	56.68	56.10	55.32
	10	39.88	41.51	40.94	40.82	40.53	41.30	40.40
	0	19.34	23.06	23.81	23.94	22.74	24.75	**24.41**

4.4. Conclusiones

En este capítulo, hemos presentado un nuevo método de parametrización para la tarea de clasificación de eventos acústicos, motivado por el estudio de las características espectrales de los sonidos de naturaleza distinta al de la voz. Primero, hemos realizado un estudio empírico del contenido espectral de diferentes eventos acústicos, concluyendo que las frecuencias medias y altas son especialmente importantes para discriminar entre sonidos que no corresponden a la voz. Segundo, a partir de este estudio, hemos propuesto un nuevo esquema para AEC, que es una extensión de la parametrización MFCC y está basado en el filtrado paso alto de la señal de audio. En la práctica, el esquema propuesto consiste en la modificación del banco de filtros auditivo en escala Mel mediante de la eliminación explícita de un cierto número de filtros de baja frecuencia.

El esquema propuesto ha sido probado en condiciones limpias y ruidosas y comparado con los MFCCs convencionales en una tarea de clasificación de eventos acústicos. Los resultados muestran el hecho de que el filtrado paso alto de la señal de audio es en términos generales beneficioso para el sistema, de forma que la eliminación de frecuencias por debajo de $100 - 275Hz$ en el proceso de parametrización en condiciones limpias y por debajo de $400 - 500Hz$ en condiciones ruidosas, mejora significativamente el funcionamiento del sistema con respecto al experimento base.

Capítulo 5

Integración temporal de características acústicas basada en NMF para CEA

En la sección 2.3.1 del capítulo 2 se mencionó que para la tarea de clasificación de eventos acústicos, habitualmente se utilizan características segmentales o a largo plazo que tratan de describir de forma compacta las propiedades más relevantes de la señal de audio en ventanas temporales de varios segundos. Dichas características segmentales se obtienen a partir de parámetros acústicos extraídos a corto plazo o a nivel de trama (típicamente sobre ventanas de $20 - 30ms$) y agregados mediante un método determinado de integración de características. Mientras que en el capítulo 4 se utilizó la técnica basada en estadísticos, en este capítulo profundizaremos en la basada en coeficientes de banco de filtros (FC, *Filterbank Coeficients*), que fue inicialmente propuesta para la clasificación de géneros musicales y audio en general [Meng et al., 2007], [McKinney and Breebaart, 2003], [Arenas Garcia et al., 2006], y que más recientemente ha sido experimentada en la tarea de CEA con resultados prometedores [Mejia Navarrete et al., 2011].

En contraste a las características segmentales basadas en estadísticos, los coefi-

cientes de banco de filtros permiten capturar la estructura dinámica de los parámetros a corto plazo. Esta técnica consiste básicamente en resumir la información contenida en los periodogramas de cada dimensión de los parámetros a corto plazo en varios valores de potencia calculados en las bandas de frecuencia determinadas por un banco de filtros predefinido, como el propuesto en [McKinney and Breebaart, 2003]. Sin embargo, como se señala en [Arenas Garcia et al., 2006], este banco de filtros fijo puede no ser lo suficientemente general, puesto que la importancia de las distintas características dinámicas de los parámetros a corto plazo para clasificación puede ser dependiente de la tarea. En este contexto, en [Arenas Garcia et al., 2006] se presenta un método supervisado para el aprendizaje automático de un banco de filtros óptimo para clasificación de géneros musicales.

Por contra, en este capítulo, presentamos un método no supervisado basado en el algoritmo NMF para el diseño de un banco de filtros para la extracción de características FC más adecuado para la tarea de clasificación de eventos acústicos. A lo largo del capítulo mostraremos que con el método propuesto se superan los resultados obtenidos con el correspondiente experimento base en condiciones limpias y ruidosas. Además, otra ventaja de este esquema es su versatilidad, en el sentido de que no es específico para CEA y, por lo tanto, puede ser aplicado a otras tareas de clasificación de voz y audio.

5.1. Integración temporal de características basada en coeficientes de banco de filtros

En la figura 5.1 se representa el diagrama de bloques del proceso de extracción de características usando la técnica FC. Consiste de dos etapas principales: extracción de características a corto plazo e integración temporal de características mediante FC. A continuación, describimos los dos principales módulos del parametrizador FC.

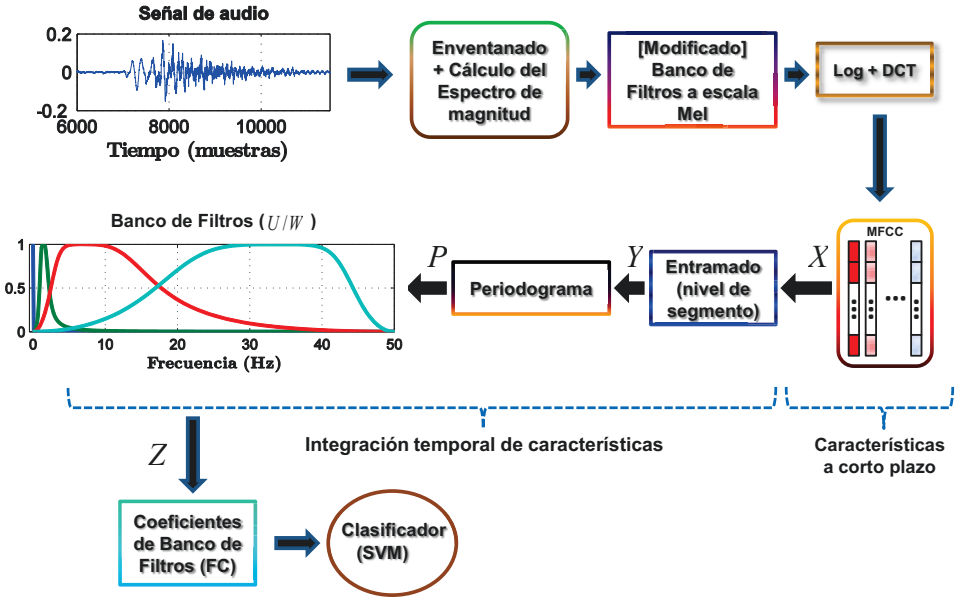


Figura 5.1: Diagrama de bloque del proceso de extracción de características FC.

5.1.1. Extracción de características a corto plazo

En este capítulo, hemos considerado dos tipos de parámetros acústicos diferentes como características a corto plazo: los MFCC convencionales y su extensión, denotada como MFCC_HPN, que incluye el filtrado paso alto de la señal de audio implementado mediante la modificación del banco de filtros auditivo, tal y como se describe en el capítulo 4 de la presente tesis.

En ambos casos, MFCC y MFCC_HPN, la log-energía de cada trama y sus primeras derivadas (cuando se indica) se calculan y concatenan a los coeficientes cepstrales estáticos.

5.1.2. Integración temporal de características mediante FC

Una vez extraídos los coeficientes cepstrales, se aplica integración temporal sobre los segmentos de audio de una longitud dada con la finalidad de obtener un conjunto de vectores de características a una escala de tiempo más lar-

ga. En este capítulo, nos enfocamos en la aproximación denominada coeficientes de banco de filtros (FC) [Meng et al., 2007], [McKinney and Breebaart, 2003], [Arenas Garcia et al., 2006], que ayuda a capturar el comportamiento temporal de los parámetros a corto plazo.

En primer lugar, la secuencia de T parámetros a corto plazo de dimensión D_x, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ se divide en K segmentos, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_K\}$ de la siguiente manera,

$$\mathbf{y}_k = f\left(\mathbf{x}_{k \cdot H_s}, \mathbf{x}_{k \cdot H_s + 1}, ..., \mathbf{x}_{k \cdot H_s + L_s - 1}\right) \tag{5.1}$$

donde L_s es el tamaño del segmento, H_s es el desplazamiento del segmento, ambos definidos en tramas y f es la función que combina los parámetros a nivel de trama en parámetros a nivel de segmento..

En segundo lugar, se determina el periodograma de cada dimensión de parámetros a corto plazo contenidos en el k-simo segmento $\mathbf{y}_k$ y la información contenida en dicho periodograma se resume mediante la obtención de la potencia en las diferentes bandas de frecuencia del banco de filtros predefinido ($\mathbf{U}$),

$$\mathbf{z}_k = \mathbf{P}_k \mathbf{U} \tag{5.2}$$

donde $\mathbf{P}_k$ representa los periodogramas de la secuencia de coeficientes a corto plazo perteneciente al k - simo segmento, $\mathbf{U}$ es la magnitud de la respuesta en frecuencia del banco de filtros predefinido y $\mathbf{z}_k$ es el vector de características final. Las dimensiones de $\mathbf{P}_k$, $\mathbf{U}$ y $\mathbf{z}_k$ son, respectivamente, D_x x D_p, D_p x n_f y D_x x n_f, donde D_p es la dimensionalidad de cada periodograma individual y n_f es el número de filtros del banco de filtros. Los coeficientes de banco de filtros, $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_K\}$ son la entrada al sistema de clasificación de eventos acústicos.

Habitualmente, dicho banco consta de cuatro filtros correspondiente a las siguientes bandas de frecuencia [McKinney and Breebaart, 2003]:

- Filtro 1: 0 Hz. Filtro DC
- Filtro 2: 1 - 2 Hz. Energía de modulación
- Filtro 3: 3 - 15 Hz. Energía de modulación
- Filtro 4: 20 - 43 Hz. Aspereza (*roughness*) perceptual

Como la importancia de las diferentes componentes dinámicas de las características a corto plazo para clasificación puede ser dependiente de la tarea, se puede argumentar que este banco de filtros fijo ($\mathbf{U}$) no es óptimo para todos los problemas de clasificación de audio. En otras palabras, algunas frecuencias de modulación pueden ser relevantes para distinguir entre, por ejemplo, diferentes géneros musicales, y no entre eventos acústicos. En la siguiente sección, presentamos un método no supervisado para diseñar un banco de filtros FC adaptado para la tarea de CEA.

5.2. Diseño del banco de filtros FC basado en NMF

Nuestra meta es desarrollar un método no supervisado para encontrar el banco de filtro óptimo de tal manera que los parámetros FC resultantes $\mathbf{Z}$ contengan la información más significativa acerca de la estructura temporal de las características a corto plazo. Este problema puede formularse como la descomposición de los periodogramas $\mathbf{P}$ en sus componentes principales (por ejemplo, en sus bandas de frecuencia más relevantes).

Como ya se ha mencionado previamente, la factorización en matrices no - negativas establece una forma de descomponer una señal en la combinación convexa de bloques de construcción no - negativos (llamados vectores espectrales base) mediante la minimización de una función de coste dada (en nuestro caso, la divergencia KL). Como el espectro de potencia de los parámetros a corto plazo (en este caso, MFCCs) y la respuesta en frecuencia de los elementos del banco de filtros son inherentemente positivos, NMF puede ofrecer una adecuada solución al problemática de determinar

el banco de filtros FC óptimo, como se explicará en las siguientes subsecciones. A lo largo del resto de este capítulo, denotamos al banco de filtro obtenido por NMF como **W** con la finalidad de distinguirlo del banco de filtro fijo **U**.

5.2.1. Construcción del banco de filtro FC con NMF

En este caso, la matriz a descomponer en sus componentes principales está formada por los periodogramas de las características a corto plazo. Como se aprende un único banco de filtros para todas las componentes de los parámetros acústicos, la matriz **P** es la concatenación por fila de los periodogramas D_x de los parámetros a corto plazo extraídos de las señales de audio del conjunto de entrenamiento. Por lo tanto, la dimensión de **P** es (D_x x n_s) x D_p, donde n_s es el número total de segmentos en el conjunto de entrenamiento.

Una vez que esta matriz es transpuesta ($\mathbf{P^T}$), sus correspondientes matrices factorizadas **WH** se obtienen usando las reglas de aprendizaje de NMF dadas en la ecuación 2.4 del capítulo 2. Las dimensiones de **W** y **H** son, respectivamente, D_p x n_f y n_f x (D_x x n_s). Finalmente, la matriz resultante **W** contiene los SBVs que representan las bases del espectro de potencia de las características a corto plazo, y dado que se verifica que $\mathbf{P^T} \approx \mathbf{WH}$ dichos vectores base pueden interpretarse como las respuestas en frecuencia de los filtros del banco de filtros FC requerido.

Con la finalidad de calcular los parámetros FC, la ecuación 5.2 se aplica sustituyendo el banco de filtros fijo **U** por **W**.

5.2.2. Experimentos y resultados

5.2.3. Base de datos y sistema base

La base de datos para los experimentos es la misma que se utilizó en el capítulo 4 de la presente tesis.

El protocolo experimental que se siguió fue el mismo que para los experimentos del capítulo 4, de modo que se utilizó un esquema de validación cruzada de 6 grupos

(*6-fold*), de modo que los resultados que se presentan corresponden con la media de los 6 subexperimentos realizados.

Asimismo, el sistema de clasificación de eventos acústicos sobre el que se realiza la experimentación es el mismo que el del capítulo 4 que utiliza una SVM con configuración uno contra uno y kernel RBF y ha sido desarrollado usando el software LIBSVM [Chih-Chung and Chih-Jen, 2011].

5.2.4. Extracción de características

En este capítulo, se han considerado dos tipos diferentes de características a corto plazo, MFCC y MFCC_HPN, correspondiendo estos últimos a la parametrización propuesta en el capítulo 4. Recordamos que la diferencia entre ellas es que para MFCC_HPN se eliminan los N primeros filtros de baja frecuencia en el banco de filtros auditivo en escala Mel, de modo que, dependiendo del valor de N, no se consideran las frecuencias por debajo de un cierto valor en el cálculo de los coeficientes cepstrales. En ambas parametrizaciones, se extraen 12 parámetros cepstrales cada $10ms$ usando una ventana de análisis de Hamming de $20ms$ de longitud y un banco de filtros en escala Mel compuesto de 40 y $(40 - N)$ bandas triangulares para MFCC y MFCC_HPN, respectivamente. También se calculan la log - energía de cada trama y sus primeras derivadas (cuando se indica) y se concatenan a los coeficientes cepstrales estáticos, dando lugar a un vector de parámetros a corto plazo de dimensión $D_x = 13$ (o 26 cuando se usa la primera derivada).

Para la integración temporal de características se consideran segmentos de audio de $2s$ de longitud con un solape de $1s$. Los periodogramas de cada dimensión de las características a corto plazo se calculan sobre estos segmentos y posteriormente se filtran usando el banco de filtros **U** definido en la sección 5.1.2, para los parámetros FC del sistema base y el banco de filtros **W** obtenido con el método NMF para los parámetros FC basados en NMF.

Los filtros **U** son filtros Butterworth de segundo orden. Por el contrario, en el método basado en NMF, para cada subexperimento, el banco de filtros **W** se obtiene

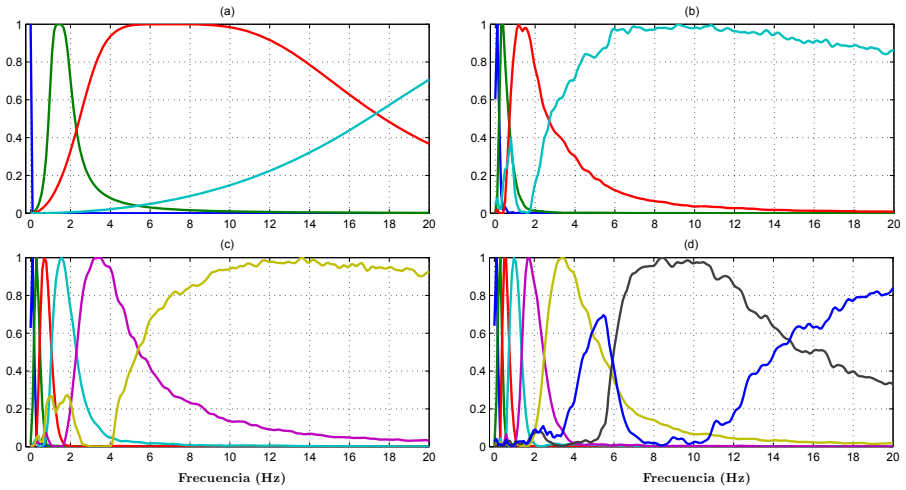


Figura 5.2: Respuesta en frecuencia de los bancos de filtro usados en el proceso de integración temporal de características. (a) Banco de filtros fijo (**U**), 4 filtros; Bancos de filtro determinados por NMF (**W**): (b) 4 filtros; (c) 6 filtros; (d) 8 filtros.

aplicando el método descrito en la sección 5.2 sobre el conjunto de entrenamiento correspondiente. En todos los subexperimentos, NMF se inicializa con el método usual, es decir, generando 10 matrices aleatorias (**W** y **H**), de tal manera que la factorización que produce la distancia euclídea menor entre $\mathbf{P^T}$ y (**W H**) es la que se escoge para inicialización. Posteriormente, estas matrices iniciales se refinan usando las reglas de actualización multiplicativas dadas en la ecuación (2.4) considerando como distancia, la divergencia KL y como criterio de parada, un máximo de 200 iteraciones. Después de todo este proceso, la matriz **W** resultante contiene la respuesta en frecuencia de los filtros del banco de filtros FC.

Las figuras 5.2 (b), (c) y (d) representan el banco de filtros obtenidos por NMF (**W**) para el primer subexperimento usando el procedimiento descrito previamente para $n_f = 4$, 6 y 8 filtros, respectivamente. Con el propósito de facilitar la comparación, el banco de filtros del sistema base **U** también se representa en la figura 5.2 (a). Nótese que, aunque la máxima frecuencia de modulación es $50Hz$ (las características a corto plazo se extraen cada $10ms$), para mejorar la visualización de las figuras se

Tabla 5.1: Tasa de clasificación [%] para diferentes conjuntos de características.

Características a corto plazo	Integración temporal	Tasa clasif. (segmento) [%]	Tasa clasif. (evento) [%]
MFCC	FC	$65{,}68 \pm 1{,}06$	$70{,}59 \pm 1{,}94$
MFCC_HP2	FC	$70{,}91 \pm 1{,}01$	$76{,}39 \pm 1{,}81$
MFCC_HP2	FC_NMF	$74{,}39 \pm 0{,}97$	$79{,}29 \pm 1{,}73$
MFCC + Δ	FC	$67{,}92 \pm 1{,}04$	$71{,}75 \pm 1{,}92$
MFCC_HP2 + Δ	FC	$72{,}36 \pm 0{,}99$	$76{,}39 \pm 1{,}81$
MFCC_HP2 + Δ	FC_NMF	$76{,}15 \pm 0{,}95$	$80{,}15 \pm 1{,}70$

representan solo las frecuencias por debajo de $20Hz$. A partir de la comparación de las figuras 5.2 (a) and (b), se puede observar que los filtros 1 y 2 de **U** aparecen aproximadamente en **W**. El filtro de frecuencia más alta en **W** presenta un ancho de banda elevado y cubre las frecuencias de modulación de los filtros base 3 y 4. Finalmente, el filtro 4 de **U** es sustituido por un filtro de baja frecuencia en **W**, sugiriendo que, para describir la estructura temporal de los MFCCs, las bajas frecuencias de modulación son más relevantes que las altas. Puede llegarse a la misma conclusión a partir de las figuras 5.2 (c) y (d), donde se puede observar que, cuando el número de filtros se incrementa, NMF tiende a situar más filtros en las bajas y medias frecuencias de modulación que en las altas. Por otra parte, es importante mencionar que los filtros resultantes no difieren mucho entre subexperimentos.

