Universidad Autónoma de Madrid
Escuela politécnica superior

Proyecto fin de carrera

SEGMENTACIÓN DE AUDIO MEDIANTE CARACTERÍSTICAS CROMÁTICAS EN FICHEROS DE NOTICIAS

Ingeniería de Telecomunicación

Elena Gómez Rincón

Junio 2015

SEGMENTACIÓN DE AUDIO MEDIANTE CARACTERÍSTICAS CROMÁTICAS EN FICHEROS DE NOTICIAS

AUTOR: Elena Gómez Rincón
TUTOR: Javier Franco Pedroso

Área de Tratamiento de Voz y Señales
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio 2015

Índice general

Índice de figuras

Índice de tablas

Resumen

Resumen

Este proyecto se centra en la implementación y análisis de diversas técnicas de extracción de características para segmentación de audio en ficheros de noticias. En él se distinguen dos tipos de características, tímbricas y cromáticas, si bien el objetivo principal es investigar en mayor profundidad el uso de estas segundas, cuya introducción viene motivada por una mayor capacidad potencial para distinguir entre voz y música.

En la primera fase del proyecto se optimiza un sistema basado en características tímbricas MFCC-SDC, que además de servir como sistema de referencia, ha permitido al grupo ATVS participar en la evaluación Albayzín 2014 de segmentación de audio. Dicho sistema se basa en tres detectores GMM-UBM, cada uno de los cuales se diseña para detectar la presencia de cada una de las clases acústicas consideradas en la evaluación: voz, música y ruido. La base de datos proporcionada por la organización para el desarrollo de los sistemas de segmentación ha servido además como marco experimental para desarrollar y evaluar las nuevas técnicas de extracción de características propuestas en este proyecto.

En la segunda fase, el proyecto se ha centrado en el uso de características cromáticas para la segmentación de audio. En primer lugar, se ha actualizado y adaptado un sistema previo de segmentación de audio basado en estadísticos de la entropía cromática, comparando su rendimiento con el sistema basado en características tímbricas (MFCC-SDC) en el mismo marco experimental (base de datos de desarrollo de la evaluación Albayzín 2014). Posteriormente, se han implementado dos nuevos extractores de características cromáticas: uno basado en la agrupación de la energía por subbandas y otro en la agrupación por octavas. Ambas características han sido empleadas de forma conjunta para el desarrollo de un nuevo sistema de segmentación de audio, evaluado sobre el mismo marco experimental (Albayzín 2014).

Finalmente, se ha analizado la complementariedad de los sistemas de segmentación de audio basados en los distintos tipos de características mediante su combinación tanto a nivel de características como a nivel de puntuaciones.

Palabras Clave

Segmentation Error Rate (SER), Reconocimiento de patrones, Gaussian Mixture Model- Universal Background Model (GMM-UBM), segmentación de audio, evaluaciones *Albayzin*, Equal Error Rate (EER), *Accuracy*, *k-means*, Maximum Likelihood (*ML*), Maximum A Posteriori (*MAP*).

Abstract

This project focuses on the implementation and analysis of different feature extraction techniques for segmenting audio news files. In it, two types of characteristics, timbric and chromatic are distinguished, although the main objective is to further investigate the use of these latter, whose introduction is motivated by a greater potential to distinguish between voice and music capacity.

In the first phase of the project a system based on timbral characteristics MFCC-SDC is optimized. Also considered as a reference system, as it allowed ATVS group to participate in the 2014 Albayzín audio segmentation competition. This system is based on three detectors GMM-UBM, each of which is designed to detect the presence of each of the acoustic classes considered in the evaluation: speech, music and noise. The database provided by the organization for the development of segmentation systems has also served as an experimental framework to develop and evaluate new features extraction techniques proposed in this project.

In the second phase, the project has focused on the use of chromatic characteristics for audio segmentation. First, a segmentation system based on statistics of chromatic entropy has been updated and adapted, comparing their performance with the system based on timbral characteristics (MFCC-SDC) in the same experimental framework (database development system from Albayzín 2014 evaluation). Subsequently, two new extractors based on chromatic features have been implemented: one based on the grouping of the energy in sub-bands and one in the group by octaves. Both features have been employed together for the development of a new audio segmentation system, evaluated on the same experimental setting (Albayzín 2014).

Finally, the compatibility of audio segmentation systems based on different types of features has been analyzed, by combining both in terms of features and in terms of ratings.

Key words

Segmentation Error Rate (SER), Pattern recognition, Gaussian Mixture Model - Universal Background Model (GMM-UBM), audio segmentation, *Albayzin*, Equal Error Rate (EER), *Accuracy*, *k-means*, Maximum Likelihood (*ML*), Maximum A Posteriori (*MAP*).

Agradecimientos

En primer lugar mi agradecimiento es de doble vía, por un lado, estoy muy agradecida con Daniel Ramos por confiar en mí y ofrecerme un proyecto a la altura de mis inquietudes, por compartir conmigo sus conocimientos, inquietudes musicales, por su cercanía, por su empatía y por sacarme una sonrisa cada mañana con sus bromas. Por otro lado, muchas gracias a mi tutor Javier Franco por todo su tiempo, por su dedicación, su paciencia en mis momentos espesos, por ayudarme a descubrir el encanto de línux, por transmitirme tantos conocimientos y por guiarme a lo largo del proyecto. Gracias por estar siempre disponible y por su cercanía.

En segundo lugar, no puedo olvidarme de la persona que confió en mí desde el principio y sembró en mí ese interés por el ATVS y el reconocimiento de patrones ya desde hace algún año. Gracias Javier Ortega por la cercanía que me has ofrecido durante los últimos años de carrera, por tus sabios consejos y por la experiencia de aprendizaje en la cátedra.

En tercer lugar, y siguiendo en el terreno universitario, gracias a las personas que me habéis acompañado estos seis años. Por los momentos buenos y por los duros momentos que nos curten y nos hacen aprender. Gracias Ana por haber estado siempre siempre a mi lado, gracias a Alex y Alicia por vuestro amistad y apoyo, y gracias Pascu por tu amistad y por ser siempre un modelo a seguir durante estos cinco años, doy gracias por que nos hayamos encontrado en el camino. Gracias a los profesores que han transmitido enseñanza con entusiasmo y entrega, y a todos aquellos que me han ayudado a aprender, y a ver realizado el proyecto de ser Ingeniera en Telecomunicaciones.

Gracias a todos los compañeros del ATVS, por acogerme como una más y compartir ratos de descanso, las comidas, etc. Gracias por compartir inquietudes y empatizar en los retos y dificultades que han ido surgiendo en el proyecto. Gracias a Ester, a Alicia, a Aythami y a Rubén I, II y III por su cercanía durante estos meses.

Y pensando en las personas que más han hecho por mí en mi vida, gracias a mi madre y mi hermana por hacer tanto por la persona en la que me he convertido, muchos de mis logros en la vida no los podría haber conseguido sin su apoyo. Gracias a mi abuela por su cariño y por compartir conmigo su deseo de tener una nieta ingeniera. Gracias a Juan, por quererme cada día por la persona que soy, por ayudarme a sacar lo mejor de mí misma y por compartir conmigo éxitos y fracasos.

Gracias a todas esas personas que se van quedando en el camino, pero que marcan tu vida y me han dejado huella. A los amigos con los que comparto ocio e inquietudes de vida, a mis amigos del instituto y a los que han ido viniendo después.

Exitos y fracasos modelan nuestra vida, pero la vida es un regalo que hay que disfrutar cada día.

Elena Gómez Rincón Junio 2015

Felicidad no es hacer lo que uno quiere sino querer lo que uno hace.
Jean Paul Sartre

1

Introducción.

1.1. Motivación del proyecto.

Actualmente, el acceso a los archivos multimedia con contenidos de audio está al alcance de casi cualquiera. Gracias a los teléfonos móviles y el abaratamiento de la electrónica de consumo en general, se facilita la grabación de contenidos de audio en cualquier ubicación (y no sólo en estudios, donde las condiciones acústicas son controladas) por lo que la variabilidad del sonido del que se puede disponer es muy alta. En el caso de una emisión de radio, por ejemplo, se pueden encontrar contenidos de audio, de voz aislada, fragmentos de habla con ruido de fondo, etc. En este contexto, aparecen los sistemas de segmentación de audio, los cuales proporcionan información relativa a la naturaleza del contenido acústico. Dicha información supone una herramienta de mejora en los sistemas de tratamiento de señales que trabajan con clases aisladas, por ello la presencia de cada una de estas clases debe ser correctamente detectada y etiquetada. Esto explica la necesidad de desarrollar buenos sistemas de detección y segmentación de audio en el área de las tecnologías del habla.

La tarea de segmentación puede ser considerada como una fase de pre-procesado en sistemas de tratamiento de audio de un tipo concreto. En algunas aplicaciones del área es considerada una tarea muy importante, como es el caso de los sistemas de detección de actividad de voz (*VAD*). Estos sistemas se apoyan en las técnicas de clasificación de audio para facilitar la tarea de sistemas de reconocimiento de voz, de locutor o sistemas de segmentación de locutor, como por ejemplo los sistemas que se encargan de llevar a cabo el subtitulado automático de películas y programas de televisión (reconocimiento de voz) previa detección de contenidos de voz sobre el audio (segmentación de audio). Otro ejemplo a destacar son los sistemas de indexación de contenidos, los cuales especifican el contenido de los ficheros en función de meta-información que puede ser extraída de forma automática, como es el caso de los sistemas de segmentación de audio. No obstante, los sistemas de segmentación de audio ya constituyen una aplicación en sí misma, considerando por ejemplo, la selección de acceso a contenidos específicos de música dentro de un fichero de audio en general.

La realización de este proyecto en el grupo *ATVS* ha contribuido a generar los resultados que han sido presentados en la tercera convocatoria de la evaluación *Albayzin* de segmentación

de audio, la cual ha tenido lugar en octubre de 2014. En esta competición se evalúa la eficacia de diferentes sistemas de segmentación de audio en un marco común de referencia. El artículo generado a raíz de los resultados obtenidos [6] se encuentra publicado en el acta del congreso *IberSpeech 2014* que tuvo lugar en las Palmas de Gran Canaria.

En este proyecto se plantea el uso de características cromáticas y se compara el rendimiento de los sistemas basados en éstas frente al uso de las características tímbricas en el mismo marco experimental. Si bien las características tímbricas como los MFCC constituyen el estado del arte desde hace décadas en varias aplicaciones de las tecnologías del habla, el uso de características cromáticas en segmentación de audio es un tema que plantea aún muchos interrogantes y líneas de investigación, y por ello se busca con el presente proyecto desarrollar sistemas basados en características cromáticas, y estudiar cómo complementan a un sistema basado en características tímbricas (concretamente, MFCC-SDC).

1.2. Objetivos

La realización de este trabajo busca cumplir los siguientes objetivos, entre los que cabe destacar:

- Participar en la implementación de un sistema de segmentación de audio (basado en características tímbricas, en concreto coeficientes MFCC-SDC) el cual será presentado en la evaluación *Albayzin 2014 de segmentación de audio* [6].
- A partir de la base de datos proporcionada por dicha evaluación, analizar la eficacia de un sistema previamente desarrollado en el grupo *ATVS* basado en características cromáticas (estadísticos de la entropía cromática), y estudiar la fusión de dicho sistema con el presentado en la Evaluación *Albayzin 2014*.
- Desarrollar un nuevo sistema de extracción de características cromáticas basado en la información extraída directamente del cromagrama en contraposición al cálculo de la entropía cromática y posteriormente sus estadísticos, y estudiar su eficacia en un marco común de referencia.
- Contrastar el rendimiento de los diferentes sistemas por separado y al ser fusionados (a nivel de características y de *scores*) en la tarea de segmentación de audio.

1.3. Contribuciones

Dentro del área de las tecnologías del habla, el presente Proyecto Fin de Carrera ha supuesto las siguientes contribuciones:

- La propuesta de este PFC ha impulsado la participación por parte del grupo *ATVS* en la evaluación *Albayzin 2014* de segmentación de audio, generando un artículo que describe el sistema propuesto y forma parte de las actas del congreso *Iberspeech 2014*.[6]
- Se han revisado y optimizado algoritmos de segmentación de audio desarrollados con anterioridad en el grupo *ATVS*, contrastando su eficacia con el estado del arte actual.

- Se ha desarrollado un nuevo algoritmo de extracción de características basado en nuevas características cromáticas, cuyos resultados mejoran para algunas clases los presentados en la evaluación.
- Se ha realizado un análisis detallado del rendimiento de cada uno de los sistemas basados en distintos tipos de características aplicados a la tarea de segmentación de audio.
- Se han llevado a cabo experimentos de fusión de sistemas a nivel de características y de scores, quedando los resultados reflejados en esta memoria.

1.4. Organización de la memoria

Esta memoria consta de los siguientes capítulos:

- **Capítulo 1: Introducción.**
  Este capítulo presenta la motivación para el desarrollo de este proyecto, los objetivos a cumplir durante el desarrollo del mismo, sus consecuentes contribuciones al grupo *ATVS*, y por último, la estructura de dicha memoria.
- **Capítulo 2: Segmentación de audio.**
  Este capítulo ofrece un acercamiento a las diferentes naturalezas del audio, detalla en qué consiste y en qué ámbitos se aplica la segmentación de audio, y las diferentes medidas de error que se pueden emplear para evaluar el rendimiento de los sistemas. Asímismo se presenta una de las evaluaciones competitivas internacionales relacionadas con dicha tarea.
- **Capítulo 3: Estado del Arte**
  Este capítulo detalla las distintas tareas que debe realizar un sistema de segmentación de audio, desde la extracción de parámetros hasta la clasificación de segmentos entre los distintos tipos de audio considerados. Para cada tarea destacada se explican algunos de los algoritmos más utilizados en el estado del arte actual, y se contempla el modo de llevar a cabo diferentes tipos de fusiones de sistemas con el objetivo de mejorar el rendimiento de los sistemas individuales.
- **Capítulo 4: Marco experimental y sistema de referencia**
  En este capítulo se presenta el sistema de segmentación de audio basado en características tímbricas propuesto para la evaluación *Albayzin* 2014 detallando la técnica de segmentación y los parámetros empleados. A continuación, se presentan los resultados de este sistema.
- **Capítulo 5: Sistemas basados en información cromática**
  Es este capítulo se detallan cada uno de los sistemas basados en características cromáticas desarrollados en este proyecto, así como los sistemas fusionados, bien a nivel de scores, bien a nivel de características, con el sistema de referencia. Una vez detallados y explicados, se incluye un apartado de comparativa de sistemas.
- **Capítulo 6: Conclusiones y trabajo futuro**
  En este capítulo se presentan por un lado las conclusiones obtenidas en el proyecto a raíz de desarrollar y/o evaluar cada uno de los sistemas de segmentación de audio con diferentes extractores de características tímbricas y cromáticas. Por otro lado, se enumeran las posibles vías de trabajo futuro que quedan abiertas en la línea de segmentación de audio tras finalizar la investigación con este proyecto.

2

Segmentación de audio.

2.1. Introducción

En este capítulo se define el problema de la segmentación de audio, empezando por una propuesta de clasificación de los tipos de audio, seguida de un encuadre algo más detallado de dicho problema, y terminando con un análisis de posibles medidas de rendimiento aplicadas a dicha tarea.

2.2. Naturaleza de las diferentes clases de audio

El audio es una señal analógica que pretende representar a una señal sonora [7]. Las ondas sonoras que percibe el oído humano se distinguen principalmente por tres características: nivel de intensidad acústica, tono y timbre, las cuales se relacionan con las propiedades físicas de intensidad, frecuencia fundamental y forma de onda, respectivamente.

- **Nivel de intensidad**: viene determinado por la amplitud del movimiento oscilatorio. En términos matemáticos, es la potencia transferida por una onda por unidad de área $I = P/A$, y usualmente se mide en (W/cm^2) . No obstante, al medir ondas sonoras, la intensidad se suele expresar en *decibelios (dB)*, $I(db) = 10 * logI/I_o$, donde $I_o = 10^{-10} mW/cm^2$ es considerado como umbral de audición del audio humano.

- **Tono o frecuencia**: mide el número de oscilaciones por segundo de una onda. Su unidad de medida son los *Hertzios (Hz)*. Este parámetro permite distinguir entre sonidos graves (baja frecuencia) y agudos (alta frecuencia). En este contexto, se entiende como frecuencia fundamental a la frecuencia de apertura y cierre de los pliegues vocales (en los seres humanos) lo que se corresponde con frecuencia de afinación en los instrumentos musicales.

- **Timbre**: es la cualidad que permite diferenciar dos sonidos con igual intensidad y tono. El timbre es característico para cada instrumento musical y para cada voz humana, y se caracteriza principalmente por la intensidad de los armónicos presentes en una onda sonora.

Puesto que los sonidos pueden provenir de diferentes fuentes (voz humana, voz cantada, música instrumental, ruido ambiente, etc.), también el audio podrá ser clasificado en función de su naturaleza. En este proyecto, se va a trabajar con tres clases concretas de audio: voz, música y ruido.

2.2.1. Voz

La señal de voz es el resultado de codificación del lenguaje hablado [8]. Este proceso de codificación del lenguaje, que es diferente en cada persona, facilita que seamos capaces de reconocer a una persona por su voz. Del mismo modo, este mecanismo humano de producción de voz es capaz de generar infinidad de sonidos diferentes, lo que facilita el proceso de comunicación entre las personas.

No obstante, se pueden extraer rasgos comunes de la señal de voz que permiten distinguirla del resto de señales, como por ejemplo, el rango de frecuencias. El rango común de frecuencias de la voz oscila entre los 87 y los 1175 Hz [1] ampliándose en algunas personas desde los 60 hasta las 7000 kHz (e incluyendo voz cantada),por lo que se puede decir que la señal de voz solamente abarca una parte del espectro auditivo, que abarca desde los 20 a los 20000 Hz aproximadamente. Nuestros oídos están preparados para oír más allá de lo que la voz alcanza, como por ejemplo, señales de ruido o música instrumental.

La señal de voz puede considerarse pseudo-estacionaria a corto plazo (en fragmentos de voz del orden de decenas de ms). Por este motivo, las técnicas de análisis y procesado de la señal de voz trabajan con segmentos de duración limitada (decenas ms), lo que se conoce como *tramas* de voz. Se puede contrastar este carácter pseudo-estacionario de la voz comparando una locución de voz de varios segundos (figura 2.1) con el análisis localizado de la señal de voz sobre unas pocas tramas (figura 2.2). A partir de estas figuras se puede ver cómo en el primero de los casos la señal de onda no presenta a simple vista ciclos periódicos, o repeticiones, mientras que en la segunda figura ya se puede apreciar cierta periodicidad de la señal, la señal se presenta con un carácter estacionario.

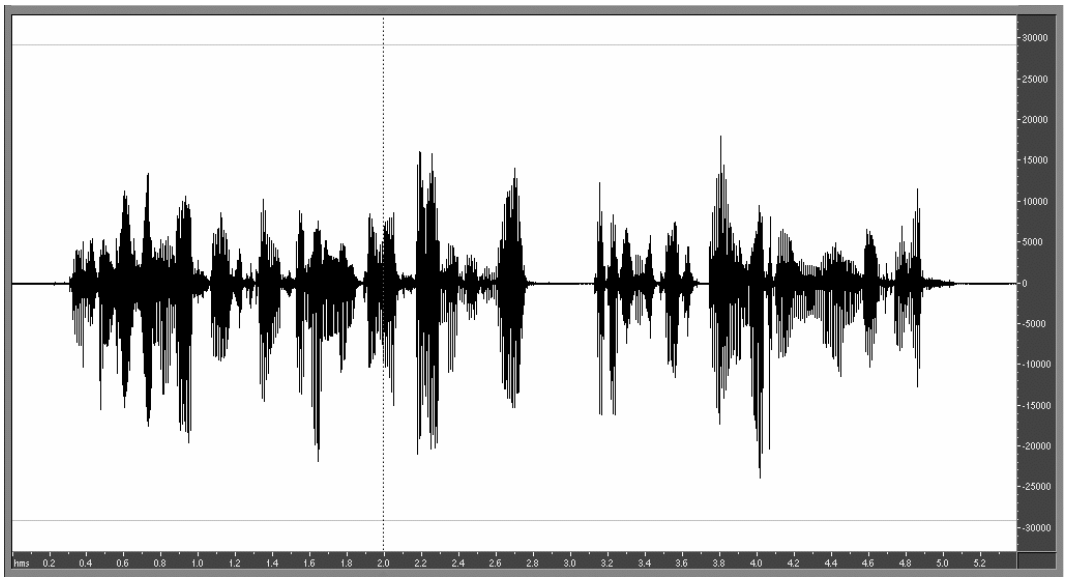


Figura 2.1: Locución de 5.4 segundos de voz. Fuente: [1].

[1] NIST Special Publications 559 - Time and Frequency Users Manual

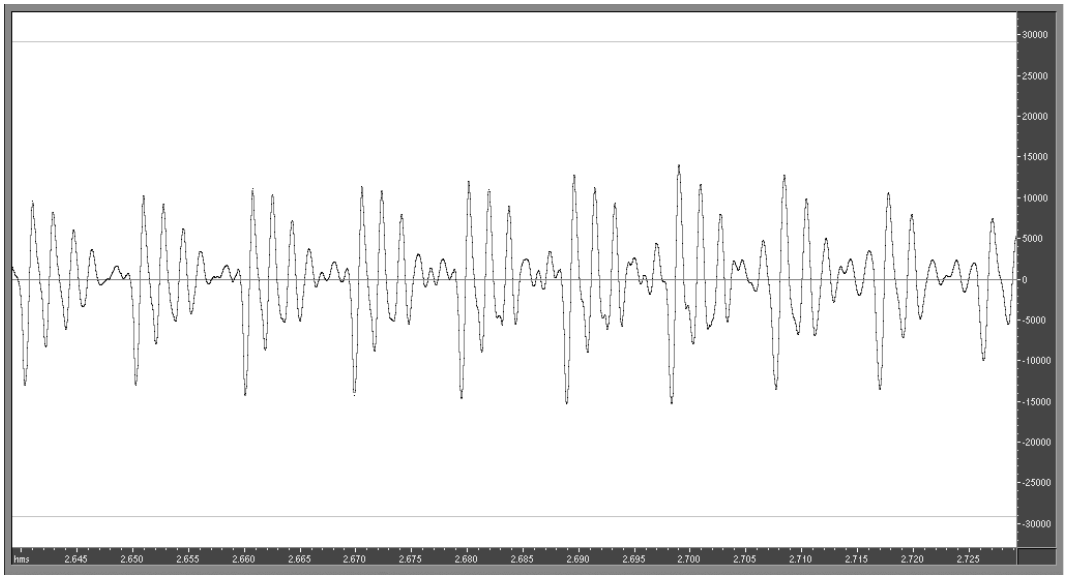


Figura 2.2: Locución de decenas de ms. Fuente: [1].

2.2.2. Música

Se puede definir la música como el arte de combinar sonidos en una secuencia temporal, atendiendo a las leyes de la armonía, la melodía y el ritmo, o de producirlos con instrumentos musicales. Estos conceptos se describen a continuación.

- **Melodía**: está constituida por una serie de eventos lineales sonoros no simultáneos pero con cambios de algún tipo de modo que permiten en conjunto percibir la melodía como una sola entidad. Se denominan canciones melódicas a aquellas que ponen el mayor énfasis en la sucesión lineal de notas, como es el caso de la música pop; e instrumentos melódicos, a aquellos que sólo pueden hacer sonar un sonido detrás de otro, como es el caso de la flauta, por ejemplo.
- **Armonía**: es el arte de combinar sonidos de forma simultánea, produciendo un resultado agradable al oído humano. Existen instrumentos capaces de generar varios sonidos simultáneamente, como es el caso del piano o la guitarra. Con esta técnica, se pueden generar diferentes sensaciones (o climas), como de alegría, tristeza, tensión, etc.
- **Ritmo**: es un atributo sonoro marcado por la sucesión regular de eventos de diferente naturaleza, que resulta característico para cada estilo musical. Por ejemplo, estilos como el *Rock and Roll* se identifican por un ritmo rápido, mientras que estilos de balada se caracterizan por ser de ritmo lento.

Estos rasgos diferencian claramente a la música de cualquier ruido o voz hablada. No obstante, lo que a veces es muy claro para el hombre, no siempre es fácil de enseñar a una máquina o autómata.

La señal musical abarca un amplio abanico de frecuencias y ofrece diversos estilos. Cada instrumento abarca un rango determinado de frecuencias, como se muestra en la figura 2.3. Como se puede observar en dicha imagen, los instrumentos musicales pueden generar ondas sonoras con frecuencias tan bajas como los 28 *Hz*, en términos de frecuencia fundamental, y

rozar los 20 *KHz* a nivel de armónicos. Esto supone que, de media, un oído humano es capaz de escuchar cualquier sonido producido por un instrumento, si bien solamente un oído especialmente dotado y entrenado (lo que en música se conoce como oído absoluto) sería capaz de reconocer cada una de las notas o acordes que produce dicho instrumento. En esta línea, el tratamiento de señales en el dominio frecuencial supone una herramienta clara para poder reconocer frecuencias, es decir, notas o acordes.

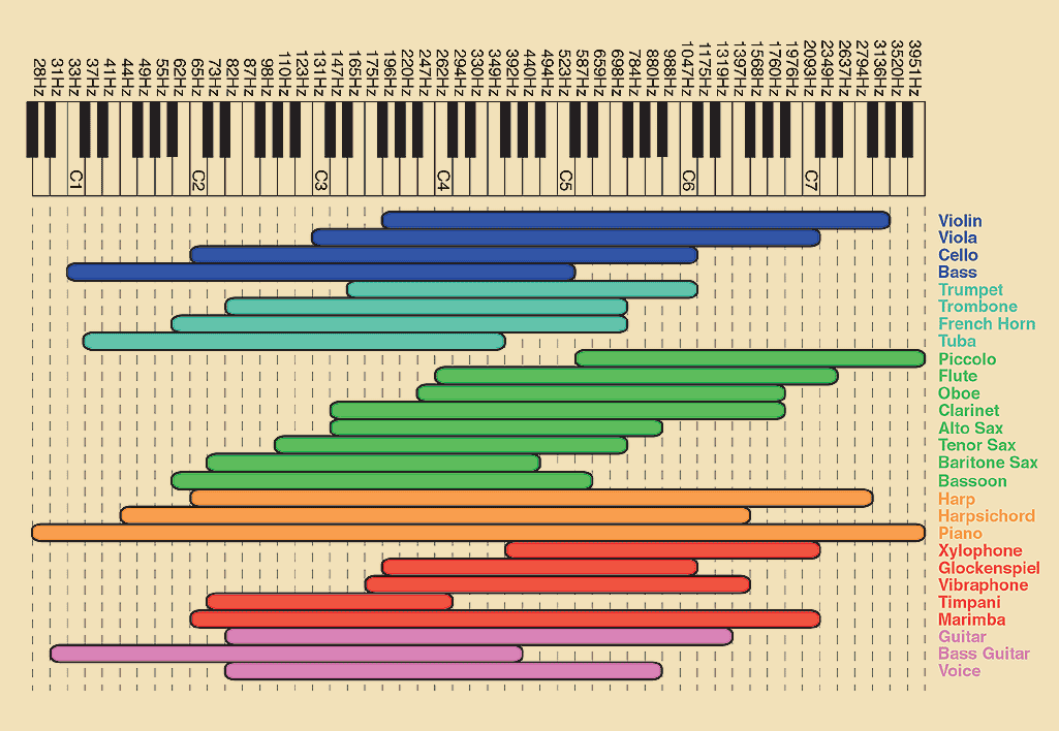


Figura 2.3: Frec. fundamental de instrumentos musicales (http://www.punksunidos.com).

Al igual que la voz humana, la música tiene una naturaleza de carácter pseudo-estacionaria a corto plazo, por lo que el procesado de estas señales sobre pequeños segmentos de audio facilita el análisis de las ondas sonoras.

2.2.3. Ruido

En el área de tratamiento de señales el ruido es generalmente considerado como un fenómeno no deseado y en muchas ocasiones inevitable. El ruido existente en una señal puede venir de diversas fuentes, como por ejemplo el ruido que proviene del canal (generado en el proceso de comunicación) o el ruido de cuantificación (resultado del proceso de digitalización de una señal). Se puede encontrar ruido en casi cualquier frecuencia, lo que implica que los seres humanos sólo distinguimos ciertos tipos de ruido, esto es, los que se encuentran dentro del rango audible.

Por otro lado, si bien el ruido puede ser considerado como una señal, su naturaleza usualmente aleatoria lo hace muy difícil de tratar. Aunque un oído humano es capaz de distinguir con relativa facilidad toda señal considerada como ruido, resulta más complicado generar este aprendizaje en una máquina autómata ya que el ruido carece de patrones debido a su naturaleza. No obstante, esta falta de patrones es la que podría permitir al ruido diferenciarlo de otras clases acústicas, como la música o la voz, y poder así ser detectado por un sistema de segmentación de audio.

2.3. Segmentación de audio: definición y objetivos

Como ya se introdujo en el apartado 1.1, segmentar audio consiste en identificar por regiones el contenido acústico de un fichero, asignando a cada segmento la etiqueta de la clase a la que pertenece y dotando de este modo a los ficheros de audio de información textual relativa al contenido acústico de los mismos.

Un sistema de segmentación de audio permite definir regiones acústicas partiendo de un fichero de audio cualquiera, que más adelante puedan ser aplicadas a diferentes sistemas de tratamiento de audio (como un reconocedor de voz). Por ejemplo, en el caso de un fichero de audio de un fragmento del telediario, éste contendrá generalmente contenidos muy variados (fragmentos de música, habla de diferentes locutores, ruido de fondo en algunas ocasiones, etc). En este marco, si se quiere detectar cuando habla una persona concreta, el primer objetivo consiste en detectar todos los tramos que contienen voz (sistema de segmentación de audio). A continuación, se identifican y agrupan los tramos donde hablan distintas personas (sistema de seguimiento de locutor) y a continuación se identifican cuales de éstos corresponden al presidente (sistemas de reconocimiento de locutor). Finalmente, se puede reconocer el contenido de los fragmentos detectados (sistemas de reconocimiento de voz) e incluso traducir y sintetizar el contenido en otro idioma. En este caso intervienen todos los sistemas, pero para ello, lo primero es la segmentación de audio.

El audio que se presenta en los ficheros de la base de datos responde a tres naturalezas diferenciadas: música, voz y ruido. Estas tres naturalezas del audio (utilizadas para el desarrollo de este proyecto) se pueden encontrar de manera aislada (cuando percibimos solamente música, o voz o ruido) o mezcladas (cuando simultáneamente escuchamos música y voz, o voz con ruido de fondo por ejemplo). De este modo, sobre un segmento de audio que presente más de una clase acústica, se pueden obtener una o varias etiquetas (según el modo elegido de generar dichas etiquetas). Así pues, las clases o etiquetas podrían definirse como:

- Homogéneas (clases no solapadas): una etiqueta por cada tipo de audio encontrado, de modo que puedan aparecer varias etiquetas asociadas a una misma región de audio.
- Heterogéneas o mixtas (clases solapadas): una única etiqueta por cada región de audio.

En la figura 2.4 se ejemplifican ambos métodos de etiquetado. Para cualquiera de los casos, el conjunto de clases sobre las que trabajar puede variar. Si el objetivo es aislar una clase concreta del resto, como por ejemplo la clase de música, un posible conjunto de clases serían música y otros. Si por el contrario el objetivo es reconocer cada una de las clases que están presentes en un fichero de audio (ya sea de manera aislada o no), se puede trabajar con tantas etiquetas como clases. En este caso, un posible conjunto estaría formado por las clases de música, voz y ruido. Este caso, en particular, es el objetivo propuesto en la evaluación *Albayzin 2014* de segmentación de audio.

Si bien la tarea de segmentación de audio encuentra aplicaciones en sí misma, a continuación se resumen algunos de los sistemas que se benefician de este tipo de sistemas para mejorar su rendimiento, además de los que ya se mencionaron en la sección 1.1:

- Sistemas de seguimiento de locutores, cuyo objetivo es separar las regiones en las que hablan distintas personas en una conversación, y agrupar las regiones que pertenecen a la misma persona en función de las características acústicas. En este contexto, es importante poder distinguir una señal de ruido u otra naturaleza de la voz hablada del locutor para poder localizarle correctamente.

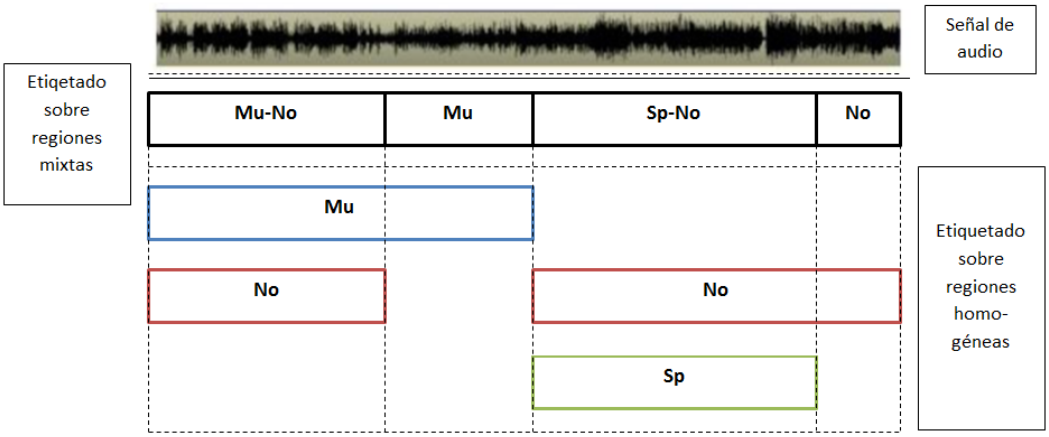


Figura 2.4: Ejemplo de etiquetado de señal de audio.

- Sistemas de traducción simultánea, los cuales integran un módulo de reconocimiento y otro de síntesis de la señal a procesar. Estos sistemas trabajan con señales de voz, por lo que la presencia de música o ruido podría perturbar el rendimiento del sistema.
- Sistemas de recuperación de información musical enmarcados en el área de *Music Information Retrieval* (MIR), los cuales tratan de extraer información de género, ritmo, etc. de los ficheros para recomendación o recuperación de música, para lo cual la inclusión de otro tipo de señales acústicas que no pertenezcan al género musical (como la presencia de ruido) dificultaría enormemente el trabajo a realizar. Aplicaciones como Spotify se enmarcan en este tipo de sistemas.

2.4. Medidas de rendimiento

Cualquier sistema de reconocimiento de patrones debe incluir en su desarrollo una fase de evaluación que permita analizar la bondad del sistema y medir sus capacidades. Existen diferentes técnicas para medir el rendimiento del sistema, adaptadas a los diferentes objetivos que se deseen alcanzar. No se medirá el error del mismo modo en un sistema cuya prioridad sea filtrar en la medida de lo posible fragmentos de ruido en una grabación de un programa de radio, que en un sistema cuyo objetivo sea detectar con el mayor grado de acierto todas las clases presentes en el mismo fichero. Una buena medida de error para el primer caso vendría dada por la tasa de falso rechazo, mientras que en el segundo caso las matrices de confusión darían una visión más genérica de la precisión del sistema para cada una de las clases.

Las medidas de rendimiento empleadas en segmentación de audio (y reconocimiento de patrones en general) se pueden distinguir en función de su utilidad.

- Errores de detección para optimizar el detector de una clase particular (p.ej. el detector de voz).
- Exactitud o *Accuracy* para evaluar de forma global el sistema de segmentación.
- Matrices de confusión para evaluar la precisión por clase y tener una información detallada de qué clases se confunden entre sí.

Por otro lado, cabe destacar el error de segmentación (SER) empleado en las evaluaciones *Albayzin*[9].

2.4.1. Errores de detección

La tarea de segmentación de audio se puede afrontar a partir de varios detectores individuales de cada clase homogénea, como hacemos en este PFC. Se puede entender como detector al sistema que, dada una trama de audio y un modelo determinado de una clase (voz, música, ...) proporciona una puntuación (score) que es más alta cuanto más alto es el apoyo del sistema a que dicha clase está presente en esa trama de audio.

A partir de los *scores* obtenidos se puede elegir un umbral de detección θ para cada clase, a partir del cual determinar para cada uno de los segmentos si pertenecen o no a cada una de las clases. Una vez establecido el umbral sobre cada detector (que puede ser o no ser el mismo para los diferentes detectores), todas las tramas cuya puntuación quede por encima de dicho valor serán asignadas a la clase sobre la que están siendo evaluados y viceversa, es decir, tramas cuyas puntuaciones no superen el umbral serán rechazadas.

Cuando un sistema detecta si una trama corresponde o no a esa clase, el resultado puede ser exitoso o erróneo. En caso de que el sistema falle, se expresa dicho fenómeno bajo el error de falso rechazo y el error de falsa aceptación, tal y como se detalla a continuación:

- **Error de Falso rechazo (FR)**: se produce cuando una trama se acepta como perteneciente a una determinada clase, siendo dicha asignación errónea. Esto quiere decir que la puntuación ha quedado por debajo del umbral cuando la trama realmente pertenecía a la clase.

- **Error de falsa aceptación o falsa alarma (FA)**: mide el caso opuesto al error de falso rechazo, es decir, se produce cuando tramas que no pertenecen a una clase determinada son aceptadas como pertenecientes a dicha clase, lo que produce una asignación errónea de las mismas.

Las curvas DET [10] representan en los ejes x e y las tasas de falso rechazo y de falsa aceptación, respectivamente, para diferentes valores del umbral θ. En esta línea, la figura 2.5 muestra un ejemplo en el que se representa la curva DET de cada uno de los detectores de un sistema de segmentación en el que se modelan las tres clases acústicas con las que se trabaja en este proyecto (música, ruido y voz), con su respectivo EER. Como se puede apreciar con este ejemplo, dichas curvas resultan muy útiles para poder ver de forma visual el poder discriminativo de un sistema. Cuanto más se acerca dicha curva al origen de coordenadas, mayor es el poder discriminativo de dicho sistema. Por otro lado, el cruce entre la curva DET con la bisectriz de los ejes de la gráfica se corresponde con el EER (*equal error rate*), punto de la gráfica en el que ambas probabilidades (FA y FR) se igualan. Proporcionar el EER es una forma muy utilizada para resumir en un solo valor el rendimiento del sistema (en lugar de dar FA y FR).

Un sistema de segmentación de audio puede construirse a partir de varios detectores, cada uno de los cuales será el encargado de etiquetar la presencia de la clase que representa (p.ej. música, voz o ruido); por lo tanto, cuanto mejor sea el rendimiento de cada uno de los detectores por separado mejor funcionará el sistema de segmentación construido con dichos detectores.

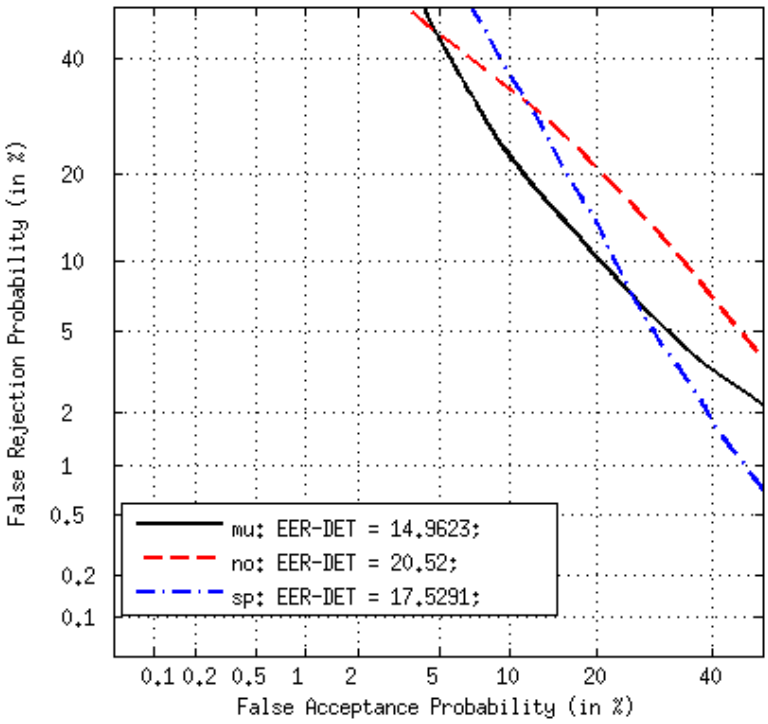


Figura 2.5: Ejemplo de curva DET con su respectivo EER para cada detector de clase (música, voz y ruido).

2.4.2. Matrices de confusión y *Accuracy*

Se entiende como matriz de confusión a la tabla de valores o matriz que permite visualizar la precisión de un sistema en cuanto a nivel de aciertos y asignación incorrecta entre clases. De este modo se puede ver en el caso de los errores, no solamente la cantidad de segmentos que han sido erróneamente asignados, sino también a qué clase han sido incorrectamente asignados, lo que permite reconocer la confusión *intraclases* generada. Se trata por lo tanto de matrices cuadradas de dimensión $n * n$, donde n representa el número de clases analizadas en el sistema. A partir de la siguiente tabla 2.1 se ilustra la información que representan las matrices de confusión:

Referencia \ Asignación	Música	Ruido	Total
Música	60 %	20 %	80 %
Ruido	15 %	5 %	20 %
Total	75 %	25 %	**65 %**

Tabla 2.1: Ejemplo de matriz de confusión sencilla.

En este caso, la matriz de confusión ha sido aplicada a un sistema de segmentación de audio en dos clases, música y ruido. La lectura de la tabla por columnas muestra el total de muestras que han sido experimentalmente asignadas a cada clase, mientras que la lectura por filas determina el número de muestras reales pertenecientes a cada una de las clases. En el caso del ejemplo, una lectura por columnas nos permite apreciar que el 75 % de los segmentos han

sido asignados a la clase música, y un 25 % a la clase ruido. Sin embargo, una lectura realizada por filas muestra que en total son un 80 % los segmentos reales de música y solamente el 20 % los segmentos de ruido. Las casillas situadas en la diagonal de la tabla contienen el total (en porcentaje) de segmentos que han sido correctamente asignados, ofreciendo en la esquina inferior derecha el porcentaje total de aciertos (nivel de exactitud), que en este ejemplo es de un 65 %. En un caso óptimo, la suma de los valores de la diagonal debiera coincidir con el número de muestras analizadas, reflejando de este modo que todas las asignaciones realizadas por el sistema han sido exitosas.

A partir de estas matrices se pueden obtener varias medidas de acierto. En este proyecto se ha trabajado en concreto con el nivel de exactitud global (*accuracy*) y con el nivel de exactitud por clases. El *accuracy* mide la proporción de muestras que han sido correctamente asignadas respecto del total. Dicha medida ofrece una visión general del poder discriminativo del sistema. No obstante, es importante recalcar que la probabilidad de que un detector de clase funcione bien aumenta con el número de muestras de dicha clase con las que ha sido entrenado el sistema. Como en cualquier sistema de reconocimiento de patrones, cuanto mayor sea el conjunto de datos de entrenamiento, mayor es la capacidad de generalización. Para poder evaluar de manera aislada cómo de preciso es cada uno de los detectores sobre cada conjunto, se calcula la exactitud por clase a partir del número de muestras que han sido correctamente asignadas sobre el total de muestras de cada conjunto. Dicha medida de error, acompañada de un diagrama de distribución del contenido de audio por clases, ofrece una visión precisa de cómo se está comportando el sistema. Este detalle es bastante importante, ya que tomando como referencia la tabla anterior, si se presta atención se puede comprobar cómo a pesar de que el total de segmentos correctamente asignados (nivel de *accuracy*) es de 65 sobre 100 (65 %), el detector de ruido es especialmente inexacto, ya que sólo clasifica correctamente un 5 % sobre el total de 20 % que pertenecen a esta clase.

En la figura 2.6 se muestra un ejemplo de una matriz de confusión que evalúa el rendimiento de un segmentador de audio de ocho clases (ruido, voz con ruido, voz, voz con música, voz con música y ruido, música, música con ruido u silencio), con su respectivo nivel de exatitud global mostrado al pie de la gráfica y la exactitud por clases codificada en escala de grises. En este ejemplo (y en cualquiera de los resultados presentados en el proyecto) se representa la información de la matriz de confusión en porcentaje respecto al total de tramas de cada clase, codificado con escala de grises (cuanto más oscuro mayor porcentaje). De este modo, en primer lugar cada celda se ha normalizado sobre el total de segmentos que pertenecen a cada clase, y se multiplica por 100, generando así un valor normalizado entre 0 y 100. A partir de estos valores, se intercambian los números por tonalidades del gris, donde el valor 0 se corresponde con el blanco y el 100 con el negro. Esta visualización facilita la comprensión de los resultados de manera global y sobre cada uno de los detectores por separado.

2.4.3. SER

El SER o *segmentation error rate* es la medida de error de la evaluación *Albayzin* 2014 de segmentación de audio utilizada en las evaluaciones NIST (*National Institute of Standards and Technology*) de seguimiento de locutor. Esta tasa de error mide y promedia la fracción de tiempo en el que una clase no ha sido correctamente asignada, respecto del tiempo total [9]. Considera tres tipos de errores: *miss*, *false alarm* y *class*.

Sin embargo, para el presente proyecto el SER solamente se ha medido sobre el sistema de referencia presentado a la evaluación, ya que las otras métricas consideradas dan mucha más

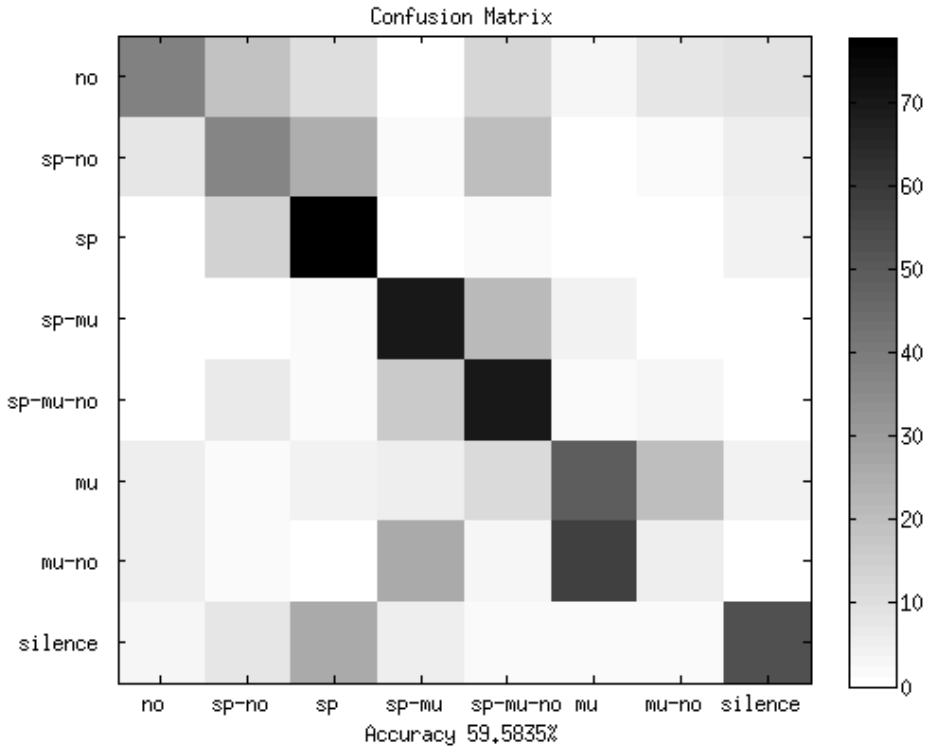


Figura 2.6: Ejemplo de matriz de confusión.

información, tanto para optimizar los detectores individuales (EER por detector) como para evaluar el sistema de segmentación en su conjunto (matrices de confusión y accuracy).

2.5. Evaluaciones Albayzin de segmentación de audio

Las evaluaciones *Albayzin* de segmentación de audio son campañas competitivas llevadas a cabo por la Red Temática en Tecnologías del Habla (RTTH) en las que se comparan distintas propuestas de sistemas de segmentación de audio sobre un marco común. La primera evaluación de esta índole tuvo lugar en el año 2010. En esta primera convocatoria participó el grupo ATVS con un sistema basado en *HMM's* [11], y hasta la fecha se han celebrado dos convocatorias más en este área en los años 2012 y 2014. Cada año se publican los resultados en las actas de las Jornadas de Tecnologías del Habla, conferencia en la que se presentan, además de los resultados de ésta y otras evaluaciones Albayzín, diversas contribuciones científicas en el ámbito de las tecnologías del habla y el cual recibe el nombre de *IberSpeech* desde el año 2012. Los resultados de la evaluación correspondiente al año 2014, entre los cuales se incluye el artículo que describe el sistema propuesto por el grupo de reconocimiento biométrico ATVS [6], han sido publicados en las actas de la conferencia *Iberspeech* 2014.

Al igual que en las evaluaciones de 2010 y 2012, el objetivo de la evaluación de 2014 consiste en presentar sistemas que sean capaces de segmentar ficheros de audio provenientes de una base de datos de noticias indicando con una etiqueta independiente sobre cada segmento (imagen 2.3) la presencia de voz, música y/o ruido. Esto quiere decir que más de una clase puede ser encontrada simultáneamente, por lo que el sistema a desarrollar debe ser capaz de detectar si una, dos o tres clases están presentes en un instante de tiempo. En esta evaluación de 2014 se considera que la voz está presente cada vez que una persona está hablando, pero no cuando se oye de fondo. La música se entiende en un contexto general, sin hacer diferencias de estilos o procedencia, y el ruido engloba cualquier contenido acústico que no pueda ser considerado como voz o música.

La base de datos ofrecida para esta evaluación está formada por una combinación de tres bases de datos:

- El primer conjunto está formado por archivos del canal de televisión catalana *3/24* (tv3.cat), utilizados para la evaluación de 2010. Esta base de datos fue compuesta por el grupo de investigación TALP de la Universidad Politécnica de Cataluña (UPC) en 2009 bajo el proyecto de *Tecnoparla* financiado por la generalitat de Cataluña. Se compone de 87 horas de grabaciones en las cuales predomina el género de voz, el cual está presente en un 92 % de los segmentos, de los cuales un 40 % están acompañados de ruido de fondo y en un 15 % acompañados de música.
- El segundo conjunto de archivos proviene de la Radio de Aragón, en concreto, de la *Corporación Aragonesa de Radio y Televisión* (CARTV), los cuales fueron empleados para la evaluación de 2012. En este caso, no se especifica la distribución de los contenidos.
- El tercer conjunto de datos lo forman sonidos ambientales provenientes de diversas fuentes, como *Freesound.org* y *HuCHorpus*.

Estos datos forman una base de datos compuesta por 35 grabaciones de audio de una duración aproximada de 60 minutos cada una. Los datos fueron distribuidos del siguiente modo:

- Las 20 primeras grabaciones fueron distribuidas con antelación (junto con sus etiquetas) a los participantes para desarrollar y testear sus sistemas.
- Las 15 últimas se entregaron a los participantes sin el correspondiente etiquetado para el desarrollo de la evaluación, y fueron empleados por la organización para comparar de

forma objetiva el rendimiento de los sistemas propuestos por los distintos participantes. Una vez publicados los resultados, fueron liberadas las etiquetas de estos ficheros de audio completando así la base de datos para desarrollar y evaluar futuros sistemas de segmentación.

Las etiquetas proporcionadas junto con los ficheros de audio continen una división del contenido por clases, marcando los tiempos de inicio y la duración de cada uno de los segmentos obtenidos.

3

Estado del arte.

3.1. Introducción

En el siguiente capítulo se presenta el estado del arte de los diferentes algoritmos y herramientas que se aplican en el proceso de segmentación de audio. En primer lugar, se encuentran los algoritmos de **extracción de características**, los cuales se centran en obtener información relativa a las muestras. A continuación, se encuentran los algoritmos de clasificación, los cuales tratan de buscar similitudes entre las características de la muestra a clasificar y los correspondientes modelos o plantillas considerados en la fase de entrenamiento. En la figura 3.1 se muestra un esquema general de un sistema de reconocimiento de patrones, distinguiendo entre una primera fase de entrenamiento y una segunda fase de testeo:

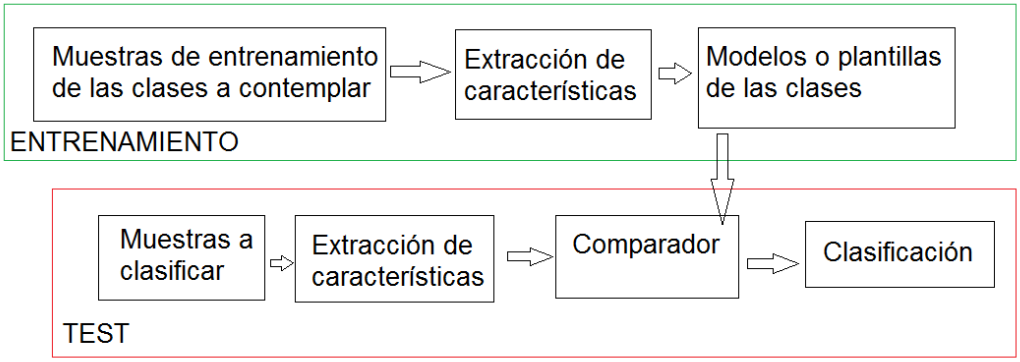


Figura 3.1: Esquema general de reconocimiento de patrones.

3.2. Extracción de características

La tarea de extracción de características, común a cualquier sistema de reconocimiento de patrones, busca extraer un conjunto de rasgos característicos de las clases a las que pertenecen las muestras a clasificar. Dicha tarea es requerida tanto en la fase de entrenamiento como en la de testeo. En la primera fase los resultados serán empleados para la obtención de modelos que representen a dichas clases. Una vez el sistema ha sido entrenado generando así los modelos requeridos, se extraen las características de las muestras de test, las cuales se compararán con los modelos previamente obtenidos.

Como se comentó en la sección 2.2, debido a la naturaleza pseudo-estacionaria a corto plazo de las señales de audio, las técnicas de procesado de audio trabajan con segmentos de duración limitada (tramas). El mecanismo que permite trabajar con tramas consecutivas de la señal se denomina *enventanado*. La técnica de enventanado consiste en multiplicar (en el dominio temporal) la señal de audio completa sobre una función limitada en el tiempo, también denominada ventana, produciendo así una nueva señal de audio de la misma duración que la original pero con valor nulo fuera del intervalo definido por la ventana. Esta operación se puede expresar del siguiente modo:

$$x(m) = s(n) \times w(m-n) \tag{3.1}$$

siendo *s(n)* la señal completa de audio original, *w(m-n)* la ventana deslizante aplicada y *x(m)* la nueva señal de audio con valor nulo fuera del intervalo $n \in [m-N+1, m]$ siendo N la duración en número de muestras de la ventana. Esta multiplicación en el dominio temporal (fruto del enventanado) se convierte en una convolución frecuencial, lo que supone la generación de réplicas de la señal desplazadas. La respuesta en frecuencia del filtro al que equivaldría el enventadado muestra un lóbulo principal y lóbulos secundarios residuales. En un caso ideal, el lóbulo principal sería rectangular y los secundarios nulos. No obstante, este tipo de filtro no es realizable, por lo que como resultado del enventanado, se introduce cierta distorsión que se visualiza bajo la presencia de lóbulos secundarios de diferente amplitud y anchura. Por este motivo, la respuesta en frecuencia de las distintas técnicas de enventanado presenta un compromiso entre el lóbulo principal (que introduce un suavizado de la señal) y los lóbulos secundarios (que introducen distorsión). En las señales de voz, la inserción de distorsión es más perjudicial que el efecto de suavizado, motivo por el cual se emplea típicamente la ventana de tipo *Hamming* de 20 ms de duración, la cual se caracteriza por tener lóbulos secundarios bajos.

La ventana tipo *Hamming*, que presenta una estructura de coseno alzado en sus valores no nulos, atiende a la siguiente ecuación:

$$\omega(n) = \begin{cases} 0{,}54 - 0{,}46 * \cos\frac{2\pi n}{N-1}, & 0 \leq n \leq N-1 \\ 0, & \text{en caso contrario} \end{cases} \tag{3.2}$$

Debido al efecto de suavizado que se acentúa en los extremos de la señal, generalmente se aplican ventanas solapadas al 50 % del tamaño de la muestra, tal y como se muestra en la figura 3.2.

Una vez realizado el enventanado sobre la señal de audio, se obtienen los segmentos de trabajo sobre los que se aplica la extracción de parámetros. A continuación se detallan algunas de las técnicas más usadas en detección de audio [9], así como las desarrolladas en este proyecto. Todas ellas obtienen los coeficientes del dominio espectral, por ser el que se ha comprobado que mejor representa las características acústicas de las señales de audio.

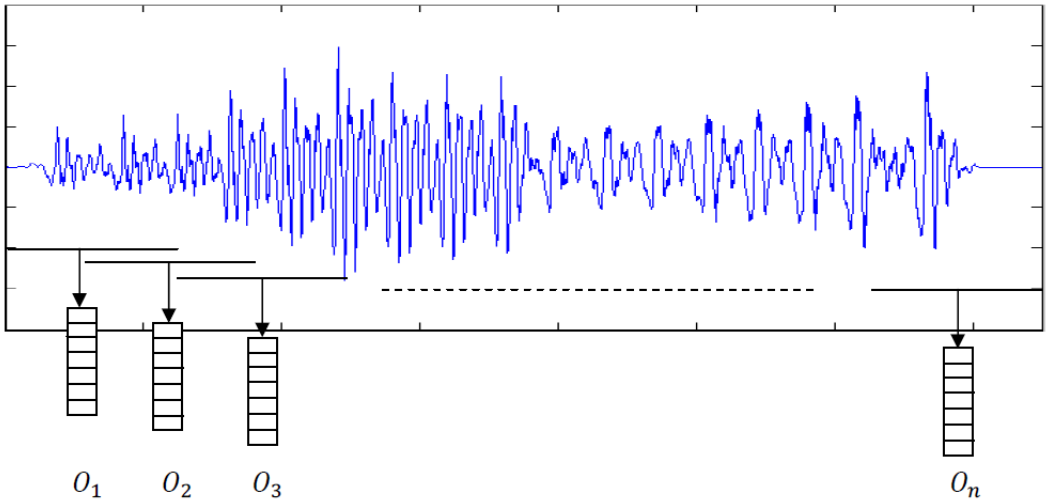


Figura 3.2: Aplicación de ventanas solapadas sobre la señal de audio. Fuente: [1].

3.2.1. Características tímbricas

El término de características tímbricas hace alusión a la característica que permite diferenciar dos sonidos que tienen la misma intensidad percibida y el mismo tono. La envolvente espectral mide la relación entre armónicos de una señal, y el nivel de estos armónicos es el principal responsable de el timbre, por lo que las características basadas en envolvente espectral (como es el caso de los MFCC) se denominan también tímbricas.

3.2.1.1. Coeficientes MFCC (*Mel-Frequency Cepstral Coefficients*)

Los coeficientes Mel-Frequency Cepstral Coefficients (MFCC) fueron introducidos por Davis y Mermelstein en 1980 [12] originalmente para ser empleados en tareas de reconocimiento de habla, y posteriormente adaptados para otras tareas relacionadas con el procesamiento de señales de audio, como el reconocimiento de locutor, de voz o la segmentación de audio [13] [14]. Anteriores a estos, ya se usaban coeficientes basados en el dominio *cepstral* como los LPCCs (*Linear Prediction Cepstral Coefficients* [15]).

Se trata de coeficientes que basándose en la percepción auditiva humana trabajan sobre las frecuencias de la escala *Mel* [1] en el dominio *cepstral*[2]. El proceso de obtención de dichos coeficientes se puede resumir en los siguiente pasos:

1. Se calcula la transformada de Fourier (DFT) para poder trabajar en el dominio frecuencial.
2. A continuación se le aplican a la señal una serie de filtros basados en la escala MEL, la cual mantiene la siguiente relación con la frecuencia objetiva de la señal (frecuencia lineal f_L):

$$f_{MEL} = 1127{,}01 * ln(1 + f_L/700) \tag{3.3}$$

[1] La escala MEL es una escala en el dominio frecuencial cuya distribución está relacionada con el mecanismo de percepción subjetiva.

[2] El dominio cepstral se define como la transformada inversa de Fourier del logaritmo del módulo espectral, entendiéndose dicho módulo como la multiplicación de la respuesta del tracto vocal por la excitación glotal [1].

El objetivo de este filtrado es el de aproximar la resolución espectral del oído humano (basada en el concepto de bandas críticas [3]) a las señales de audio.