5.2.5. Experimentos con parámetros FC basados en NMF en condiciones limpias

La tabla 5.1 muestra los resultados logrados en términos de la tasa de clasificación promedio a nivel de segmento (porcentaje de segmentos correctamente clasificados) y a nivel de evento acústico (porcentaje de eventos correctamente clasificados) así como los intervalos de confianza del 95 % para las diferentes parametrizaciones consideradas. FC y FC_NMF indican, respectivamente, el uso del banco de filtros fijo y el

basado en NMF, ambos compuestos de 4 filtros, en el proceso de integración temporal de características. El sufijo $+\Delta$ indica que el conjunto de características a corto plazo incluye la primera derivada de los coeficientes cepstrales.

Primero de todo, se puede observar que, en general, el uso de los parámetros Δ mejoran los resultados de clasificación con respecto al caso en el que Δ no se considera, aunque estas diferencias no son estadísticamente significativas. De todos modos, ambos casos siguen la misma tendencias. De hecho, para ambos casos (sin o con Δ), cuando se compara MFCC con MFCC_HP2, cuando ambos usan el banco de filtros base $\mathbf{U}$ (FC), se puede observar que MFCC_HP2 logra los mejores resultados, siendo la diferencia en funcionamiento con respecto a MFCC estadísticamente significativa con una confianza del 95 %. Este resultado sugiere que el filtrado paso alto de la señal de audio antes del cálculo de los coeficientes cepstrales es útil para obtener características más discriminativas, y por lo tanto, para mejorar los resultados finales. Obsérvese que esta misma conclusión se extrajo en el capítulo 4 para el caso de integración temporal basada en estadísticos.

Con respecto al uso del banco de filtros extraídos por el procedimiento NMF en combinación con las características a corto plazo MFCC_HP2 (MFCC_HP2 + FC_NMF), se puede observar que esta parametrización supera al banco de filtros fijo (MFCC_HP2 + FC). En este caso, la reducción de error relativo con respecto a MFCC_HP2 + FC es de alrededor el 12,0 % a nivel de segmento y 12,3 % a nivel de evento cuando los parámetros Δ no se consideran y alrededor de 13,7 % a nivel de segmento y 15,9 % a nivel de evento cuando los Δ se incluyen. Además, en este último caso, las diferencias en rendimiento son estadísticamente significativas. Este resultado muestra que los filtros aprendidos por NMF son capaces de capturar la estructura dinámica de los coeficientes cepstrales, produciendo un banco de filtros FC más adecuado para CEA que el banco de filtros fijo.

Tabla 5.2: Tasa de clasificación [%] para diferente número de filtros en el banco de filtros FC obtenido con NMF.

Características a corto plazo	Número de filtros	Tasa clasif. (segmento) [%]	Tasa clasif. (evento) [%]
MFCC_HP2	4	$74{,}39 \pm 0{,}97$	$79{,}29 \pm 1{,}73$
	6	$74{,}02 \pm 0{,}97$	$79{,}19 \pm 1{,}73$
	8	$73{,}69 \pm 0{,}98$	$78{,}99 \pm 1{,}74$
	10	$73{,}65 \pm 0{,}98$	$79{,}09 \pm 1{,}73$
MFCC_HP2 + Δ	4	$76{,}15 \pm 0{,}95$	$80{,}15 \pm 1{,}70$
	6	$75{,}51 \pm 0{,}96$	$78{,}37 \pm 1{,}76$
	8	$74{,}70 \pm 0{,}97$	$76{,}20 \pm 1{,}82$
	10	$73{,}99 \pm 0{,}98$	$74{,}36 \pm 1{,}86$

5.2.6. Experimentos con diferente número de filtros en el banco de filtros FC basado en NMF en condiciones limpias

Para los dos tipos de conjuntos de características, MFCC_HP2 y MFCC_HP2 + Δ, se realizaron experimentos considerando 4, 6, 8 y 10 bandas en el banco de filtros FC basado en NMF. La tabla 5.2 contiene las correspondientes tasas de clasificación así como los correspondientes intervalos de confianza al 95 %.

Para MFCC_HP2, los resultados varían con el número de filtros, aunque las diferencias son pequeñas y no estadísticamente significativas. Sin embargo, para MFCC_HP2 + Δ, la tasa de clasificación disminuye según aumenta el número de bandas de frecuencia, sugiriendo que 4 filtros son suficientes para representar el comportamiento temporal de las características a corto plazo (especialmente para los parámetros Δ).

5.2.7. Experimentos con parámetros FC basados en NMF en condiciones ruidosas

Al igual que en el capítulo 4, sección 4.3.3, en este capítulo también se ha realizado el estudio del impacto que este nuevo esquema de parametrización basado en coeficientes de banco de filtros produce en el rendimiento del sistema de CEA frente a la presencia de ruido de ambiente, para lo cual se realizaron diversos experimentos usando seis tipos de ruido diferentes: (*aeropuerto*, *voces*, *restaurante*, *tren*, *sala de exposiciones* y *metro*) a SNRs desde $0dB$ hasta $20dB$ con pasos de $5dB$. Por brevedad, solo se muestran en esta sección los resultados para el caso en que se usan los parámetros estáticos junto con sus correspondientes derivadas.

En la figuras 5.3 y 5.4 se representan el promedio de la reducción del error relativo para cada tipo de ruido con respecto al experimento base (condiciones ruidosas sin filtrado paso alto de la señal de audio y banco de filtros FC en la integración de características) calculado sobre todas las SNRs consideradas en función del número de filtros de baja frecuencia eliminados a nivel de segmento y evento acústico con las configuraciones FC y FC_NMF, respectivamente. También se incluye la media de la reducción del error relativo sobre todos los tipos de ruido y SNRs.

Al igual que en los resultados en condiciones ruidosas mostrados en el capítulo 4, observamos una disminución en las tasas de clasificación siendo el impacto menos dañino para algunos tipos de ruido (*aeropuerto*, *voces*, *restaurante* y *tren*) que para otros (*sala de exposiciones* y *metro*).

Para el esquema FC, a partir de los resultados mostrados en la figura 5.3 se puede observar que para los tipos de ruido *aeropuerto*, *voces*, *restaurante* y *tren*, las tasas de clasificación a nivel de segmento y evento mejoran considerablemente cuando el número de filtros eliminados se incrementa, para todas las SNRs. En este caso, los valores óptimos se obtienen cuando no se consideran las frecuencias por debajo de $200 - 400Hz$ en el cálculo de los parámetros a corto plazo, lo que corresponde con la eliminación de los $5 - 10$ primeros filtros de baja frecuencia. Para los tipos de ruido

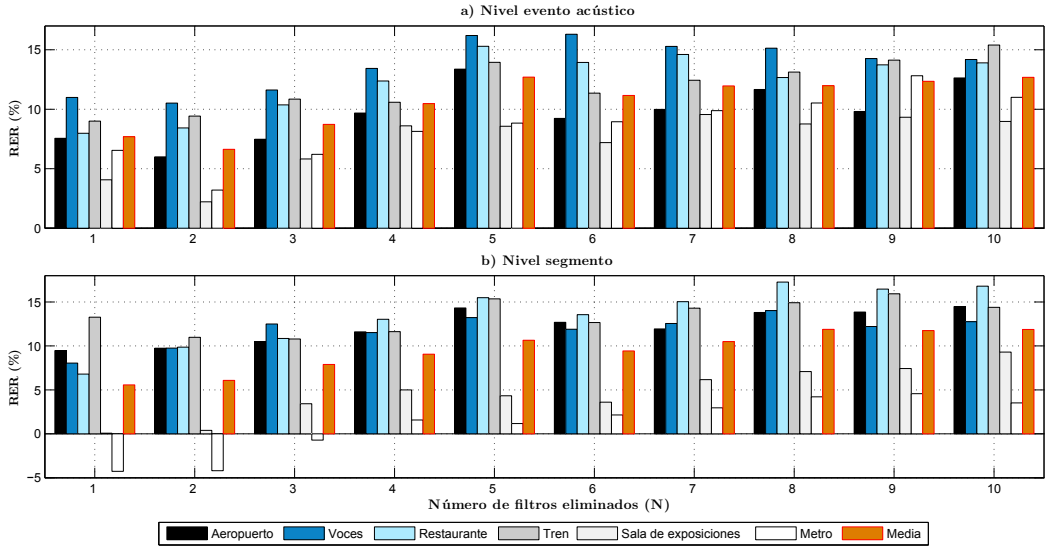


Figura 5.3: Reducción del error relativo [%] con respecto al experimento base para la parametrización MFCC_HPN + Δ + FC: (a) a nivel de evento acústico; (b) a nivel de segmento.

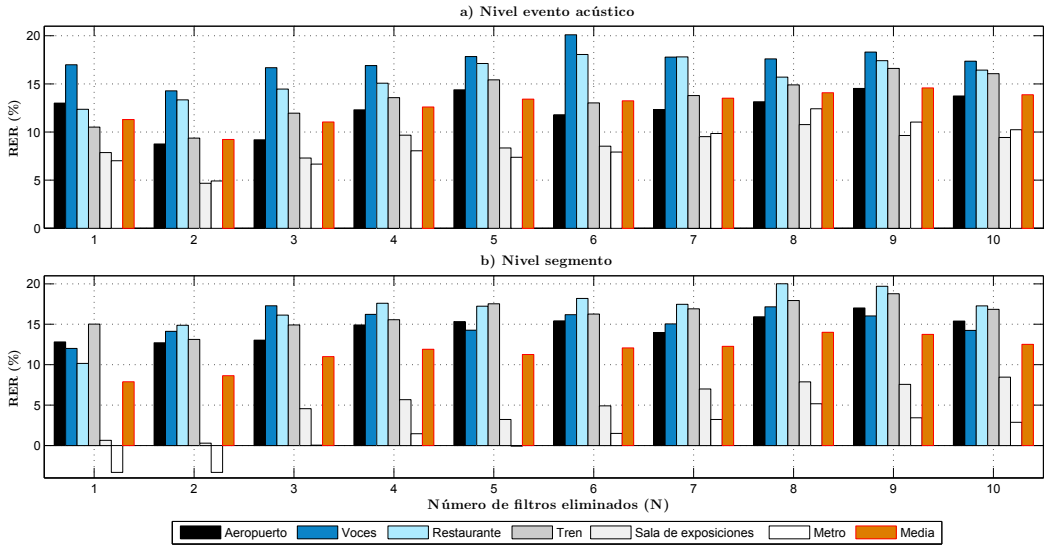


Figura 5.4: Reducción del error relativo [%] con respecto al experimento base para la parametrización MFCC_HPN + Δ + FC_NMF: (a) a nivel de evento acústico; (b) a nivel de segmento.

sala de exposiciones y *metro*, a diferencia de los resultados obtenidos en la sección 4.3.3 del capítulo 4, los resultados a nivel de segmento y evento mejoran de manera significativa cuando el número de filtros eliminados se incrementa, obteniendo los mejores resultados cuando se suprimen los $9-10$ primeros filtros de baja frecuencia.

En promedio, cuando se eliminan los primeros 9 filtros de baja frecuencia, se obtienen reducciones del error relativo con respecto al experimento base de aproximadamente 11,7 % a nivel de segmento y 12,3 % a nivel de evento acústico, respectivamente (ver figura 5.3).

Para el caso del esquema FC_NMF, los resultados se muestran en la figura 5.4. Nuevamente observamos que las tasas de clasificación a nivel de segmento y evento acústico mejoran considerablemente cuando el número de filtros eliminados se incrementa, para todas las SNRs para los tipos de ruido *aeropuerto*, *voces*, *restaurante* y *tren*. En este caso, los valores óptimos se obtienen cuando no se consideran frecuencias por debajo de $200-400Hz$, que corresponde con la eliminación de los $5-10$ primeros filtros de baja frecuencia. Para los tipos de ruido *sala de exposiciones* y *metro*, los resultados a nivel de segmento mejoran de manera significativa cuando el número de filtros eliminados se incrementa, en este caso corresponde a la eliminación de los $8-10$ primeros filtros de baja frecuencia; mientras que a nivel de evento, el número de filtros suprimidos óptimo es 8.

En promedio cuando se eliminan los primeros 9 filtros, se obtienen reducciones del error relativo con respecto al experimento base de aproximadamente 13,8 % a nivel de segmento y 14,6 % a nivel de evento, respectivamente (ver figura 5.4).

De acuerdo a los resultados de tasa de clasificación, promediado sobre todos los valores SNR considerados cuando son eliminados los primeros 9 filtros de baja frecuencia (MFCC_HP9), mostrados en las tablas 5.3 y 5.4, se muestra que el esquema FC_NMF supera a FC, siendo esta diferencia estadísticamente significativa para los tipos de ruido *aeropuerto*, *voces*, *restaurante* y *tren*; sin embargo esta diferencia no es significativa para los tipos de ruido *sala de exposiciones* y *metro*. Estos resultados se observan a nivel de segmento (ver tabla 5.3) y evento acústico (ver tabla 5.4).

Tabla 5.3: Tasa de clasificación promedio [%] a nivel de segmento, promediado sobre todos los valores SNR considerados con MFCC_HP9.

MFCC_HP9	**Esquemas de Parametrización**	
Ruido	**FC**	**FC_NMF**
Aeropuerto ($N1$)	$52{,}14 \pm 0{,}5$	$53{,}89 \pm 0{,}5$
Voces ($N2$)	$51{,}69 \pm 0{,}5$	$53{,}79 \pm 0{,}5$
Restaurante ($N3$)	$52{,}12 \pm 0{,}5$	$53{,}97 \pm 0{,}5$
Tren ($N4$)	$58{,}19 \pm 0{,}49$	$59{,}61 \pm 0{,}49$
Sala de exposiciones ($N5$)	$41{,}15 \pm 0{,}49$	$41{,}24 \pm 0{,}49$
Metro ($N6$)	$39{,}57 \pm 0{,}49$	$38{,}85 \pm 0{,}49$
Promedio ($N1 - N6$)	$49{,}14 \pm 0{,}20$	$50{,}23 \pm 0{,}20$
Promedio ($N1 - N4$)	$53{,}54 \pm 0{,}25$	$55{,}32 \pm 0{,}25$

Tabla 5.4: Tasa de clasificación promedio [%] a nivel de evento acústico, promediado sobre todos los valores SNR considerados con MFCC_HP9.

MFCC_HP9	**Esquemas de Parametrización**	
Ruido	**FC**	**FC_NMF**
Aeropuerto ($N1$)	$53{,}68 \pm 0{,}96$	$56{,}1 \pm 0{,}96$
Voces ($N2$)	$53{,}22 \pm 0{,}96$	$55{,}43 \pm 0{,}96$
Restaurante ($N3$)	$52{,}33 \pm 0{,}96$	$54{,}36 \pm 0{,}96$
Tren ($N4$)	$57{,}28 \pm 0{,}95$	$58{,}51 \pm 0{,}95$
Sala de exposiciones ($N5$)	$48{,}47 \pm 0{,}96$	$48{,}65 \pm 0{,}96$
Metro ($N6$)	$48{,}58 \pm 0{,}96$	$47{,}53 \pm 0{,}96$
Promedio ($N1 - N6$)	$52{,}26 \pm 0{,}39$	$53{,}43 \pm 0{,}39$
Promedio ($N1 - N4$)	$54{,}13 \pm 0{,}48$	$56{,}1 \pm 0{,}48$

5.3. Conclusiones

En este capítulo, hemos presentado un nuevo esquema de parametrización para AEC basado en la mejora de los parámetros FC mediante el uso de NMF. En particular, NMF se utiliza para el aprendizaje no supervisado del banco de filtros FC que captura el comportamiento temporal más relevante en las características a corto plazo. A partir de la respuesta en frecuencia de los filtrros obtenidos con NMF, hemos observado que las bajas frecuencias de modulación son más importantes que las altas frecuencias para distinguir entre diferentes eventos acústicos. Los experimentos han mostrado que las características segmentales obtenidas con este método logran mejoras significativas en el rendimiento de clasificación de un sistema de AEC basado en SVM en comparación con los parámetros FC obtenidos de un banco de filtros predefinido. También se realizaron experimentos en condiciones ruidosas, mostrando que al igual que en el caso de los resultados en el capítulo 4 se producen mejoras significativas cuando no se consideran frecuencias por debajo de los $200Hz$ para el cálculo de los coeficientes a corto plazo mel-cepstrales.

Capítulo 6

Parametrización basada en la selección automática de bandas espectrales para CEA

En el capítulo 4 de la presente tesis se mostró que las parametrizaciones convencionales utilizadas habitualmente para reconocimiento de habla y locutor (como los coeficientes MFCC) no son necesariamente las más apropiadas para la tarea de clasificación de eventos acústicos puesto que dichas parametrizaciones fueron diseñadas teniendo en cuenta las características espectrales de la voz, que son, en general, diferentes de las de los eventos acústicos. A partir del estudio llevado a cabo en dicho capítulo se determinó la importancia de las medias y altas frecuencias para discriminar entre diferentes eventos acústicos, lo que llevó a desarrollar un nuevo esquema de extracción de características acústicas basado en el filtrado paso alto de las señales de audio con el que se logró buenos resultados en condiciones limpias y ruidosas. Este filtrado paso alto se implementó en la práctica por medio de la modificación del banco de filtros auditivo en escala Mel, de tal forma que varios filtros de baja frecuencia no fueran considerados en el cálculo de los vectores de parámetros cepstrales. La determinación del número de filtros de baja frecuencia a eliminar se

realizó empíricamente.

A diferencia de la aproximación anterior, la idea principal en este capítulo es usar métodos de selección de características (FS, *Feature Selection*) para encontrar el conjunto óptimo de bandas de frecuencia para CEA. Para ello, se han considerado diversas técnicas de selección de características, en concreto, las que utilizan la Información Mutua (MI, *Mutual Information*) como medida de similitud entre los datos.

En la literatura reciente, se pueden encontrar varios ejemplos del uso de algoritmos de selección de características para CEA. En [Zhuang et al., 2008], el conjunto final de características se seleccionó de acuerdo a un criterio bayesiano aplicado sobre un conjunto formado por los MFCCs y energías en banda decorrelados previamente mediante Análisis de Componentes Principales (PCA, *Principal Component Analysis*). En [Zhuang et al., 2010] se propuso un método basado en AdaBoost para construir el mejor conjunto de características a partir del mismo conjunto de parámetros utilizado en el trabajo anterior sin PCA. En [Butko and Nadeu, 2010] se describe el uso de un algoritmo secuencial de selección de características basado en envoltorios denominado *Forward Wrapper* aplicado a 16 log-energías en banda filtradas en frecuencia y sus primeras derivadas. Finalmente en [Kiktova et al., 2013] se usaron dos algoritmos de selección basados en información mutua sobre un conjunto de parámetros derivados de un análisis espectro-temporal. Es de destacar que en contraposición con estos trabajos, en este capítulo no se pretende utilizar la selección de características para reducción de la dimensionalidad sino para obtener una mejor representación de los eventos acústicos mediante la determinación de las bandas espectrales más relevantes y menos redundantes. De hecho, en los trabajos antes mencionados, las características seleccionadas son directamente la entrada al clasificador, mientras que en nuestra propuesta las log-energías de los filtros escogidos son decorrelados usando la DCT de forma previa a su entrada al clasificador.

6.1. Algoritmos de selección de características basados en información mutua.

EL principal objetivo de los métodos de selección de características es construir subconjuntos de parámetros que sean útiles para la clasificación [Guyon and Elisseeff, 2003]. Estos métodos se dividen en aquellos que son dependientes del clasificador (métodos basados en envoltorios o *wrappers*) y los que son independientes del clasificador (métodos filtro o *filter*) [Guyon and Elisseeff, 2003]. Los métodos filtro buscan el mejor conjunto de características mediante el cálculo de una medida de similitud sobre los datos, tales como la distancia [Bins and Draper, 2001], [Sebban and Nock, 2002] o información mutua [Peng et al., 2005], [Fernandez et al., 2009] y [Brown et al., 2012] de forma independiente del clasificador en particular que se vaya a utilizar en el sistema final y, por lo tanto, es menos probable que sufran del efecto de sobreajuste y son menos costosos computacionalmente en comparación con los métodos *wrappers*. Por estas razones, en el presente trabajo, se escogió utilizar los métodos filtro, en particular, aquellos basados en información mutua.

6.1.1. Información Mutua

La información mutua es una medida natural de la cantidad de información que dos variables aleatorias tienen en común. Es simétrica y no - negativa y es cero si y solo si las variables son independientes [Cover and Thomas, 2006]. La información mutua puede ser vista como una forma de cuantificar la relevancia de una variable aleatoria con respecto a otra. Dados $\mathbf{L}$ y $\mathbf{S}$ dos variables aleatorias discretas y $\mathbf{l}$ and $\mathbf{s}$, dos valores adoptados por respectivamente, $\mathbf{L}$ y $\mathbf{S}$, la información mutua $\mathbf{I}(\mathbf{L};\mathbf{S})$ entre $\mathbf{L}$ y $\mathbf{S}$ está dada por

$$\mathbf{I}(\mathbf{L};\mathbf{S}) = \sum_{\mathbf{l}\in\mathbf{L}}\sum_{\mathbf{s}\in\mathbf{S}} \mathbf{p}(\mathbf{l},\mathbf{s}) \log\left(\frac{\mathbf{p}(\mathbf{l},\mathbf{s})}{\mathbf{p}(\mathbf{l})\,\mathbf{p}(\mathbf{s})}\right) \tag{6.1}$$

donde $\mathbf{p}(\mathbf{l})$ y $\mathbf{p}(\mathbf{s})$ son las distribuciones de probabilidad de $\mathbf{L}$ y $\mathbf{S}$ y $\mathbf{p}(\mathbf{l}, \mathbf{s})$ es su distribución de probabilidad conjunta.

6.1.2. Criterio de selección de características basado en información mutua

Los métodos de selección de características basados en MI utilizan un cierto criterio de selección, $\mathbf{J}$, que está relacionado con la información mutua entre las características y la clase (o etiqueta) a la que pertenecen y cuantifica la utilidad de un subconjunto de características para la tarea de la clasificación. En [Brown et al., 2012] se presenta un punto de vista unificado de varias técnicas de selección de características basadas en MI existentes en la literatura, mostrando que el criterio usado en alguno de ellos puede ser expresado como una combinación lineal de MIs, como se estable en (6.2),

$$\mathbf{J}(\mathbf{L_k}) = \mathbf{I}(\mathbf{L_k}; \mathbf{S}) - \beta \sum_{\mathbf{L_j} \in \theta} \mathbf{I}(\mathbf{L_k}; \mathbf{L_j}) + \gamma \sum_{\mathbf{L_j} \in \theta} \mathbf{I}(\mathbf{L_k}; \mathbf{L_j} | \mathbf{S}) \tag{6.2}$$

donde $\mathbf{L_k}$ es la característica que va ser evaluada para su inclusión en el conjunto de parámetros y θ es el conjunto de características actualmente seleccionado. El primer término asegura la relevancia de $\mathbf{L_k}$, el segundo término esta relacionado con la redundancia de $\mathbf{L_k}$ con respecto a las características seleccionadas en θ y el tercer término, llamado *redundancia condicional*, permite la inclusión de características correladas que, sin embargo, podrían ser útiles para la tarea de clasificación [Brown et al., 2012]. Los diferentes valores de las constantes β y γ conducen a diferentes algoritmos conocidos de selección de características. En particular, en este trabajo hemos considerado los siguientes métodos:

- **Mínima - Redundancia Máxima - Relevancia (mRMR, *Minimum-Redundancy Maximum-Relevance*)** ($\beta = \frac{\mathbf{1}}{|\theta|}$, siendo $|\theta|$ el tamaño del conjunto de características actual seleccionado y el parámetro $\gamma = \mathbf{0}$). Esta variante busca escoger las características que tienen la relevancia más alta con

respecto a las clases consideradas, mientras que se minimiza la redundancia [Peng et al., 2005].

- **Información Mutua Conjunta (JMI, *Joint Mutual Information*)** ($\beta = \frac{\mathbf{1}}{|\theta|}$ y $\gamma = \frac{\mathbf{1}}{|\theta|}$). Este método incluye el término de redundancia condicional para permitir la inclusión de características correladas con información complementaria [Meyer et al., 2008].

- **Extracción de características Informativa Condicional (CIFE, *Conditional Informative Feature Extraction*)** ($\beta = \mathbf{1}$ y $\gamma = \mathbf{1}$). Al igual que el método anterior también incluye los términos de redundancia y redundancia condicional, pero con diferentes valores o pesos de ponderación que en JMI [Lin and Tang, 2006].

- **Redundancia Condicional, (CondRed, *Conditional Redundancy*)** ($\beta = \mathbf{0}$ y $\gamma = \mathbf{1}$). Este variante no tiene en cuenta el término de redundancia.

6.2. Esquema de parametrización basado en la selección de bandas espectrales

En esta sección, describiremos el proceso de selección automática de las bandas espectrales más apropiadas para AEC y el procedimiento para obtener las características acústicas a corto plazo a partir de dichas bandas.

6.2.1. Selección de bandas espectrales basada en información mutua

El espacio de parámetros de entrada para los algoritmos de selección de características consiste en las log-energías obtenidas después de aplicar un banco de filtros auditivo en escala de frecuencia Mel sobre el espectro de magnitud de las instancias de eventos acústicos pertenecientes a la partición de entrenamiento de la base

de datos. En nuestro caso, estos parámetros se extraen cada $10ms$ usando ventanas de análisis de Hamming de longitud de $20ms$ y un banco de filtros en escala Mel compuesto por 40 bandas triangulares. Este proceso ha sido implementado usando la toolbox *VOICEBOX* [Brookes, 2009].

Se han considerado los cuatro algoritmos de selección de características basados en información mutua descritos en la sección anterior (mRMR, JMI, CIFE y CondRed), que se aplican sobre estos datos usando la toolbox *FEAST* [Brown et al., 2011], de tal manera que las variables involucradas en las ecuaciones (6.1) y (6.2) son las log-energías en banda, es decir, las salidas de los filtros en escala Mel[1], $\mathbf{L} \in \mathbb{R}^N$ (siendo N el número inicial de filtros), y $\mathbf{S}$ un conjunto finito y discreto de clases de eventos acústicos. Despúes de este procedimiento, para cada método de selección de características, se obtiene un ranking de las bandas espectrales seleccionadas. Estas bandas escogidas finalmente se ordenan en forma ascendente, de forma que este proceso puede verse como la modificación del banco de filtros en escala Mel original en el que se eliminan varios filtros.

6.2.2. Extracción de características

En la figura 6.1 se representa el diagrama de bloques del esquema propuesto para CEA, que consiste en dos etapas principales: extracción de características a corto plazo e integración temporal.

En la etapa de extracción de características a corto plazo, las señales de audio se analizan cada $10ms$ usando una ventana de Hamming de $20ms$ de longitud. Para cada ventana, se obtiene el espectro de magnitud que es filtrado con el banco de filtros modificado según el correspondiente método de selección de características. De esta forma solo se calculan las log-energías de las bandas de frecuencia seleccionadas. A continuación, el vector de log-energías resultantes se rellena con ceros hasta completar el número de filtros del banco de filtros original (en nuestro caso, 40) y posteriormente

[1]Como las log-energías son valores reales, antes de la selección propiamente dicha se realiza un proceso de cuantización uniforme con 256 niveles de estos valores.

se aplica sobre ellos la transformada de coseno discreto, dando lugar a un conjunto de 12 coeficientes cepstrales (C_1 hasta C_{12}). Hay que tener en cuenta que en el caso de usar el banco de filtros en escala Mel completo (es decir, cuando no se descarta ninguna de las bandas espectrales), los coeficientes resultantes son los MFCC convencionales. Finalmente, se calculan la log-energía total de cada trama y sus primeras derivadas y se añaden a los coeficientes cepstrales (MFCC + Δ).

Una vez que estos coeficientes cepstrales son extraídos, se aplica sobre ellos la técnica de integración temporal de características llamada coeficientes de banco de filtros (FC) descrita en la sección 5.1 el capítulo 5 de la presente tesis. En este capítulo, se consideran dos bancos de filtros FC diferentes, que son los que ya se utilizaron previamente en el el capítulo 5:

- Un banco de filtros predefinido compuesto de cuatro filtros correspondiente a las siguientes bandas de frecuencia: 1) $0Hz$, 2) $1-2Hz$, 3) $3-15Hz$ y 4) $20-43Hz$. Esta aproximación se denomina *FC*.

- Un banco de filtros compuesto de cuatro filtros aprendidos automáticamente usando un método no supervisado basado en NMF. Esta aproximación se denomina *FC_NMF* y se puede encontrar explicada con mayor detalle en la sección 5.2 del capítulo 5.

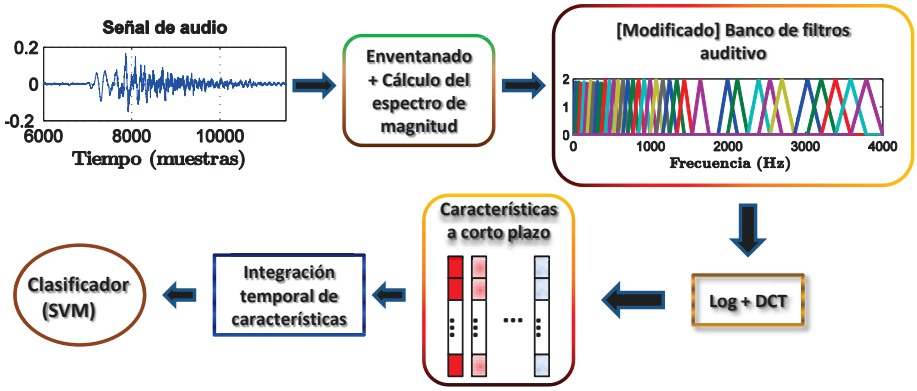


Figura 6.1: Diagrama del proceso de extracción de características.

6.3. Experimentos y resultados

6.3.1. Base de datos y sistema base

Tanto la base de datos para los experimentos, como el sistema base y el protocolo experimental son los mismos que se utilizaron en los capítulos 4 y 5 de la presente tesis y están descritos en la sección 4.3.1 del capítulo 4.

6.3.2. Bandas espectrales seleccionadas con los diferentes métodos

Para cada subexperimento, se realizó la selección de las bandas de frecuencia apropiadas para CEA, siguiendo el procedimiento descrito en la sección 6.2.1 y usando los datos de entrenamiento del subexperimento correspondiente. Las celdas de color azul en la figura 6.2 representan las 12 primeras bandas espectrales no seleccionadas determinadas por los algoritmos mRMR, JMI, CIFE y CondRed para el primer subexperimento. El número dentro de cada celda indica la posición en el ranking de las bandas descartadas (por ejemplo, para el algoritmo mRMR la primera banda que no es seleccionada es la banda 30). Las bandas descartadas no difieren mucho entre los 6 subexperimentos.

CondRed	1	2	3	4	5	6	7	8	9	10	11	12								
CIFE												11		9		7			6	
JMI													10	5	1	3	8		4	
mRMR																11			8	
# Banda	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

CondRed																				
CIFE	2	1	4	10	5		3	8	12											
JMI	2	7		11							9	6	12							
mRMR	6		4	10	2			7		1		3	9		5	12				
# Banda	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40

Figura 6.2: Bandas eliminadas por diferentes métodos de selección de características para el conjunto de entrenamiento del primer subexperimento.

A partir de esta figura, se puede observar el siguiente comportamiento de los métodos de selección de características. El método CondRed descarta los primeros filtros de baja frecuencia (es decir, es el método más equivalente al filtrado paso alto de los capítulos 4 y 5). Las bandas no seleccionadas por JMI están situadas en dos regiones de frecuencia diferentes, la primera desde $530Hz$ hasta $1530Hz$ y la segunda desde $2125Hz$ hasta $2685Hz$. El método mRMR descarta varias bandas no adyacentes en la región del espectro entre $920Hz$ hasta $3200Hz$. Finalmente, CIFE no selecciona bandas en una región casi continua entre $920Hz$ hasta $2110Hz$ y de manera dispersa entre $650Hz$ hasta $825Hz$.

6.3.3. Resultados en condiciones limpias

En esta subsección, presentamos los experimentos llevados a cabo con la finalidad de evaluar el rendimiento del esquema propuesto en comparación con el sistema base (es decir, cuando se usa el banco de filtros en escala Mel completo). Se han considerado dos diferentes parametrizaciones, FC and FC_NMF, tal y como se ha explicado en la subsección 6.2.2. Por brevedad, sólo se muestran los resultados a nivel de evento acústico.

La tasa de clasificación promedio a nivel de evento (porcentaje de eventos acústicos correctamente clasificados) para los sistemas de referencia es 71,75 % para FC

y 73,15 % para FC_NMF. En la figura 6.3 a) y b) se representan, respectivamente, las Reducciones de Error Relativo con respecto a los respectivos sistemas base FC y FC_NMF en función del número de bandas eliminadas por los cuatro algoritmos de selección de características considerados [2]: mRMR, JMI, CIFE y CondRed.

Como se puede observar, para la parametrización FC, considerar solo las bandas espectrales más importantes para el cálculo de las características a corto plazo siempre supera al experimento base, especialmente cuando el número de las bandas no seleccionadas está en el rango entre 6 y 12. Con respecto al funcionamiento de las diferentes técnicas de selección de características, el método CondRed produce la mejora más pequeña en comparación con los restantes algoritmos, mientras que los métodos mRMR y JMI logran resultados más similares. El método CIFE es el que produce los mejores resultados con respecto al experimento base, consiguiendo una reducción de error relativo entre 16 % y 19 % cuando se descartan más de 5 bandas.

En términos generales, FC_NMF sigue una tendencia similar a FC, aunque las reducciones de error relativo son más notables. Otra vez, las mejoras menores se obtienen con el método CondRed. Sin embargo, en este caso, JMI produce los mejores resultados con RERs por encima del 26 % en el rango desde 7 hasta 9 bandas espectrales no seleccionadas. De cualquier modo, en ambas parametrizaciones, los algoritmos de selección de características que exhiben mejor funcionamiento son aquellos en los que se tienen en cuenta los términos de redundancia y redundancia condicional (JMI y CIFE). En estos casos, las bandas de frecuencia no consideradas en el proceso de extracción de características pertenecen principalmente a la región media del espectro.

En la tabla 6.1 se muestran las tasas de clasificación a nivel de evento con sus

[2]Se han probado otros criterios basados en combinaciones lineales de la información mutua, tales como MIFS (*Mutual Information Feature Selection*) [Battiti, 1994] y combinaciones no lineales, tales como CMIM (*Conditional Mutual Information Maximization*) [Fleuret and Guyon, 2004] y DISR (*Double Input Symmetrical Relevance*) [Meyer and Bontempi, 2006]. Con estos métodos no se lograron mejores resultados que con los otros cuatro métodos considerados y, por brevedad, no hemos incluido los resultados correspondientes en la sección experimental.

correspondientes intervalos de confianza al 95 % obtenidos con FC y FC_NMF, para sus respectivos experimentos base y la mejor configuración de los diferentes métodos de selección de características. Para ambas parametrizaciones, la selección de bandas espectrales produce una mejora significativa con respecto a los sistemas de referencia. Para el esquema FC, el método CIFE obtiene los mejores resultados con 12 bandas espectrales descartadas, mientras que para FC_NMF, la tasa de clasificación más alta corresponde a JMI con 7 bandas no seleccionadas. En ambos casos, la mejora es similar con respecto a sus experimentos base respectivos (alrededor del 5 % absoluto). Finalmente, comparando las tasas con la mejor configuración, se puede observar que el esquema FC_NMF superan a FC, siendo las diferencias de funcionamiento estadísticamente significativas.

Tabla 6.1: Tasas de clasificación a nivel de evento acústico [%] para diferentes métodos de selección de características en condiciones limpias.

Método	FC		FC_NMF	
	Tasa. Classif. [%]	No. bandas espectrales	Tasa Classif. [%]	No. bandas espectrales
Base	71,75 ± 1,92	-	73,15 ± 1,89	-
mRMR	76,05 ± 1,82	11	79,53 ± 1,72	8
JMI	76,48 ± 1,81	7	**81,02 ± 1,67**	7
CIFE	**77,11 ± 1,79**	12	80,30 ± 1,70	12
CondRed	75,18 ± 1,84	4	79,00 ± 1,74	8

6.3.4. Resultados en condiciones ruidosas

En esta sección evaluamos el rendimiento del esquema propuesto en condiciones ruidosas usando los mismos tipos de ruido usados que en los capítulos 4 y 5 (*aeropuerto*, *voces*, *restaurante*, *tren*, *sala de exposiciones* y *metro*) en SNRs desde $0dB$ hasta $20dB$ con pasos de $5dB$. Por brevedad, sólo se muestran los resultados a nivel de evento acústico.

En la figura 6.4 a) y b) se representan, respectivamente, las reducciones de error

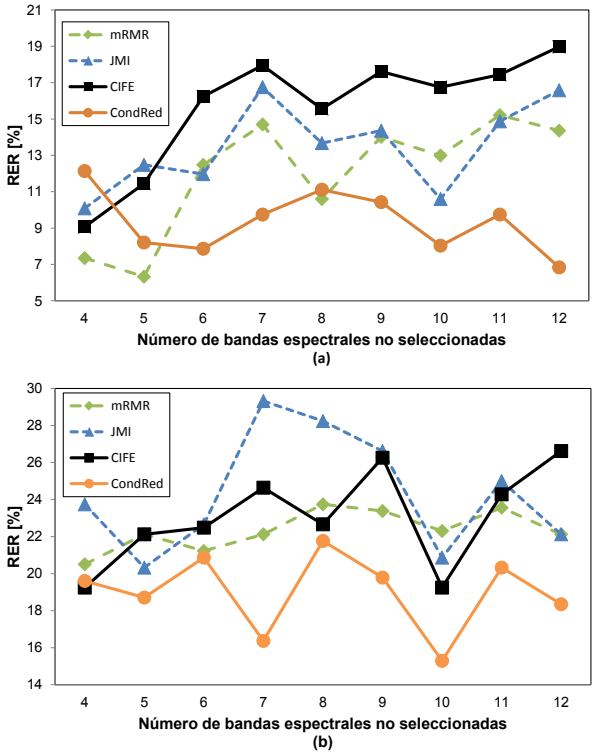


Figura 6.3: Reducción de error relativo [%] con respecto a sus respectivos experimentos base (a nivel de evento acústico) en condiciones limpias: (a) Parametrización FC; (b) Parametrización FC_NMF.

relativo con respecto a sus respectivos sistema base FC y FC_NMF en función del número de bandas eliminadas para los cuatro algoritmos de selección de características considerados (mRMR, JMI, CIFE y CondRed).