3. A partir del banco de filtros generado, se calcula la energía por cada banda.

4. Finalmente se aplica la transformada discreta del coseno (DCT) al logaritmo de la energía en cada una de las bandas generadas, obteniéndose tantos coeficientes como bandas. No obstante, habitualmente se trabaja con un subconjunto de coeficientes, ya que la DCT comprime la energía en los primeros coeficientes de modo que los últimos suelen ser muy próximos a cero.

El proceso de obtención de características MFCC que incluye la tarea de enventanado se resume en la siguiente figura (3.3):

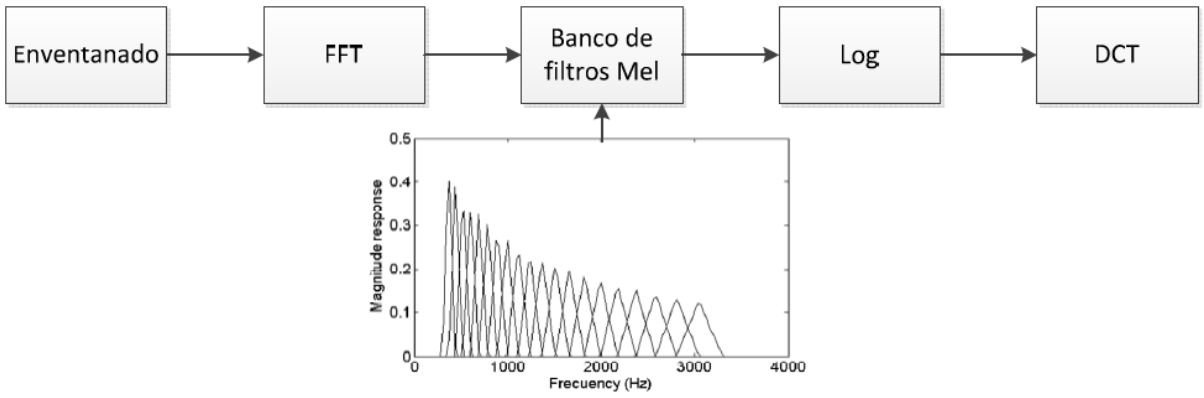


Figura 3.3: Proceso de extracción de característicass MFCC. Fuente: [2].

3.2.1.2. Parámetros dinámicos derivados de los MFCC

Para incorporar información dinámica a los coeficientes MFCC (que aportan información del contenido cepstral sólo de la ventana sobre la que se calculan) se introducen este tipo de coeficientes, que en el caso de los delta ($\triangle$) y delta-delta ($\triangle\triangle$) son, respectivamente, la diferenciación de primer y segundo orden en la ventana actual [16].Por otro lado, los coeficientes Shifted Delta Cepstra (SDC) vienen especificados por cuatros parámetros (figura 3.4), **N-d-P-k**, donde **N** es el número de coeficientes cepstrales en cada trama, **d** representa el desplazamiento en tiempo para el cálculo de deltas en número de tramas, **P** es el desplazamiento entre bloques consecutivos y **k** es el número de bloques que serán concatenadas para formar el vector final. Una configuración típica a la hora de emplear este tipo de parametrización es 7-1-3-7, la cual ha sido empleada en este proyecto para la parte correspondiente a características MFCC-SDC.

Estos parámetros dinámicos empezaron usándose para identificación de idioma [17] ya que capturan la dependencia temporal latente entre tramas consecutivas, y después se han aplicado en otros ámbitos como la segmentación de audio [11]. Una vez calculados, los coeficientes derivados se concatenan a los estáticos (MFCC).

[3]Este concepto surge del hecho de que el oído humano enmascara las frecuencias que no superan cierto nivel de amplitud o que se encuentran muy próximas entre sí, por lo que su sensibilidad frecuencial es limitada.

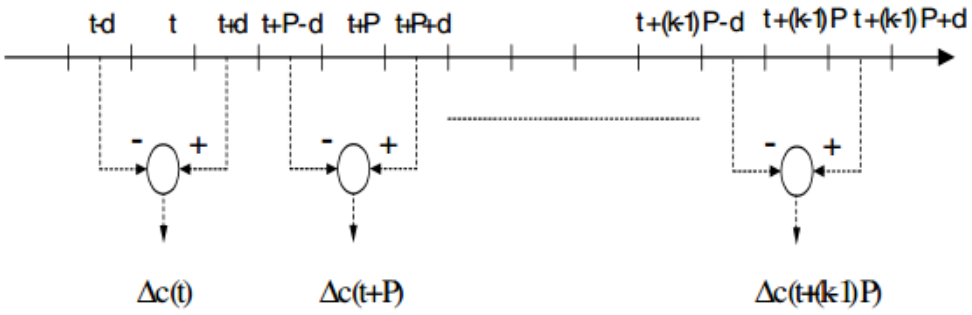


Figura 3.4: Computación del vector de características SDC en un tiempo t para los parámetros N-d-p-k. Fuente: [3].

3.2.1.3. Técnicas de compensación de canal

En cualquier sistema de comunicación, el canal empleado tiende a introducir efectos indeseados de la señal. No obstante, la perturbación que pueden introducir los canales de comunicación es variable y depende de ciertos factores; es por esto que un sistema de segmentación de audio robusto deberá actuar con la misma eficacia independientemente del canal que ha sido empleado para transmitir la señal. En este contexto surgen las técnicas de compensación de canal, las cuales permiten que un determinado segmento de una clase A (por ejemplo, voz) no se confunda con otro de clase B (por ejemplo, ruido) por efecto de la perturbación del canal.

Con el objeto de poder reducir la influencia de estas perturbaciones sobre la señal de audio se han desarrollado diferentes técnicas de compensación, algunas de las cuales se detallan a continuación:

- **Normalización por media cepstral** (CMN o *Cepstral Mean Normalization*): substrae el valor medio de la señal con objeto de eliminar el ruido aditivo introducido por el canal [18] [19].
- **RASTA Filtering** : este tipo de filtrado se aplica idealmente a sistemas de reconocimiento de voz o de locutor por los objetivos que persigue. Se trata de un filtrado que busca eliminar por un lado, los cambios lentos observados en la señal transmitida (producidos por el canal) y por el otro, los cambios rápidos de la señal que raramente habrían sido producidos por un hablante [20].
- **Feature warping**: este tipo de técnica ecualiza la distribución de coeficientes para ajustarla a una normal, compensando así los efectos de canal. Resultan por ello muy adecuados para sistemas que modelan la distribución de características mediante componentes gaussianas, como es el caso de los GMM's. No obstante, hay que tener en cuenta que estas técnicas también arrastran el inconveniente de perder carácter discriminativo en el sistema al dar un carácter homogéneo a los datos [21].

3.2.2. Características cromáticas

Quizás, la palabra croma sea más comúnmente empleada por una de sus dos acepciones del castellano, esto es, relativo al color. El otro contexto en el que se puede definir esta palabra es el de la música, en el cual la palabra **croma** se encuentra ligada a la escala musical y la transición

entre semitonos o tonos (un tono está compuesto por dos semitonos). Cada una de las notas que componen la escala musical posee nombre propio [4] y tiene asociada una frecuencia. La mínima transición frecuencial que existe entre dos notas se denomina semitono, y cada agrupación de doce semitonos consecutivos recibe el nombre de **escala cromática**. A modo de ejemplo, en la figura 3.5 se muestra una escala cromática representada en un pentagrama en orden ascendente y en orden descendente junto con sus valores en frecuencia [5].

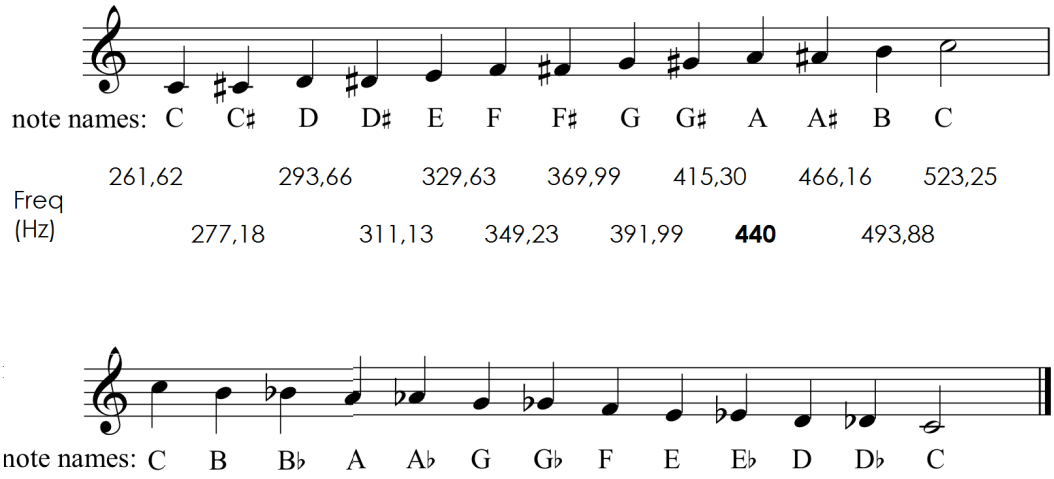


Figura 3.5: Escala cromática en orden ascendente y descendente. Fuente: *www.musiccrashcourses.com*.

Desde el punto de vista de teoría de la señal, se podría decir que la música es una composición de frecuencias agrupadas con un orden lógico (sección 2.2.2). Si denominamos **características cromáticas** a aquellas que aportan información musical relativa a este orden lógico de la señal de audio, cobra sentido realizar un estudio de las mismas en búsqueda de patrones que sirvan para desarrollar un sistema de segmentación de audio.

Como se citó en el apartado 3.2, el dominio frecuencial parece el más adecuado para representar la señal de voz y de audio en general (lo que también incluye el género musical). Las técnicas más clásicas están basadas en envolvente espectral (como es el caso de los MFCC) y también se las denomina características **tímbricas** [22], ya que el timbre está íntimamente ligado con la envolvente espectral. No obstante, las señales ofrecen información muy variada no sólo relativa al timbre, sino también relativa a la intensidad, la armonía, el rango de frecuencias predominantes, etc., lo que ha motivado recientemente [22] el análisis de otro tipo de características musicales. En este contexto aparecen las denominadas características **cromáticas** o ligadas a la escala musical. En este apartado se van a detallar dos algoritmos de obtención de características cromáticas: el primero de ellos está basado en estadísticos de la **entropía cromática** [23], mientras que el segundo estudia nueva información cromática de la música relacionada con la distribución de energía a lo largo de la escala cromática.

[4] En la notación al castellano las notas que componen la escala musical son: do-re-mi-fa-sol-la-si, mientras que en el sistema inglés se denotan con las letras del alfabeto, empezando a nombrar por la nota *la*:A-B-C-D-E-F-G-H.

[5] Las transiciones entre las 7 notas de la escala musical se representan con sostenidos y bemoles. A excepción de la transición de mi a fa, y de si a do, la cual es ya de un semitono, dos notas consecutivas distan un tono entre sí, lo que genera un total de 12 semitonos en cada escala cromática.

3.2.2.1. Estadísticos de la entropía cromática

Relativo al área de la teoría de la información [24], la entropía mide el grado de información que aporta una muestra, lo que así mismo se podría expresar como la medida del grado de incertidumbre o imprevisibilidad de una fuente de información.

Calculando la entropía sobre el módulo normalizado de la FFT se obtiene la entropía espectral, de modo que al dividir el espectro en L sub-bandas se pueda calcular la entropía espectral relativa a la distribución de energía en las subbandas. Concretamente, si la relación frecuencial entre estas bandas guarda relación con la relación frecuencial existente entre dos notas consecutivas de la escala cromática, entonces se habla de entropía cromática [25]. Para obtener estas bandas cromáticas, en primer lugar se realiza una asignación del espectro de potencia en la escala de frecuencias de Mel y se divide en doce sub-bandas, coincidiendo la frecuencia central f_k de cada banda con cada uno de los doce semitonos de la escala musical. De este modo, a partir de una frecuencia central fija (la más baja a evaluar) se obtienen las frecuencias centrales de las L sub-bandas restantes:

$$f_k = 1127{,}01 * ln(1 + f_0 * r^k/700), \quad k = 0, 1, 2, ..., L-1 \tag{3.4}$$

donde L es el número de subbandas que se generan desde f_0 hasta la mitad de la frecuencia de muestreo, y f_k es la frecuencia central obtenida para cada una de las L subbandas. Por otro lado r toma el valor de $\sqrt[12]{2}$ y es la razón que existe entre dos semitonos consecutivos, por lo que a partir de una frecuencia de referencia es posible obtener la frecuencia de cualquier otra nota si se conoce su posición relativa en la escala. Finalmente se obtiene la energía de cada subbanda y se normaliza por el total de energía.

El método más sencillo de trabajar con la entropía cromática es obteniendo la media estadística de la energía de varias muestras consecutivas. Trabajando con parámetros estadísticos de esta entropía cromática sobre muestras consecutivas, la media estadística resulta el parámetro más sencillo de obtener (orden uno). No obstante, obtener una sóla característica de cada conjunto de muestras puede no resultar un dato suficientemente representativo de la misma. Por ello, en trabajos previos de análisis de la entropía cromática aplicada a segmentación de audio [23] se han propuesto otras tres medidas estadísticas que proporcionan en conjunto una información más completa sobre la forma de la distribución de la entropía cromática: la varianza, el *skewness* y la *kurtosis*.

- **Media muestral *(mean)***: también denominada estadístico de primer orden, es el valor promedio aplicado a un conjunto N de muestras de una variable aleatoria $X_1, X_2, X_3, ..., X_N$, y se define como:

$$\overline{X} = \frac{X_1 + X_2 + X_3 + ... + X_N}{N} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{3.5}$$

- **Varianza muestral*(variance)***: también denominada estadístico de segundo orden, se define como la esperanza E del cuadrado de la desviación del conjunto de datos respecto de su media μ:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2 \tag{3.6}$$

- ***Skewness* muestral**: el *skewness* o estadístico de tercer orden constituye el grado de asimetría de una distribución. Si una distribución gausiana presenta un extremo más prolongado hacia la derecha que hacia la izquierda se dice que el *skewness* es positivo y

viceversa. Dicha medida se calcula a partir de los momentos de orden uno y de orden dos tal y como se muestra a continuación:

$$b_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^3}{(\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2)^{3/2}} \tag{3.7}$$

- ***Kurtosis* muestral**: este estadístico de cuarto orden se encarga de medir la forma de la distribución, en cuanto a la proporción de datos concentrados en torno a la media o en torno al valor de la varianza. Análogamente al *skewness* se obtiene a partir de los momento de orden uno y de orden dos:

$$g_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^4}{(\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2)^2} - 3 \tag{3.8}$$

3.2.2.2. Nueva información cromática

Como se describió en el apartado 2.2.2 el concepto de armonía se encuentra ligado al nivel de agrado que experimenta una persona al escuchar un fragmento de audio. Generalmente se puede decir que los sonidos tienden a resultar agradables al oído si están afinados, es decir, han sido producidos de acuerdo a unas normas de afinación. Una pieza musical tocada por instrumentos desafinados o cantada por una voz que no entona resultará desagradable al oído.

La voz hablada es melódica (en cuanto a que para un instante de tiempo t los seres humanos no son capaces de producir dos notas simultáneas), por lo que producida de manera individual no guarda relación directa con la armonía (sobre un mismo tiempo t). Esta relación armónica cobra sentido cuando dos cantantes cantan un dueto o un coro canta una pieza musical a diferentes voces. En este contexto, las características armónicas que presentan las clases de música, ruido y voz pueden ser contrastadas. En este proyecto se van a estudiar dos nuevos modos de extracción de características cromáticas inspirados en la disposición frecuencial (y en algún caso armónica) de las clases acústicas: agrupando la energía en las doce subbandas de la escala cromática, o creando filtros de energía que abarcan el rango frecuencial de una octava.

3.2.2.2.1. Energía por subbandas (o cromagrama)

Cada nota (o semitono) de la escala musical tiene asociada una frecuencia, y conociendo la posición relativa de las notas entre sí es posible calcular la frecuencia de cada una de éstas tan sólo conociendo la frecuencia de una de éstas notas y la posición relativa con respecto a cada una de las otras (número de semitonos que distan entre sí). Concretamente, la relación frecuencial que existe entre dos notas separadas por una octava (es decir, doce semitonos) responde a una relación exponencial de base dos, de tipo 2^k, siendo k el número de octavas enteras que hay entre ambas notas. A continuación se muestra un ejemplo de las frecuencias de la nota *La* en un rango de 7 octavas calculadas a partir de la frecuencia central de dicha nota, es decir, 440 Hz:

$$frecuencia_{LA}(Hz) = 440 \times 2^k, \quad k = [-4, -3, -2, -1, 0, 1, 2] \tag{3.9}$$

Por otro lado, lo que provoca que dos instrumentos que tocan la misma nota suenen diferentes se relaciona con el nivel y la cantidad de armónicos que dicha nota genera en el dominio frecuencial (timbre). Estos armónicos son (por lo general) señales de menor intensidad que se generan a partir de la frecuencia fundamental. Por ejemplo, los armónicos de un *La* en la octava central, es decir, a 440Hz, se podrían encontrar en cualquiera de los submúltiplos y múltiplos de dicha frecuencia, como por ejemplo: 220Hz, 110Hz, 880Hz, 1760Hz, etc.

En la agrupación en primer lugar, se suman todas las energías de cada nota para cada instante de tiempo, a partir de una escala cromática cualquiera (tomada como referencia), sobre un rango de frecuencias determinado. Dicho rango de frecuencias límite vendrá fijado por las limitaciones del oído humano, y en el caso de tratar con señales digitales por la frecuencia de muestreo. Para este proyecto se ha propuesto un esquema de trabajo de 10 octavas donde cada octava comprende desde la frecuencia de *Do* al *Si* inmediatamente superior en sentido ascendente, que abarca hasta los 8KHz aproximadamente (desde los 8.17 Hz hasta los 7.9 KHz). En el anexo A se detallan las frecuencias empleadas.

La frecuencia de cada una de estas notas será considerada la frecuencia central de cada uno de los filtros que permiten obtener la energía por banda frecuencial. Una vez que se tienen localizadas las frecuencias centrales de todos los filtros, se calcula la energía de la señal por tramas producida en cada frecuencia y se asigna sobre las diferentes bandas frecuenciales de trabajo o filtros. En este proyecto se trabaja con un total de diez octavas, lo que se corresponde con $10 * 12 = 120$ filtros de energía. Estos filtros serán agrupados en 12 subbandas del siguiente modo.

$$E_k = \sum_{i=1}^{L} e_{k+12(i-1)}, \quad k = 1, ..., 12 \qquad e_j =< e_1, e_2, e_3, ..., e_{12L} > \tag{3.10}$$

siendo L el número de octavas de trabajo, E_k la energía total para la banda k de cada una de las 12 subbandas, e_j la energía obtenida sobre cada uno de los filtros, y k cada una de las 12 subbandas de trabajo.

k	1	2	3	4	5	6	7	8	9	10	11	12
notación española	Do	Do#	Re	Re#	Mi	Fa	Fa#	Sol	Sol#	La	La#	Si
notación inglesa	C	C#	D	D#	E	F	F#	G	G#	A	A#	B

Tabla 3.1: Correspondencia entre número de subbanda y nota musical

3.2.2.2.2. Energía por octavas

Como se ha mencionado con anterioridad, el rango de frecuencias que presenta la música es más amplio que el de la voz hablada. En el caso del ruido la definición de este rango es más compleja, ya que el rango de frecuencias que abarca y su distribución dependen de la naturaleza del ruido. Por ejemplo, en caso de tratar con ruido blanco, el espectro se mantiene constante en todo el rango, lo que implica que abarca todas las frecuencias. Otro tipo de ruido medible es el ruido rosa, cuyo nivel sonoro es inversamente proporcional a la frecuencia, es decir, que su amplitud decae según aumenta la frecuencia. Es decir, que este ruido decrece 3 dB por octava, intervalo en el cual el ancho de banda se duplica. Es por ello que este tipo de ruido se utiliza para analizar el comportamiento del sonido en salas, altavoces y equipos de sonido.

Tomando el mismo patrón de subbandas del apartado anterior, se puede obtener y analizar la cantidad de energía de una señal de audio presente en cada octava. A partir de la idea de que los rangos de frecuencia para cada clase puedan variar, se propone un esquema de agrupación de energía por frecuencias consecutivas (a diferencia de la agrupación anterior por múltiplos frecuenciales). Dado que se está trabajando sobre la idea de armonía de la música y su relación con la escala cromática, esta agrupación se propone a nivel de octavas, esto es, diseñar diferentes filtros que abarquen cada uno de ellos un rango de frecuencias equivalente al de una escala cromática. Siguiendo los mismos criterios del apartado anterior, el límite de frecuencias se establece en 10 octavas (abarcando hasta los 8 KHz aproximadamente).

Si bien la agrupación por subbandas viene motivada por la búsqueda de patrones de armonía sobre la escala cromática, este segundo conjunto se inspira en la capacidad para abarcar frecuencias de cada una de las clases acústicas. A partir de la energía obtenida para cada una de las 12 subbandas de cada octava de trabajo, se agrupa la energía del siguiente modo:

$$E_l = \sum_{i=1}^{12} e_{i+12(l-1)}, \quad l = 1, ..., 10$$
$$e_j =< e_1, e_2, e_3, ..., e_{12L} > \tag{3.11}$$

siendo L el número total de octavas, l cada una de las octavas y e_j el vector que contiene la energía de cada uno de los filtros.

La representación gráfica que resume la estrategia de agrupación seguida en cada uno de los casos se muestra en las figuras 3.6 y 3.7.

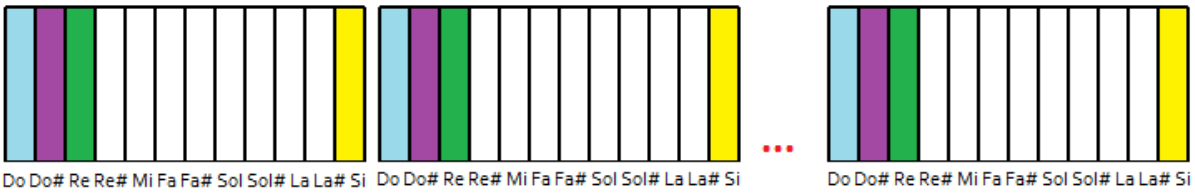


Figura 3.6: Agrupación de la energía por subbandas.

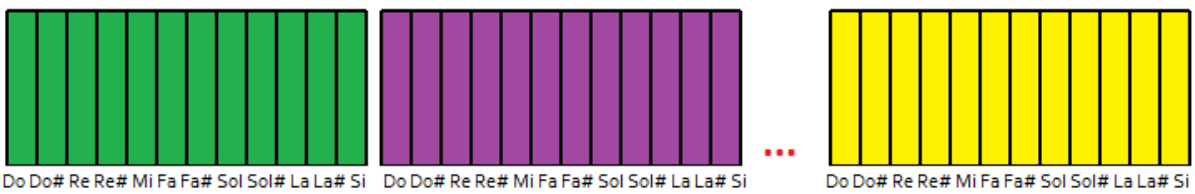


Figura 3.7: Agrupación de la energía por octavas.

3.3. Sistemas de segmentación de audio

Una vez que los vectores de características han sido extraídos sobre la señal de audio (siguiendo alguna de las técnicas expuestas en la sección anterior u otras), los datos deben ser clasificados en base a su contenido. Dicha clasificación en función de la naturaleza del audio

da lugar a los modelos de clases o patrones de referencia, los cuales servirán de referencia para clasificar los contenidos de nuevos ficheros de audio que necesitan ser segmentados y etiquetados. En la literatura científica pueden encontrarse principalmente dos aproximaciones principales al problema de la segmentación de audio [4]: por una parte, la segmentación basada en distancia, en la cual los sistemas utilizan medidas de distancia entre conjuntos de muestras para determinar la presencia de un punto de cambio entre clases acústicas y posteriormente estos segmentos se clasifican; y por otro, la segmentación basada en modelos, en la cual se segmenta y clasifica en un mismo paso mediante la comparación trama a trama frente a los modelos de las clases contempladas.

3.3.1. Segmentación basada en distancia

Los sistemas de segmentación de audio basados en distancia, trabajan con las tramas o vectores de características obtenidos en la fase de extracción analizando su contenido en busca de similitudes que permitan agrupar estas tramas en segmentos, lo que expresado con otras palabras viene a decir que buscan puntos de cambio entre las tramas que les permita identificar cambios de clases acústicas. Para ello, se mide la distancia o disimilitud estadística entre dos bloques de tramas consecutivas, y si ésta supera un umbral determinado se considera que hay un cambio de clase acústica en ese instante temporal.

3.3.1.1. Detección de puntos de cambio

Las diversas técnicas de detección se resumen en algoritmos basados en distancia que realizan un análisis sobre un conjunto de datos en búsqueda del punto de cambio que les permita separar segmentos [4]. La medida para obtener esta distancia se obtiene de una función matemática que va tomando medidas sobre los vectores de características del audio en bloques de tramas. Mediante una ventana deslizante se van seleccionando los bloques de tramas de trabajo, si bien el modo de desplazar esta ventana y el tamaño de estos bloques de tramas depende concretamente de cada una de las técnicas de detección de puntos de cambio a emplear.

Con el conjunto de estas medidas tomadas, se crea una curva que expresa las diferencias encontradas entre los distintos bloques analizados, de tal modo que los máximos de dicha curva corresponden con los candidatos a punto de cambio. En la literatura, se pueden encontrar muy diversas técnicas de detección de puntos de cambio tales como la Distancia Euclídea [26], *The Bayesian information criterion* (BIC) [26] [27], *The Kullback Leibler KL2 distance* [27] o The *Generalized Likelihood Ratio* [26] [27]. En la figura 3.8 se presenta un esquema general de detección de puntos de cambio.

3.3.1.2. Etapa de clasificación

Una vez determinados los puntos de cambio, los cuales definen distintos segmentos (que en teoría contienen una única clase acústica), éstos se clasifican entre las clases consideradas (pudiendo unirse segmentos consecutivos que inicialmente se consideraban clases acústicas diferentes), contrastando cada uno de estos segmentos con los modelos de clases generados. Para ésto puede usarse cualquier técnica de clasificación (como SVMs, GMM-UBM, i-vector [28] , etc) algunas de las cuales se detallarán en la siguiente sección.

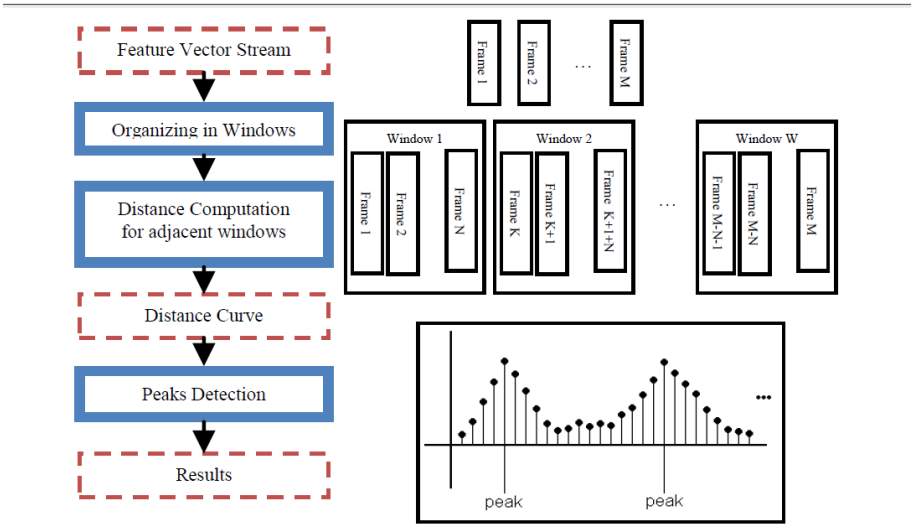


Figura 3.8: Esquema de detección de puntos de cambio. Fuente:[4].

3.3.2. Segmentación basada en modelos

Otra de las formas más comunes de enfocar un sistema de segmentación de audio consiste en clasificar directamente cada una de las tramas obtenidas por un extractor de características, asignando la información obtenida en cada trama a la clase correspondiente. Tanto en el sistema de segmentación basada en modelos, como en el sistema de segmentación basada en distancia, el conjunto de técnicas que se emplean para la clasificación puede coincidir, ya que la diferencia principal entre ambos de cara a la clasificación reside en la longitud de los segmentos (en el primer caso de duración variable, y en este segundo de duración fija y limitada). Entre las técnicas más empleadas destacan las basadas en Modelos de Mezclas de Gaussianas, las cuales pueden usarse tanto trama a trama como para clasificar un segmento homogéneo (como es el caso de la segmentación basada en distancia). Adicionalmente, cabe destacar los Modelos Ocultos de Markov HMMs (*Hidden Markov Models*) como una técnica de decodificación, la cual dada un flujo contínuo de datos, asigna cada trama a su clase correspondiente.

3.3.2.1. Modelos Ocultos de Markov (HMMs)

Los modelos ocultos de Markov son modelos utilizados originalmente en reconocimiento de voz, que también se aplican en algunos sistemas de segmentación de audio [28]. Estos modelos, los cuales modelan estadísticamente la acústica de la voz, sustituyeron en la década de los 80 a otras técnicas de modelado determinista como es el caso de los *Dynamic Time Warping* (DTW). El modelo oculto de Markov es un modelo de Markov de primer orden y discreto en el que las salidas observables se correponden de forma probabilística con los estados del sistema. El orden del sistema determina la dependencia con las últimas n salidas del mismo, por lo que un modelo de primer orden depende únicamente de la salida anterior. Los elementos que caracterizan un HMM son los siguientes [29]:

1. El número de estados en el modelo, que se identifica con la letra N. En un modelo ergódigo todos los estados se encuentran interconectados entre sí, pero también hay combinaciones que sólo permiten transiciones hacia adelante (como el caso de HMM's de Bakis). Cada

estado se denomina como S_j, siendo q_t el instante de tiempo t en el que se encuentra dicho estado.

2. El número de símbolos que se pueden observar en cada estado (M), que se corresponde con las diferentes salidas del sistema que está siendo modelado. Cada uno de los símbolos se denota como V_j , y la observación de un símbolo en un instante de tiempo dado O_t.

3. La matriz de probabilidades de transición entre estados $A = \{a_{ij}\}$, que resume la posibilidad de que estando en un estado S_i en un instante de tiempo t el sistema se mueva a otro estado S_j en el instante *t+1*. Este comportamiento se resumen en la siguiente notación:

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i], \quad 1 \leq i, j \leq N \tag{3.12}$$

4. A continuación se determina la distribución de probabilidad $B = \{b_j(k)\}$, de observar un símbolo en cada estado S_j, esto es:

$$b_j(k) = P[v_k(t) | q_t = S_j], \quad 1 \leq j \leq N, \quad 1 \leq k \leq M \tag{3.13}$$

5. Por último, se define la probabilidad inicial de ocupación de cada estado $\pi = \{\pi_i\}$ como:

$$\pi_i = P[q_1 = S_i], \quad 1 \leq i \leq N \tag{3.14}$$

Por lo tanto, el sistema de modelos ocultos de Markov queda determinado por los valores de M, N, y las tres medidas de probabilidad A, B y π. Por conveniencia, estos parámetros quedan resumidos en la siguiente expresión: $\lambda = (A, B, \pi)$. En la literatura, se encuentran ejemplos de uso de HMMs para sistemas de segmentación de audio, como en [11], en cuyo caso cada clase acústica se modela mediante un GMM que constituye uno de los estados del HMM.

3.3.2.2. Modelos de mezclas de Gaussianas (GMMs)

A diferencia de otros sistemas como los HMMs la técnica básica de Modelos de Mezclas de Gausianas (GMM) resulta un algoritmo más intuitivo y sencillo de aplicar. Los sistemas basados en GMMs [5] son modelos estadísticos empleados en tratamiento de audio que ofrecen resultados claramente notorios en el área de las tecnologías del habla. En concreto, en el área de reconocimiento de locutor independiente de texto han resultado la aproximación más exitosa [5] durante bastantes años. Sin embargo, no solamente se emplean para tareas de reconocimiento de locutor sino que también constituyen una de las técnicas de referencia en las tareas de procesado de voz y audio [22]. Los GMMs definen la función densidad de probabilidad de las observaciones x dado un modelo (λ), determinado por una serie de parámetros (pesos, medias y covarianzas). Los modelos de un GMM se componen de un número finito de gaussianas multivariadas, cada una de las cuales queda definida por su media y su varianza. En resumen, los GMMs definen una función de densidad de probabilidad la cual queda expresada del siguiente modo:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x) \tag{3.15}$$

donde M representa el número de gaussianas, w_i los pesos de cada una de éstas (bajo la restricción de que $\sum_{i=1}^{M} w_i = 1$) y p_i es la función densidad de probabilidad gaussiana d-variada (d-dimensiones, tantas como número de características componen un vector):

$$p_i(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} e^{-(1/2)(x-\mu_i)^T \Sigma_i (x-\mu_i)^{-1}} \quad i = 1, ..., M \tag{3.16}$$

En este caso μ_i es el vector de medias de la gaussiana i-ésima y Σ_i la matriz de covarianza de la misma. Normalmente estas matrices de covarianza se definen diagonales, lo que implica trabajar con vectores en lugar de matrices reduciendo así el coste computacional y la cantidad de datos necesarios para su estimación robusta. Resultados encontrados en la literatura [3] han demostrado que el uso de matrices diagonales produce resultados casi idénticos a si se usase la matriz completa, motivo por el cual generalmente se trabaja con esta simplificación.

El número de gaussianas que forman cada modelo mantiene por lo general cierta relación con la cantidad de datos de entrenamiento y con cómo se distribuye la densidad de probabilidad en éstos. Dado que el conjunto de gaussianas del modelo tiene como objetivo adaptarse lo mejor posible a los datos de entrada, el número de gaussianas (mezclas) empleadas para generar cada modelo será decisivo para un ajuste preciso. A mayor número de mezclas mayor será la precisión del ajuste, pero también mayor es el coste computacional además de que un número alto de mezclas también puede ocasionar sobreajuste a los datos de entrenamiento, por lo que la elección de este parámetro presenta un compromiso coste-precisión que debe ser ajustado experimentalmente. La figura 3.9 muestra un ejemplo de función densidad de probabilidad de un GMM generado con 128 mezclas sobre un modelo bi-dimensional (dos características):

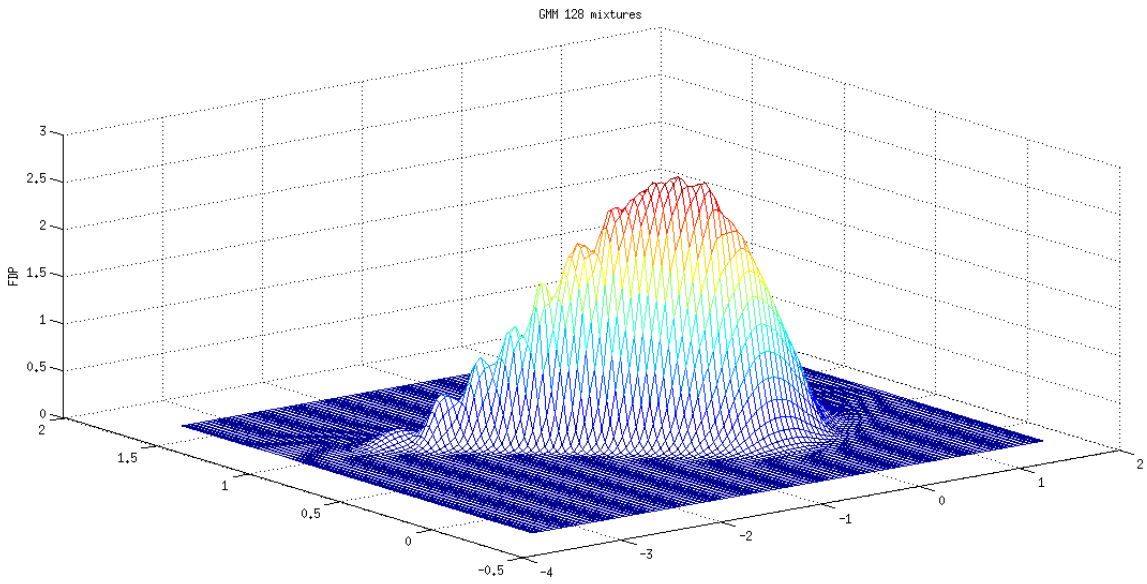


Figura 3.9: GMM de 128 gaussianas.

La aproximación más utilizada para calcular los parámetros del modelo es la estimación de máxima probabilidad (*maximum likelihood* o ML), la cual busca modificar los parámetros del modelo (pesos, medias y covarianzas) de modo que aumente la probabilidad de los datos de entrenamiento dado el modelo. Este proceso se lleva a cabo mediante operaciones iterativas del algoritmo Expectation Maximization (EM) [30],de modo que el valor de similitud entre el modelo y los datos en la iteración *k+1* sea mayor que en la iteración *k-ésima*, es decir: $p(X|\lambda(k+1)) > p(X|\lambda(k))$. A partir de un modelo inicial λ se realizan modificaciones sobre el mismo hasta que el valor de verosimilitud converge o se alcanza un número de iteraciones previamente establecido.

El modelo se puede inicializar a partir del algoritmo *k-means* [31]. Este método de agrupa-

miento particiona el total de datos en k regiones con sus respectivos k centroides [6], reduciendo el número de iteraciones necesarias para la convergencia del modelo. De este modo, los k centroides dan valor a cada uno de los vectores de medias del GMM, las covarianzas del modelo se obtienen del conjunto de vectores asignados a cada centroide y el porcentaje de vectores que son asignados a cada centroide se corresponde con el vector de pesos del GMM. Este proceso de entrenar cada una de las clases directamente a partir de los datos de entrenamiento (en contraposición a la adaptación MAP a partir del UBM) se conoce como GMM-ML, de maximum likelihood). Una vez generados los modelos de cada una de las clases acústicas a partir de los datos de entrenamiento, se calcula (mediante la ecuación 3.15) la probabilidad a posteriori de un fragmento de audio dado un modelo.

3.3.2.3. GMM-UBM

A la hora de entrenar GMMs se pueden combinar diferentes técnicas que optimicen el rendimiento del sistema en función de los requisitos del mismo y la base de datos disponible. Por ejemplo, en el caso de trabajar con escasez de datos de entrenamiento resulta de gran utilidad adaptar el GMM de cada clase a partir de un modelo universal común a todas las clase (obtenido a partir de datos de entrenamiento) conocido como UBM (*universal background model*). Las ventajas de emplear este método de entrenamiento combinado no sólo radican en compensar la escasez de datos, sino que además proporciona mecanismos para normalizar los resultados obtenidos en función de lo representativas que sean las características contrastadas con el modelo universal generado. Cuando surge el problema de entrenar modelos con escasez de datos, la técnica GMM-UBM resulta una propuesta efectiva de sistema que mejora la precisión de un sistema GMM-ML. Para trabajar sobre esta técnica, por un lado se obtienen los parámetros del modelo UBM (al igual que en el caso del GMM) a partir del algoritmo de *expectation maximitation* el cual estará formado por n mezclas. De este modo se obtienen las características comunes a todas las clases de audio, lo que ofrece un conocimiento a priori muy útil en la fase de modelado por clases. Por otra parte, cada modelo de clases particular se obtiene de adaptar el modelo del UBM generado con los vectores de características de cada una de las clases. En la fase de segmentación de audio, una vez que los modelos han sido generados, las puntuaciones de verosimilitud de los vectores de características de test van a ser contrastadas no solo con cada modelo de clases sino también con el modelo universal, lo que permite normalizar los resultados. De este modo, una puntuación alta no sólo se obtiene si el vector de características se asemeja en gran medida a uno de los modelos de clases, sino que además es preciso que dicho vector sea muy poco semejante al modelo universal. En la práctica, la puntuación final de un segmento de datos sobre cada modelo se obtiene de normalizar la puntuación obtenida sobre el modelo universal de una clase respecto al UBM del siguiente modo (lo que en escala logarítmica equivale a restar ambos scores en lugar de ser divididos):

$$score\ clase_i = \frac{p(x|\lambda_{GMM_i}}{p(x|\lambda_{GMM_i}} \tag{3.17}$$

donde i representa cada una de las clases de audio.

Cada uno de los modelos del sistema GMM-UBM se obtienen de adaptar los parámetros del modelo universal con los datos pertenecientes a cada una de las clases. Uno de los modos de adaptación típicamente usados es la adaptación MAP, la cual permite adaptar todos los

[6] El centroide en un espacio d-dimensional formado por n muestras se define como la intersección entre los diferentes hiperplanos que dividen al conjunto de muestra, lo que equivale al vector más representativo de dichas muestras en un espacio d-dimensional.

parámetros del UBM, esto es, vectores de medias μ, matrices de covarianza Σ y pesos w a pesar de que en algunos casos la adaptación por medias es suficiente y reduce en gran medida el coste computacional del sistema. En el caso del reconocimiento de locutor se ha demostrado [5] que la adaptación de medias es suficiente para generar buenos resultados, ya que teniendo muy pocos datos de entrenamiento de un locutor para reestimar de forma robusta todos los parámetros, adaptar sólo medias permite tener diferencias entre locutores sin desajustar el resto de parámetros. Para la tarea de segmentación de audio no existen documentos que afirmen la bondad de un sistema de adaptación de medias frente a la adaptación completa, por lo que en este proyecto se han contrastado experimentalmente ambas posibilidades, resultando satisfactoria la adaptación de sólo medias al igual que en el caso del locutor.

En la adaptación MAP el UBM ($\lambda = \{w_k, \mu_k, \Sigma_k\}_{k=1}^{K}$) es utilizado como un modelo de partida, el cual se adapta al conjunto de datos de entrenamiento $X = \{x_1, x_2, ..., x_T\}$, dando lugar a los nuevos modelos del GMM-UBM. Las nuevos valores que definirán cada gaussiana se obtienen del siguiente modo:

$$w_i' = [\frac{\alpha_i n_i}{N_x} + (1 - \alpha_i) w_i]\eta \tag{3.18}$$

$$\mu_i' = \alpha_i E_i(x) + (1 - \alpha_i)\mu_i \tag{3.19}$$

$$\Sigma_i' = \alpha_i E_i(x^2) + (1 - \alpha_i)(\Sigma_i + \mu_i^2) - \mu_i^2 \tag{3.20}$$

Los parámetros que construyen las fórmulas recientemente expuestas se explican a continuación:

- w_i' representa el nuevo peso de la gaussiana *i-ésima*, μ_i' el nuevo vector de medias y Σ_i' la matriz de covarianzas adaptada.
- $\alpha_i = \frac{\eta_i}{\eta_i + r}$ es el cociente de adaptación que controla la influencia de los datos de entrenamiento de la clase a adaptar con respecto del UBM. Se regula a partir del factor de relevancia r en función de la cantidad de audio disponible para el entrenamiento de dicha clase. Cuanto mayor sea el factor de relevancia, mayor será la influencia de los datos de entrenamiento sobre el modelo adaptado y viceversa. Un factor de relevancia alto es deseable cuando se dispone de una gran cantidad de datos de entrenamiento, y uno bajo en el caso contrario. Cuando los datos sobre una clase son muy escasos, es interesante que el modelo de dicha clase se asemeje en mayor medida al modelo universal ya que una adaptación estricta podría resultar en sobreajuste de los datos sobre dicha clase.
- η es el factor de normalización para el vector de pesos de cada clase.
- Finalmente, los estadísticos de orden cero (n_i), uno ($E_i(x)$) y dos ($E_i(x^2)$) se obtienen a partir de la *probabilidad de ocupación gaussiana* ($\gamma_i(t)$), que determina la probabilidad de observación de x dada la gaussiana i del siguiente modo:

$$\gamma_i(t) = \frac{w_i p_i x_t}{\Sigma_{j=1}^{M} w_j p_j x_t} \tag{3.21}$$

$$n_i = \Sigma_{t=1}^{N} x \gamma_i(t), \quad i = 1, ..., M \tag{3.22}$$

$$E_i(x) = \frac{1}{n_i} \Sigma_{t=1}^{N} x \gamma_i(t) x_t, \quad i = 1, ..., M \tag{3.23}$$

$$E_i(x^2) = \frac{1}{n_i}\Sigma_{t=1}^{N} x\gamma_i(t)x_t^2, \quad i = 1, ..., M \tag{3.24}$$

La figura (3.10) resume de manera gráfica el proceso de adaptación MAP para un modelo de 4 mezclas bi-dimensional:

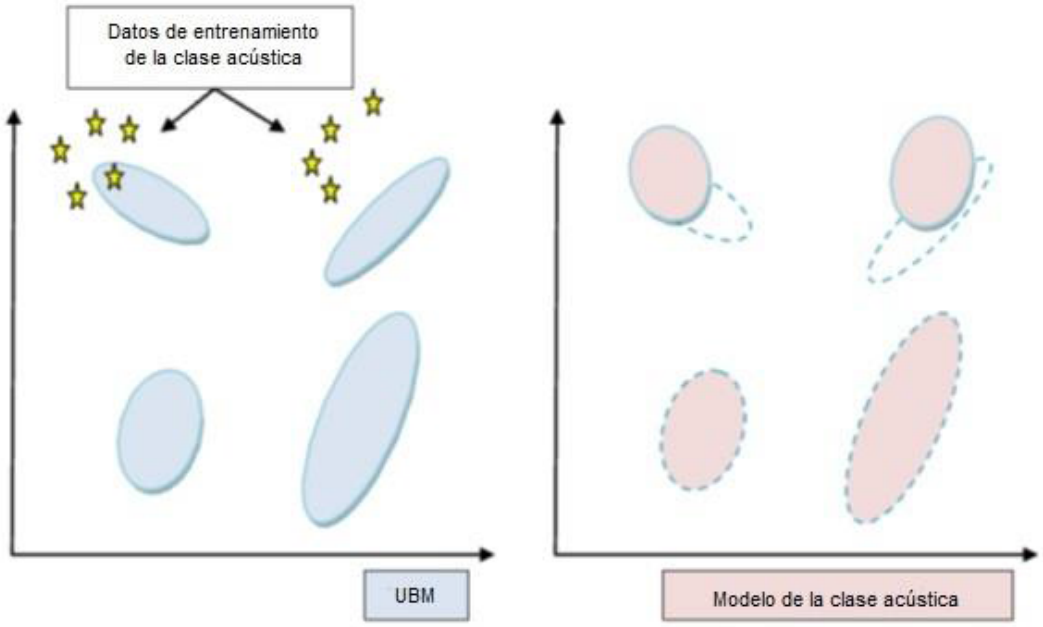


Figura 3.10: Proceso de adaptación MAP con 4 gaussianas. Fuente: [5].

En la gráfica de la izquierda, los puntos estrellados representan la posición de los vectores de características de entrenamiento de una clase concreta, y los óvalos azules cada una de las 4 gaussianas del UBM. En este caso se puede ver cómo los nuevos datos se encuentran próximos a las dos gaussianas superiores del UBM, por lo que los vectores de dicha clase guardan mayor relación de similitud con ellas. Esto implica que dichas gaussianas sufrirán cambios en el modelo de la clase acústica, es decir, que sus parámetros se adaptarán a los datos de entrenamiento (gráfica de la derecha), mientras que las dos gaussianas de abajo permanecen iguales a las del UBM. El proceso de adaptación puede constar de varias iteraciones MAP, aunque en algunos casos una sóla iteración es suficiente. En el caso de que solamente se adapten las medias del UBM, cada una de las gaussianas que componen los modelos compartirá la misma forma que las del UBM (matrices de covarianza y pesos serán los mismos), siendo la posición de las mismas (media de la gaussiana) lo que marcará la diferencia entre el modelo universal y el adaptado.

Una mejora notable del uso de modelos adaptados del UBM frente al modelo de GMM-ML es que permite aplicar el concepto de las *mezclas más pesadas*, lo que permite reducir cálculos en la obtención de los valores de densidad de probabilidad $p(x|\lambda)$ ya que sólo se usan las I gaussianas con mayor densidad de probabilidad $p_i(x)$, que suelen acumular casi toda la $p(x|\lambda)$ y son las mismas en el UBM y en los modelos adaptados.

Estas gaussianas de mayor probabilidad a posteriori (*mezclas más pesadas*) normalmente representan un número de gaussianas muy bajo con respecto del total. Por ejemplo, en un sistema de 1024 mezclas puede ser suficiente seleccionar tan sólo las 5 mezclas más pesadas que acumulan la mayor densidad de probabilidad y despreciar las 1019 restantes (lo que aligera en gran medida los cálculos). Esta técnica ha sido empleada durante prácticamente la totalidad del

proyecto para adaptar los modelos de las clases acústicas, y los valores escogidos se detallan en secciones siguientes.

Basadas en esta técnica de GMM-UBM, se han desarrollado en el ámbito de reconocimiento de locutor, pero también han sido aplicadas a segmentación de audio, otras técnicas más complejas que intentan modelar y compensar la variabilidad de canal y locutor, como son Joint Factor Análisis (JFA) [32] y total variability [33].

3.3.2.4. GMM-SVM

Las máquinas de vectores soporte (SVM - *Support Vector Machines*) son clasificadores muy potentes que se caracterizan por su capacidad discriminativa. Formalmente, un SVM es un clasificador de binario de patrones que separa los datos objetivo y los no objetivo (*target* y *non-target*) mediante un hiperplano [7]. El uso de este técnica se aplica a diferentes áreas del procesado de señales, entre las que se encuentra la segmentación de audio [34] y la identificación de idioma y reconocimiento de locutor [35]. Dicha ténica puede usarse para clasificar tanto vectores de características directamente [36] como *supervectores* [37], que son los vectores formados por la concatenación de los vectores de medias de un GMM.

3.4. Calibración

La puntuación (score) de una observación o un conjunto de ellas respecto a una clase representa la similitud de aquéllas frente a éste. Si bien de manera genérica se puede establecer que a mayor puntuación mayor será la similitud de la muestra con el modelo y viceversa, esa puntuación carece de significado si no se establece además un umbral a partir del cual realizar la decisión de pertenencia o no pertenencia al modelo. Este umbral puede ser fijado de manera experimental, buscando el valor que disminuye el error de segmentación que arroja el sistema tras colocar diferentes umbrales de detección, o bien, de una manera precisa, mediante la calibración del sistema.

La calibración [38] es una transformación que se aplica a los scores por la cuál éstos se transforman a valores con significado en sí mismos, conocidos como relaciones de verosimilitud (likelfihood ratios, LRs), los cuales indican la relación entre la probabilidad de que las observaciones correspondan a la clase frente a que no correspondan. Por lo tanto, un LR $>$ 1 (logLR $>$ 0) indica que es más probable que las observaciones correspondan al modelo frente a que no correspondan, y lo contrario para un LR $<$ 1 (logLR $<$ 0). Así, en la propia transformación está implícito el establecimiento del umbral en LR = 1 (logLR = 0). Mientras que en el caso de usar scores directamente el umbral óptimo del sistema puede cambiar de un conjunto de datos a otro, con el uso de LRs, si el sistema está bien calibrado, esto no sucede, ya que el LR indicará siempre lo mismo (la relación entre las hipótesis target y non-target). Esta transformación se entrena a partir de un conjunto de scores, de los que se sabe si corresponden a comparaciones target o non-target, procedente de un conjunto de desarrollo. En el caso de este proyecto, dicha calibración se ha realizado mediante regresión logística lineal [39], proceso por el cual se obtienen los coeficientes de una transformación lineal que *mapea* scores a LRs.

El problema que corrige la calibración se representa en la siguiente imagen 3.11, la cual

[7] un hiperplano es una extensión del concepto de plano aplicado a cualquier dimensión. En un espacio unidimensional el hiperplano es un punto, en un espacio bi-dimensional se corresponde con una recta, etc.

muestra los resultados o *scores* obtenidos de un sistema sin calibrar (figuras de la izquierda) con un sistema calibrado (figuras de la derecha).

En esta imagen puede observarse en primer lugar que la calibración no modifica el poder discriminativo del sistema (las distribuciones se solapan en la misma proporción), y en segundo lugar, que se minimiza el porcentaje de comparaciones non-target con logLR > 0 y el de comparaciones target con logLR < 0.

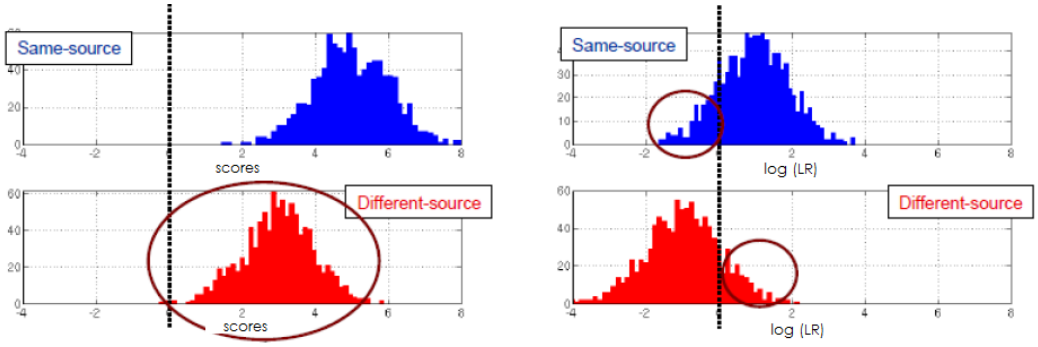


Figura 3.11: Evidencia de la necesidad de calibrado.

Adicionalmente, la calibración resulta especialmente útil cuando se trabaja con la fusión de *scores*, ya que si el umbral de trabajo es el mismo para ambos, los logLRs pueden combinarse directamente mediante reglas simples (fusión suma, promedio, etc.) que no necesitan datos de entrenamiento adicionales.

3.5. Fusión de sistemas

La fusión de sistemas de reconocimiento de patrones consiste en combinar dos o más sistemas con el objeto de aumentar el rendimiento. De las diversas técnicas, destacan la fusión a nivel de características y la fusión a nivel de *scores*

3.5.1. Fusión a nivel de características

La fusión a nivel de características es la fusión a más bajo nivel, en la cual se combinan los vectores de características obtenidos de aplicar diferentes sistemas de extracción de características sobre el mismo fichero de audio.

Para poder llevar a cabo este tipo de fusión ambos sistemas de extracción de características deben trabajar con segmentos de la misma duración y el mismo valor de solapamiento, de modo que el vector *i-ésimo* de cada sistema represente la información correspondiente a la misma región temporal. Ambos vectores obtenidos de los diferentes extractores de características se concatenan, de modo que el resultado de fusionar un vector de n características con otro de m características será un vector de dimensión $n+m$. Este método resulta de utilidad cuando los sistemas de segmentación de audio que se basan en cada una de las diferentes técnicas de extracción de características emplean no sólo la misma técnica de entrenamiento de patrones sino los mismos parámetros de ajuste. En este caso, los vectores obtenidos por cada una de las tramas podrían contener más información, lo que podría permitir a su vez obtener una representación más completa de los ficheros de audio. Puesto que existen diversas enfoques para

extraer las características de la señal de audio esta fusión puede resultar especialmente útil cuando la información que se combina tiene diferente naturaleza, como por ejemplo, coeficientes MFCC-SDC con estadísticos de la entropía cromática. Esta técnica no debe confundirse con la obtención de coeficientes derivados, como es el caso de los SDC, los cuales ofrecen información, también complementaria, pero dependiente de los primeros coeficientes (MFCC).

Puesto que esta fase genera nuevos vectores de características, se debe aplicar a todos los datos que van a ser tratados en la tarea de segmentación, tanto los datos de entrenamiento como los datos de testeo y ajuste del sistema.

3.5.2. Fusión a nivel de *scores*

El objetivo de este tipo de fusión es combinar las puntuaciones obtenidas para cada trama por diferentes sistemas de segmentación. Al igual que en el método anterior, la frecuencia de trama debe ser igual entre los sistemas a combinar, para que dicha combinación de resultados sea coherente. Sin embargo, a diferencia de la fusión a nivel de características, los sistemas que explotan por separado los distintos tipos de información pueden optimizarse individualmente, en lugar de tener que buscar una única configuración que funcione bien para los vectores que combinan distintos tipos de información. En este punto, todas las combinaciones de algoritmos de entrenamiento son posibles siempre que se respete la misma frecuencia de trama. Puede ser aplicada a dos o más sistemas, y la obtención de resultados es más inmediata que en el caso anterior, ya que se trabaja solamente con las puntuaciones finales obtenidas sobre cada sistema, y no es necesario entrenar nuevos sistemas u obtener nuevos vectores de características como en el caso anterior. En la fusión a nivel de scores cabe distinguir la fusión basada en reglas fijas, y la fusión basada en reglas entrenadas.

Las operaciones que caracterizan una fusión basada en reglas fijas son sencillas, tal es el caso de la fusión suma, el uso de la semi-suma, el producto, o la elección de máximos y mínimos. Por otro lado, la fusión basada en reglas entrenadas consiste en emplear los *scores* de cada uno de los sistemas de segmentación como patrones de entrada para un nuevo clasidicador, por lo que la fusión pasa a ser considerada un problema de clasificación de patrones en el que cada uno de los scores obtenidos sobre cada segmento con cada uno de los sistemas pasará a ser una característica del nuevo vector de datos que va a ser entrenado.

3.5.3. Fusión basada en decisiones categóricas

La fusión basada en decisiones categóricas se realiza a nivel de etiquetas, es decir, en la última fase del proceso de segmentación de audio (una vez que se han obtenido los scores, se han calibrado y se les ha asignado una clase).

En este contexto de clasificación, se pueden emplear operaciones lógicas sencillas como es el caso del operador AND (entendido bajo el criterio de unanimidad) o el operador OR (en el cual un voto a la clase a tratar es suficiente para seguir considerando que el segmento pertenece a dicha clase en el etiquetado final). Cuando se trata de fusionar más de dos sistemas, cobra sentido la técnica del máximo voto. En este proyecto se han podido generar scores para cada sistema individual por lo que este tipo de fusión (quizás menos precisa) no se ha llevado a cabo.

4

Marco experimental y sistema de referencia

4.1. Introducción

En este capítulo se va a describir en primer lugar el marco experimental, empleado para evaluar el sistema de referencia presentado a la evaluación Albayzin (basado en características MFCC-SDC) así como los demás sistemas de segmentación de audio presentados en el siguiente capítulo. A continuación, se describe el sistema de segmentación desarrollado para la evaluación Albayzin 2014 y se expondrán los resultados obtenidos de la detección así como los diferentes fases de ajuste de parámetros llevadas a cabo.

4.2. Marco experimental

4.2.1. Contenido de la base de datos

La base de datos con la que se trabaja a lo largo del proyecto, para desarrollar y medir el rendimiento de cada uno de los sistemas desarrollado en este proyecto, es la proporcionada por la evaluación Albayzin 2014, la cual contiene (como ya se mencionó en la sección 2.5) ficheros de audio mayoritariamente de noticias entre los que se pueden encontrar presentes las clases de voz, música y ruido.

4.2.2. Organización de la base de datos

Dicha base de datos está formada en primer lugar por 20 grabaciones (*tracks 1 al 200*) de audio etiquetados de aproximadamente 60 minutos, de los cuales se utilizan los tracks 1 al 15 para desarrollo (1 al 10 para entrenar, 11 al 15 para ajustar los parámetros del sistema y obtener puntuaciones para la calibración del mismo), y los 5 últimos para testear. Adicionalmente, en segundo lugar se emplearon los tracks 21 a 35 para evaluar el sistema de referencia (basado en características MFCC-SDC) en el marco de la evaluación *Albayzin*.

Cada uno de los tracks viene asociado con un fichero de texto *RTTM* (Rich Text Transcription Mark [1]) cuyo formato se ha ajustado en base a la herramienta empleada en las evaluaciones NIST de Diarización de locutores [2]. Concretamente, estos ficheros contienen *metda-datos* con los elementos presentes en cada grabación (entendiendo elementos como cada una de las posibles clases que se pueden encontrar en esta base de datos) y su correspondencia temporal en el fichero de audio. Las tres clases contempladas son música, voz y ruido. Cada uno de estos ficheros contiene tantas líneas como segmentos de las distintas clases se hayan destacado, donde la información de cada línea detalla (en el siguiente orden): track al que pertenece, tiempo de inicio, tiempo final, y clase acústica asociada a ese segmento.

A partir de estas etiquetas, se pueden detectar los momentos en que varias clases se presentan en el mismo instante de tiempo así como los fragmentos de audio en los que sólo se presenta una clase acústica, o ninguna (en caso de producirse silencio). De este modo, la presencia combinada o aislada de estas tres clases se puede entender como un sistema de ocho clases: silencio (silence), música con ruido (mu-no), música aislada (mu), voz con música y ruido (sp-mu-no), voz con música (sp-mu), voz aislada (sp), voz con ruido (sp-no) y ruido aislado (no). El contenido de la base de datos en referencia a este desglose detallado por clases se presenta en la figura 4.1.