Como se puede observar, para ambas parametrizaciones FC y FC_NMF, al igual que en los resultados en condiciones ruidosas mostrados en el capítulo 4, en todos los casos se incrementa la tasa de clasificación cuando se descartan una o varias bandas espectrales. De la comparación de las diferentes técnicas de selección entre sí, podemos concluir que el método CondRed es el que produce mejores resultados, obteniendo con RERs con respecto a su respectivo experimento base entre 11 % y 13 % para la parametrización FC y entre 13 % y 16 % para FC_NMF cuando se

descartan más de 5 bandas. Los algoritmos restantes, mRMR, JMI y CIFE, producen resultados similares y su funcionamiento tiende a empeorar cuando se incrementa el número de bandas espectrales no seleccionadas.

Por otra parte, a diferencia de los resultados en condiciones limpias mostrados en la sección 6.3.3, en ambas parametrizaciones, el algoritmo de selección de características que exhibe mejor rendimiento es CondRed, en el que no se tiene en cuenta el término de redundancia y para la selección de bandas sólo se consideran los términos de relevancia y de redundancia condicional. En este caso, las bandas de frecuencia que se descartan se sitúan principalmente en la región baja del espectro, siendo este método el más similar a la aproximación empírica de filtrado paso alto desarrollada en los capítulos 4 y 5 de la presente tesis.

En la tabla 6.2 se muestran las tasas de clasificación promedio a nivel de evento acústico sobre todos los tipos de ruido y SNRs con sus correspondientes intervalos de confianza al 95 % obtenidos con FC y FC_NMF, para los respectivos experimentos base y la mejor configuración de los diferentes métodos de selección. Para ambas parametrizaciones, el hecho de descartar ciertas bandas espectrales da lugar a una mejora significativa con respecto a los respectivos sistemas base. Para FC, el método CondRed obtiene los mejores resultados con 5 bandas espectrales descartadas, mientras que para FC_NMF con 9 bandas espectrales eliminadas. En el caso del esquema FC, la mejora con respecto a su experimento base es alrededor del 7 % absoluto; mientras que en el esquema FC_NMF es de alrededor el $8,5\,\%$. Finalmente al comparar entre ambos esquemas, podemos observar que la parametrización FC_NMF produce resultados ligeramente mejores que FC con JMI y CIFE, siendo la diferencia más evidente con mRMR y CondRed.

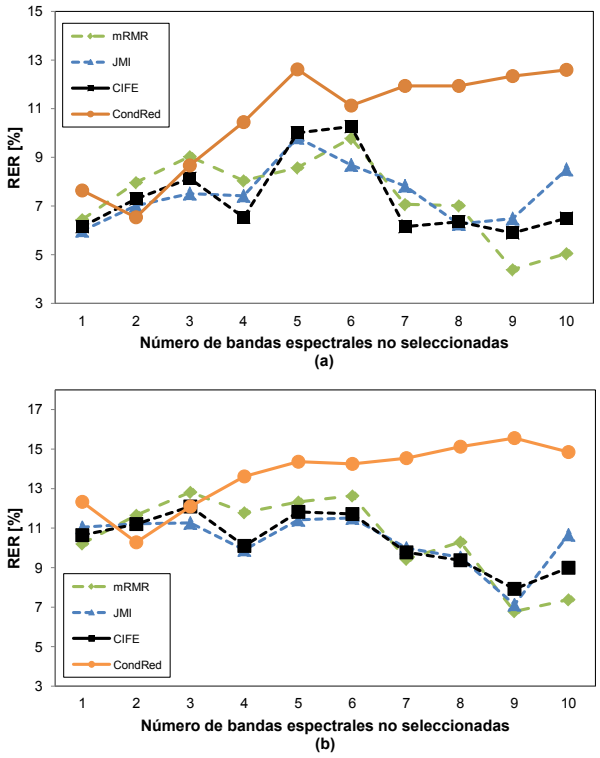


Figura 6.4: Reducción de error relativo [%] con respecto a sus respectivos experimentos base (a nivel de evento acústico) en condiciones ruidosas: (a) Parametrización FC; (b) Parametrización FC_NMF.

Tabla 6.2: Tasas de clasificación promedio a nivel de evento acústico sobre todos los tipos de ruido y SNRs [%] para diferentes métodos de selección de características en condiciones ruidosas.

Método	FC		FC_NMF	
	Tasa Clasif. [%]	No. bandas espectrales	Tasa Clasif. [%]	No. bandas espectrales
Base	$45{,}54 \pm 0{,}39$	-	$44{,}85 \pm 0{,}39$	-
mRMR	$50{,}87 \pm 0{,}39$	6	$51{,}92 \pm 0{,}39$	3
JMI	$50{,}98 \pm 0{,}39$	5	$51{,}20 \pm 0{,}39$	6
CIFE	$51{,}13 \pm 0{,}39$	6	$51{,}37 \pm 0{,}39$	5
CondRed	$\mathbf{52{,}41 \pm 0{,}39}$	5	$\mathbf{53{,}43 \pm 0{,}39}$	9

6.4. Conclusiones

En esta capítulo hemos presentado un nuevo módulo de parametrización para clasificación de eventos acústicos basado en la selección automática de bandas espectrales. Esta selección ha sido llevada a cabo por medio de la aplicación de varios algoritmos de selección de características basados en información mutua (mRMR, JMI, CIFE y CondRed) aplicados sobre las log - energías en banda en escala Mel. Una vez que las log-energías de los filtros seleccionados se calculan, se aplica sobre ellos la DCT produciendo un conjunto de coeficientes a corto plazo, que son finalmente combinados en una escala temporal más larga mediante dos técnicas de integración de características diferentes, FC and FC_NMF.

Los métodos de selección de características que logran mejores resultados son CIFE y JMI para, respectivamente, las parametrizaciones FC y FC_NMF en condiciones limpias. En ambos casos, las diferencias de funcionamiento con respecto a sus respectivos experimentos base (cuando se consideran todas las bandas de frecuencia) son estadísticamnete significativas, obteniendo reducciones de error relativo de alrededor del 19 % para FC y 29 % para FC_NMF. Estos resultados muestran que la selección de bandas de frecuencia es beneficiosa para CEA, siendo las bandas situadas en bajas y altas frecuencias las más relevantes y menos redundantes. Sin embargo en condiciones ruidosas, el mejor rendimiento se obtiene con el método de selección de características CondRed, obteniendo reducciones de error relativo de alrededor del 13 % para FC y 16 % para FC_NMF. En este caso, las bandas descartadas corresponden con las de baja frecuencia, por lo que esta técnica está estrechamente relacionada con la parametrización basada en filtrado paso alto desarrollada en los capítulos 4 y 5.

Capítulo 7

Parametrización basado en los Coeficientes de Activación NMF para CEA

La Factorización de Matrices No Negativa (NMF), es un método no supervisado que descompone una matriz de datos no - negativa como por ejemplo el espectro de magnitud de una señal de audio **V** como el producto de dos matrices no - negativas (**W** y **H**); donde las columnas de **W** contiene los vectores espectrales base (SBVs) que representan al espectro de la señal de audio y las filas de **H** contienen los coeficientes de activación o ganancia de los SBVs. La factorización se logra a través de la minimización de una determinada función de coste (por ejemplo, la divergencia de Kullback - Leibler, KL) usando un esquema iterativo con reglas de actualización multiplicativa. Para tal fin, se requiere de un adecuado proceso de inicialización, que normalmente se realiza usando matrices aleatorias con distribución uniforme, encontrandose con el problema de múltiples mínimos locales, afectando la convergencia del algoritmo y conduciendo a una factorización inapropiada, por lo que no sería adecuada para una aplicación de clasificación.

Con la finalidad de mitigar este problema se ha desarrollado una versión su-

pervisada del algoritmo NMF, que consiste en mantener constante la matriz que contiene los vectores espectrales base (**W**), actualizando solo la matriz de los coeficientes de activación **H** en el proceso de factorización [Cyril and Bjorn, 2012], [Yong-Choon et al., 2003], [Schuller and Weninger, 2010], [Cotton and Ellis, 2011]. Los coeficientes de activación NMF en combinación con las características mel-cepstrales han mostrado ser buenos parámetros acústicos para distintas tareas tales como clasificación de audio [Cyril and Bjorn, 2012], [Yong-Choon et al., 2003], reconocimiento robusto del habla [Schuller et al., 2010], discriminación de voz y vocalizaciones no linguísticas [Schuller and Weninger, 2010] y en la detección de eventos acústicos [Cotton and Ellis, 2011].

Esta variante supervisada del método NMF se presentó en el capítulo 3 para la mejora de la calidad de la señal de voz, que consiste en aprender un modelo acústico a partir de datos de voz y ruido. De tal manera que, para el proceso de mejora, el modelo acústico se mantiene constante, y solo se actualizan los coeficientes de activación NMF, usando restricciones de dispersión (*sparsity*), mostrando mejoras significativas en esta tarea.

En este capítulo, se muestra como los coeficientes de activación NMF son usados como características acústicas para la tarea de clasificación de eventos acústicos (CEA). A diferencia de otras aproximaciones, aplicamos la transformada del coseno discreto (DCT) al logaritmo de la matriz de los coeficientes de activación **H**, logrando una mejor representación de la estructura espectro - temporal de la señal de audio. Estos parámetros acústicos basado en NMF permiten extraer información complementaria y de robustez frente al ruido, lo que produce mejoras significativas en el rendimiento de clasificación.

Los experimentos muestran que las características de activación NMF combinado con el nuevo esquema de parametrización basado en el filtrado paso alto de las coeficientes mel-cepstrales presentado en el capítulo 4, logran mejoras significativas en el rendimiento de la tasa de clasificación de un sistema basado en una Máquina de Vectores Soporte (SVM) en condiciones limpias y ruidosas.

7.1. Extracción de características basada en NMF

En esta sección se describe el proceso de extracción de características a partir de los coeficientes de activación NMF en combinación con las características mel-cepstrales.

7.1.1. Aprendizaje de modelos acústicos basado en NMF

En esta sección, se describe la variante supervisada del algoritmo NMF para el aprendizaje de modelos acústicos de los diferentes eventos acústicos. En esta variante, solo la matriz de coeficientes de activación se actualiza iterativamente, manteniendo fija la matriz de los vectores espectrales base. En este caso buscamos aprender un modelo acústico para cada una de las clases correspondientes a los diferentes eventos acústicos, algo parecido al procedimiento desarrollado en el capítulo 3, donde se encontró un modelo para la voz y ruido a partir de datos de voz limpio y de ruido en la tarea de eliminación de ruido. Del mismo modo, los modelos acústicos se encuentran aplicando el algoritmo NMF sobre datos de audio limpios correspondiente al conjunto de entrenamiento. Primero se cálcula el espectro de magnitud de cada uno de los conjuntos de muestras correspondientes a cada clase de eventos acústicos por ejemplo, $|V_1|$, $|V_2|$, etc correspondiente a la clase 1, clase 2, etc respectivamente. Después se minimiza la divergencia de Kullback - Leibler entre su espectro de magnitud y sus correspondientes matrices factorizadas (W_1H_1, W_2H_2, etc.) usando reglas de aprendizaje dadas en la ecuación 2.4 del apartado 2.1 (capítulo 2). Debido a que NMF es un algoritmo iterativo, es importante realizar un adecuado proceso de inicialización de las matrices factorizadas. En este caso el proceso de inicialización se realiza usando el procedimiento de multi-inicio dado en la sección 2.1.2 del capítulo 2. Los vectores espectrales base contenidos en W_1, W_2, etc son usados como modelos acústicos para la clase 1, clase 2, etc respectivamente. En la figura 7.1 se muestra el proceso de determinación de los modelos acústicos usando NMF.

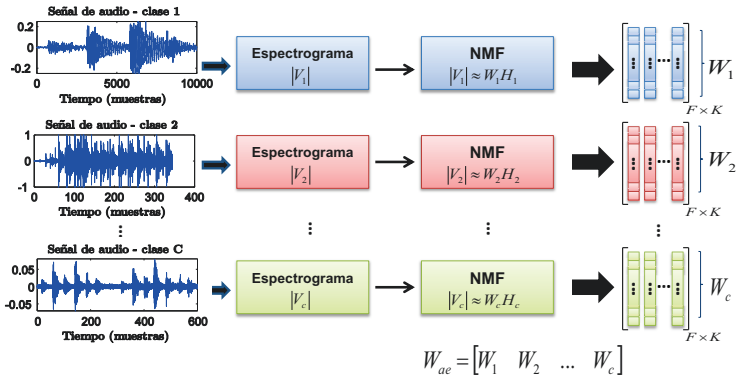


Figura 7.1: Modelo acústico basado en NMF.

7.1.2. Extracción de las características a corto plazo basado en NMF

Se asume que W_1, W_2, etc. encontradas en la sección anterior (7.1.1) son buenas funciones espectrales base que describen a las diferentes clases de eventos acústicos. Estas bases se mantienen fijas y luego son concatenadas para formar un único conjunto de SBVs llamado W_{ae}. Dado el espectro de magnitud de una señal de audio (evento acústico) $|V_{\text{test}}|$, calculamos su factorización $|V_{\text{test}}| \approx W_{\text{ae}} H_{\text{ae}}$ minimizando la divergencia KL entre $|V_{\text{test}}|$ y $(W_{\text{ae}} H_{\text{ae}})$, actualizando solo la matriz de coeficientes de activación H_{ae} con las reglas de aprendizaje dadas en la ecuación 2.4 del apartado 2.1.

En la figura 7.2 se representa el diagrama de bloques del proceso de extracción de características acústicas propuesto para CEA. Consiste de dos etapas principales: Extración de características a corto plazo basada en los coeficientes de activación NMF en combinación con las características a corto plazo a escala de frecuencia Mel con filtrado paso alto (MFCC_HPN), donde N indica el número de filtros de baja frecuencia eliminados y la técnica de Integración temporal de características, mostrado en el capítulo 5.

En la etapa de extracción de características a corto plazo, las señales de audio son analizadas cada $10ms$ usando una ventana de Hamming de $20ms$ de longitud. Las

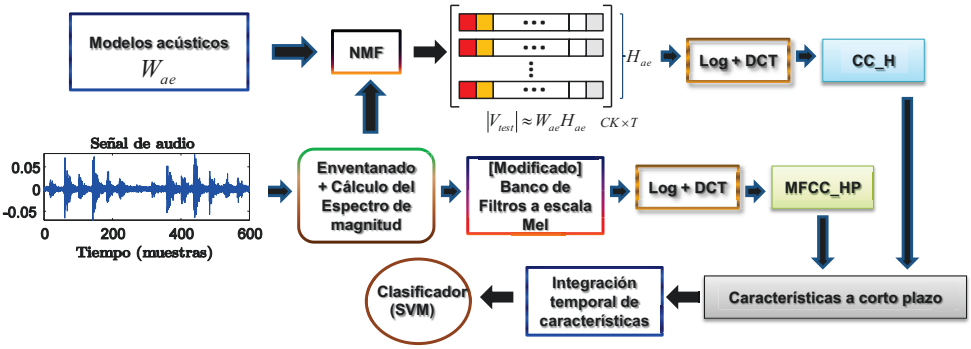


Figura 7.2: Diagrama de bloques del esquema combinado propuesto para la tarea AEC.

características basada en los coeficientes de activación NMF se encuentran aplicando la transformada del coseno discreto (DCT) sobre el logaritmo de la matriz de coeficientes de activación actualizada H_{ae}, lo que sería equivalente aplicar transformación cepstral sobre H_{ae}, generándose coeficientes cepstrales basado en los coeficientes de activación NMF llamada CC_H.

7.1.3. Extracción de características acústicas

Una vez que son extraídos los coeficientes cepstrales basado en los coeficientes de activación NMF (CC_H), se cálculan los coeficientes cepstrales a escala de frecuencia Mel con filtrado paso alto (MFCC_HPN), motivado por los buenos resultados mostrados en los capítulos 4, 5 y 6. Estos dos vectores de coeficientes cepstrales (CC_H y MFCC_HPN) son combinados para generar un único vector de características a corto plazo (MFCC_HPN + CC_H). Tener en cuenta que en el caso de usar un banco de filtros completo (en nuestro caso, 40) a escala Mel (por ejemplo, cuando no se descarta ninguna de las bandas espectrales ($N = 0$) en el cálculo de los coeficientes cepstrales), los coeficientes cepstrales corresponden a los MFCC convencionales. Finalmente, para la parametrización CC_H, se agrega el coeficiente de ganancia máxima NMF (G_NMF), de tal manera que para alguna trama t, la ganancia máxima de H_{ae} se cálcula de la siguiente manera:

$$g_t = \underset{k}{argmax}\left(H_{ae_{kt}}\right), k \in \{1, ..., KC\} \tag{7.1}$$

Donde K es el número de componentes NMF por clase (en nuestro caso se establece $K = 4$) y C es el número total de clases de eventos ($C = 12$). El coeficiente G_NMF ha sido motivado porque al parecer permite una buena caracterización del timbre del sonido y por su robustez frente al ruido [Schuller et al., 2010], [Cyril and Bjorn, 2012].

Para la parametrización MFCC_HPN, se adiciona la log - energía de cada frame.

Al vector de características a corto plazo combinado, se aplica la técnica de integración temporal de características, para obtener un conjunto de vectores de características acústicas en una escala de tiempo más larga donde se pueda capturar el comportamiento temporal de los parámetros a corto plazo. El procedimiento es como sigue: la secuencia de coeficientes a corto plazo (MFCC_HPN + CC_H) + sus primeras derivadas (cuando es indicado) se dividen en segmentos de $2s$ de longitud con un solape de $1s$; a partir de esta segmentación, se aplica el método de integración temporal de características. En este capítulo hemos considerado dos métodos de integración temporal de características, en el primero, se cálculan los estadísticos (media, varianza y simetría) de cada uno de segmentos (para más detalle ver capítulo 4). A esta parametrización se le denomina características acústicas basado en estadísticos y en el segundo, se cálcula sobre cada segmento los periodogramas de cada dimensión de las características a corto plazo y, entonces, son resumidos a través del cálculo de la potencia en diferentes bandas de frecuencia usando un cierto banco de filtros. En este caso, se consideran dos bancos de filtros diferentes (**U** y **W**), los mismos que son usados en la sección 5.1 (capítulo 5). A esta parametrización se le denomina características acústicas basada en coeficientes de banco de filtros (FC, si se usa el banco de filtros **U** y FC_NMF, si se usa el banco de filtros **W** aprendido usando el algoritmo NMF.

Tabla 7.1: Tasa de Clasificación [%] para diferentes configuraciones de características a corto plazo basado en NMF.