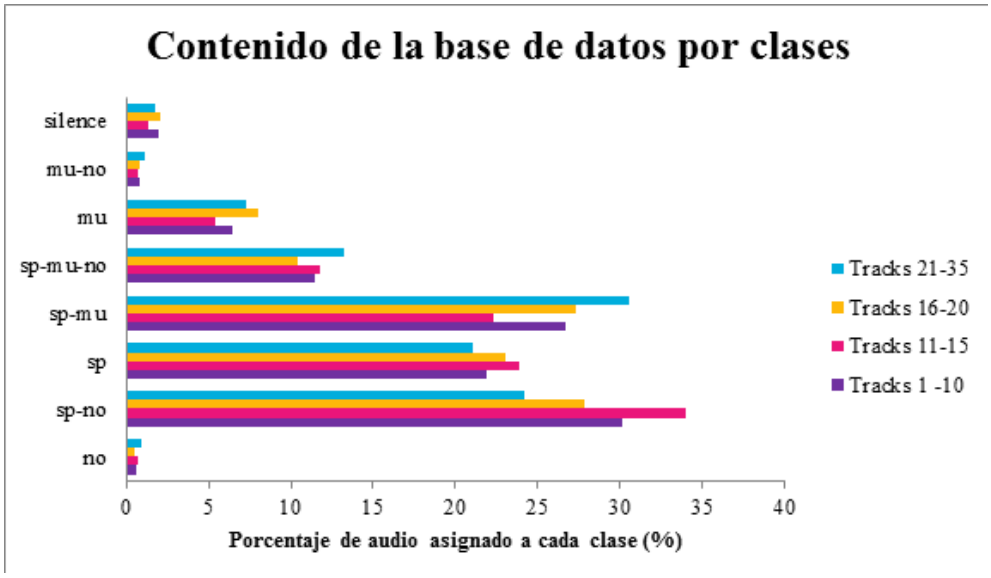


Figura 4.1: Contenido de la base de datos.

Tras contemplar esta distribución de datos, se pueden destacar las siguientes observaciones:

- La proporción entre las distintas clases es muy similar en los diferentes subconjuntos de tracks. Por ejemplo, para cualquiera de los cuatro conjuntos el porcentaje de voz aislada sobre el total se sitúa entre 20 y 25 %
- Por otro lado, se ve claramente cómo la cantidad de audio no es proporcional entre las clases, lo que pudiera dificultar la generación de un detector de audio preciso que ha sido entrenados con pocos datos de dicha clase. Tal es el caso del ruido aislado, que en general no supera el 2 % del total del contenido de la base de datos. No obstante, la presencia de esta clase simultánea con la clase voz si es muy significativa.

[1] http://www.itl.nist.gov/iad/mig/tests/rt/2003-fall/docs/RTTM-format-v13.pdf
[2] http://www.itl.nist.gov/iad/mig/tests/rt/

- Del desglose en ocho clases, se aprecia un pequeño porcentaje en el que las tres clases están ausentes, esto es, pequeños fragmentos de silencio.
- La presencia de la clase de voz es siempre la más significativa, ya que si se suma el porcentaje de tiempo en el que está presente dicha clase (ya sea de manera aislada -sp- o en cualquiera de sus combinaciones -sp-no, sp-mu o sp-mu-no) se obtiene un valor de más del 60 %.
- Puesto que la probabilidad a priori de cada una de las clases es diferente, a la hora de medir el rendimiento final de los detectores no es lo mismo no detectar los segmentos que aparezcan de ruido aislado (con muy poca presencia en la base de datos) que los de voz (más numerosos y/o largos), siendo esto último más perjudicial para el rendimiento del sistema.

4.3. Sistema de referencia (*Albayzin 2014*)

El sistema de referencia en el que se apoya el proyecto está enmarcado en el estado del arte de los sistemas de reconocimiento de patrones. Se trata de un sistema compuesto por 3 detectores GMM-UBM basados en características tímbricas MFCC-SDC, capaz de etiquetar segmentos de ficheros de audio por clases acústicas según su naturaleza. Las clases con las que trabaja dicho sistema son tres: música, ruido y voz. Cualquier segmento analizado puede pertenecer a una, dos o tres clases, por lo que un mismo segmento pueden corresponderle hasta tres etiquetas.

Este sistema se encuadra como sistema de referencia ya que ha sido presentado y evaluado en una competición de segmentación de audio [6] durante la realización de este proyecto, y sus resultados han sido contrastados con otros sistemas de segmentación de audio de actualidad. Este sistema de referencia proporciona etiquetas en formato RTTM como las proporcionadas en la evaluación Albayzin, en las que cada clase acústica (música, ruido y voz) se etiqueta de manera independiente (por lo que pueden aparecer solapadas o no en el tiempo con alguna de las otras clases), lo que produce que segmentos acústicos asignados a determinados instantes de tiempo se encuentren contenidos en diferentes líneas del fichero de texto (figura 2.4) .

4.3.1. Estructura del sistema y diagrama de bloques

Las fases por las que pasa el sistema de segmentación de audio de referencia se detallan a continuación:

1. En primer lugar, los ficheros de audio de entrada han sido analizados en segmentos de 20 ms, bajo un sistema de ventana deslizante Hamming con un 50 % de solapamiento entre ellos, obteniendo de cada segmento un vector de 56 características MFCC-SDC. De estas 56 características, las 7 primeras son coeficientes MFCC y las 49 restantes son características temporales (SDC) derivados de estos primeros. Las características SDC responden a una configuración 7-1-3-7, que se corresponde con los parámetros **N-d-p-k** expuestos en el apartado 3.2.1.2. En este apartado, cabe destacar que debido a la disposición del contenido de la base de datos, los vectores de características obtenidos se han almacenado desde dos enfoques diferentes: por un lado, se han agrupado todos los vectores de características por *tracks* de audio completos sin tener en cuenta su naturaleza acústica (esto es, sin emplear las etiquetas) generando un fichero de características para cada track; por otro lado, se han agrupado los segmentos de audio atendiendo al contenido de las etiquetas de cada track,

esto es, generando un fichero de características por cada segmento perteneciente a una misma clase acústica. De este modo, por cada *trackXX* analizado se obtendrá un fichero con los vectores de características del *trackXX* completo, y tantos ficheros de duración variable correctamente etiquetados como fragmentos de clases contenga el fichero de etiquetas de dicho *track*. En esta fase, solamente se trabaja con los tracks de audio de la base de datos asignados para entrenamiento del sistema (1 al 10). Del total de audios, sólo algunos son empleados para entrenar y otros para calibrar y testear el sistema. Evaluar sobre los propios datos de entrenamiento puede crear sobreajuste, por lo que parte de los datos etiquetados de los que se dispone deben ser reservados para la fase de prueba.

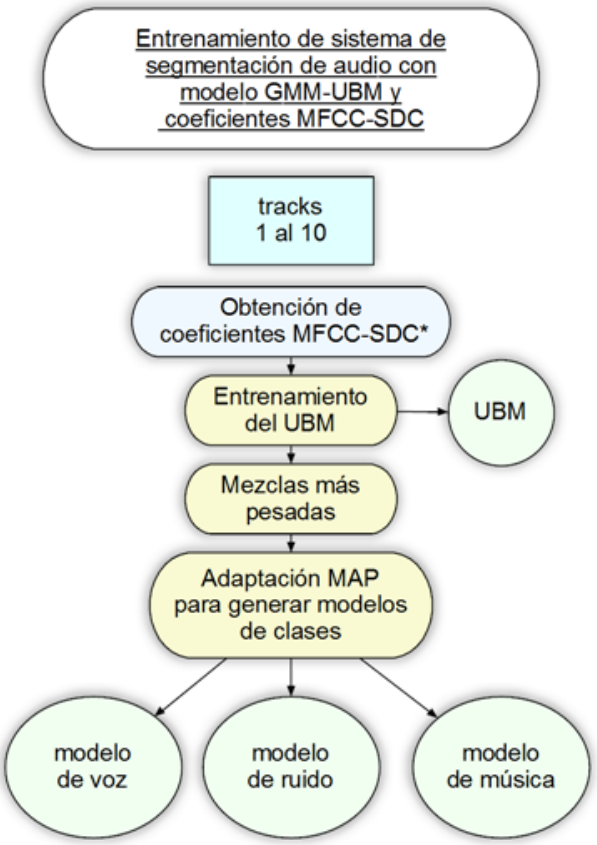


Figura 4.2: Diagrama de bloques de entrenamiento de sistema de referencia.

2. En segundo lugar, se genera un modelo universal (UBM) a partir de los ficheros de características obtenidos de cada uno de los *tracks* completos, dado que en esta fase todavía no se contempla la separación del contenido por clases. El número de gaussianas que se han empleado sobre el sistema final es de 1024, por ser el modelo que proporciona mejores resultados en el conjunto de desarrollo. Este número de gaussianas del UBM es una configuración típica en otras tareas, como identificación de idioma en la que también se trabaja con las mismas características (tipo y número), si bien otros valores como 512 también se han propuesto y descartado por ofrecer peores resultados en la fase de evaluación. El modelo ha sido inicializado con una iteración del algoritmo *K-means* seguida de 5 iteraciones del del algoritmo EM (Expectation-Maximization)

3. En tercer lugar, con el objeto de reducir el coste computacional para la adaptación de modelos de clases en la siguiente etapa, se buscan, para los vectores de características de

cada clase, las *mezclas más pesadas* del UBM (es decir, aquellas que acumulan el mayor porcentaje de la probabilidad frente al modelo). En este sistema se ha elegido un valor de 5 mezclas, por ser un número suficientemente bajo (en comparación con las 1024) pero que ofrece un resultado prácticamente idéntico en la adaptación de modelos de las clases acústicas reduciendo significativamente el número de operaciones necesarias para generar el modelo final.

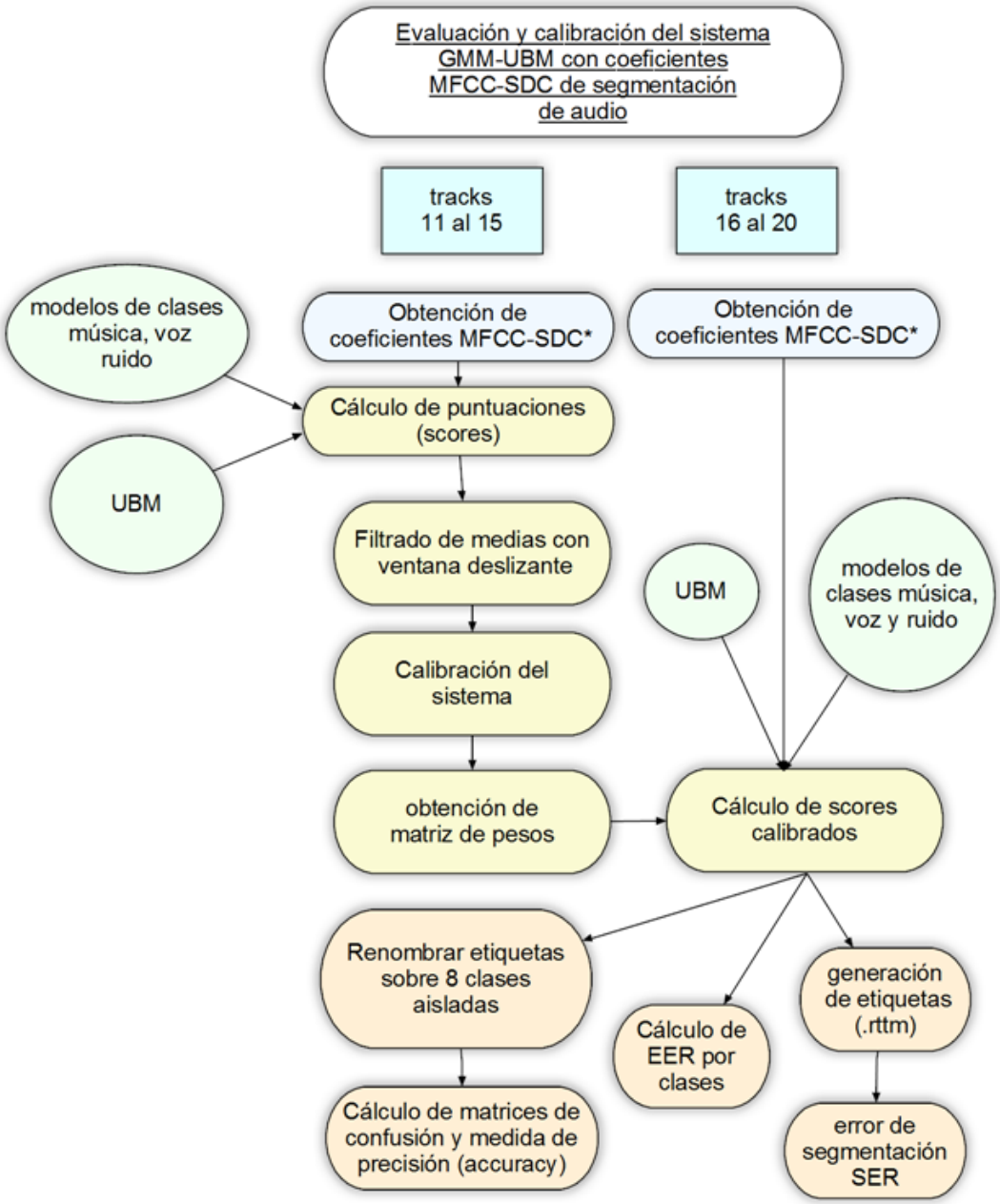


Figura 4.3: Diagrama de bloques de las etapas de desarrollo y test del sistema de referencia.

4. En cuarto lugar, se generan los modelos de clases a partir de los vectores de características, de la información obtenida en la fase tres y del UBM generado. En este punto, se probaron dos combinaciones de datos: por un lado, se probó a dividir los datos de entrenamiento en tracks para entrenar el UBM y tracks para entrenar el modelo por clases, de modo que

el modelo de clases se calculase sobre unos datos diferentes a los empleados en el modelo universal; y, por por otro lado, se probó a trabajar con el mismo conjunto de datos para la obtención del modelo universal y del modelo por clases. La primera de las opciones dejaba muy pocos datos para la adaptación de los modelos de cada clase, por lo que se escogió la segunda opción para la generación de modelos. Cada uno de estos modelos por clases se obtiene de adaptar el modelo universal mediante el algoritmo *maximum a posteriori* (sección 3.3.2.3). El proceso de adaptación se ha realizado solamente sobre las medias del UBM, manteniendo para cada modelo por clase los parámetros de covarianzas y pesos del modelo universal, lo que implica que la forma de cada gaussiana se mantiene pero cambia la posición de cada una respecto del modelo universal. Asímismo, del resto de parámetros que es preciso especificar en este proceso de adaptación cabe resaltar que se ha realizado una sóla iteración del algoritmo MAP, con un factor de relevancia de 16 (r = 16). El resultado de esta fase es la generación de tres modelos de clase que permitirán al sistema detectar cada una de las clases acústicas (música, voz y ruido) presentes en un fichero de audio cualquiera.

5. Una vez obtenidos los modelos, la siguiente tarea se enmarca dentro de la fase de desarrollo del sistema. Esta tarea comienza por calcular la relación de verosimilitud, o score, de cada vector de características del fichero de audio de test respecto a cada modelo de clase acústica. Dicha relación se obtiene normalizando la verosimilitud del vector de test dado el modelo de clase respecto a la verosimilitud dado el UBM.

6. El valor del score trama a trama (por cada vector de características) presenta una gran variabilidad, al evaluarse duraciones muy cortas (20 ms por trama). Para reducir dicha variabilidad, se aplica un filtro de medias de longitud determinada sobre las puntuaciones, cuyo objeto es conseguir valores de puntuación frente a cada clase más estables a lo largo del tiempo. Dado que la longitud óptima de este filtro es desconocida y depende de la naturaleza del audio, se ha calculado experimentalmente para cada clase, obteniendo un valor diferente para cada una de éstas. Normalmente la longitud del filtro escogido (ventana óptima) se obtiene de realizar un barrido de valores de longitud para el filtro de media que abarca desde 400 tramas o vectores de características a 2000, escogiendo los valores que minimizan el error de detección en cada clase evaluado mediante el EER. Dado que una trama abarca 10 ms de audio (segmentos de 20 ms con 50 % de solape), las duraciones temporales sobre las que se calcula la longitud de filtro óptimo oscilan entre 4 y 20 segundos: $400 segmentos * 10 ms/segmento = 4s, 2000 segmentos * 10 ms/segmento = 20s$.

7. Una vez que las puntuaciones han sido filtradas tiene lugar una fase de calibración, en la cual se obtiene una función de *mapeo* de scores de cada clase a partir de las puntuaciones obtenidas. Dicha calibración va a ser empleada junto con el valor de ventana óptimo para estimar la pertenencia de los segmentos a cada una de las clases con el mayor acierto posible. El proceso de calibración empleado en este sistema y en los siguientes expuestos en el siguiente capítulo está basado en el algoritmo de regresión logística [39]. En la gráfica 4.4 se puede ver un ejemplo de los scores obtenidos para la clase música calibrados.

8. La penúltima tarea consiste en probar experimentalmente el resultado del sistema de segmentación sobre un conjunto de datos diferente al usado para entrenamiento y desarrollo. Para ello, los vectores de características obtenidos se comparan con cada clase y con el modelo universal obteniendo unas puntuaciones. Dichas puntuaciones se filtran bajo los valores de ventana óptimos y se calibran. En este caso, un segmento será considerado como perteneciente a una clase si la puntuación final obtenida sobre dicha clase supera el umbral. Gracias al proceso de calibración, puede utilizarse directamente un umbral igual a 0 para

tomar la decisión de clase detectada o no detectada. Una vez clasificados los segmentos, se van a generar las etiquetas (*.rttm*) correspondientes a los ficheros de test evaluados.

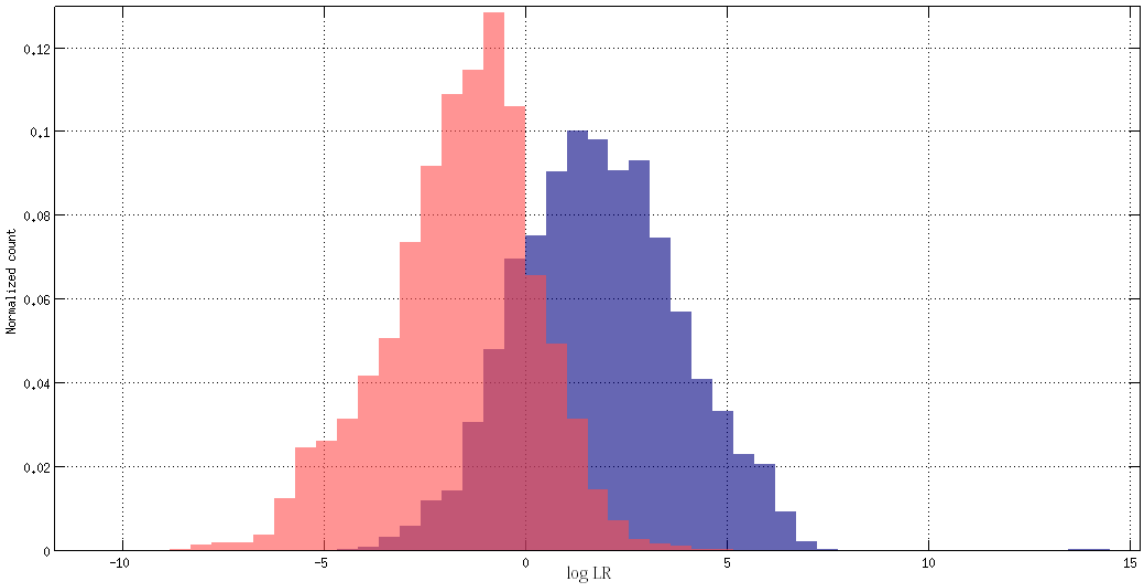


Figura 4.4: Resultados de la calibración scores target (morado) frente a non target (rosa claro) para clase música.

9. Finalmente se estudia la bondad del sistema de clasificación en base a diferentes medidas de error. En primer lugar, se evalúa el rendimiento mediante la tasa de igual error (EER).

   En segundo lugar, las matrices de puntuaciones obtenidas con este sistema son adaptadas a un modelo de 8 clases considerando el etiquetado de regiones homogéneas de la figura 2.4. Las nuevas matrices de resultados obtenidas son la base para medir el error de confusión del sistema haciendo especial hincapié en la bondad del sistema tanto para cada clase acústica (voz, música, ruido) de forma aislada como con las posibles combinaciones.

Una vez que el sistema ha sido desarrollado, calibrado y evaluado, el comité organizativo de la evaluación entregó a cada concursante un conjunto de datos sin etiquetar formado por 15 tracks de audio, a partir de los cuales cada concursante debió generar las etiquetas experimentales con su sistema y enviarlas a la evaluación. No obstante, ni este sistema ni los desarrollados en este proyecto han usado estos tracks de evaluación, sino que se han considerado siempre los datos de la base de datos inicial (tracks 1 al 20) para desarrollar los sistemas.

El proceso de segmentación de audio del sistema de referencia recientemente detallado se resume en los siguientes diagramas de bloques: en la figura 4.2 se detallan los pasos a seguir para la generación de modelos de clases, mientras que en la figura 4.3 se detallan los pasos a seguir para la calibración de puntuaciones y evaluar el sistema de forma experimental.

4.3.2. Análisis de resultados

Una vez aplicado el sistema de segmentación desarrollado sobre los tracks de test (16-20), se procede a la fase de evaluación mediante varias de las métricas expuestas en la sección 2.4. En cualquiera de los sistemas se ha tratado con el EER, las matrices de confusión y el error de

precisión (o *accuracy)*. Adicionalmente para este sistema se ofrecen los resultados obtenidos de participación en la evaluación Albayzin.

4.3.2.1. Valores óptimos del filtro de medias

Para encontrar el tamaño óptimo de ventana se filtran los scores obtenidos en los tracks 11 al 15 con valores de ventana comprendidos entre 400 y 2000 tramas a intervalos de 50 tramas (esto es, 40 ms). Para cada clase, se calcula el EER de su detector GMM-UBM obtenido con cada valor de ventana, y se resume en una gráfica que relaciona cada tamaño con el valor de error obtenido por clase. Los resultados obtenidos se muestran en la gráfica 4.5 y se resaltan en la tabla 4.1.

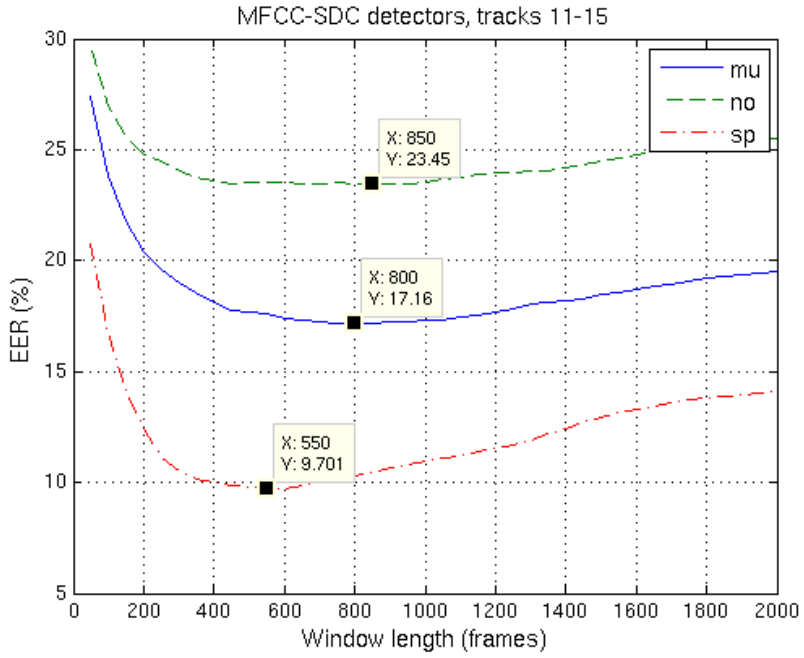


Figura 4.5: Análisis de EER para distintos valores del filtro de medias sobre tracks 11 al 15 sistema de referencia.

Longitud óptima del filtro (número de tramas)	800	850	550
EER	17.16	23.45	9.71
Clase acústica	música (mu)	ruido (no)	voz (sp)

Tabla 4.1: Valores de ventana óptimos sobre el sistema de referencia.

Estos valores obtenidos de ventana van a ser los empleados para evaluar el sistema (con los tracks 16 al 20) y segmentar cualquier fichero de audio del que se desee conocer la distribución de clases. Sobre este conjunto de datos la clase que mejor se detecta es la voz, seguida de la música y del ruido. No obstante, dado que estos datos se enmarcan en la etapa de desarrollo u optimización (tracks 11 - 15) el EER obtenido no se contrasta más en profundidad con otros sistemas.

4.3.2.2. Rendimiento de detección

Una vez obtenidas las puntuaciones finales ya calibradas, y escogido un tamaño de filtro de medias óptimo, se calculan las puntuaciones sobre el conjunto de datos asignados para test, esto es, de los tracks 16 al 20. Se representan los errores de detección mediante curvas DET, que no sólo muestran el valor de EER sino además la relación entre la tasa de falso rechazo y falsa aceptación para diferentes puntos de trabajo del detector. Adicionalmente, se realiza un barrido de valores de ventana para comprobar si el valor de ventana previamente escogido sobre los datos de desarrollo se asemeja al que sería el óptimo en este conjunto de datos de test.En la gráfica 4.6 se refleja la curva DET de cada uno de los tres detectores y en la figura 4.7 se muestra el resultado de error para diferentes valores de ventana.

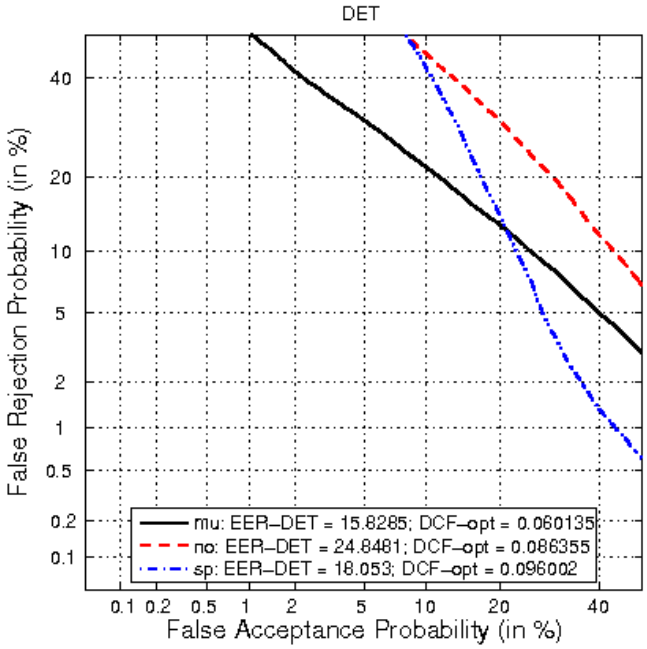


Figura 4.6: Curvas DET correspondientes a los 3 detectores del sistema de referencia sobre los tracks 16-20.

En primer lugar, cabe resaltar que la clase que mejor se detecta es la clase música, seguida de la voz y por último el ruido. A partir de las gráficas se puede observar cómo el valor de EER obtenido para las clases de música y ruido sobre el conjunto de datos de evaluación (tracks 16 al 20) se asemeja al obtenido sobre los tracks 11 al 15, con una variación máxima de 2 puntos de EER. No obstante, los valores de EER obtenidos de los tracks 11 al 15 se han empleado para escoger ventana óptima, por lo que la comparativa de resultados en cuanto a niveles de EER no es el motivo de este punto, como sí lo es el medir un posible sobreajuste de ventana. Si calculamos en valor absoluto la diferencia de EER obtenido de los tracks 16 al 20 con el valor de ventana predeterminado, con respecto del que resultado óptimo del nuevo barrido se puede apreciar que las desviaciones son insignificantes, además de que se aprecia cómo el rendimiento del detector de voz parece muy dependiente del conjunto de datos de test.

Con todo ello, es importante destacar para este caso la alta similitud entre los valores de ventana escogidos y los que serían óptimos sobre los datos de evaluación.

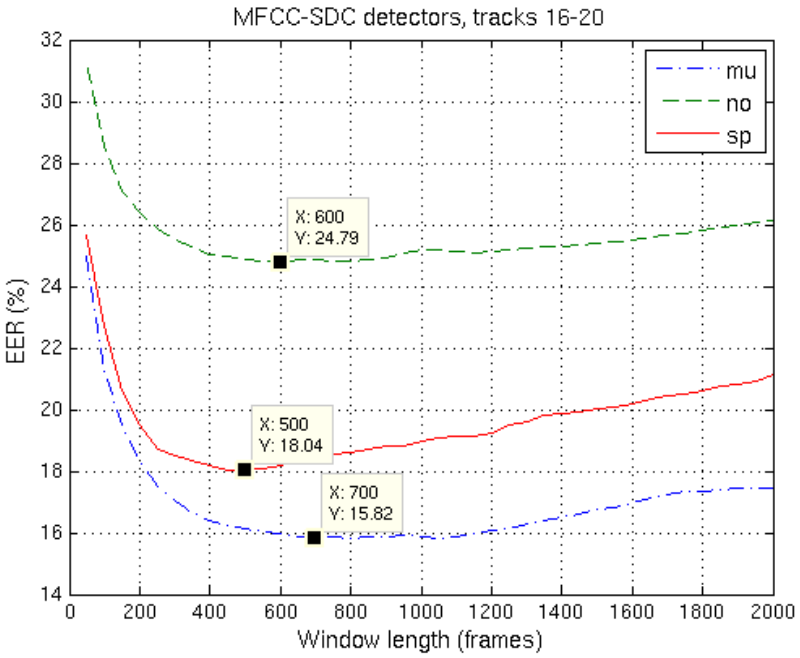


Figura 4.7: Análisis de EER para distintos valores del filtro de medias sobre tracks 16 al 20.

Comparativa de EER sobre tracks 16 al 20	**mu**	**no**	**sp**
EER con valor de ventana obtenido en desarrollo (tracks 11-15)	15,82 (800)	24, 84 (850)	18,053 (550)
EER con valor optimo	15,82 (700)	24,79 (600)	18,04 (500)

Tabla 4.2: Comparativa de EER para estimar posible sobreajuste de ventana.

4.3.2.3. Matrices de confusión

A continuación se muestra la gráfica 4.8 con la matriz de confusión obtenida para el sistema sobre los tracks de test 16 al 20. En dicha gráfica, se puede apreciar a simple vista qué clases de referencia son detectadas con mayor acierto así como el total de segmentos que han sido correctamente asignados respecto del total, esto es, el valor de *accuracy* que aparece al pie de la gráfica. En segundo lugar se adjunta una tabla (4.3) que refleja la bondad de aciertos sobre cada clase en valor porcentual.

De esta primera gráfica se observa que un poco más de la mitad de las muestras son bien clasificadas. Además, es interesante ver que la clase de silencio (para la cual no hay asignada un modelo concreto, sino que se considera la ausencia de relación con ninguno de los tres modelos), se confunde mayoritariamente con la voz. Por otro lado, destaca la clasificación de la clase voz y la de voz con música, que guarda relación con el total de segmentos asociados a estas clases que hay en la base de datos.

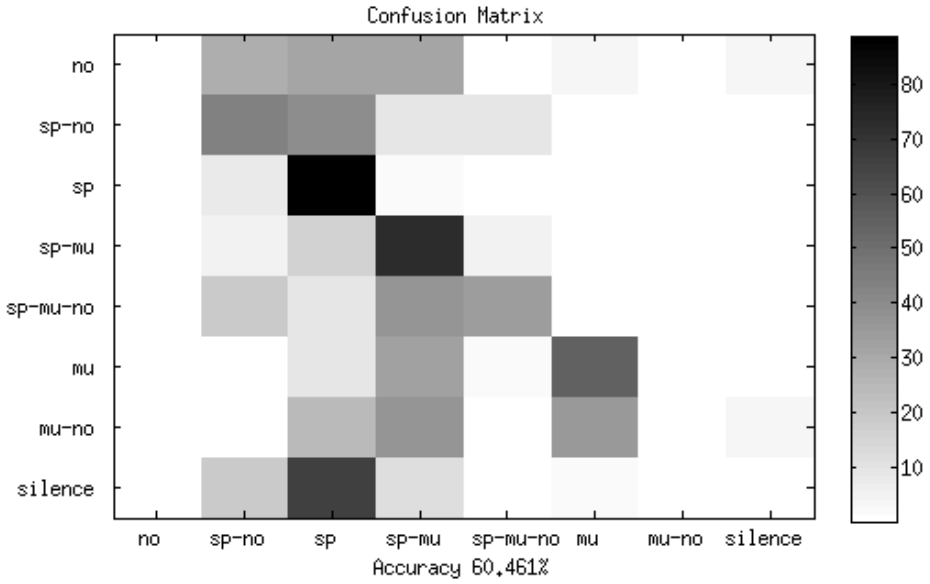


Figura 4.8: Matriz de confusión del sistema de referencia sobre tracks 16 al 20.

Clase	**no**	**sp-no**	**sp**	**sp-mu**	**sp-mu-no**	**mu**	**mu-no**	**silence**
Precisión (%)	0.35	43.45	88.84	72.98	34.57	54	0	0.32

Tabla 4.3: Valores de precisión por clases sobre sistema de referencia.

Los resultados de la matriz de confusión (figura 4.8) permiten contrastar a simple vista que las clases de voz aislada y de música aislada son las que mejor se detectan con un porcentaje alto de acierto. No obstante, dado que el valor de precisión se calcula sobre el total de segmentos sin tener en cuenta el total de datos sobre cada clase, el valor de precisión obtenido total apenas supera el 60 %. Si se contrasta la gráfica con los valores de la tabla, se puede apreciar cómo la precisión de detección de la clase voz casi alcanza el 90 % seguida de la clase de música y voz con más de un 70 %. Resulta curioso destacar que la detección de ruido, de silencio y de música con ruido es prácticamente nula, lo que pone de manifiesto la imprecisión del sistema sobre estos conjuntos. Si se acude a la gráfica que resume el porcentaje de audio que hay en cada clase (figura 4.2.2) se puede observar cómo la presencia de datos sobre estos tres conjuntos es muy reducida y significativamente menor que las otras cinco clases. Reconocer un tipo de datos concreto con un sistema de segmentación que ha sido entrenado con muy pocos datos de esa misma índole justifica los malos resultados obtenidos con este sistema sobre dicho conjunto. No obstante, esta justificación de los resultados es válida sólo hasta cierto punto ya que la clase de voz y ruido (sp-no) es de las que más datos de entrenamiento tiene y sin embargo la precisión obtenida es significativamente menor (43.45 %) que la de las clases voz y voz con música que se entrenan con una cantidad similar de datos y sin embargo han sido detectadas con gran precisión.

4.3.2.4. Resultados de la evaluación

Como resultado de la evaluación de este sistema de segmentación de audio basado en características tímbricas se obtuvo una posición tercera sobre cuatro participantes. El resultado de la evaluación se llevó a cabo mediante la medición del SER de los resultados generados por

cada sistema sobre el conjunto de tracks 21 a 35. Adicionalmente, el nivel de SER empeoró considerablemente para este conjunto de datos con respecto a los datos de test (tracks 16 al 20), pasando de un 21 % de SER a un 30.67 %, por lo que una vez finalizado el concurso y obtenidas las etiquetas de este conjunto de datos, se realizó un análisis de los resultados a nivel de EER para intentar comprender los resultados obtenidos.

En primer lugar, se calculó el EER sobre los valores de ventana óptimos y los resultados se muestran en la gráfica 4.9. A partir de esta gráfica se aprecia cómo el EER para la clase de música pasa de un 15.82 % sobre los tracks 16 al 20 a un 27.99 % en el conjunto de datos de la evaluación (tracks 21 a 35). Para el caso del ruido el error aumenta en unos 4 puntos de EER y sólamente para la clase de voz los resultados se mantienen similares (menos de un punto de EER de variación). Los resultados obtenidos sobre este conjunto de datos de evaluación han resultado en general peor de lo que se esperaba a raíz de los resultados del primer conjunto de datos de desarrollo, lo que ha motivado un análisis del sistema para detectar si un posible sobreajuste de datos o algún tipo de error se hubiese cometido, o si simplemente el suceso se debía a que los datos de este último conjunto eran notablemente diferentes a los datos de entrenamiento.

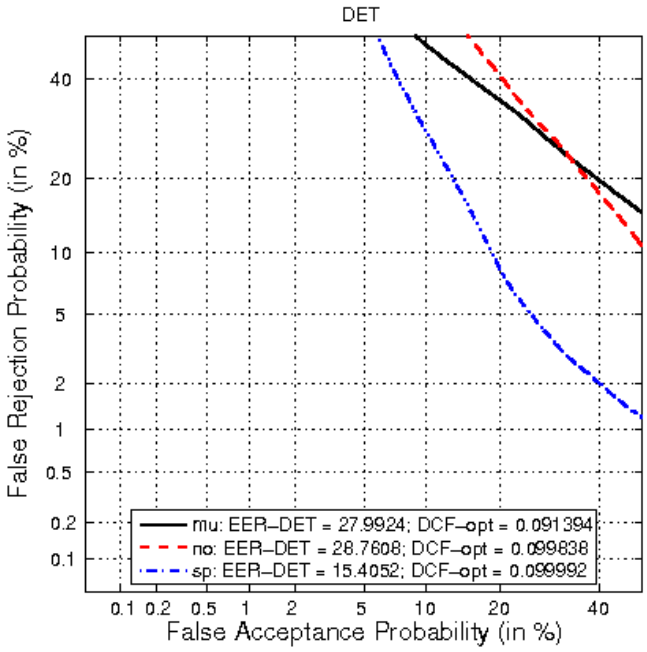


Figura 4.9: Medida de EER y curva DET sobre tracks 21 a 35 en sistema de referencia.

Por lo tanto, en segundo lugar se ha calculado el EER para un amplio conjunto de valores del filtro de medias (en lugar de sólo para el óptimo según los datos de desarrollo) y los resultados se muestran en la gráfica 4.10). Los valores de ventana óptimos para este nuevo conjunto han sido 850, 750 y 400 (para las clases de música, ruido y voz respectivamente), mientras que los valores obtenidos sobre los datos de desarrollo fueron de 800, 850 y 550. Además, en el caso de los tracks 21 a 35, el EER adquiere cierta estabilidad en el intervalo que rodea al valor óptimo, por lo que se puede concluir que en el tamaño de ventana escogido, no ha habido sobreajuste en la elección, ya que los valores de ventana óptimos del conjunto de tracks 21 a 35 son muy similares a los empleados. Por este motivo, se descarta que el valor de ventana haya sido el causante de la diferencia de resultados entre ambos conjuntos de datos.

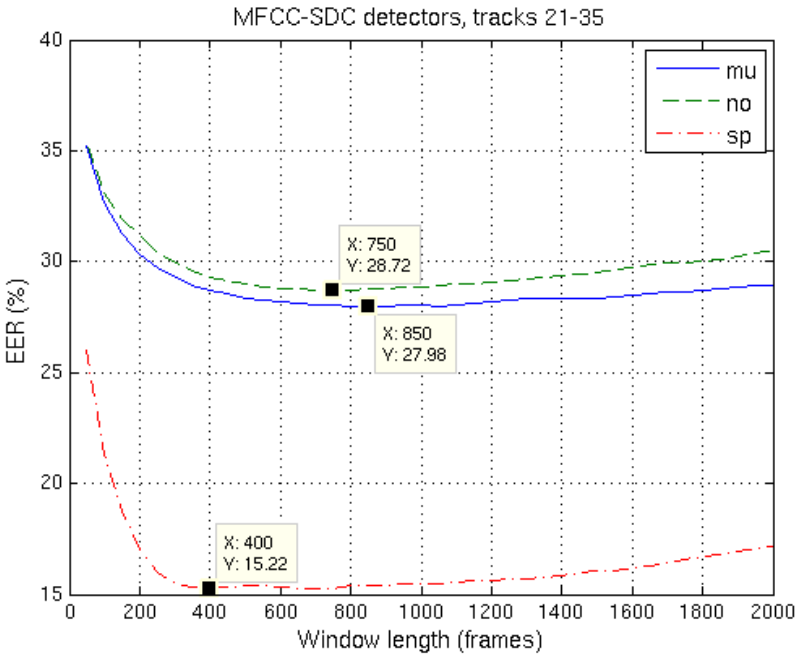


Figura 4.10: Valores de EER sobre tracks 21 a 35.

Por otro lado, contrastando el contenido de la base de datos (figura 4.2.2) se puede apreciar cómo en general la proporción de audio existente para cada una de las clases sobre ambos conjuntos de datos es muy similar, es decir, la longitud de las barras azules del diagrama (que se corresponden con los tracks 16 al 20) para cada una de las clases se asemeja a la longitud de las barras amarillas del mismo diagrama (tracks 21 a 35). Por lo tanto, también se descarta que la diferencia de rendimiento se deba a una diferente proporción de clases presentes entre los tracks de evaluación (tracks 21 a 35) y los de test (tracks 16 a 20). Con todo esto, quizás esta diferencia de rendimiento sí se deba a un sobreajuste de los datos empleados para entrenamiento que no hayan sido suficientes para generalizar, y generar por lo tanto unos modelos representativos.

5

Sistemas basados en información cromática

En el presente capítulo se presentan dos sistemas de segmentación de audio basados en características cromáticas. El primero de ellos, obtiene las características de estadísticos de la entropía cromática (media, varianza, *skewness* y *kurtosis*). El segundo de ellos, obtiene nuevas características cromáticas de los ficheros de audio basadas en la distribución de energía espectral. El sistema de segmentación de audio desarrollado a partir de ellas está basado en detectores GMM-UBM al igual que el sistema de referencia (expuesto en el capítulo anterior). Además de su rendimiento de forma aislada, se estudia la fusión de cada uno de estos sistemas con el sistema de referencia tanto a nivel de *scores* como de características, y los resultados se detallan en el presente capítulo.

5.1. Sistema basado en estadísticos de la entropía cromática

A continuación se detalla el sistema de segmentación de audio basado en estadísticos de la entropía cromática. Las características cromáticas, si bien son características frecuenciales al igual que los MFCC-SDC, están basadas en conceptos de la distribución frecuencial de las notas, por lo que al menos para la clase música se espera que la fusión con un sistema basado en características MFCC-SDC mejore los resultados de este último por sí sólo.

5.1.1. Estructura del sistema

El extractor de estadísticos de entropía cromática forma parte de los algoritmos desarrollados en el grupo ATVS [23]. Este sistema fue inicialmente desarrollado con la base de datos de la evaluación *Albayzin 2010* de segmentación de audio, parte de la cual está contenida en la base de datos de la evaluación de 2014. Tomando como referencia el primer sistema desarrollado para este proyecto, se ha estudiado el rendimiento del sistema basado en cuatro estadísticos de la entropía cromática (media, varianza, *skewness* y *kurtosis*) generando nuevos modelos acústicos con la nueva base de datos disponible. Como objetivos principales de trabajar con este sistema se plantea por un lado estudiar una posible mejora en la detección de música con respecto del

sistema de referencia, y por el otro estudiar la posibilidad de ser fusionado con el sistema de referencia propuesto (motivado por el hecho de que cada uno trabaja con un tipo de características de naturaleza diferente) en busca de un sistema de segmentación de audio más robusto y con mejores prestaciones.

El esquema propuesto para el sistema de segmentación de audio basado en estadísticos de la entropía cromática solamente difiere del sistema de referencia en el tipo de características que se obtienen, por lo que se asumen las mismas fases de entrenamiento, desarrollo y evaluación del sistema detalladas en el capítulo anterior (4.3). No obstante, difieren algunos de los parámetros empleados para la generación del modelo: en concreto, el UBM que ofrece resultados óptimos está inicializado con 5 iteraciones del algoritmo *k-means*, seguido de 10 iteraciones del algoritmo ML y trabaja con 128 mezclas. El número de mezclas más pesadas se mantiene en 5, y en la fase de adaptación MAP se mantiene el coeficiente de adaptación a 16 pero el número de iteraciones se establece en 10. Estos parámetros se han tomado del sistema de partida de la segmentación de audio con estadísticos de la entropía cromática [23]. Una vez generados los modelos, se procede al testeo del sistema utilizando los mismos ficheros de datos que en el sistema de referencia y siguiendo el mismo procedimiento (scoring GMM-UBM por trama, selección de filtro de medias óptimo, y calibración de scores) y los resultados se muestran en la siguiente sección 5.1.3. Por otro lado, se estudia la fusión con el sistema de referencia a nivel de características y a nivel de scores. Para la fusión a nivel de scores se emplean para el sistema basado en estadísticos de la entropía los parámetros del sistema de partida reciéntemente mencionados; sin embargo, para la fusión a nivel de características, los parámetros que se seleccionan para el entrenamiento son los empleados en el sistema de referencia, esto es, 1024 mezclas, 1 iteración del algoritmo *k-means* y 5 de ML (para el UBM), y 1 iteración MAP. Finalmente se evalúa el rendimiento del sistema con el EER por detector y las matrices de confusión aplicadas a 8 clases.

5.1.2. Análisis de las características y los modelos

Una peculiaridad de este sistema es que trabaja con vectores de tan sólo 4 características, un número muy pequeño en comparación con las 56 del sistema basado en MFCC-SDC. Este número tan bajo anima a realizar un análisis de los vectores de características obtenidos para visualizar de manera gráfica si los modelos entrenados se adaptan bien a los vectores de características de cada clase, y la forma que toma el UBM en comparación de los modelos. Si bien este análisis es siempre de utilidad para cualquier sistema, cobra especial sentido en este contexto donde el número de características es manejable y puede ser analizado gráficamente.

De este análisis destaca la forma que toman la distribución de cada uno de los vectores de características de los cuatro estadísticos de la entropía cromática. Dado que los modelos de las clases acústicas se componen de mezclas de gaussianas, es recomendable que la distribución que generan los vectores de características se asemeje a la de una distribución normal. Tomando como ejemplo el histograma de la clase música (figura 5.1), se puede apreciar cómo la media estadística sigue una distribución de estas características, al igual que sucede con el skewness. Sin embargo, la naturaleza de la varianza no genera una distribución de este estilo, ya que como se puede ver en el histograma los valores de varianza más frecuentes se sitúan en torno a cero, y complementado este hecho con que no tiene cabida que la varianza tome valores negativos (es un valor cuadrático), la distribución final se asemeja más a una log-normal, con lo que aplicando el logaritmo a estas características el resultado será una distribución de forma más aproximada a una normal. Dicha técnica también va a ser empleada para el caso de la kurtosis, ya que debido a su naturaleza (sección 3.2.2.1) únicamente toma valores positivos. Los resultados de aplicar esta transformación se ejemplifican para el caso de la varianza en las figuras 5.2 y 5.3.

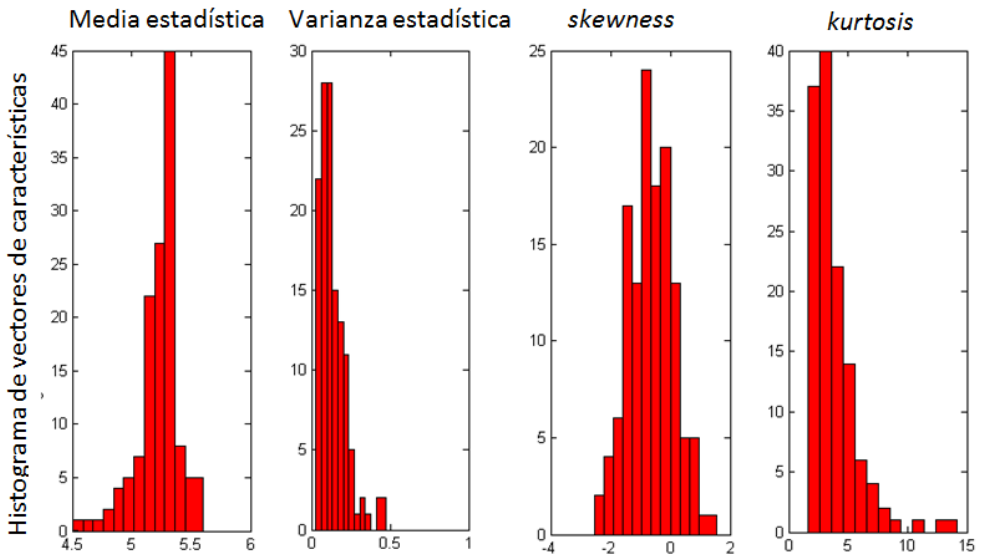


Figura 5.1: Histograma de los estadísticos de la entropía sobre el modelo de música.

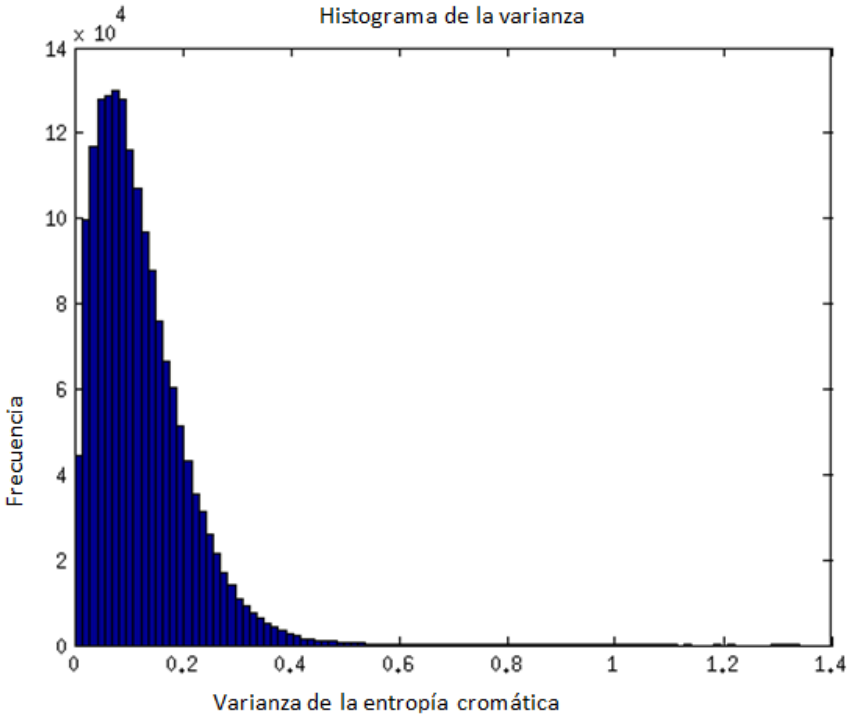


Figura 5.2: Distribución estadística de la varianza sobre datos de música.

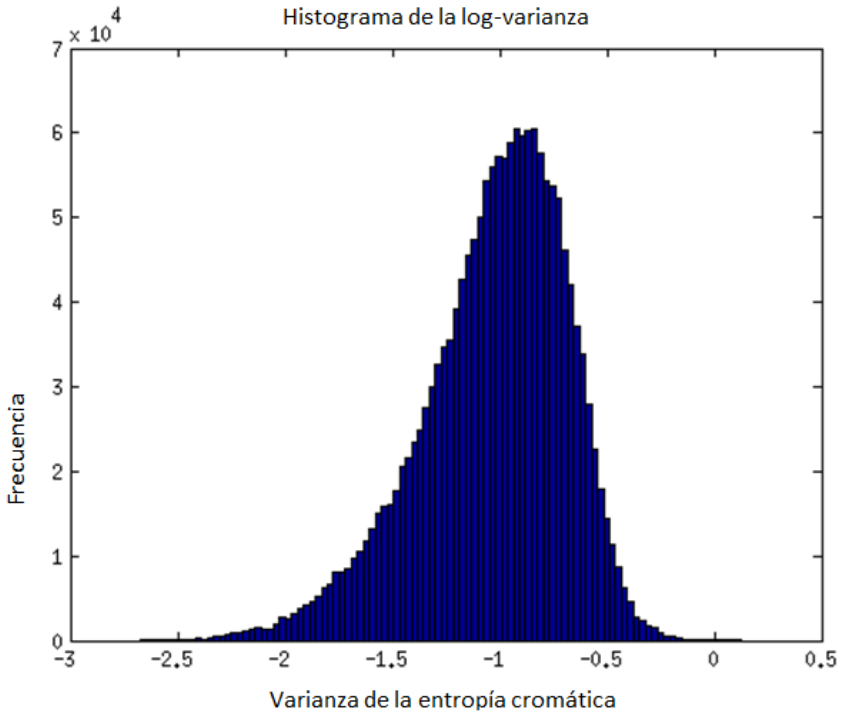


Figura 5.3: Distribución estadística del logaritmo de la varianza sobre datos de música.

Asímismo, experimentalmente se comprueba cómo los resultados del sistema que utilizan el logaritmo de ambas variables reducen globalmente el EER sobre cada clase, por lo que los resultados que se han seleccionado en este capítulo se han obtenido trabajando directamente con vectores de características formados por los siguientes parámetros estadísticos de la entropía: media, logaritmo decimal de la varianza, *skewness* y logaritmo decimal de la *kurtosis*.

5.1.3. Análisis de resultados

A continuación se analiza el rendimiento de este sistema por separado, es decir, antes de ser considerado como parte de una fusión de sistemas. Se obtiene el valor de ventana óptimo para el filtro de *scores* a partir del EER aplicado a los tracks 11 al 15, se calcula el EER de los tracks 16 al 20 con el valor de ventana óptimo y se calculan asimismo las matrices de confusión y los valores de precisión por clase, tal y como se ha expuesto para el sistema de referencia.

5.1.3.1. Valores óptimos del filtro de medias

De manera análoga al sistema anterior, se busca el valor óptimo de ventana para el filtro de medias como aquel que minimiza el EER por detector sobre el subconjunto de desarrollo (tracks 11 al 15). En este caso, como se ha avanzado en la sección 5.1.1 se estudian dos sistemas que siguen el mismo esquema pero con diferentes parámetros para el UBM y la adaptación MAP. La gráfica resultante para cada uno de estos se presenta en las figuras 5.4 y 5.5.

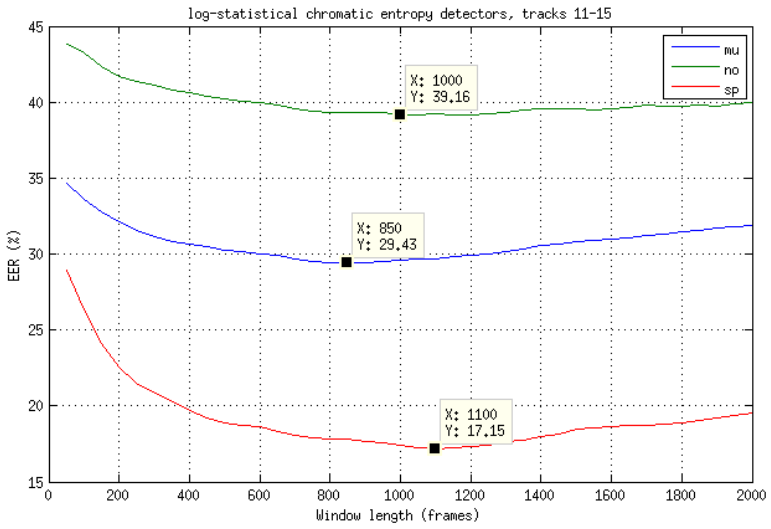


Figura 5.4: Análisis de EER sobre tracks 11 al 15, sistema basado en entropía cromática con parámetros del sistema de partida.

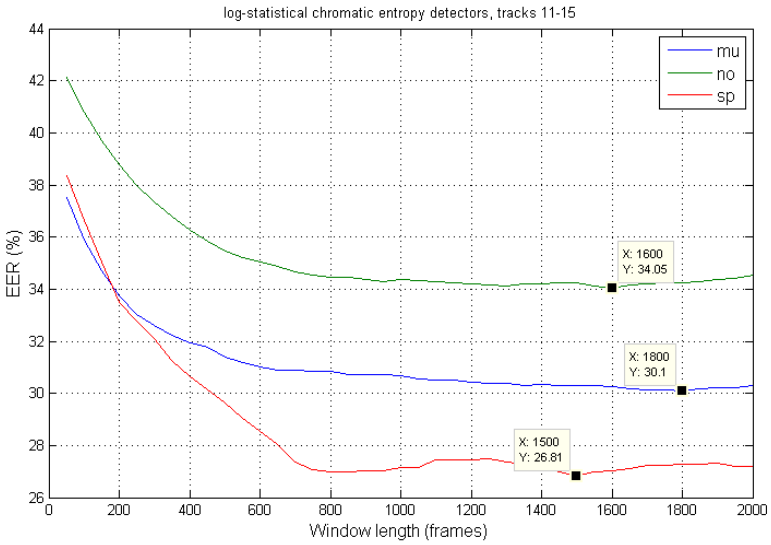


Figura 5.5: Análisis de EER sobre tracks 11 al 15, sistema basado en entropía cromática con parámetros del sistema de referencia.

A partir de las gráficas obtenidas se selecciona el valor de ventana óptimo, el cual se recoge para cada caso en las siguientes tablas (5.1 y 5.2):

Longitud optima del filtro de medias (número de tramas)	850	1000	1100
EER	29.43	39.16	17.15
Clase acústica	música (mu)	ruido (no)	voz (sp)

Tabla 5.1: Valores de ventana óptimos sobre el sistema basado en entropía cromática con parámetros del sistema de partida.

Longitud óptima del filtro de medias (número de tramas)	1800	1600	1500
EER	30.1	34.05	26.61
Clase acústica	música (mu)	ruido (no)	voz (sp)

Tabla 5.2: Valores de ventana óptimos sobre el sistema basado en entropía cromática con parámetros del sistema de referencia.

De estos resultados obtenidos destaca los valores de ventana óptimos obtenidos para el sistema entrenado con parámetros del sistema de referencia, ya que un filtro de 1800 tramas sobre este sistema implica aplicar filtros de 18 segundos. Por otro lado, cabe destacar que el EER obtenido para la clase de voz en el caso del sistema entrenado con parámetros del sistema de partida es casi 10 puntos menor que en el otro caso.

5.1.3.2. Rendimiento de detección

Una vez seleccionada la longitud óptima del filtro se calcula el EER de los tracks de test de manera análoga a como se ha trabajado con el sistema de referencia. Por ello, no solamente se analiza el EER final obtenido con el valor de ventana fijado en desarrollo, sino que en el caso del conjunto de test(tracks 16 al 20) se analiza cual sería el valor óptimo de filtro sobre este conjunto de datos y si el mejor valor de EER distaría mucho del resultado obtenido. Se muestran en primer lugar los resultados de trabajar con los parámetros del sistema de partida, esto es, un UBM generado con 5 iteraciones *k-means*, 10 de ML y modelos de clases adaptados con 10 iteraciones MAP. Para facilitar la comprensión de los resultados estos primeros parámetros del sistema basado en estadísticos de la entropía cromática se considerará que constituyen el sistema A mientras que el sistema ajustado con los parámetros de entrenamiento del sistema de referencia se va a mencionar como sistema B. Las DET que se muestran son los resultados del sistema sobre los tracks de test para el tamaño de ventana fijado en desarrollo (figuras 5.6 y 5.7), mientras que como resultado del análisis del tamaño de ventana sobre estos tracks para ver si este rendimiento es parecido al que se obtendría usando el tamaño de ventana óptimo fijado en desarrollo se muestran las gráficas 5.8 y 5.9.

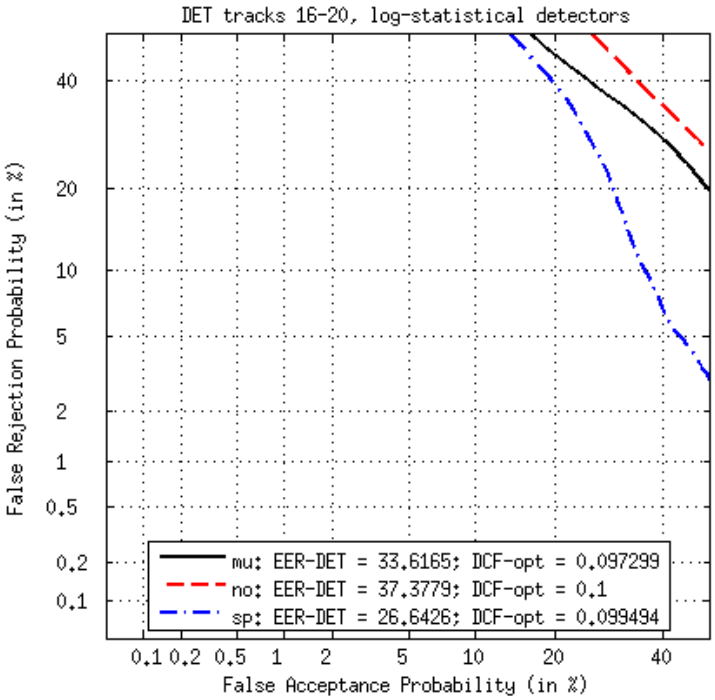


Figura 5.6: Curvas DET y EER por detector obtenido sobre tracks de test (16 al 20), sistema A.