Características a corto plazo	Tasa clasif. (segmento) [%]	Tasa clasif. (evento acústico) [%]	Número de características
CC_H	64,76 ± 1,06	74,41 ± 1,88	13
CC_H + G_NMF	68,79 ± 1,03	75,13 ± 1,86	14
Log_H	70,51 ± 1,01	79,77 ± 1,73	48
Log_H + G_NMF	71,83 ± 1,00	79,96 ± 1,72	49
MFCC + CC_H + G_NMF	79,27 ± 0,90	83,67 ± 1,59	27
MFCC + Log_H + G_NMF	76,18 ± 0,95	84,60 ± 1,56	62

7.2. Experimentos y resultados

7.2.1. Base de datos y sistema base

La base de datos usada para la expermientación, el sistema base y el protocolo experimental son los mismos que se utilizaron en los capítulos 4, 5 y 6 de la presente tesis y están descritos en la sección 4.3.1 del capítulo 4.

7.2.2. Experimentos en condiciones limpias

En esta subsección, presentamos los experimentos llevados a cabo con la finalidad de evaluar el rendimiento del esquema propuesto en condiciones limpias (cuando ningún ruido se adiciona a los archivos de audio original).

Para tal fin, se ha llevado a cabo una experimentación preliminar usando características segmentales basado en los estadísticos de las características a corto plazo. Estos resultados se muestran en la tabla 7.1, donde se puede observar que el rendimiento en la tasa de clasificación, con las características basadas en el logaritmo de los coeficientes de activación NMF (Log_H), es superior a las características obtenidas aplicando transformación cepstral sobre los coeficientes de activación NMF (CC_H). Sin embargo, cuando se realiza la combinación con las características MFCC, notamos que la configuración cepstral de los coeficientes de activación NMF es mejor que

los parámetros Log_H, a nivel de segmento siendo esta mejora estadísticamente significativa. A nivel de evento acústico no se muestra una diferencia estadísticamente significativa entre ambas configuraciones. Cuando se agrega el término de ganancia G_NMF, se produce una mejora en ambas configuraciones, siendo mayor con los parámetros CC_H. De acuerdo a estos resultados y teniendo en cuenta el número de características usadas mostradas en la tabla 7.1, la configuración que se utiliza a lo largo de este capítulo es CC_H.

A continuación describimos la parametrización usada:

- Para las características CC_H: se extraen 13 coeficientes cepstrales ($C1$ al $C13$). También fue calculado y adicionado a los coeficientes cepstrales el término de ganancia máxima NMF (G_NMF) y sus primeras derivadas (donde es indicado).

- Para las características MFCC_HPN: se extraen 12 coeficientes cepstrales ($C1$ al $C12$) a partir de eliminar ciertas bandas de baja frecuencia dado por el valor de N, en el banco de filtros auditivo a escala de frecuencia Mel. Cuando no se elimina ninguna banda de frecuencia (en nuestro caso de 40 bandas espectrales) en el banco de filtros auditivo, el cálculo de las características MFCC_HPN corresponde a los MFCC convencionales. También fueron calculados y adicionados a los coeficientes cepstrales MFCC_HPN, la log-energía de cada trama (en vez del coeficiente de orden cero $C0$) y sus primeras derivadas (donde es indicado).

Finalmente, para el caso de la parametrización basada en los estadísticos, los vectores de características acústicos finales consisten de los estadísticos (media, desviación estándar y simetría) de la combinación de ambas características MFCC_HPN y CC_H. Mientras que para la parametrización basada en los coeficientes de banco de filtros, cuando a los segmentos de características (MFCC_HPN + CC_H) se le aplica un banco de filtros fijo ($\mathbf{U}$), los vectores acústicos final reciben el nombre de MFCC_HPN + CC_H + FC. Cuando el banco de filtros usado es $\mathbf{W}$, aprendido

Tabla 7.2: Tasa de clasificación promedio [%] (segmento) en condiciones limpias.

Param.	Número de filtros eliminados (N)												
	MFCC	MFCC_HPN + CC_H											
	Base	0	1	2	3	4	5	6	7	8	9	10	11	12
CC	75.10	79.27	80.1	79.65	80.05	80.48	80.46	79.78	80.12	79.56	80.37	80.57	79.65	79.92
CC+ΔCC	77.57	78.83	80.21	80.37	79.97	80.08	80.35	80.17	80.55	80.35	80.5	79.74	79.49	80.58

por NMF, los vectores acústicos final reciben el nombre de MFCC_HPN + CC_H + FC_NMF.

7.2.2.1. Experimentos con la parametrización basada en los estadísticos de las características a corto plazo

Las tablas 7.2 y 7.3 muestran respectivamente los resultados logrados en términos de la tasa de clasificación promedio a nivel de segmento (porcentaje de segmentos correctamente clasificados) y a nivel de evento acústico (porcentaje de eventos acústicos correctamente clasificados), para la parametrización basada en estadísticos, variando el número de bandas de baja frecuencia eliminadas (N) en el banco de filtro auditivo. Los resultados para el sistema base (cuando ninguna banda de frecuencia es eliminada) también esta incluida. Ambas tablas contienen las tasas de clasificación para dos diferentes conjunto de parámetros acústicos, CC (Coeficientes cepstrales + log-energía) y CC + ΔCC (coeficientes cepstrales + log-energía + sus primeras derivadas).

Para la parametrización CC, el esquema combinado (MFCC_HPN + CC_H) supera al sistema base cuando no se elimina ningún filtro de baja frecuencia en el banco de filtros auditivo (MFCC_HP0), siendo este resultado estadísticamente significativo a nivel de segmento con una reducción del error relativo alrededor del 16,75 %. Con respecto a nivel de evento acústico se produce una mejora con respecto al sistema base con una reducción del error relativo alrededor del 13,74 %. Del mismo modo, se produce una mejora en el rendimiento del sistema combinado con respecto al siste-

Tabla 7.3: Tasa de clasificación promedio [%] (evento acústico) en condiciones limpias.

Param.	Número de filtros eliminados (N)													
	MFCC	MFCC_HPN + CC_H												
	Base	0	1	2	3	4	5	6	7	8	9	10	11	12
CC	81.07	83.67	84.11	83.58	84.89	84.26	84.16	84.21	84.84	84.45	84.45	84.11	85.03	84.02
CC+ ΔCC	81.41	82.81	83.92	84.11	84.55	84.4	84.31	84.21	84.02	84.11	84.4	83.82	82.57	83.39

ma base cuando se realiza el filtrado paso alto de la señal del evento acústico, siendo esta mejora más notable cuando el número de filtros eliminados varía desde 3 hasta 11. A partir de la figura 4.3 mostrada en la sección 4.2 del capítulo 4, se puede observar que estos rangos de filtros eliminados corresponden a frecuencias debajo de los aproximadamente 100 - 460Hz. Además podemos observar que dentro de este rango de filtros de baja frecuencia, las variaciones en el rendimiento en las tasas de clasificación son muy pequeñas, de tal manera que para efectos de comparación con los resultados obtenidos en la sección 4.3.2 capítulo 4, podemos considerar que cuando no son considerados en el cálculo de los coeficientes cepstrates los primeros siete filtros de baja frecuencia (MFCC_HP7), se obtienen mejoras significativas en el rendimiento de clasificación. En este caso, la diferencia en rendimiento a nivel de segmento con respecto al sistema base es estadísticamente significativa a un nivel de confianza del 95 %, alcanzando reducciones de error relativo de alrededor del 20,16 % a nivel de segmento y 19,92 % a nivel de evento acústico. Lo que produce una mejora significativa comparado con los resultados encontrados en el capítulo 4 en condiciones limpias.

Para la parametrización CC + ΔCC se produce una mejora en el rendimiento de clasificación con respecto al sistema base cuando no se elimina ningún filtro de baja frecuencia, a nivel de segmento y evento acústico, aunque esta mejora no es estadísticamente significativa. Al igual que en la parametrización CC, los mejores resultados se obtienen cuando las bajas frecuencias (por debajo de $100 - 460Hz$) no

son considerados en el proceso de extracción de características. Cuando se compara con CC para el caso de los 7 primeros filtros paso banda eliminados, se puede observar que CC + ΔCC logra una mejora de aproximadamente 0,43 % absoluto a nivel de segmento y una disminución alrededor de 0,82 % absoluto a nivel de evento acústico sobre CC. Sin embargo estas diferencias no son estadísticamente significativas.

Con la finalidad de realizar un análisis más detallado acerca del rendimiento del sistema CEA en base a este nuevo esquema de parametrización, hemos analizado las matrices de confusión producido por el sistema base y el esquema propuesto. Como ejemplo, las figuras 7.3(a) y (b) muestran las matrices de confusión a nivel de segmento para los parámetros CC + ΔCC del sistema base y la versión modificada de esta parametrización cuando los primeros 7 filtros son eliminados. En ambas tablas, las columnas corresponden a la clase correcta, las filas son la clase hipotetizada y los valores dentro de ellas son calculados sobre el promedio de estos 6 subexperimentos. Como se puede observar, en el esquema propuesto, la tasa de reconocimiento de todas las clases acústicas se incrementan con la excepción de la clases *teclear* y *tintineo llaves*. La principal mejora en el esquema propuesto se debe a que las clases *aplausos*, *movimiento sillas*, *tocar puerta*, *arrugar papel*, *timbre telefónico* y *tintineo cuchara/taza* reducen significamente su cantidad de confusiones en comparación al sistema base con una diferencia mayor al 4 % absoluto. En comparación con los resultados mostrados en el capítulo 4, el esquema combinado MFCC_HP7 + CC_H incrementa significativamente el rendimiento de clasificación para una cantidad mayor de clases.

7.2.2.2. Experimentos con la parametrización basada en los coeficientes de banco de filtros

En esta subsección evaluamos el esquema combinado propuesto usando la parametrización basada en los coeficientes de banco de filtros. En la tabla 7.4 se muestran los resultados logrados en términos de la tasa de clasificación promedio a nivel de segmento y a nivel de evento acústico, así como los intervalos de confianza al 95 %

	1	2	3	4	5	6	7	8	9	10	11	12
1	90,40	0,51	0,00	0,25	0,53	0,12	0,44	0,54	0,00	0,15	0,00	1,07
2	0,00	67,95	3,09	3,56	1,60	2,04	7,06	0,54	0,41	1,04	6,67	0,00
3	0,62	1,79	65,73	2,54	2,67	0,54	1,43	0,76	0,68	5,06	3,23	0,71
4	0,00	0,51	2,25	68,19	3,73	1,26	0,11	0,11	0,27	5,95	0,22	0,71
5	0,00	0,51	0,56	5,85	75,20	0,72	0,33	0,76	1,08	1,49	1,51	0,89
6	0,00	3,08	4,49	1,27	4,27	80,54	2,76	7,34	2,71	4,02	6,24	2,84
7	8,05	23,33	4,78	4,07	4,80	0,66	84,23	2,37	4,74	1,64	3,44	2,31
8	0,62	1,03	2,25	0,51	1,33	11,71	1,54	80,58	3,11	4,32	1,29	12,08
9	0,31	0,26	1,40	0,00	1,33	0,42	1,32	1,83	84,57	0,45	3,66	1,95
10	0,00	0,00	12,36	12,21	3,20	0,96	0,22	1,73	0,14	74,26	5,81	2,13
11	0,00	1,03	0,56	0,00	0,80	0,42	0,11	0,76	1,22	1,04	64,95	2,13
12	0,00	0,00	2,53	1,53	0,53	0,60	0,44	2,70	1,08	0,60	3,01	73,18

(a)

	1	2	3	4	5	6	7	8	9	10	11	12
1	97,83	3,08	0,28	2,80	1,60	0,18	1,32	1,83	1,35	0,45	0,65	0,89
2	0,00	70,51	3,09	1,27	0,53	1,68	4,96	0,54	0,41	1,04	4,30	0,71
3	0,31	0,77	73,31	2,54	1,87	0,96	1,76	0,22	0,00	3,42	0,86	0,71
4	0,00	1,03	1,40	78,12	5,33	2,58	0,11	0,43	0,14	5,65	0,00	0,36
5	0,00	0,26	1,12	6,62	77,60	1,44	0,33	0,43	0,27	0,74	0,22	0,18
6	0,00	1,54	3,37	0,25	1,33	79,52	0,33	5,39	1,62	2,38	5,38	4,09
7	0,62	20,77	5,62	3,05	5,33	1,02	87,65	0,76	2,71	2,38	6,45	2,66
8	1,24	1,03	3,37	0,00	1,87	9,97	1,54	84,90	1,62	5,06	1,08	11,72
9	0,00	0,26	0,84	0,00	1,33	0,78	0,55	1,08	89,04	0,30	5,81	2,13
10	0,00	0,00	5,90	4,33	1,87	0,60	0,22	0,76	0,14	77,38	1,51	2,49
11	0,00	0,26	0,00	0,76	0,53	0,48	0,66	0,11	2,03	0,45	70,54	2,84
12	0,00	0,51	1,69	0,25	0,80	0,78	0,55	3,56	0,68	0,74	3,23	71,23

(b)

Figura 7.3: Matrices de confusión [%] a nivel de segmento para la parametrización CC+ ΔCC: (a) Base; (b) Esquema con los 7 primeros filtros de baja frecuencia eliminados con la parametrización basada en los estadísticos del esquema combinado.

para el esquema combinado (MFCC_HPN + CC_H) con la parametrización basada en coeficientes de banco de filtros FC y FC_NMF mostrado en el capítulo 5. Donde FC y FC_NMF indican, respectivamente, el uso del banco de filtros fijo ($\mathbf{U}$) y el basado en NMF ($\mathbf{W}$), ambos compuestos de 4 filtros paso banda. El sufijo $+\Delta$ indica que el conjunto de características a corto plazo incluye la primera derivada de los coeficientes cepstrales y el término MFCC_HP2, indica que se han eliminado dos bandas espectrales de baja frecuencia ($N = 2$) en el banco de filtros auditivo. El sistema base esta formado por los parametrización segmental basado en coeficientes de banco de filtros FC y coeficientes MFCC (MFCC + FC).

De acuerdo a los resultados mostrados en la tabla 7.4, podemos observar que se produce una mejora en el rendimiento de la tasa de clasificación usando el esquema combinado (MFCC + CC_H) con respecto al sistema base. Esta mejora, a nivel de segmento, es estadísticamente significativa, logrando en este caso una reducción de

Tabla 7.4: Tasa de Clasificación [%] para diferentes conjuntos de características.

Características a corto plazo	Integración temporal	Tasa clasif. (segmento) [%]	Tasa clasif. (evento acústico) [%]
MFCC	FC	$65{,}68 \pm 1{,}06$	$70{,}59 \pm 1{,}94$
MFCC + CC_H	FC	$72{,}76 \pm 0{,}99$	$78{,}17 \pm 1{,}76$
MFCC_HP2 + CC_H	FC	$72{,}76 \pm 0{,}99$	$79{,}24 \pm 1{,}73$
MFCC + CC_H	FC_NMF	$71{,}95 \pm 1{,}00$	$79{,}33 \pm 1{,}73$
MFCC_HP2 + CC_H	FC_NMF	$73{,}68 \pm 0{,}98$	$80{,}35 \pm 1{,}69$
MFCC + Δ	FC	$67{,}92 \pm 1{,}04$	$71{,}75 \pm 1{,}92$
MFCC + CC_H + Δ	FC	$71{,}05 \pm 1{,}01$	$73{,}01 \pm 1{,}89$
MFCC_HP2 + CC_H + Δ	FC	$72{,}6 \pm 0{,}99$	$75{,}13 \pm 1{,}84$
MFCC + CC_H + Δ	FC_NMF	$72{,}10 \pm 1{,}00$	$75{,}33 \pm 1{,}84$
MFCC_HP2 + CC_H + Δ	FC_NMF	$74{,}78 \pm 0{,}97$	$77{,}45 \pm 1{,}78$

error relativo de alrededor del 20,63 % para FC y 18,27 % para FC_NMF sin considerar los parámetros Δ y alrededor del 9,76 % para FC y 13,03 % para FC_NMF con los parámetros Δ. Mientras que a nivel de evento acústico, la mejora es estadísticamente significativa solo cuando no son considerados los parámetros Δ, con una reducción de error relativo de alrededor de 25,77 % para FC y 29,72 % para FC_NMF.

Cuando se eliminan dos bandas de baja frecuencia en el banco de filtros auditivo (mostrados como MFCC_HP2 en la tabla 7.4), se produce en general una mejora estadísticamente significativa en el rendimiento del sistema con respecto al sistema base (MFCC + FC). En este caso a nivel de segmento, la reducción de error relativa es alrededor del 20,63 % para FC y 23,31 % para FC_NMF sin considerar los parámetros Δ y alrededor del 14,59 % para FC y 21,38 % para FC_NMF con los parámetros Δ. Mientras que a nivel de evento acústico las reducciones de error relativo es alrededor del 29,41 % para FC y 33,19 % para FC_NMF con los parámetros CC, siendo esta mejora estadísticamente significativa; cuando son considerados los parámetros Δ, la reducción de error relativa es alrededor del 11,96 % para FC y 20,18 % para FC_NMF. Además en este último caso las diferencias en rendimiento son estadísticamente significativas.

Cuando se compara el esquema combinado con las configuraciones FC y FC_NMF, se puede observar que la parametrización FC_NMF, supera a la parametrización FC

con una reducción de error relativa de alrededor del 3 % a nivel de segmento y 5 % a nivel de evento acústico cuando los parámetros Δ no son considerados y alrededor del 8 % a nivel de segmento (siendo en este caso estadísticamente significativo) y 9 % a nivel de evento acústico cuando el parámetro Δ es incluido.

Estos resultados muestran que al igual que en los experimentos realizados en el capítulo 5, los filtros aprendidos por NMF son más adecuados para la tarea CEA que el banco de filtros fijo. Del mismo modo realizar un filtrado paso alto a la señal del evento acústico a través de eliminación de bandas de baja frecuencia en el banco de filtros auditivo también resulta beneficioso al sistema combinado propuesto.

7.2.3. Experimentos en condiciones ruidosas

En esta subsección se realiza el estudio del impacto de ambientes ruidosos sobre el rendimiento del sistema combinado MFCC_HPN + CC_H, para tal fin, varios experimentos fueron llevados a cabo usando los mismos tipos de ruido considerados en los capítulos 4, 5 y 6 (*aeropuerto*, *voces*, *restaurante*, *tren*, *sala de exposiciones* y *metro*) en SNRs desde $0dB$ hasta $20dB$ con pasos de $5dB$. Por brevedad, solo reportamos en esta subsección resultados para el sistema base (banco de filtros completo) y para el esquema combinado propuesto en el caso de parámetros CC + ΔCC.

7.2.3.1. Experimentos con la parametrización basada en los estadísticos de las características a corto plazo

En la figura 7.4 se representa el promedio de la reducción del error relativo para cada tipo de ruido con respecto al sistema base (condiciones ruidosas sin filtrado paso alto de la señal de audio) calculada a través de las SNRs consideradas ($0dB$ hasta $20dB$ con pasos de $5dB$) como una función del número de filtros de baja frecuencia eliminados (N) a nivel de segmento y evento acústico. La media de la reducción del error relativo sobre todos los tipos de ruido y SNRs son también indicados.