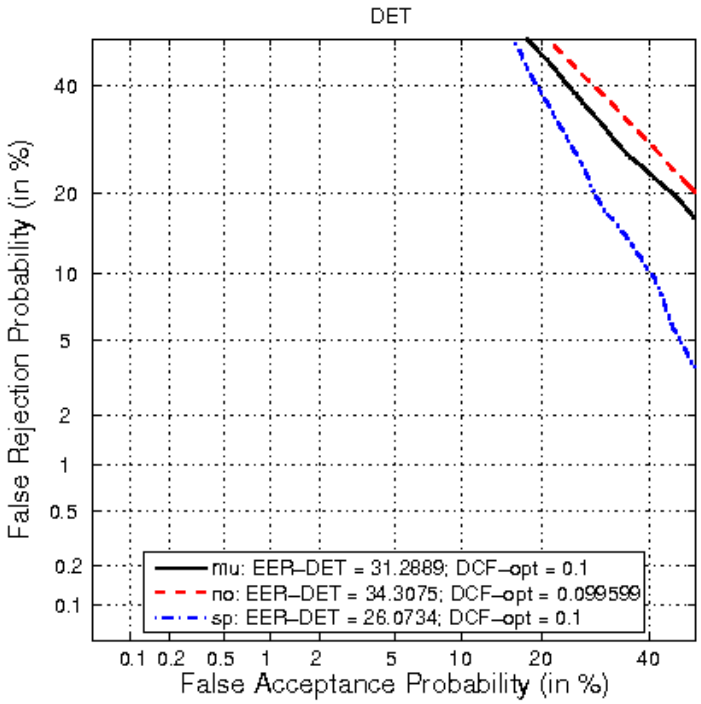


Figura 5.7: Curvaa DET y EER por detector obtenido sobre tracks de test (16 al 20), sistema B.

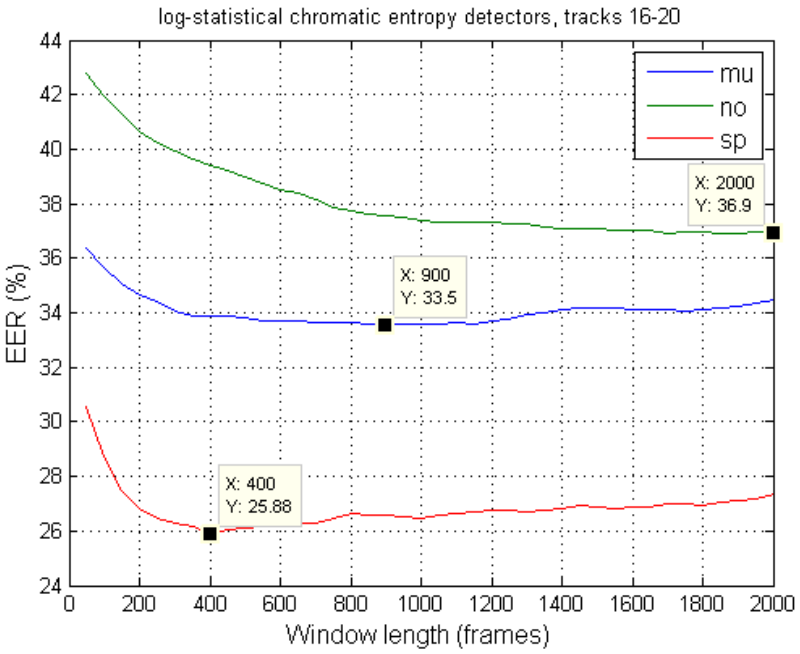


Figura 5.8: Análisis EER por detector para distintos valores del filtro de medias sobre tracks 16 al 20, sistema A.

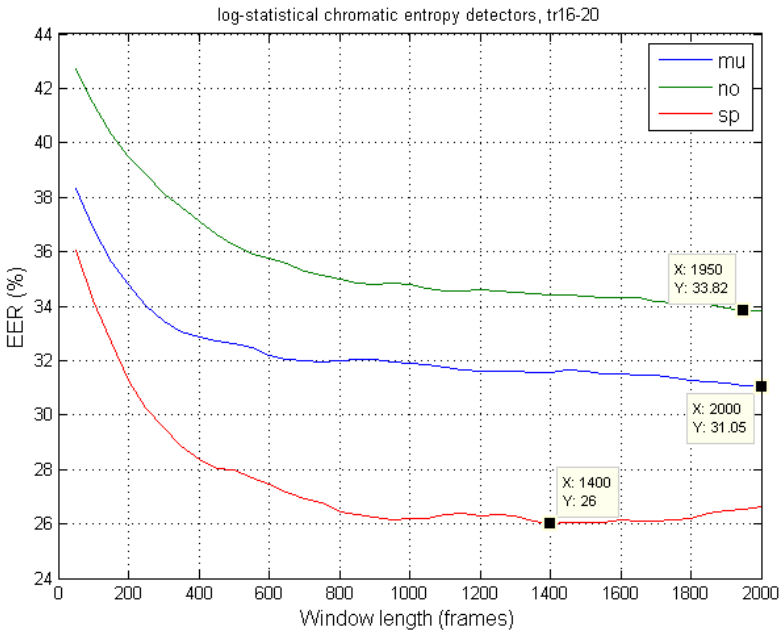


Figura 5.9: Análisis EER por detector para distintos valores del filtro de medias sobre tracks 16 al 20, sistema B.

De estos resultados cabe destacar que en ambos sistemas la clase que mejor se detecta es la voz (es la que menor EER tiene), seguida de la música y del ruido. No obstante, los resultados obtenidos sobre cada una de las tres clases muestran en general, que la precisión del

sistema basado en estadísticos de la entropía cromática no supera los obtenidos con el sistema de referencia.

5.1.3.3. Matrices de confusión

A continuación se refleja el rendimiento del sistema a nivel de matrices de confusión y valor de precisión global y por clases, para lo cual se hace uso de un etiquetado homogéneo (figura 2.4) lo que supone trabajar con ocho clases (figura 4.1). Al igual que se ha detallado para el sistema de referencia, se muestra la matriz de confusión resultante en escala de grises (figura 5.10 para el sistema A y figura 5.11 para el sistema B) y la tabla que refleja la precisión para cada una de las clases (tabla 5.3 para el sistema A y tabla 5.4 para el sistema B).

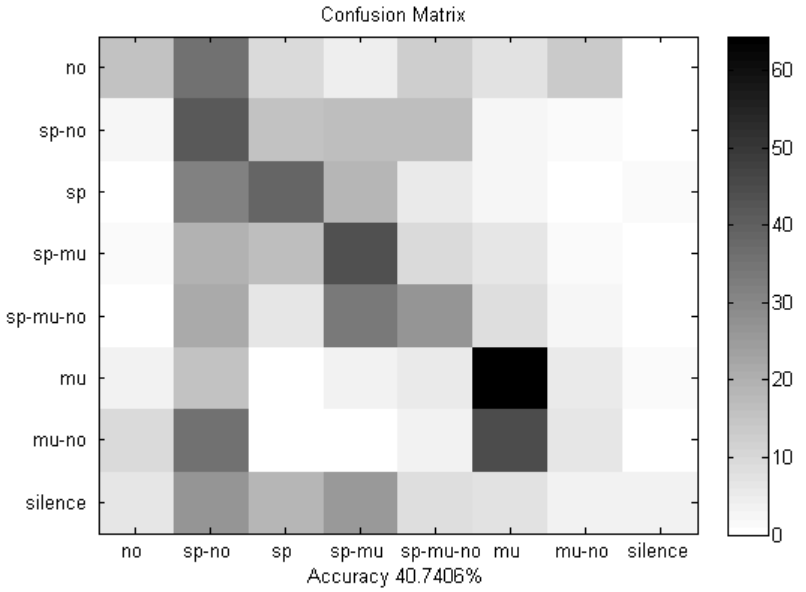


Figura 5.10: Matriz de confusión del sistema A (estadísticos de la entropía) sobre tracks 16 al 20.

En líneas generales, los resultados demuestran que el sistema de segmentación basado en estadísticos de la entropía que mejores resultados ofrece es aquel que toma los mismos parámetros que el sistema de referencia, esto es mayor número de mezclas pero con menos adaptaciones del modelo. La diferencia a nivel de EER no es muy significativa, ya que en el peor de los casos se da una variación de 2.7 puntos de EER para las clases de música y ruido, y la desviación en el nivel de precisión *accuracy* obtenido de las matrices de confusión es menor del 5 %. Por otro lado, el análisis detallado de precisión por clases permite comprender (tal y como era el objetivo) un poco mejor cómo está funcionando el sistema, y cuales son las clases que mejor detecta y las que peor. En este caso, cabe destacar que si bien la precisión total del sistema A apenas alcanza el 40 %, los resultados de detección de la clase música son mucho más satisfactorios, siendo la única clase cuyos resultados son notablemente mejores que el valor de precisión total. En cuanto al sistema B, sucede algo similar aunque en este caso el mayor valor de precisión, que también se da para la clase música, apenas supera el 53 %.

Clase	**no**	**sp-no**	**sp**	**sp-mu**	**sp-mu-no**	**mu**	**mu-no**	**silence**
Precisión (%)	16.05	42.20	38.7	43.94	26.4	64.36	6.6	0.32

Tabla 5.3: Valores de precisión del sistema A.

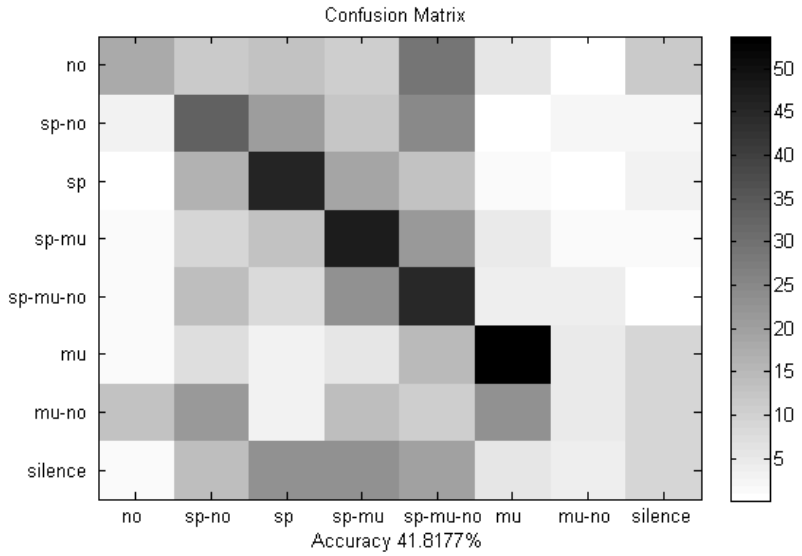


Figura 5.11: Matriz de confusión del sistema B (estadísticos de la entropía) sobre tracks 16 al 20.

Clase	**no**	**sp-no**	**sp**	**sp-mu**	**sp-mu-no**	**mu**	**mu-no**	**silence**
Precisión (%)	18.07	32.89	45.53	47.42	44.62	53.76	4.9	8.95

Tabla 5.4: Valores de precisión del sistema B.

5.1.4. Fusión con el sistema de referencia

Como ya se introdujo previamente, uno de los principales objetivos de tratar con el sistema de características cromáticas era el de fusionarlo con el sistema de referencia, aunque cabe destacar que en el estudio previo del sistema basado en estadísticos de la entropía cromática se han encontrado ciertas mejoras (como es el caso de emplear el logaritmo para dos características) que mejoran los resultados con respecto de un sistema básico, entendiendo como básico que toma las cuatro características iniciales (media, varianza, *skewness* y *kurtosis*) sin aplicar logaritmo. De entre los diferentes tipos de fusión expuestos en el capítulo de estado del arte, en este proyecto se ha escogido trabajar con la fusión a nivel de características y a nivel de scores.

5.1.4.1. Resultados de fusionar a nivel de características

La fusión a nivel de características implica utilizar un único sistema basado en vectores de características compuestos por características de distinta naturaleza (por ejemplo, tímbrica y cromática). Antes de fusionar los sistemas es especialmente importante analizar el rendimiento

por separado de cada sistema para una configuración igual , ya que si uno de ellos funciona muy mal para esa configuración, la fusión seguramente no ofrezca ninguna ganancia. Por lo tanto, es importante encontrar una configuración común para la cual el rendimiento de ambos sistemas por separado no esté muy lejos de su óptimo. Por este motivo, se calculó en el apartado anterior el rendimiento de un sistema basado en estadísticos de la entropía cromática entrenado no sólo con los parámetros del sistema de partida sino también con los mismos parámetros del sistema de referencia.

Para lograr la fusión, se han combinado los vectores de características obtenidos en cada uno de los dos sistemas anteriores. Una vez que la fusión se ha realizado con éxito, el sistema se entrena y evalúa siguiendo el mismo procedimiento que se ha seguido para cada uno de los dos sistemas anteriores, y los resultados obtenidos se exponen a continuación.

5.1.4.1.1. Valores óptimos del filtro de medias

Dado que se trata de un sistema con un nuevo conjunto de características, se entrena un nuevo UBM y se adapta en función de los vectores de características de cada clase, dando lugar a los modelos de cada clase. Una vez obtenidos los *scores* de detección de audio de vectores de características sobre los modelos, se calcula experimentalmente el valor de ventana óptimo para aplicar a los datos de test (tracks 16 a 20) en la fase de filtrado (suavizado de los resultados) de la tarea de segmentación. La gráfica resultante del análisis de ventana (figura 5.12) y la tabla que detalla la localización de dichos valores (tabla 5.12) se muestran a continuación:

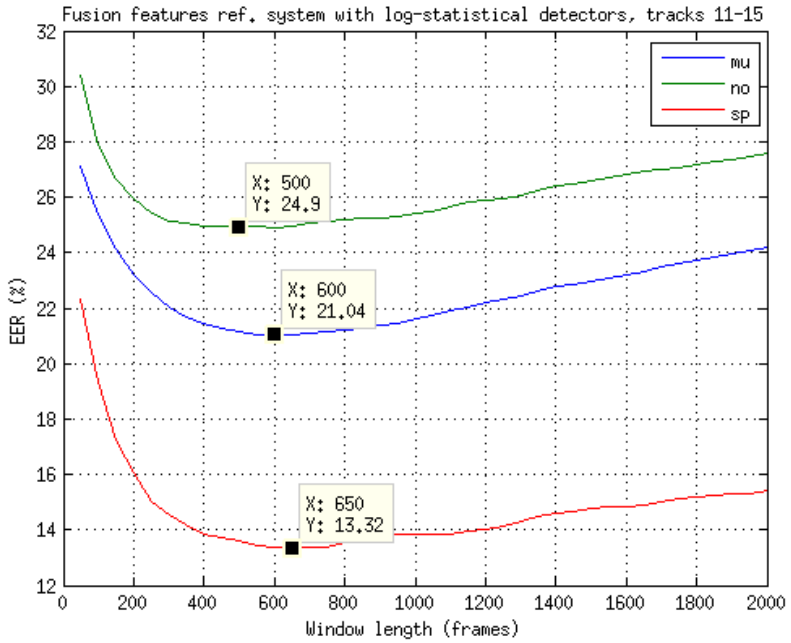


Figura 5.12: Análisis de EER sobre tracks 11 al 15, sistema de fusión con estadísticos de la entropía.

Longitud óptima del filtro de medias (número de tramas)	600	500	650
EER	21.04	24.9	13.32
Clase acústica	música (mu)	ruido (no)	voz (sp)

Tabla 5.5: Valores de ventana óptimos sobre el sistema de fusión con estadísticos de la entropía.

En este caso, los valores óptimos de ventana para cada una de las clases se encuentran en 6, 5 y 6.5 segundos. Si bien los valores de EER mínimos que permiten obtener el valor de ventana óptimo se aventuran mejores que los obtenidos para el sistema basado en características de la entropía cromática, para una comparación rigurosa se analizarán los resultados obtenidos sobre el conjunto de datos de test.

5.1.4.1.2. Rendimiento de detección

Una vez seleccionada la longitud óptima del filtro se calcula el EER de los tracks de test de manera análoga a como se ha trabajado en los casos anteriores. En la figura 5.13 se muestra el valor de EER obtenido sobre el conjunto de test con el valor de ventana escogido sobre los datos de desarrollo así como en la figura 5.14 se ofrecen los resultados de analizar el EER para diferentes valores de ventana.

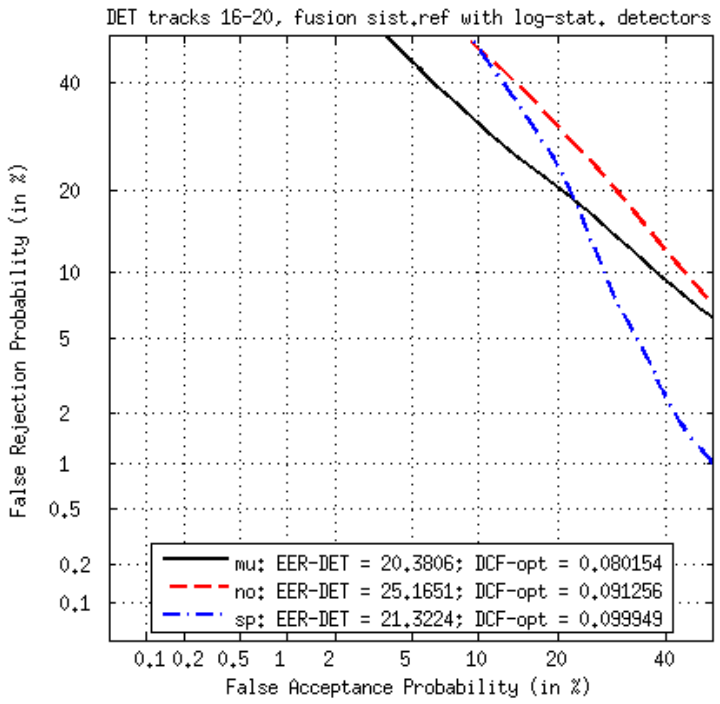


Figura 5.13: Curvas DET y EER obtenidos sobre tracks de desarrollo (16 al 20).

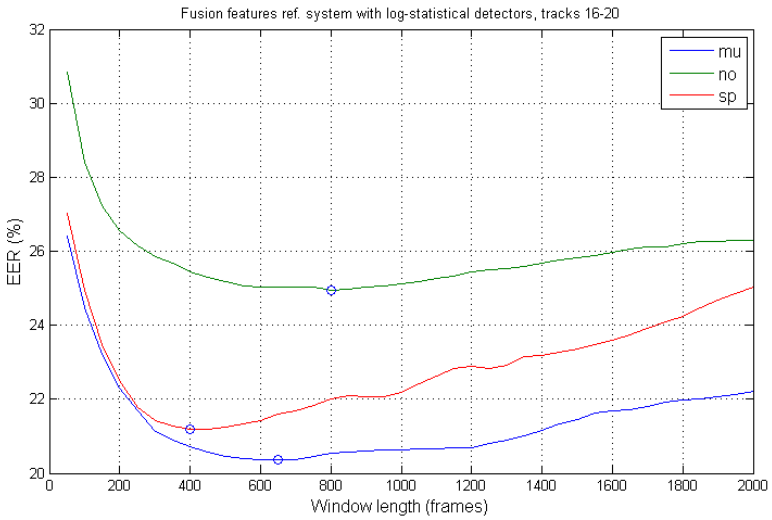


Figura 5.14: Análisis EER para distintos valores del filtro de medias sobre tracks 16 al 20.

En términos de EER, los resultados obtenidos del sistema de fusión evaluado sobre los tracks 16 al 20 son mejores que los obtenidos sobre el sistema cromático pero peores que los del sistema de referencia, por lo que en esta línea no se ha conseguido una mejora de resultados. No obstante, esto podría ser de algún modo un resultado esperado ya que el rendimiento de ambos sistemas era bastante diferente, siendo uno de ellos significativamente mejor que el otro, por lo que se puede argumentar que partiendo de sistemas con eficiencias tan diferentes, el resultado en rendimiento sea un valor aproximado a la media de ambos. No obstante, este análisis de EER no es del todo exhaustivo por lo que se continúa la evaluación del sistema con las matrices de confusión.

5.1.4.1.3. Matrices de confusión

En el siguiente apartado se refleja el rendimiento del sistema evaluado con matrices de confusión y el nivel de precisión de cada clase. Por un lado, la figura 5.15 muestra la matriz de confusión resultante junto con la precisión global. Por otro lado, la tabla 5.6 refleja la precisión del sistema por cada una de las ocho clases.

A partir de los resultados se puede apreciar que el valor de precisión obtenido es mejor que el obtenido con el sistema basado en estadísticos de la entropía pero algo peor que el resultado obtenido con el sistema de referencia, al igual que pasaba al medir en términos de EER. Además, no existe ninguna clase cuyo valor de precisión supere a los obtenidos con cada uno de los sistemas por separado. El sistema basado en estadísticos de la entropía en general funciona peor que el sistema basado en coeficientes MFCC-SDC, por lo que si se comparan los valores de precisión obtenidos en la fusión para cada una de las clases con respecto a los obtenidos en el sistema basado en estadísticos de la entropía se puede ver una notable mejoría en los resultados de la fusión.

Sin embargo, la comparativa que se puede llevar a cabo con el sistema de referencia resulta interesante. Por un lado, se puede ver cómo la precisión obtenida en las clases de *sp* y *sp-mu* es peor en la fusión en comparación al sistema de referencia (88 % frente a 70 % en la clase voz y 73 % frente a un disminuido 52 % en la clase voz-música respectivamente). Por otro lado, el valor de precisión para el resto de clases (que en el caso del sistema de referencia no ofrecían una precisión muy elevada como las dos clases anteriores) ha mejorado significativamente.

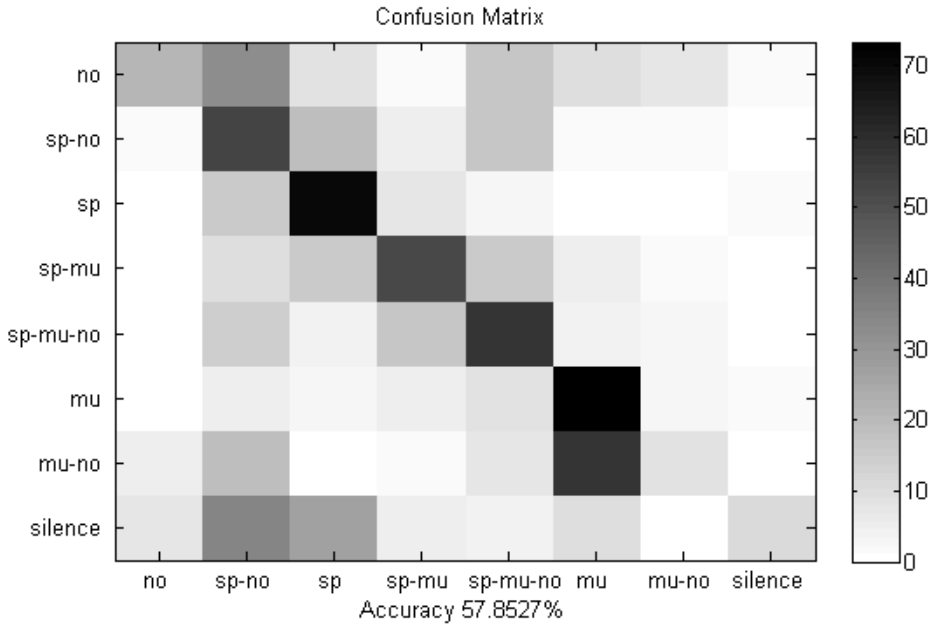


Figura 5.15: Matriz de confusión de la fusión con estadísticos de la entropía.

Clase	**no**	**sp-no**	**sp**	**sp-mu**	**sp-mu-no**	**mu**	**mu-no**	**silence**
Precisión (%)	21.06	53.7	70.95	52.44	57.54	73.30	8.41	3.36

Tabla 5.6: Valores de precisión de la fusión de caract. del sistema de referencia con los estadísticos de la entropía.

5.1.4.2. Resultados de fusionar a nivel de scores con reglas fijas

La fusión a nivel de scores resulta de gran utilidad cuando las características no son tan compatibles entre sí. En este caso, la estrategia a seguir consiste en combinar los scores obtenidos en cada sistema para cada una de las tramas, y con estos nuevos resultados realizar la segmentación del audio a partir de un umbral. Dado que los scores de cada sistema por separado han sido calibrados (transformando los scores de forma que el umbral esté siempre en 0), esta fusión de scores de tipo suma reforzará valores que resulten positivos en ambos casos, y viceversa con los negativos; y, en el caso de segmentos que hayan recibido una puntuación negativa en un sistema y positiva en otra, la fusión ofrecerá un resultado final positivo o negativo (lo que marca la pertenencia o no a cada clase) en función del sistema cuyo resultado haya sido más alto en valor absoluto. En esta línea cabe destacar que cada una de las puntuaciones (previas a la fusión) se calibra individualmente para cada detector y es en el último paso antes de segmentar el audio cuando ambos resultados se combinan. Por lo tanto, para realizar esta tarea sólo es preciso añadir una fase al esquema de trabajo en la etapa de evaluación. Tomando como referencia el esquema mostrado para el sistema de referencia sobre las fases de desarrollo y test (figura 4.3), se precisaría incluir una etapa de proceso adicional sobre dicho esquema, tal y como se muestra en la figura 5.16.

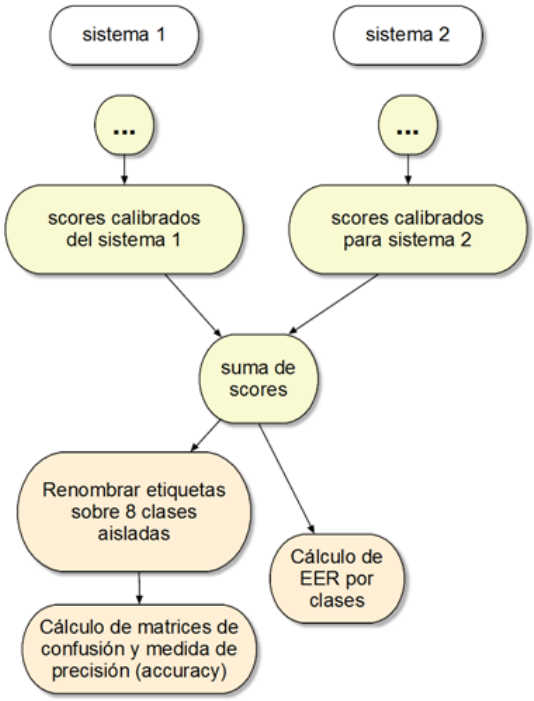


Figura 5.16: Fragmento de diagrama del sistema de fusión de scores.

En este caso, los valores del filtro de medias empleados para cada sistema son los obtenidos sobre el conjunto de tracks de desarrollo para cada caso. Finalmente, se muestran los resultados a nivel de matrices de confusión y precisión.

5.1.4.2.1. Matrices de confusión

En este apartado se va a medir el rendimiento del sistema fusionado a nivel de scores con matrices de confusión y nivel de precisión global y por clases, para lo cual se hace uso de un etiquetado homogéneo (figura 2.4) lo que supone trabajar con ocho clases (figura 4.1). La figura 5.17 muestra la matriz de confusión resultante, en la cual se puede apreciar a pie de figura cómo el nivel de precisión global ha mejorado con respecto al obtenido con cualquiera de los dos sistemas por separado y con respecto a la fusión a nivel de características. Por otro lado, la tabla 5.7 refleja la precisión del sistema para cada una de las ocho clases.

Contrastando los resultados a nivel de precisión, en líneas generales se puede ver cómo la fusión a nivel de *scores* ofrece mejores resultados que la fusión a nivel de características. En primer lugar, el valor de precisión total ha mejorado con respecto al obtenido para el sistema de referenca (60.46 %) y alcanzado el valor máximo hasta el momento (un 61,12 %). Por otro lado, el valor de precisión por cada clase ha mejorado en líneas generales en comparación a los resultados obtenidos para el sistema de fusión a nivel de características, y sigue siendo significativamente mayor (en la mayoría de las clases) que para cada uno de los dos sistemas por separado. Por lo tanto, cabe concluir que el mejor modo de combinar estos dos sistemas cuyos vectores de características son significativamente diferentes, es entrenar cada sistema por separado y combinar finalmente las puntuaciones obtenidas (una vez calibradas) para tomar la decisión final. En este caso, se aprecia cómo la fusión de scores permite optimizar cada sistema por separado, mientras que para la fusión de características es preciso buscar una nueva

configuración óptima, la cual puede ser difícil de encontrar debido a que la complejidad del modelo aumenta al aumentar el número de características.

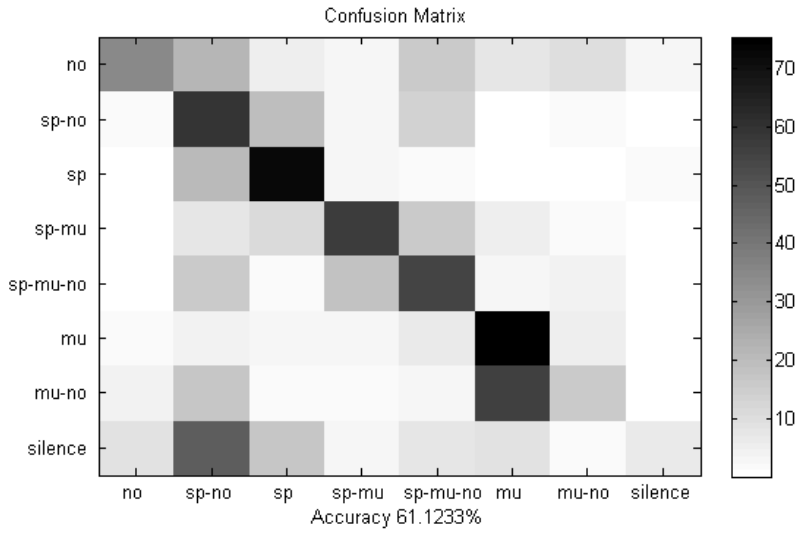


Figura 5.17: Matriz de confusión de la fusión de scores con el sistema de referencia.