Con la finalidad de observar con mayor detalle el comportamiento del esquema combinado con respecto a todos los tipos de ruido y SNRs, en la tabla 7.5 muestra

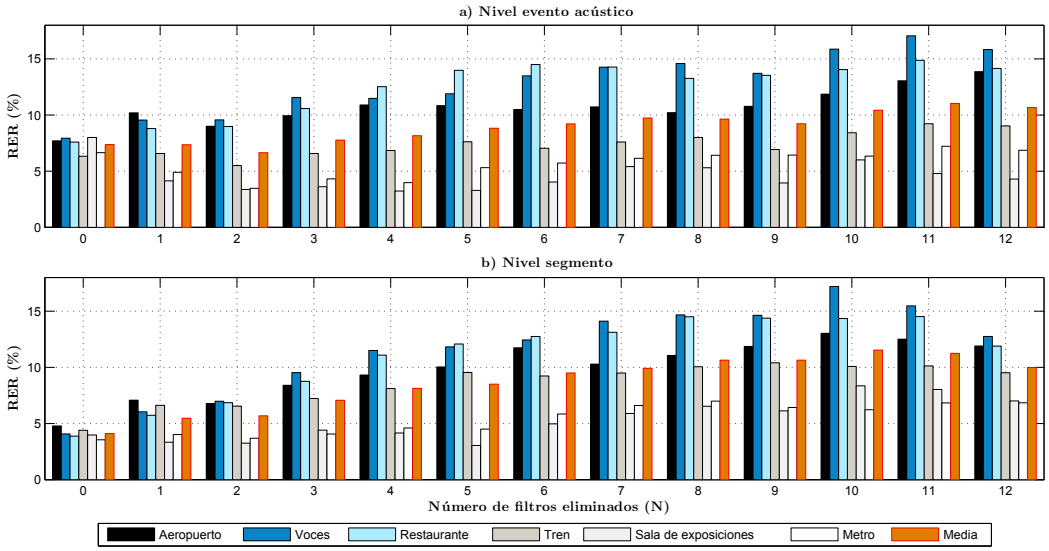


Figura 7.4: Reducción del error relativa [%] con respecto al sistema base para la parametrización CC+ΔCC, la escala Mel y en condiciones ruidosas: (a) a nivel de evento acústico; (b) a nivel de segmento.

las tasas de clasificación a nivel de segmento para el sistema base y para el esquema propuesto en varias SNRs seleccionadas (20, 10 y $0dB$) para los seis tipos de ruido considerados y el rango del número de filtros eliminados desde 7 hasta 12.

A diferencia de los resultados mostrados en el capítulo 4, se produce una mejora significativa en el rendimiento de la tasa de clasificación para aquellos tipos de ruido (*sala de exposiciones* y *metro*) donde debido a su distribución espectral (medias y altas frecuencias) considerablemente enmascara la estructura espectral fundamental de los eventos acústicos. Estos resultados muestran la característica de robustez del algoritmo NMF frente al ruido, mejorando de esta manera el rendimiento de la tasa de clasificación.

A partir de los resultados mostrados en la figura 7.4 se puede observar que con respecto al sistema base, el rendimiento en la tasa de clasificación del esquema combinado propuesto (MFCC_HPN + CC_H), para todos los tipos de ruido considerados (*aeropuerto*, *voces*, *restaurante*, *tren*, *sala de exposiciones* y *metro*), mejora conside-

rablemente cuando el número de filtros eliminados (N) se incrementa, para todo SNR considerado (ver las correspondientes filas etiquetadas como $0dB$, $10dB$ y $20dB$ en la tabla 7.5). A nivel de segmento, los valores óptimos se obtienen cuando las frecuencias debajo de $400 - 500Hz$ no son consideradas en el cálculo de las características cepstrales, que corresponde a la eliminación de los $10 - 11$ primeros filtros de baja frecuencia. Observaciones similares se pueden extraer analizando los resultados a nivel de evento acústico.

Sin embargo, en promedio, el esquema combinado propuesto (MFCC_HPN + CC_H), cuando $N = 11$ filtros son eliminados, se obtienen reducciones del error relativo con respecto al sistema base (ver figura 7.4) alrededor del 11 % a nivel de segmento y de evento acústico. Nuevamente podemos observar una mejora significativa que con respecto a los resultados en condiciones ruidosas mostrados en el capítulo 4.

7.2.3.2. Experimentos con la parametrización basada en los coeficientes de banco de filtros

En el capítulo 5, sección 5.2.7, se describe como los parámetros FC basado en NMF supera significativamente al esquema FC en condiciones ruidosas. Con la finalidad de evaluar el rendimiento del esquema combinado con la parametrización basada en los coeficientes de banco de filtros (FC y FC_NMF) en ambientes ruidosos, se realiza una serie de experimentos, que fueron llevados a cabo con los mismos tipos de ruido (*aeropuerto*, *voces*, *restaurante*, *tren*, *sala de exposiciones* y *metro* en SNRs desde $0dB$ hasta $20dB$ con pasos de $5dB$). Por brevedad, solo reportamos en esta sección al igual que en la sección anterior, resultados para el sistema base y para el esquema propuesto (MFCC_HPN + CC_H) en el caso de parámetros CC + ΔCC.

En la figuras 7.5 y 7.6 se representan el promedio de la reducción del error relativo para cada tipo de ruido con respecto al sistema base (condiciones ruidosas sin filtrado paso alto de la señal de audio) calculada a través de las SNRs consideradas ($0dB$ hasta $20dB$ con pasos de $5dB$) como una función del número de filtros de baja

Tabla 7.5: Tasa de clasificación promedio [%] (segmento) para la parametrización CC + ΔCC y diferentes tipos de ruido y SNRs.

Ruido	**SNR (dB)**	**Número de filtros eliminados (N)**						
		MFCC	**MFCC_HPN + CC_H**					
		Base.	**7**	**8**	**9**	**10**	**11**	**12**
Aeropuerto	**20**	66.51	71.73	72.28	72.46	71.97	72.01	71.63
	10	49.92	56	55.75	56.25	57.42	57	56.61
	0	29.01	33.47	33.77	34.41	35.52	35.37	34.95
Voces	**20**	67.09	72.14	72.02	72.24	72.32	72.01	71.55
	10	52.27	59.85	60.14	60.11	61.07	60.52	59.01
	0	27.59	36.53	36.99	36.77	39.72	38.36	35.44
Restaurante	**20**	67.43	71.74	72.06	72.23	71.88	71.79	71.51
	10	53.09	59.68	60.52	60.45	59.87	59.97	58.97
	0	25.65	34.43	35.54	35.12	36.06	35.81	32.89
Tren	**20**	71.18	73.97	73.91	73.85	73.55	73.9	73.98
	10	58.69	62.59	62.8	62.86	62.51	62.76	62.45
	0	40.46	45.53	46.4	47.15	46.93	47.06	46.36
Sala de exposiciones	**20**	58.00	62.67	63.14	63.24	63.57	62.97	62.95
	10	42.66	46.63	46.71	47.04	48.3	47.77	47.21
	0	22.00	23.74	24.03	23.48	25.06	25.77	24.75
Metro	**20**	56.90	60.57	60.66	60.99	61.22	60.91	60.45
	10	39.88	44.3	44.71	44.32	43.78	44.32	44.43
	0	19.34	22.49	22.99	21.92	21.73	23.07	**23.14**

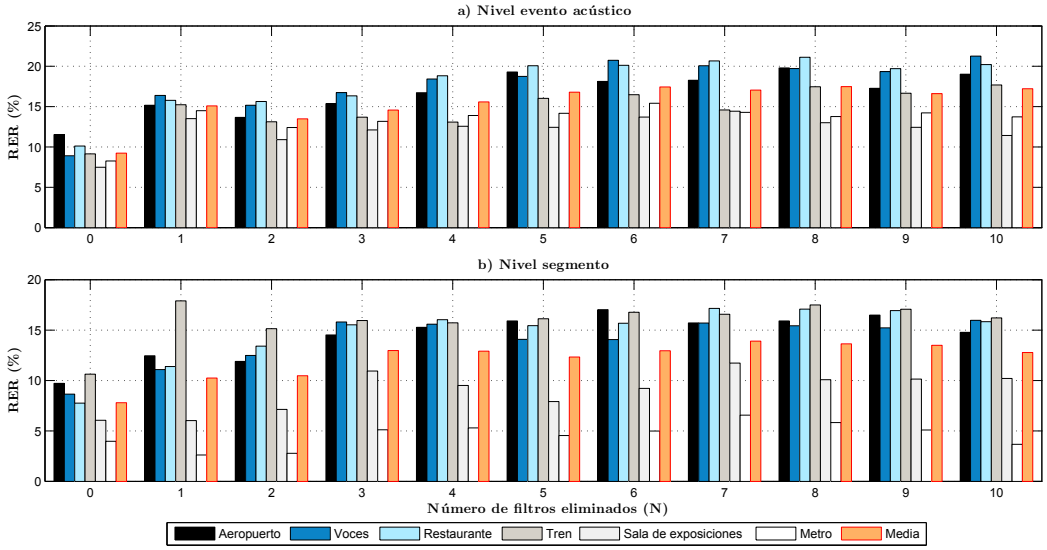


Figura 7.5: Reducción del error relativa [%] con respecto al sistema base para la parametrización FC + CC + ΔCC y la escala Mel en condiciones ruidosas: (a) a nivel de evento acústico; (b) a nivel de segmento.

frecuencia eliminados a nivel de segmento y evento acústico para la configuración FC y FC_NMF, respectivamente. La media de la reducción del error relativo sobre todos los tipos de ruido y SNRs son también indicados.

Para la parametrizaciones FC y FC_NMF, a partir de los resultados mostrados en las figuras 7.5 y 7.6 se puede observar que para los todos los tipos de ruido considerados (*aeropuerto*, *voces*, *restaurante*, *tren*, *sala de exposiciones* y *metro*), la tasa de clasificación a nivel de segmento y evento acústico mejoran considerablemente cuando el número de filtros eliminados (N) se incrementa, para todo SNR. En este caso los valores óptimos se obtienen cuando las frecuencias debajo de $200 - 400Hz$ no son considerados en el cálculo de las características cepstrales, que corresponde a la eliminación de los $5 - 10$ primeros filtros de baja frecuencia.

Para FC, en promedio cuando se eliminan los primeros 9 filtros, se obtienen reducciones del error relativo con respecto al sistema base (ver figura 7.5) de aproximadamente 13,5 % a nivel de segmento y 16,6 % a nivel de evento acústico, respecti-

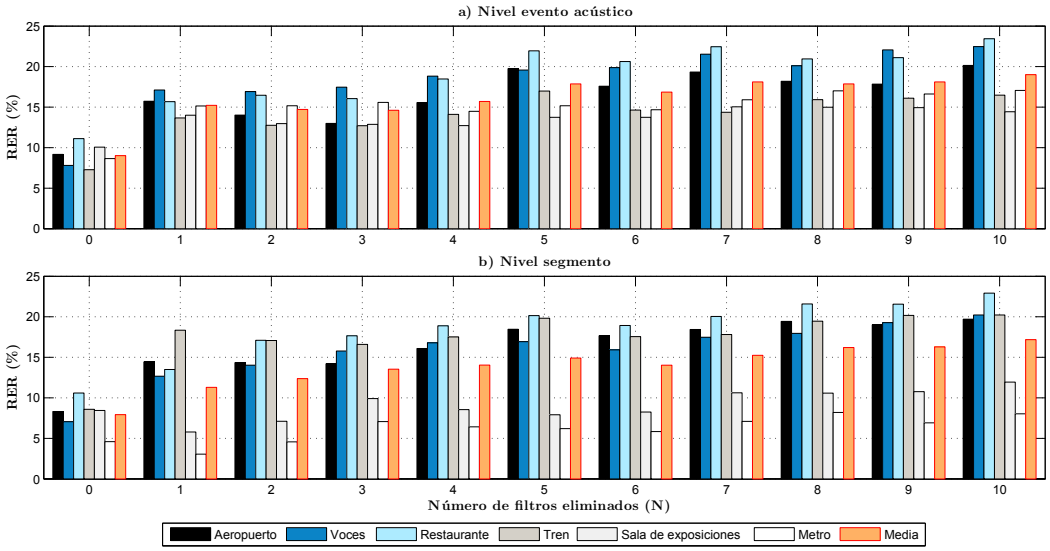


Figura 7.6: Reducción del error relativa [%] con respecto al sistema base para la parametrización FC_NMF + CC + ΔCC y la escala Mel en condiciones ruidosas: (a) a nivel de evento acústico; (b) a nivel de segmento.

vamente.

Para el caso del esquema FC_NMF, en promedio cuando 9 filtros son eliminados, se obtienen reducciones del error relativo con respecto al sistema base (ver figura 7.6) de alrededor del 16,3 % a nivel de segmento y 18,1 % a nivel de evento acústico, respectivamente.

De acuerdo a los resultados de tasa de clasificación promediado sobre todos los valores SNR considerados cuando se eliminan los primeros 9 filtros paso banda de baja frecuencia (MFCC_HP9), mostrados en las tablas 7.6 y 7.7, se puede observar que en general el esquema FC_NMF supera al esquema FC a nivel de segmento y evento acústico, siendo esta mejora estaísticamente significativa para los ruidos *aeropuerto*, *voces*, *restaurante* y *tren* a nivel de segmento y para nivel de evento acústico esta mejora no es estadísticamente significativa. Este hecho, nuevamente muestra el beneficio en el rendimiento de la tasa de clasificación cuando se utiliza un banco de filtros basado en NMF, capturando mucha de la información temporal de

Tabla 7.6: Tasa de clasificación promedio [%] a nivel de segmento, promediado sobre todos los valores SNR considerados con MFCC_HP9.

MFCC_HP9	**Esquemas de Parametrización**	
Ruido	**FC**	**FC_NMF**
Aeropuerto ($N1$)	$53{,}6 \pm 0{,}5$	$55 \pm 0{,}5$
Voces ($N2$)	$53{,}35 \pm 0{,}5$	$55{,}58 \pm 0{,}49$
Restaurante ($N3$)	$52{,}39 \pm 0{,}5$	$55{,}04 \pm 0{,}5$
Tren ($N4$)	$58{,}76 \pm 0{,}49$	$60{,}3 \pm 0{,}49$
Sala de exposiciones ($N5$)	$42{,}88 \pm 0{,}49$	$43{,}27 \pm 0{,}49$
Metro ($N6$)	$39{,}91 \pm 0{,}49$	$41{,}06 \pm 0{,}49$
Promedio ($N1 - N6$)	$50{,}15 \pm 0{,}2$	$51{,}71 \pm 0{,}2$
Promedio ($N1 - N4$)	$54{,}53 \pm 0{,}25$	$56{,}48 \pm 0{,}25$

los parámetros acústicos a corto plazo conjuntamente con las características CC_H.

7.2.4. Conclusiones

En este capítulo, hemos presentado un nuevo esquema de parametrización para la tarea de clasificación de eventos acústicos basado en la combinación de las características a corto plazo MFCC con filtrado paso alto (MFCC_HPN), motivado por los buenos resultados obtenidos en los capítulos 4, 5 y 6 y las características basado en los coeficientes de activación NMF (CC_H). Los experimentos han mostrado que los parámetros CC_H brindan importante información complementaria, mejorando de esta manera el rendimiento del sistema de clasificación especialmente en condiciones ruidosas, mostrando su robustez frente al ruido.

Los experimentos fueron llevados a cabo en condiciones limpias y ruidosas usando dos métodos de integración temporal de características basados: en los estadísticos de las características a corto plazo y coeficientes de banco de filtros fijo ($\mathbf{U}$) y aprendidos

Tabla 7.7: Tasa de clasificación promedio [%] a nivel de evento acústico, promediado sobre todos los valores SNR considerados con MFCC_HP9.

MFCC_HP9	**Esquemas de Parametrización**	
Ruido	**FC**	**FC_NMF**
Aeropuerto ($N1$)	$57{,}51 \pm 0{,}95$	$57{,}8 \pm 0{,}95$
Voces ($N2$)	$55{,}99 \pm 0{,}96$	$57{,}48 \pm 0{,}95$
Restaurante ($N3$)	$55{,}63 \pm 0{,}96$	$56{,}4 \pm 0{,}96$
Tren ($N4$)	$58{,}54 \pm 0{,}95$	$58{,}26 \pm 0{,}95$
Sala de exposiciones ($N5$)	$50{,}24 \pm 0{,}96$	$51{,}66 \pm 0{,}96$
Metro ($N6$)	$49{,}41 \pm 0{,}96$	$50{,}82 \pm 0{,}96$
Promedio ($N1 - N6$)	$54{,}59 \pm 0{,}39$	$55{,}40 \pm 0{,}39$
Promedio ($N1 - N4$)	$56{,}92 \pm 0{,}48$	$57{,}49 \pm 0{,}48$

por NMF ($\mathbf{W}$).

Para condiciones limpias, las diferencias de rendimiento con respecto al sistema base (cuando no se elimina ninguna banda de frecuencia) son estadísticamnete significativas. Para la método basada en estadísticos, cuando son eliminados los primeros 7 filtros de baja frecuencia, se obtienen reducciones de error relativo de alrededor del 20,16 % y 19,92 % a nivel de segmento y evento acústico, respectivamente para la parametrización CC. Para el método basado en coeficientes de banco de filtros, el uso del banco de filtros basado en NMF (FC_NMF), supera al banco de filtros fijo (FC) logrando reducciones de error relativo de alrededor del 8 % y 9 % a nivel segmento y evento acústico, respectivamente para la parametrización CC + ΔCC, mostrando que el filtro basado en NMF es más adecuado para la tarea AEC, permitiendo capturar el comportamiento temporal más relevante en las características a corto plazo.

Para condiciones ruidosas, con el método basado en estadísticos, las mejoras en el rendimiento de clasificación son estadísticamente significativas, logrando reducciones

de error relativo de alrededor de 11 % a nivel de segmento y evento objetivo para la parametrización CC + ΔCC, en este caso al no considerar 11 bandas espectrales de baja frecuencia. Mientras que con el método basado en coeficientes de banco de filtros, para la parametrización FC, se logra una reducción del error relativo de alrededor del 13,5 % a nivel de segmento y 16,6 % a nivel de evento acústico cuando 9 filtros de baja frecuencia son eliminados y para la parametrización FC_NMF, se logra una reducción del error relativo del 16,3 % a nivel de segmento y 18,1 % a nivel de evento acústico cuando se eliminan los primeros 9 filtros de baja frecuencia. Al comparar entre los dos esquemas FC_NMF supera en general a la parametrización FC.

Capítulo 8

Conclusiones y líneas futuras

En este último capítulo, se resumen las principales conclusiones y contribuciones de la presente tesis doctoral. Así mismo se mencionan y describen varias lñeas futuras de investigación que pueden desarrollarse a partir del trabajo realizado en esta tesis.

8.1. Conclusiones y contribuciones

Las conclusiones y contribuciones se van a referir a las dos grandes líneas tratadas en la presente tesis: eliminación de ruido de señales de voz con aplicación a la mejora o realce de la voz y el reconocimiento automático del habla (RAH) y la clasificación de eventos acústicos.

8.1.1. Eliminación de ruido para la mejora de la señal de voz y el reconocimiento automático del habla

En esta parte del trabajo se ha desarrollado un método para la supresión del ruido de señales de voz afectadas por ambientes acústicos adversos (condiciones ruidosas), que puede aplicarse tanto para mejorar la calidad de la voz como etapa de preprocesamiento de un reconocedor de habla. Dicho método está basado en la factorización

en matrices no negativas y presenta dos contribuciones novedosas con respecto a trabajos anteriores: en primer lugar, no necesita información explícita acerca del ruido, ya que puede estimarlo a partir de los segmentos de ruido/silencio de la propia elocución a procesar, que han sido previamente determinados por un detector de actividad vocal; en segundo lugar, el método combina el uso de la divergencia de Kullback-Leibler con restricciones de dispersión sobre la matriz de coeficientes de activación o ganancia de algoritmo NMF.

Los experimentos realizados muestran que NMF puede utilizarse de forma satisfactoria para este tipo de tareas. En concreto, se ha comprobado que resulta beneficioso realizar un control explícito del grado de dispersión en las descomposición realizada por NMF ya que produce el refuerzo de las componentes que son relevantes (voz) y la atenuación de las que no lo son (ruido). Se ha evaluado la técnica propuesta en diversas condiciones de ruido, obteniendo mejoras significativas con respecto a la substracción espectral convencional, especialmente en valores SNRs bajos y medio, tanto en calidad de voz como en tasa de reconocimiento.

En el capítulo 3 de la presente tesis se pueden encontrar más detalles sobre esta contribución, un resumen de la cual corresponde con la siguiente publicación:

1. Ludeña - Choez, J., Gallardo - Antolín, A. (2012). Speech denoising using non - negative matrix factorization with kullback - Leibler divergence and sparseness constraints. In *Advances in Speech and Language Technologies for Iberian Languages (IberSpeech 2012), CCIS - 328*, pp. 207 – 216, Madrid, Spain.

8.1.2. Clasificación de eventos acústicos

La hipótesis de partida de esta parte de trabajo ha sido que que los parámetros acústicos usados habitualmente para clasificación de eventos acústicos no son necesariamente los más apropiados puesto que usualmente vienen heredados de diferentes tareas de procesado de voz y las características espectrales de la

voz y los eventos acústicos son, en general, muy diferentes. Por este motivo, se han desarrollado distintas parametrizaciones más adecuadas a la clasificación de sonidos diferentes de la voz.

Esta hipótesis ha sido corroborada a través del análisis espectral de los eventos acústicos basado en NMF realizado en el capítulo 4, del que se ha podido concluir que, aparte de que el contenido y estructura espectral de los diferentes eventos acústicos es distinto del de la voz, dicho contenido presenta principalmente componentes relevantes en las zonas medias y altas del espectro, lo que indica que dichas frecuencias son las más adecuadas para la discriminación entre los diferentes sonidos.

A partir de este estudio se ha desarrollado un nuevo módulo de extracción de características para la tarea CEA, consistente en una extensión de la parametrización MFCC convencional y se basa en el filtrado paso alto de la señal de audio. En la práctica, el esquema propuesto se ha implementado modificando el banco de filtros auditivo en escala de frecuencia Mel, mediante la eliminación explícita de un cierto número de filtros de baja frecuencia. Los resultados obtenidos en condiciones limpias y ruidosas muestran que el filtrado paso alto es, en términos generales, beneficioso para el sistema. En particular, la eliminación de frecuencias por debajo de $100 - 275Hz$ en condiciones limpias y por debajo de $400 - 500Hz$ en condiciones ruidosas, mejora significativamente el funcionamiento del sistema con respecto a parámetros los MFCCs convencionales.

Esta aportación corresponde con el capítulo 4 de la presente tesis y las siguientes publicaciones:

2. Ludeña - Choez, J., Gallardo - Antolín, A. (2013). NMF - based spectral analysis for acosutic event classification tasks. In *Advances in Nonlinear Speech Processing (NOLISP 2013), LNCS*, pp. 9 – 16, Mons, Belgium.

3. Ludeña - Choez, J., Gallardo - Antolín, A. (2015). Feature extraction based on the high - pass filtering of audio signals for acoustic event classification. In *Computer Speech and Language, Elsevier*, vol. 30, no. 1, pp. 32 – 42.

   La segunda de las parametrizaciones desarrolladas se enmarca en la denominada técnica de integración temporal de caracter'sticas basada en coeficientes de banco de filtros. En este caso, la factorización en matrices no negativas se utiliza para el aprendizaje no supervisado de un banco de filtros FC más adaptado a los características dinámicas de los eventos acústicos, en contraposición con trabajos previos en los se usaba un banco de filtros FC predeterminado e independiente de la tarea. Los experimentos realizados muestran que las características obtenidas con este nuevo esquema en combinación con el filtrado paso alto logran mejoras significativas en la tasa de clasificación (especialmente cuando se incluyen los parámetros de primera derivada de las características a corto plazo (Δ)) tanto en condiciones limpias como ruidosas, en comparación con los parámetros FC del sistema de referencia (con el banco de filtros predeterminado). La causa de este buen funcionamiento parece ser la mejor representación de la estructura temporal de las características a corto plazo que NMF obtiene, en la que se enfatizan las bajas frecuencias de modulación sobre las altas.

   Se pueden encontrar más detalles sobre este método, así como los resultados más relevantes en condiciones limpias y ruidosas en el capítulo 5. Por otra parte, los resultados en condiciones limpias están publicados en:

4. Ludeña - Choez, J., Gallardo - Antolín, A. (2014). NMF - based temporal feature integration for acoustic event classification. In *Proc. of the 14th annual Conference of the International Speech Communication Association (INTERSPEECH - 2013), ISCA*, pp. 2924 – 2928, Lyon, France.

En las aproximaciones anteriores, el filtrado paso alto de la señal de audio se realizaba mediante la supresión de cierto número de bandas de baja frecuencia elegido de forma empírica. Con objeto de encontrar un modo automático de realizar la selección de las bandas espectrales más apropiadas para la discriminación entre diferentes eventos acústicos, se ha propuesto la utilización de técnicas de selección de características basadas en información mutua (en concreto, se ha experimento con mRMR, JMI, CIFE y CondRed). Los experimentos realizados muestran que la selección automática de bandas usando cualquiera de estos métodos incrementa la tasa de clasificación del sistema, siendo las bandas localizadas en bajas y altas frecuencias las más relevantes y menos redundantes en condiciones limpias. Sin embargo, en condiciones ruidosas, los mejores resultados corresponden a la eliminación de las bandas situadas en bajas frecuencia. El capítulo 6 de la presente tesis está dedicado a este temática.

Finalmente, se ha desarrollado un nuevo esquema de extracción de características a corto plazo basado en los coeficientes de activación o ganancia obtenidos mediante la aplicación de NMF. A diferencia de otros trabajos encontrados en la literatura en esta línea, en el método propuesto no se usan dichos coeficientes directamente si no que previamente son transformados mediante la aplicación del logaritmo y la transformada coseno discreta. De la observación de los resultados obtenidos, es posible concluir que las características basadas en NMF brindan importante información complementaria a los coeficientes melcepstrales convencionales, mejorando esta combinación el funcionamiento del sistema de clasificación, especialmente en condiciones ruidosas en comparación con el sistema de referencia, basado en MFCC. En el capítulo 7 puede encontrarse una explicación más detallada de esta parametrización, los experimentos realizados y los resultados obtenidos.

8.2. Líneas futuras de investigación

En esta sección se enumeran algunas líneas futuras de investigación para las dos tareas vistas a lo largo del desarrollo de esta tesis.

Para la tarea de eliminación de ruido con aplicación a la mejora de voz y el reconocimiento automático del habla, se describen las siguientes líneas futuras de investigación:

- Utilizar el método NMF para el aprendizaje no supervisado del banco de filtros auditivo utilizado en la extracción de características a corto plazo. En el capítulo 3 se observó que los vectores espectrales base de la señal de la voz obtenidos mediante la aplicación de NMF presentan bandas espectrales similares al banco de filtros auditivo en escala de frecuencia Mel usado habitualmente para el cálculo de los MFCCs, en el sentido de que en la región baja del espectro se concentraban muchos de estos vectores espectrales con un ancho de banda reducido, mientras que altas frecuencias, el número de vectores era menor y con un ancho de banda mayor. Esta observación nos motiva a estudiar la utilización de los vectores espectrales base de NMF como un banco de filtros auditivo que podría ayudar a mejorar el proceso de extracción de características para el reconocimiento automático del habla.

- A la vista de los buenos resultados obtenidos en el capítulo 7 en la tarea de clasificación de eventos acústicos (en especial, en condiciones ruidosas), encontrar una nueva parametrización para reconocimiento de habla basada en los coeficientes de activación de NMF.

Para la tarea de clasificación de eventos acústicos, se describen las siguientes líneas futuras de investigación:

- En el capítulo 5 se mostró que el uso de un banco de filtros FC basado en NMF en el proceso de integración temporal de caracter'sticas resulta beneficioso para el sistema de AEC. Sin embargo, una de las limitaciones del método es que el

banco de filtros basado en NMF se aplica por igual a todas las componentes de las características a corto plazo. Como trabajo futuro, se planea diseñar un banco de filtros diferente para cada una de estas componentes, esperando obtener mejoras en el funcionamiento del sistema global.

- En trabajos previos en la literatura, otras aproximaciones basadas en NMF, como el NMF convolutivo, se han utilizado para tareas de clasificación y detección de eventos acústicos [Cotton and Ellis, 2011], mostrando su robustez en presencia de ruido en comparación con los MFCCs convencionales. Esto es debido a que el NMF convolutivo permite expresar de forma más eficiente la evolución temporal de las señales de audio y entre ellas los eventos acústicos. Motivado por estos resultados, se planea desarrollar nuevas parametrizaciones consistentes en la sustitución (allí donde sea posible) de NMF por el NMF convolutivo y estudiar sus prestaciones tanto en condiciones limpias como ruidosas.

Apéndice A

Error de aproximación promedio en NMF

Como se ha indicado en la subsección 2.1.2, el proceso de inicialización es fundamental para el algoritmo NMF, debido a que no es estrictamente convexo dada una determinada función de coste. Con la finalidad de analizar la influencia de la inicialización en el resultado de la factorización, se han realizado $N = 10$ experimentos diferentes consistentes en la extracción de los vectores espectrales base de 12 eventos acústicos, los mismos que se utilizan en la tarea de clasificación de eventos acústicos de los capítulos 4, 5, 6 y 7 de la presente tesis. El algoritmo fue inicializado usando el esquema de inicialización multi - inicio [Cichocki et al., 2009], de tal manera que se generaron 10 pares de matrices aleatorias uniformes (W_c y H_c) y se escogió para inicialización la factorización que produjo la distancia euclídea menor entre V_c y ($W_c H_c$). Finalmente, estas matrices se entrenaron mediante la minimización de la divergencia KL (ecuación 2.3) entre el espectro de magnitud V_c y sus correspondientes matrices factorizadas ($W_c H_c$) usando el esquema iterativo y las reglas de aprendizaje (ecuación 2.4) propuestas en [Lee and Seung, 1999], siendo el criterio de parada del algoritmo un número máximo de iteraciones (en nuestro caso, 200). Una vez terminados los procesos de factorización se calcularon el promedio de los errores

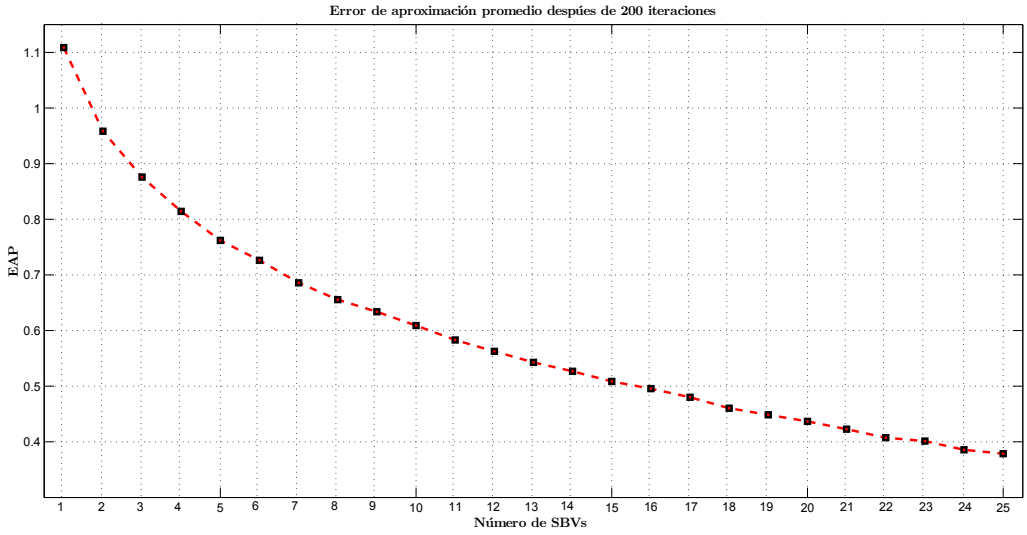


Figura A.1: Error de aproximación promedio para 12 eventos acústicos después de 200 iteraciones.

de aproximación obtenidos en cada uno de los experimentos utilizando la ecuación (A.1),

$$EAP = \frac{\sum_{n=1}^{N} \sum_{c=1}^{C} \frac{\|V_c(n) - W_c(n)H_c(n)\|_1}{\|V_c(n)\|_1}}{NC} \tag{A.1}$$

en la que $\| \ . \ \|_1$ representa la norma 1, $V_c\,(n)$ es el espectro de magnitud correspondiente al c-simo evento acústico para el n-simo experimento, W_c y H_c son las matrices que contienen los SBVs y coeficientes de activación obtenidos mediante del algoritmo NMF correspondiente al c-simo evento acústico, N, es el número de experimentos (10 en este caso) y C es el número de clases (12 en este caso).

En la figura A.1 se muestra el error de aproximación promedio para un conjunto de entrenamiento dado (en concreto, el primer grupo disjunto balanceado utilizado en los experimentos de validación cruzada de CEA). La tabla A.1 muestra la media y desviación estándar del error de aproximación promedio calculado sobre todos los eventos acústicos y los 10 experimentos para diferentes números de vectores espectrales base desde $K = 1$ hasta $K = 25$. Como se puede observar, la media del error

Tabla A.1: Media y desviación estándar del error de aproximación promedio después de 10 experimentos.

Número de SBVs (K)	**2**	**5**	**10**	**15**	**20**	**23**	**25**
Media	0.96	0.76	0.61	0.51	0.44	0.40	0.38
Desviación estándar ($\times 10^{-3}$)	2.89	3.44	1.91	2.74	2.40	1.75	1.11

de aproximación promedio disminuye cuando el número de SBVs se incrementa, tal y como cabría esperar. Por otra parte, la desviación estándar es pequeña en todos los casos, sugiriendo que el proceso de inicialización es robusto y adecuado para obtener los SBVs. Además se observa que la forma y posición de los SBVs obtenidos no difieren significativamente entre experimentos.

Apéndice B

Clasificación de eventos acústicos usando otras escalas de frecuencia

En el capítulo 4 se presentó un nueva parametrización para la clasificación de eventos acústicos. Este nuevo esquema básicamente consiste en una extensión de los coeficientes mel-cepstrales convencionales en el que se realiza un filtrado paso alto de la señal de audio mediante la eliminación explícita de un cierto número de filtros paso banda ubicados en las bajas frecuencias del banco de filtros auditivo en escala de frecuencia Mel. En este apéndice se muestra los resultados obtenidos con este mismo procedimiento aplicado en bancos de filtros en otras escalas de frecuencia, también inspiradas en el sistema auditivo humano. En concreto, nos referimos a las escalas de frecuencia Bark, ERB a las que también se añade la escala de frecuencia lineal a efectos de comparación.

Los experimentos se llevaron a cabo usando la misma base de datos usada en el capítulo 4 en condiciones limpias y con el mismo protocolo experimental. Los resultados de la tasa de clasificación a nivel de segmento se muestran en la tabla B.1 y a nivel de evento acústico en la tabla B.2.

Como se puede observar para la parametrización CC, el funcionamiento de las escalas de frecuencia Mel, ERB y Bark son bastante similares, siendo la escala de

Tabla B.1: Tasa de clasificación promedio [%] (segmento) para diferentes escalas de frecuencia.

Param.	Escala	Número de filtros eliminados												
		Base.	1	2	3	4	5	6	7	8	9	10	11	12
CC	Mel	75.10	77.47	77.66	77.58	77.63	78.16	76.95	78.11	76.87	76.12	77.23	77.23	76.10
	ERB	74.02	74.74	75.95	77.38	77.43	77.53	76.81	76.77	77.09	76.66	77.76	76.90	76.71
	Bark	74.30	77.39	77.27	77.68	76.96	77.31	76.27	77.43	76.91	76.72	77.11	76.77	76.59
	Lineal	77.29	77.30	77.62	76.84	77.26	75.52	75.33	74.96	73.88	74.43	73.36	73.22	71.83
CC+ΔCC	Mel	77.57	79.43	79.45	79.22	79.36	79.07	79.20	79.55	79.41	78.47	77.81	78.77	78.55
	ERB	76.51	77.57	78.80	79.14	79.42	78.69	79.22	79.13	79.04	78.74	79.20	78.79	78.97
	Bark	77.58	78.98	79.32	78.64	78.65	78.33	78.62	79.25	78.86	78.77	78.03	78.08	78.56
	Lineal	79.09	80.39	79.94	78.16	78.88	78.82	78.15	76.64	76.54	76.27	76.54	76.42	75.54

Tabla B.2: Tasa de clasificación promedio [%] (evento acústico) para diferentes escalas de frecuencia.

Param.	Escala	Número de filtros eliminados												
		Base.	1	2	3	4	5	6	7	8	9	10	11	12
CC	Mel	81.07	82.28	82.04	82.42	82.42	81.89	81.31	83.20	81.27	80.78	80.69	81.75	79.72
	ERB	79.43	80.73	81.46	82.09	82.57	82.52	82.71	82.42	82.28	81.46	83.29	81.51	80.73
	Bark	80.83	81.94	82.47	82.33	80.83	81.07	80.98	81.84	80.73	81.07	80.98	81.55	80.98
	Lineal	82.04	80.98	81.12	80.49	80.44	79.19	78.51	77.89	76.29	77.16	77.02	76.24	74.70
CC+ ΔCC	Mel	81.41	82.62	83.39	83.58	83.49	83.15	82.38	82.71	82.81	80.06	81.12	81.55	81.22
	ERB	80.73	80.98	82.18	82.67	83.24	82.62	82.76	81.89	82.04	81.80	82.71	81.75	82.57
	Bark	81.84	82.76	82.71	81.41	82.62	81.84	82.04	82.09	81.55	81.80	81.22	81.41	81.22
	Lineal	82.81	82.38	82.42	81.60	81.36	80.78	80.35	79.33	79.24	79.04	79.38	78.71	77.16

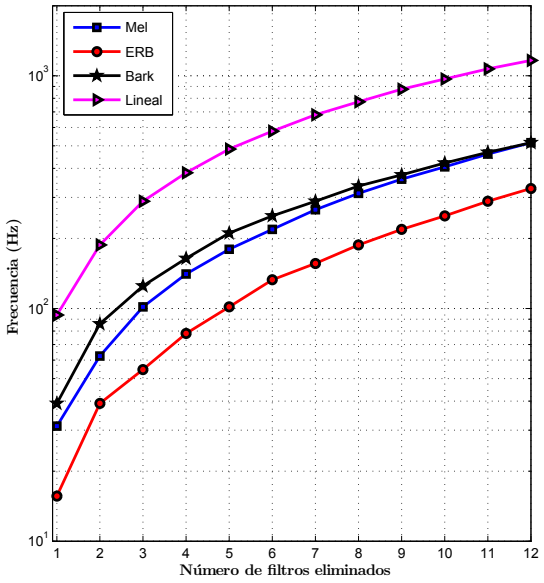


Figura B.1: Frecuencia superior de la banda de paso vs. el número de filtros eliminados para la escalas Mel, ERB, Bark y Lineal.

frecuencia Mel ligeramente mejor. El comportamiento con respecto a la eliminación de bandas de baja frecuencia sigue la misma tendencia para las tres escalas de frecuencia. En todos los casos, el filtrado paso alto supera al experimento base: para la escala Mel, la mejor tasa se logra cuando el número de filtros eliminados varía de 3 a 7, para la escala ERB, de 3 a 10 y para la escala Bark, de 2 a 7. A partir de la figura B.1, se puede obervar que estos rangos de filtros eliminados aproximadamente corresponden a un banda de paso de $0Hz$ a $100 - 275Hz$. La escala lineal supera la tasa de clasificación lograda con las otras escalas de frecuencia en el experimento base (cuando no se elimina ninguna banda de frecuencia), posiblemente debido a que, con la escala lineal, al tener todos los filtros el mismo ancho de banda, se da la misma importancia a las altas frecuencias que a las bajas en el proceso de parametrización. Sin embargo, no se obtienen mejoras cuando se eliminan varios filtros de baja frecuencia. Este resultado puede explicarse debido al mayor ancho de banda que presentan los filtros de baja frecuencia en la escala lineal con respecto a los del resto de escalas de frecuencia.

Para la parametrización $CC+\Delta CC$ se pueden extraer observaciones similares: los mejores resultados se obtiene cuando las bajas frecuencias (debajo de $100-275Hz$) no son consideradas en el proceso de extracción de características. Cuando se compara con la parametrización CC, se puede observar que $CC + \Delta CC$ logra mejoras de aproximadamente $1\,\%$ absoluto sobre CC.

Bibliografía

[Arenas Garcia et al., 2006] Arenas Garcia, J., Larsen, J., Kai Hansen, L., and Meng, A. (2006). Optimal filtering of dynamics in short-time features for music organization. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 290–295.

[Battiti, 1994] Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. on Neural Networks*, 5:537–550.

[Berouti et al., 1979] Berouti, M., Schwartz, R., and Makhou, J. (1979). Enhancement of speech corrupted by acoustic noise. In *Proc. of the Acoustics, Speech, and Signal Processing, Conference 1979*, ICASSP-79, pages 208 – 211. IEEE.

[Bertrand et al., 2008a] Bertrand, A., Demuynck, K., Stouten, V., and Van hamme, H. (2008a). Unsupervised learning of auditory filter banks using non-negative matrix factorization. In *Proc. of the Acoustics, Speech and Signal Processing IEEE International Conference*, ICASSP, pages 4713–4716, Las Vegas, NV. IEEE.

[Bertrand et al., 2008b] Bertrand, A., Demuynck, K., Stouten, V., and Van hamme, H. (2008b). Unsupervised learning of auditory filter banks using non-negative matrix factorization. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, ICASSP 2008, pages 4713 – 4716, Las vegas, NV. IEEE.

[Bins and Draper, 2001] Bins, J. and Draper, B. (2001). Feature selection from huge feature sets. In *Computer Vision Eighth IEEE International Conference on*, pages 159–165, Vancouver, BC. IEEE.

[Brookes, 2009] Brookes, M. (2009). Voicebox matlab software.

[Brown et al., 2011] Brown, G., Pocock, A., Zhao, M., and Lujan, M. (2011). Feast, a feature selection toolbox for c and matlab.

[Brown et al., 2012] Brown, G., Pocock, A., Zhao, M., and Lujan, M. (2012). Conditional likehood maximisation: A unifying framework for information theoretic feature selection. *Machine Learning Research*, 13:27–66.

[Brunet et al., 2004] Brunet, J. P., Tamayo, p., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. In *Proc. of the National Academy of Sciences*, pages 4164–4169.

[Butko and Nadeu, 2010] Butko, T. and Nadeu, C. (2010). On enhancing acoustic event detection by using feature selection and audiovisual feature-level fusion. In *Workshop on Database and Expert Systems Applications (DEXA)*, pages 271–275, Bilbao. IEEE.

[Chen et al., 2006] Chen, Z., Cichocki, A., and M Ruthowski, T. (2006). Constrained non-negative matrix factorization method for eeg analysis in early detection of alzheimer disease. In *Proc. of the IEEE International Conference on Acoustics Speed and Signal Processing*, ICASSP, pages V893–V896, Toulouse. IEEE.

[Chih-Chung and Chih-Jen, 2011] Chih-Chung, C. and Chih-Jen, L. (2011). Libsvm: A library for support vector machines. *Journal of the ACM Transactions on Intelligent Systems and Technology (TIST)*, 2:1–27.

[Chu et al., 2006] Chu, S., Narayanan, S., Jay Huo, C., and Mata, M. (2006). Where am i? scene recognition for mobile robots using audio features. In *Multimedia*

and Expo, 2006 IEEE International Conference on, ICME, 2006, pages 885–888, Toronto, Ont. IEEE.

[Cichocki et al., 2006] Cichocki, A., Zdunek, R., and Amari, S.-i. (2006). New algorithms for non-negative matrix factorization in applications to blind source separation. In *Proc. of the Acoustics, Speech, and Signal Processing, Conference 2006*, ICASSP-06, pages 621 – 625, Toulouse. IEEE.

[Cichocki et al., 2009] Cichocki, A., Zdunek, R., Huy-Phan, A., and Amari, S.-I. (2009). *Nonnegative matrix and tensor factorizations*. Ed. John Wiley and Sons, United Kingdom, UK.

[Clavel et al., 2005] Clavel, C., Ehrette, T., and Richard, G. (2005). Events detection for an audio-based surveillance system. In *Multimedia and Expo, 2005 IEEE International Conference on*, ICME, 2005, pages 1306–1309, Amsterdam. IEEE.

[Cotton and Ellis, 2011] Cotton, C. V. and Ellis, D. (2011). Spectral vs. spectro-temporal features for acoustic event detection. In *Applications of Signal to Audio and Acoustics, 2011 IEEE Workshop*, WASPAA, pages 69–72, New Paltz, NY. IEEE.

[Cover and Thomas, 2006] Cover, T. and Thomas, J. (2006). *Elements of Information Theory 2nd Edition*. Wiley Series in Telecommunications and Signal Processing, United Kingdom, UK.

[Cyril and Bjorn, 2012] Cyril, J. and Bjorn, S. (2012). Exploring nonnegative matrix factorization for audio classification: Application to speaker recognition. In *Proc. of the Speech Communication, 2012*, ITG Symposium, pages 1–4, Braunschweig, Germany. IEEE.

[Damon et al., 2013] Damon, C., Liutkus, A., Gramfort, A., and Essid, S. (2013). Nonnegative matrix factorization for single-channel eeg artifact rejection. In *Proc.*

of the Acoustics, Speech and Signal Processing, ICASSP, pages 1177–1181, Vancouver, BC. IEEE.

[Dhanalakshmi et al., 2008] Dhanalakshmi, P., Palanivel, S., and Ramalingam, V. (2008). Classification of audio signals using svm and rbfnn. *Expert Systems with Applications*, 36:6069–6075.

[Evangelista, 1993] Evangelista, G. (1993). Pitch-synchronous wavelet representations of speech and music signals. *Signal Processing, IEEE Transactions*, 41:3313–3330.

[Fernandez et al., 2009] Fernandez, R., Francois Bonastre, J., and R Calvo, J. (2009). Feature selection based on information theory for speaker verification. In *CIARP, Lecture Notes in Computer Science*, LNCS, pages 305–312. Springer.

[Fleuret and Guyon, 2004] Fleuret, F. and Guyon, I. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Reasearch*, 5:1531–1555.

[Geiger et al., 2013] Geiger, J. T., Schuller, B., and Rigoll, G. (2013). Large-scale audio feature extraction and svm for acoustic scene classification. In *Applications of Signal Processing to Audio and Acoustics, 2013 IEEE Workshop*, WASPAA, 2013, pages 1–4, New Paltz, NY. IEEE.

[Guyon and Elisseeff, 2003] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Machine Learning Research*, 3:1157–1182.

[Hirsch and Pearce, 2000] Hirsch, H. and Pearce, D. (2000). The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In *Proc. of the ASR2000-Automatic Speech Recognition, Conference 2000*, ASR2000, Paris, Francia. ISCA.

[Hlawatsch and Boudreaux-Bartels, 1992] Hlawatsch, F. and Boudreaux-Bartels, G. F. (1992). Linear and quadratic time-frequency signal representations. *Signal Processing Magazine, IEEE*, 9:21–67.

[Hu and Loizou, 2011] Hu, Y. and Loizou, P. (2011). Matlab software.

[Jingu and Haesun, 2008] Jingu, K. and Haesun, P. (2008). Sparse nonnegative matrix factorization for clustering. Technical Report 460, Georgia Institute of Technology.

[Kiktova et al., 2013] Kiktova, E., Juhar, J., and Cizmar, A. (2013). Feature selection for acoustic events detection. *Multimedia Tools and Applications*, pages 1–21.

[Kim et al., 2005] Kim, H.-G., Moreau, N., and Sikora, T. (2005). *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. John Wiley and Sons Inc, United States of America, USA.

[Kons and Toledo-Ronen, 2013] Kons, Z. and Toledo-Ronen, O. (2013). Audio event classification using deep neural network. In *Proc. of the 14th Annual Conference of the International Speech Communication Association*, INTERSPEECH-2013, pages 1482–1486, Lyon, France. ISCA.

[Kwangyoun and Hanseok, 2011] Kwangyoun, K. and Hanseok, K. (2011). Hierarchical approach for abnormal acoustic event classification in an elevator. In *Advanced Video and Signal-Based Surveillance (AVSS), 8th IEEE International Conference*, AVSS, 2011, pages 89–94, Klagenfurt. IEEE.

[Lee and Seung, 1999] Lee, D. and Seung, S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788 – 791.

[Lee and Seung, 2002] Lee, D. and Seung, S. (2002). Perceptual evaluation of speech quality (pesq), the new itu standard for end-to-end speech quality assessment. part ii. psychoacoustic model. *Audio Engineering Society*, 50:765 – 778.

[Lin and Tang, 2006] Lin, D. and Tang, X. (2006). Conditional infomax learning an integrated framework for feature extraction and fusion. In *European Conference on Computer Vision*, LNCS, pages 68–82. Springer.

[Ludena-Choez and Gallardo-Antolin, 2013] Ludena-Choez, J. and Gallardo-Antolin, A. (2013). Nmf-based spectral analysis for acoustic event classification tasks. In *Advances in Nonlinear Speech Processing (NOLISP 2013)*, Lecture Notes in Computer Science, pages 9–16, Mons, Belgium. Springer.

[McKinney and Breebaart, 2003] McKinney, M. and Breebaart, J. (2003). Features for audio and music classification. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 151–158.

[Mcloughlin, 2009] Mcloughlin, I. (2009). *Applied Speech and Audio Processing*. Cambridge University Press, United States of America, USA.

[Mejia Navarrete et al., 2011] Mejia Navarrete, D., Gallardo Antolin, A., Pelaez Moreno, C., and Valverde Albacete, Francisco, J. (2011). Feature extraction assessment for an acoustic-event classification task using the entropy triangle. In *Proc. of the 12th Annual Conference of the International Speech Communication Association*, INTERSPEECH-2011, pages 309–312, Florence, Italy. ISCA.

[Meng et al., 2007] Meng, A., Ahrendt, P., Larsen, J., and Kai Hansen, L. (2007). Temporal feature integration for music genre classification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15:1654–1664.

[Meyer and Bontempi, 2006] Meyer, P. and Bontempi, G. (2006). On the use of variable complementarity for feature selection in cancer classification. In *Applications of Evolutionary Computing*, LNCS, pages 91–102. Springer.

[Meyer et al., 2008] Meyer, P. E., Schretter, C., and Botempi, G. (2008). Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing*, 2:261–274.

[Moore and Glasberg, 1983] Moore, B. C. J. and Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, 74:750–753.

[Muller et al., 2008] Muller, C., Biel, J.-I., Kim, E., and Rosario, D. (2008). Speech-overlapped acoustic event detection for automotive applications. In *Proc. of the 9th Annual Conference of the International Speech Communication Association*, INTERSPEECH-2008, pages 2590–2593. ISCA.

[O'Shaughnessy, 2013] O'Shaughnessy, D. (2013). Acoustic analysis for automatic speech recognition. *Proceedings of the IEEE*, 101:1038–1053.

[Peng et al., 2005] Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27:1226–1238.

[Portelo et al., 2009] Portelo, J., Bugalho, M., Neto, J., Abad, A., and Serralheiro, A. (2009). Non-speech audio event detection. In *Acoustics, Speech and Signal Processing, 2009. IEEE International Conference on*, ICASSP 2009, pages 1973–1976, Taipei. IEEE.

[Rabiner, 1989] Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286.

[S0268, 2008] S0268, E. C. N. (2008). Upc-talp database of isolated meeting-room acoustic events.

[S0296, 2009] S0296, E. C. N. (2009). Fbk-irst database of isolated meeting-room acoustic events.

[Sandler and Lindenbaum, 2011] Sandler, R. and Lindenbaum, M. (2011). Nonnegative matrix factorization with earth mover distance metric for image analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 33:1590–1602.

[Scalart and Vieira, 1996] Scalart, P. and Vieira, J. (1996). Speech enhancement based on a priori signal to noise estimation. In *Proc. of the Acoustics, Speech, and Signal Processing, Conference 1996*, ICASSP-96, pages 629 – 632, Atlanta, GAY. IEEE.

[Schmidt and Olsson, 2006] Schmidt, M. and Olsson, R. (2006). Single-channel speech separation using sparse non-negative matrix factorization. In *Proc. of the Ninth Internacional Conference on Spoken Language Processing*, INTERSPEECH-06. ISCA.

[Schuller et al., 2009] Schuller, B., Lehmann, E., Weninger, F., and Eyben, F. (2009). Blind enhancement of the rhythmic and harmonic sections by nmf does it help? In *Proc. of the International Conference on Acoustics*, NAG/DAGA, pages 361–364.

[Schuller and Weninger, 2010] Schuller, B. and Weninger, F. (2010). Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization. In *Proc. of the Acoustics, Speech, and Signal Processing, Conference 2010*, ICASSP-10, pages 5054–5057, Dallas, TX. IEEE.

[Schuller et al., 2010] Schuller, B., Weninger, F., Wollmer, M., Sun, Y., and Rigoll, G. (2010). Non-negative matrix factorization as noise-robust feature extractor for speech recognition. In *Proc. of the Acoustics, Speech, and Signal Processing, Conference 2010*, ICASSP-10, pages 4562 – 4565, Dallas, TX. IEEE.

[Sebban and Nock, 2002] Sebban, M. and Nock, R. (2002). A hybrid filter/wrapper approach of feature selection using information theory. *Pattern Recognition*, 35:835–846.

[Smaragdis, 2004] Smaragdis, P. (2004). Discovering auditory objects through non-negativity constraints. In *Statistical and Perceptual Audio Processing*.

[Steeneken, 1991] Steeneken, H. J. M. (1991). Speech level and noise level measuring method. technical report. document sam-tn0-042. In *Esprit-SAM*.

[Stevens and Newman, 1937] Stevens, S. S. and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude of pitch. *The Journal of the Acoustical Society of America*, 8:185–190.

[Temko and Nadeu, 2006] Temko, A. and Nadeu, C. (2006). classification of acoustic events using svm-based clustering schemes. *Pattern Recognition*, 39:684–694.

[Vachhani and Patil, 2013] Vachhani, B. and Patil, H. (2013). Use of plp cepstral features for phonetic segmentation. In *Proc. of the Asian Language Processing (IALP)*, IALP, pages 143–146, Urumqi. IEEE.

[Vaseghi, 2007] Vaseghi, S. V. (2007). *Multimedia Signal Processing: theory and applications in speech, music and communications*. John Wiley and Sons Inc, United States of America, USA.

[Virtanen, 2007] Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. on Audio, Speech and Language Processing*, 15:1066–1074.

[Wei et al., 2003] Wei, X., Xin, L., and Yihong, G. (2003). Document clustering based non-negative matrix factorization. In *Proc. of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 03, pages 267 – 273, Toronto, Canada. ACM.

[Wilson et al., 2008] Wilson, K., Bhiksha, R., Smaragdis, P., and Divakaran, A. (2008). Speech denoising using nonnegative matrix factorization with priors. In *Proc. of the Acoustics, Speech, and Signal Processing, Conference 2008*, ICASSP-08, pages 4029 – 4032. IEEE.

[Yong-Choon et al., 2003] Yong-Choon, C., Seungjin, C., and Sung-Yang, B. (2003). Non-negative component parts of sound for classification. In *Proc. of the Signal Processing and Information Techonology, 2003*, ISSPIT-03, pages 633–636. IEEE.

[Yujin et al., 2010] Yujin, Y., Peihua, Z., and Qun, Z. (2010). Research of speaker recognition based on combination of lpcc and mfcc. In *Proc. of the Intelligent Computing and Intelligent Systems, 2010*, ICIS vol3, pages 765–767, Xiamen. IEEE.

[Zbancioc and Costin, 2003] Zbancioc, M. and Costin, M. (2003). Using neural networks and lpcc to improve speech recognition. In *Proc. of the Signals, Circuits and Systems, 2003*, SCS vol2, pages 445–448. IEEE.

[Zhang and Zhou, 2004] Zhang, Y. and Zhou, J. (2004). Audio segmentation based on multi-scale audio classification. In *Acoustics, Speech and Signal Processing, 2004. IEEE International Conference on*, ICASSP 2004, pages 349–352. IEEE.

[Zhang and Schuller, 2012] Zhang, Z. and Schuller, B. (2012). Semi-supervised learning helps in sound event classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, ICASSP, 2012, pages 333–336, Kyoto. IEEE.

[Zhuang et al., 2010] Zhuang, X., Zhou, X., Hasegawa-Johnson, M., and Huang, T. (2010). Real-world acoustic event detection. *Pattern Recognition Letters*, 31:1543–1551.

[Zhuang et al., 2008] Zhuang, X., Zhou, X., Huang, T., and Hasegawa Johnson, M. (2008). Feature analysis and selection for acoustic event detection. In *Acoustics, Speech and Signal Processing ICASSP 2008. IEEE International Conference on*, ICASSP 2008, pages 17–20, Las Vegas, NV. IEEE.

[Zieger, 2008] Zieger, C. (2008). An hmm based system for acoustic event detection. In *Multimodal Technologies for Perception of Humans*, Lecture Notes in Computer Science, pages 338–344, Baltimore, MD, USA. Springer.

[Zwicker and Terhardt, 1980] Zwicker, E. and Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *The Journal of the Acoustical Society of America*, 68:1523–1525.