Clase	**no**	**sp-no**	**sp**	**sp-mu**	**sp-mu-no**	**mu**	**mu-no**	**silence**
Precisión (%)	34.51	59.73	72.2	57.24	54.99	75.4	15.68	6.88

Tabla 5.7: Valores de precisión de la fusión de scores del sistema de referencia con los estadísticos de la entropía.

5.2. Sistema basado en agrupación por octavas y subbandas

Los resultados del sistema que se detalla a continuación trabajan con una agrupación por octavas y subbandas fusionadas a nivel de características, lo que equivale a trabajar con vectores de 22 características cromáticas: 10 por agrupación de octavas, y 12 por agrupación de subbandas.

5.2.1. Estructura del sistema

El sistema de segmentación de audio basado en agrupación por octavas y subbandas desarrollado en este proyecto presenta el mismo esquema de trabajo que el sistema de referencia, siendo la única diferencia el algoritmo de extracción de características con el que trabaja cada uno. El estudio detallado sobre ambos sistemas y los resultados que generan entrenados con diferentes ajustes del modelo (número de iteraciones MAP, número de mezclas del UBM, etc) se incluyen en el anexo B. Tras estudiar los resultados de los diferentes sistemas en términos de EER, matrices de confusión y precisión por clases se ha apreciado la complementariedad de ambos conjuntos, y es por ello que en este apartado del capítulo se van a incluir los resultados del sistema basado en 22 características, las cuales provienen de una agrupación del audio por octavas y por subbandas, tal y como se ha detallado previamente en el capítulo del estado del arte (sección 3.2.2.2).

5.2.2. Análisis de resultados

Los resultados que se exponen a continuación provienen de una selección previa de las configuraciones que mejores resultados han ofrecido en la etapa de evaluación. Sobre el conjunto de posibilidades de estudio, que es infinitamente amplio, se ha seleccionado un subconjunto de trabajo cuyos parámetros coinciden con los parámetros usados en los sistemas anteriores. En total, se han planteado cuatro escenarios diferentes que combinan diferentes parámetros de entrenamiento y de adaptación, los cuales se numeran a continuación:

- UBM generado con 1024 mezclas, inicializado con 1 iteración del algoritmo *k-means* seguida de 5 iteraciones ML, y una iteración del algoritmo de adaptación MAP (sistema A).
- UBM generado con 1024 mezclas, 5 iteraciones del algoritmo *k-means* y seguida de 10 iteraciones ML. Modelos con mayor grado de adaptación, por trabajar con 10 iteraciones MAP (sistema B).
- UBM generado con 128 mezclas, inicializado con 1 iteración del algoritmo *k-means* seguida de 5 iteraciones ML, y una iteración del algoritmo de adaptación MAP.
- UBM generado con 128 mezclas, 5 iteraciones del algoritmo *k-means* y seguida de 10 iteraciones ML. Modelos con mayor grado de adaptación, por trabajar con 10 iteraciones MAP.

De los cuatro sistemas estudiados, se ha observado que funcionan mejor aquellos que trabajan con mayor número de mezclas, en este caso, 1024, que han sido nombrados como sistemas A y B. Por lo tanto, los resultados que se van a reflejar en este capítulo van a hacer referencia a los dos sistemas entrenados con 1024 mezclas en sus diferentes combinaciones.

5.2.2.1. Valores óptimos del filtro de medias

A continuación se muestran las gráficas que reflejan el valor de ventana óptimo que habría que escoger para cada sistema así como una tabla resumen con dichos valores. De manera análoga a como se hizo en el análisis de resultados del sistema basado en estadísticos de la entropía cromática, el pie de cada una de las figuras va a hacer alusión a los sistemas A y B (ambos entrenados con 1024 mezclas) pero con diferentes parámetros de entrenamiento del UBM y la adaptación MAP.

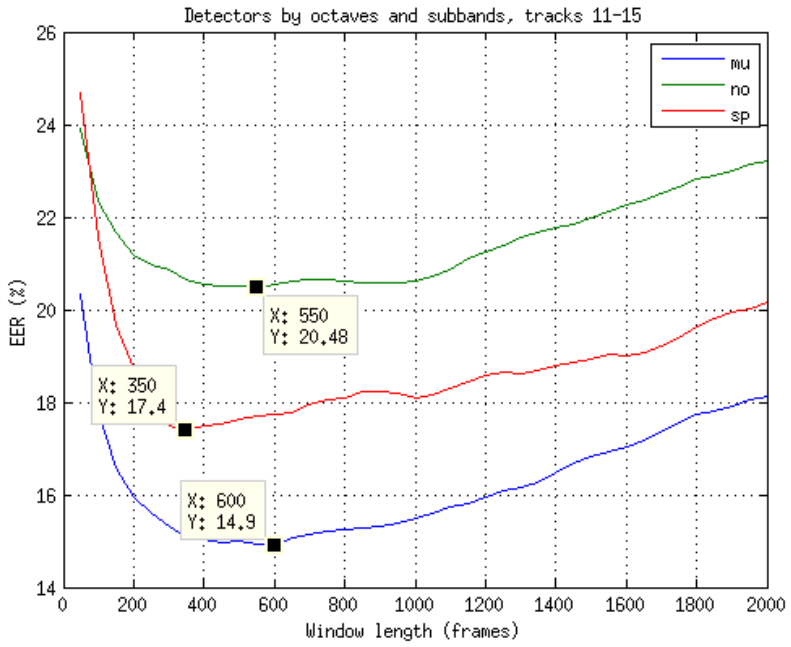


Figura 5.18: Análisis de EER sobre tracks 11 al 15 para sistema A entrenado con agrupación por octavas y subbandas

Como se ha podido observar en las gráficas 5.18 y 5.19, los valores de EER obtenidos sobre cada sistema (A y B) difieren sólo ligeramente, y sin embargo, ambos ofrecen valores de EER reducidos en comparación a los obtenidos con el sistema basado en estadísticos de la entropía cromática. Así mismo, cabe resaltar que los valores de ventana óptimos seleccionados sobre cada clase y sistema son en general valores bajos, y están situados en un valor medio de 5 segundos, valor frecuente en este proyecto. Si el tiempo de ventana es pequeño esto puedo indicar que las fluctuaciones no deseadas de la falsa presencia de clases son de muy corta duración y por lo tanto pueden ser suavizadas con pocos segmentos.

Longitud óptima del filtro de medias (número de tramas)	600	550	350
EER	14.9	20.48	17.4
Clase acústica	música (mu)	ruido (no)	voz (sp)

Tabla 5.8: Valores de ventana óptimos sobre el sistema A basado en agrupación por octavas y subbandas

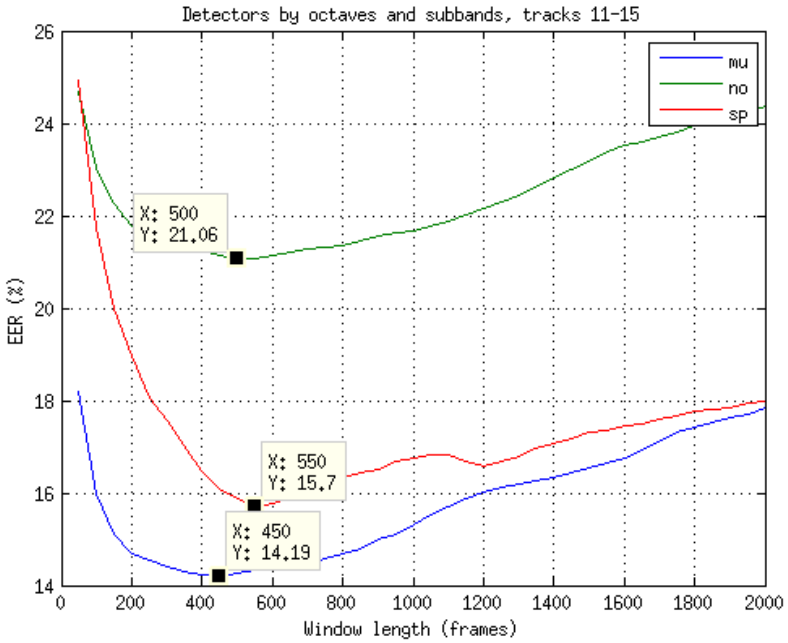


Figura 5.19: Análisis de EER sobre tracks 11 al 15 para sistema B entrenado con agrupación por octavas y subbandas

Longitud óptima del filtro de medias (número de tramas)	450	500	550
EER	14.19	21.06	15.7
Clase acústica	música (mu)	ruido (no)	voz (sp)

Tabla 5.9: Valores de ventana óptimos sobre el sistema B basado en agrupación por octavas y subbandas

5.2.2.2. Rendimiento de detección

Una vez obtenidos los valores de ventana óptimos, se mide el rendimiento del sistema generado sobre los datos de test, y los resultados se muestran a continuación para ambos sistemas A y B. Como se ha llevado a cabo con los sistemas anteriores, por un lado se muestran las curvas DET obtenidas con los valores de ventana óptimos (figuras 5.20 y 5.22), y adicionalmente se comprueba si ha habido un sobreajuste del valor de ventana escogido (figuras 5.21 y 5.23) mediante un análisis de ventana sobre el conjunto de datos de test análogo al realizado en el apartado anterior. Finalmente se contrastan los resultados.

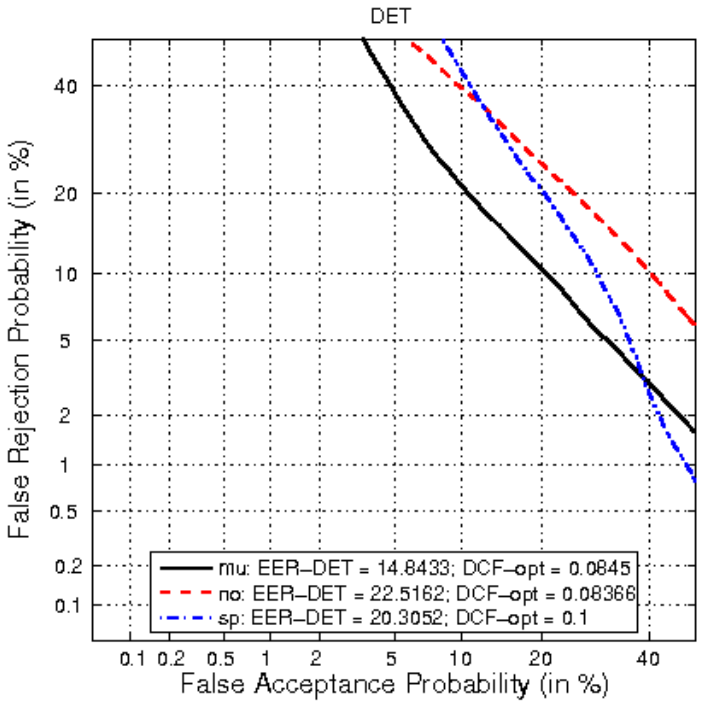


Figura 5.20: Curvas DET y EER obtenidos para los detectores del sistema A de agrupación por octavas y subbandas, tracks 16-20

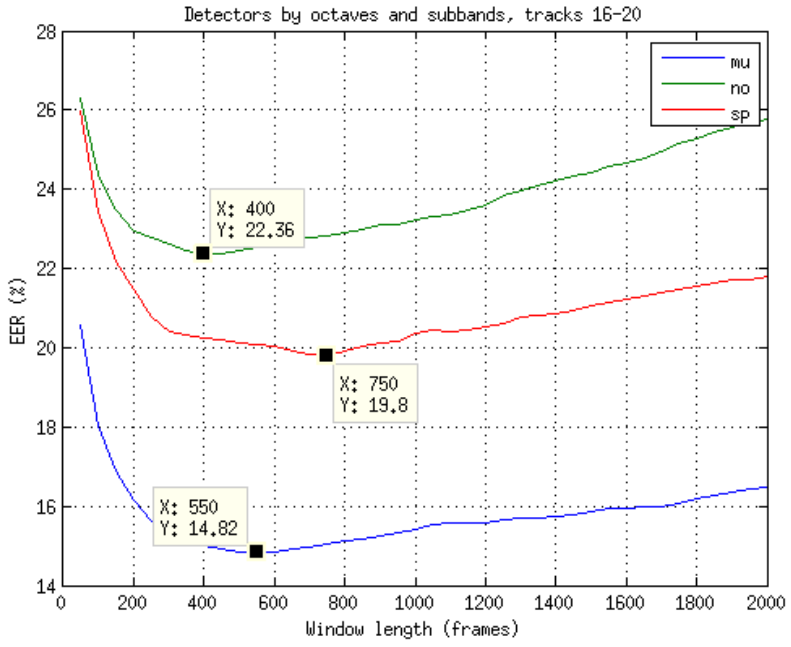


Figura 5.21: Análisis EER sobre sistema A de agrupación por octavas y subbandas, tracks 16-20

A raíz de los resultados, se puede apreciar cómo el valor de ventana escogido para la música es prácticamente coincidente con el valor óptimo del conjunto de test, por lo que el sobreajuste se

puede decir que es casi nulo. En el caso del ruido, por ejemplo, el error obtenido es casi el mismo (apenas varía en dos decimas), y de un modo similar sucede con la clase voz, lo que quiere decir que el sistema discriminará con cierta robustez. En cualquier caso, sí se puede apreciar cómo la clase que el sistema es capaz de detectar y clasificar con menor error es la música, seguida de la voz y finalmente del ruido. Como sucedía con sistemas anteriores, el ruido siempre resulta la clase más difícil de detectar, reflejo de que es complejo obtener un modelo con capacidad de generalización para la base de datos proporcionada. Sin embargo, cabe resalta que el valor de EER de música es el mínimo alcanzado hasta el momento, al igual que sucede con el de ruido.

En segundo lugar se muestran los resultados del sistema B (figuras 5.22 y 5.23) que emplean el mismo sistema GMM-UBM pero con diferentes parámetros en el entrenamiento y adaptación de modelos que los empleados en el sistema A.

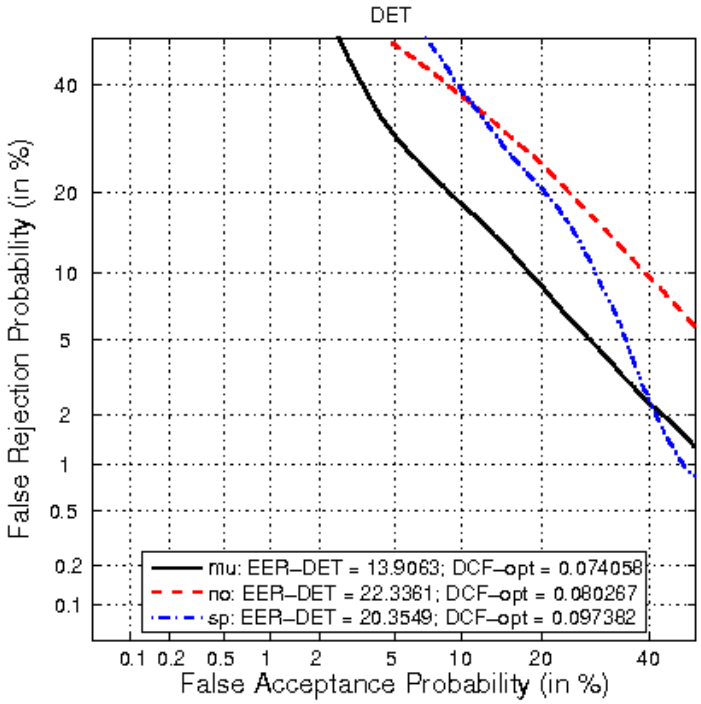


Figura 5.22: Curvas DET y EER obtenidos sobre sistema B de agrupación por octavas y sub-bandas, tracks 16-20.

De nuevo la clase que dicho sistema es capaz de detectar mejor es la música. Por otro lado, el análisis de ventana muestra poca variación de EER sobre diferentes valores de ventana por lo que a pesar de que los valores de EER obtenidos para la clase de voz sobre el conjunto de evaluación son algo peores que los obtenidos sobre los tracks 11 al 15 no tiene cabida hablar de sobreajuste, sino que podría deberse quizás a que los datos de voz contenidos en el conjunto de test son menos parecidos a los del conjunto de desarrollo y entrenamiento y por lo tanto se asemejará menos al modelo generado. En este punto, cabe resaltar cómo este fenómeno de la clase voz ya se ha producido en el sistema de referencia (el cual pasa de un 9.7 % de EER en los tracks empleados para calibrar a un 18 % sobre los datos de test) por lo que se podría generalizar que los datos de voz del conjunto 11 al 15 son demasiado similares a los datos de entrenamiento del modelo y/o que los datos de voz del conjunto 16 al 20 son mucho más diferentes y por eso el sistema los reconoce peor.

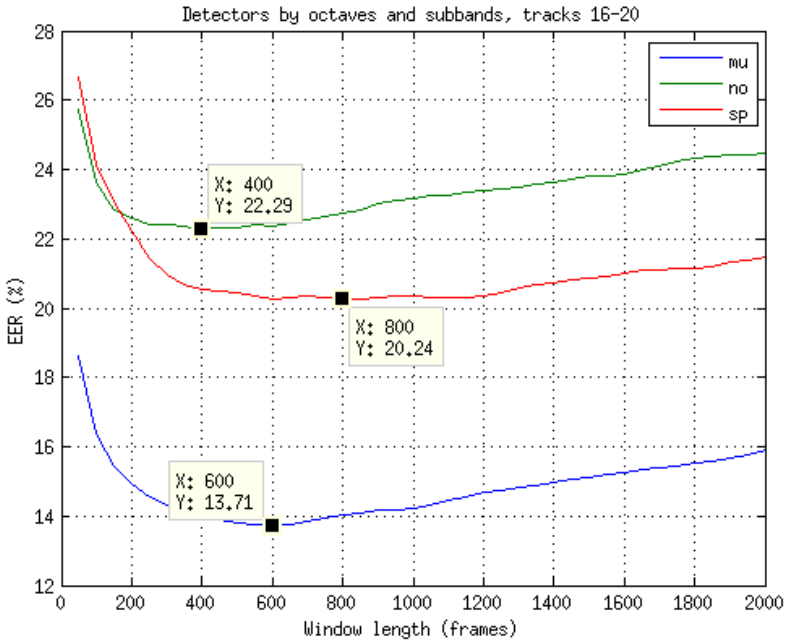


Figura 5.23: Análisis EER sobre sistema B de agrupación por octavas y subbandas, tracks 16-20.

En cualquier caso, si se comparan ambas configuraciones (sistemas A y B) obtenidas de un extractor de características de agrupación por octavas y subbandas se puede concluir que el rendimiento es casi idéntico medido en términos de EER, y bastante bueno en general en comparación al sistema basado en estadísticos de la entropía cromática. De ste modo, los resultados obtenidos confirman dos observaciones anteriormente realizadas. En primer lugar, los valores de EER obtenidos para la clase música y ruido son menores que los obtenidos para el sistema de referencia, mientras que los valores de EER para la detección de voz son mejores para el sistema basado en MFCC-SDC. Estos hechos guardan coherencia con la naturaleza con la que son entrenados cada sistema. Los sistemas basados en MFCC son ampliamente usados para reconocedores de voz por sus características óptimas que obtienen información representativa de la voz (si bien se pensó que al incluir características derivadas que miden variaciones temporales -SDC- se obtendrían mejoras en la clase música). Por otro lado, la agrupación de los vectores de características por subbandas está inspirada en la disposición armónica de la música, a la vez que la agrupación de los vectores de características por octavas vino inspirada por la disposición frecuencial (o falta de armonía) del ruido. Por lo tanto, podría seleccionarse un tipo de características distinto para cada detector en función del rendimiento conseguido para cada clase de audio: MFCC-SDC para voz, agrupación por octavas y subbandas para música y ruido.

5.2.2.3. Matrices de confusión

A continuación se evalúa el rendimiento del sistema con las matrices de confusión y el valor de precisión por clases. La figura 5.24 y la tabla 5.10 muestran los resultados del sistema A, mientras que la gráfica 5.25 y la tabla 5.11 hacen referencia al sistema B.

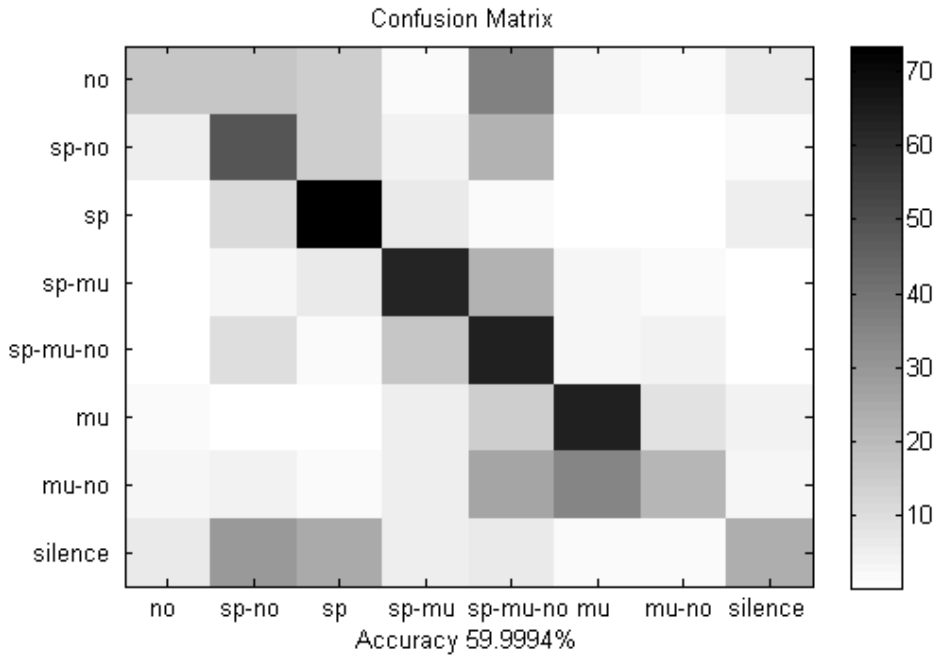


Figura 5.24: Matriz de confusión del sistema A basado en agrupación por octavas y subbandas.

Clase	**no**	**sp-no**	**sp**	**sp-mu**	**sp-mu-no**	**mu**	**mu-no**	**silence**
Precisión (%)	17.4	48.74	73.35	62.52	63.54	63.8	21.35	23.51

Tabla 5.10: Valores de precisión por clases del sistema A basado en agrupación por octavas y subbandas.

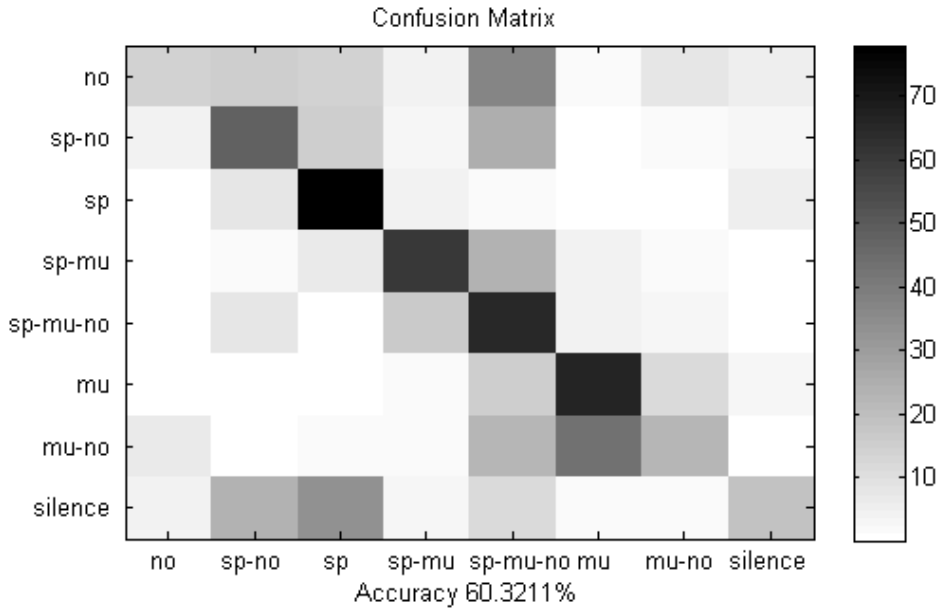


Figura 5.25: Matriz de confusión del sistema B basado en agrupación por octavas y subbandas.

Clase	**no**	**sp-no**	**sp**	**sp-mu**	**sp-mu-no**	**mu**	**mu-no**	**silence**
Precisión (%)	13.56	47.64	78.04	59.99	65.25	65.58	23.14	18.57

Tabla 5.11: Valores de precisión por clases del sistema B basado en agrupación por octavas y subbandas.

Al igual que reflejaba el EER, el rendimiento de los sitemas A y B es similar, y ambos rozan un valor de precisión del 60 %. De nuevo, las medidas de error reflejan que el sistema basado en agrupación por octavas y subbandas presenta mayor rendimiento que el sistema basado en estadísticos de la entropía cromática, en el cual, es posible que una codificación de la entropía de la cual se obtienen cuatro medidas estadísticas reduzca en gran medida la información presente en el cromagrama.

Ahora bien, si se comparan estos resultados con los obtenidos en el sistema de referencia (que trabajan con 56 características) se obtiene que con vectores mucho más cortos (de 22 características) los resultados son muy similares mientras que el coste computacional y el tiempo de cálculo se reducen. En concreto, se está comparando un valor de precisión del 60.4 % en el caso mejor del sistema de referencia con un 60.3 % en el caso mejor del sistema basado en agrupación por octavas y subbandas, empleando además el mismo número de gaussianas, esto es, 1024.

En cuanto a los resultados obtenidos por clases, la clase que mejor se reconoce con un valor de precisión mayor al 70 % en ambos sistemas A y B basados en agrupación por octavas y subbandas es la voz, al igual que sucedía en casos anteriores, y de nuevo la clase de ruido aislado, muy seguida del silencio y la clase música-ruido se detectan con muy poca precisión, siendo como ya se mencionó las clases con menos presencia en los datos de audio. En la última sección de este capítulo se realiza una comparativa más detallada.

5.2.3. Fusión con sistema de referencia

A continuación se van a combinar los dos sistemas con mejores resultados obtenidos en el presente proyecto, es decir, el sistema de segmentación de audio basado en coefientes MFCC-SDC con el sistema basado en agrupación por octavas y subbandas. Al igual que para el sistema basado en estadísticos de la entropía cromática, la fusión se va a realizar a dos niveles: de características y de scores.

5.2.3.1. Resultados de fusionar a nivel de características

Los siguientes resultados se obtienen de combinar los 56 coeficientes MFCC-SDC con las 22 características basadas en agrupación por octavas y subbandas. Con un total de 78 características por trama se ha entrenado un sistema de segmentación de audio cuyos resultados se muestran en los siguientes apartados.

En cuanto a la estructura del sistema, el modo de combinar las características sigue el mismo esquema de trabajo que la fusión de características con estadísticos de la entropía cromática (sección 5.1.4). Los vectores de características obtenidos se han usado para entrenar un sistema GMM-UBM, siendo los parámetros empleados para generar los modelos los mismos que se han empleado en el sistema de referencia, esto es:

- 1024 gaussianas por clase.

- UBM inicializado con una iteración del algoritmo k-means seguido de 5 iteraciones del algoritmo EM.
- Modelos adaptados con una iteración MAP y un valor de r=16.

5.2.3.1.1. Valores óptimos del filtro de medias

Siguiendo el mismo razonamiento expuesto en los siguientes apartados, se muestran en la figura 5.26 y en la tabla 5.12 los resultados del sistema para diferentes longitudes de ventana sobre los tracks 11 al 15.

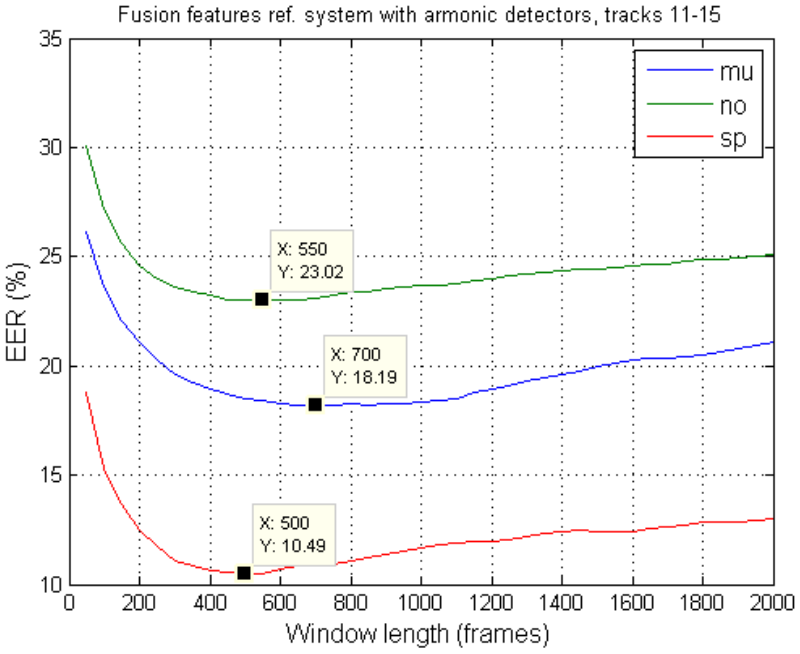


Figura 5.26: Análisis de EER sobre tracks 11 al 15 para diferentes valores de ventana, fusión a nivel de características con sistema de referencia.

Longitud óptima del filtro de medias (número de tramas)	700	550	500
Clase acústica	18.19	23.02	10.49
Clase acústica	música (mu)	ruido (no)	voz (sp)

Tabla 5.12: Valores de ventana óptimos sobre la fusión a nivel de características con el sistema de referencia.

En primer lugar, cabe destacar que los valores óptimos del filtro de medias toman valores similares a los obtenidos con otros sistemas. Por otro lado, los valores de EER que acompañan a estos valores pronostican un sistema de buenas prestaciones, similar o ligeramente mejorado respecto de cualquiera de los sistemas que constituyen los que intervienen en la fusión, por separado.

5.2.3.1.2. Rendimiento de detección

Con los valores seleccionados para la ventana del filtro de medias, se detectan cada una de las clases de los tracks 16 al 20 y se evalúa el rendimiento. Por un lado se muestran las curvas DET obtenidas en la figura 5.27, y adicionalmente se muestran todos los posibles EER obtenidos por clase sobre diferentes valores de filtrado (figura 5.28). Esta segunda gráfica permite contrastar (al igual que se ha hecho en casos anteriores) si los valores de ventana escogidos como óptimos han resultado adecuados y coherentes con este conjunto de datos.

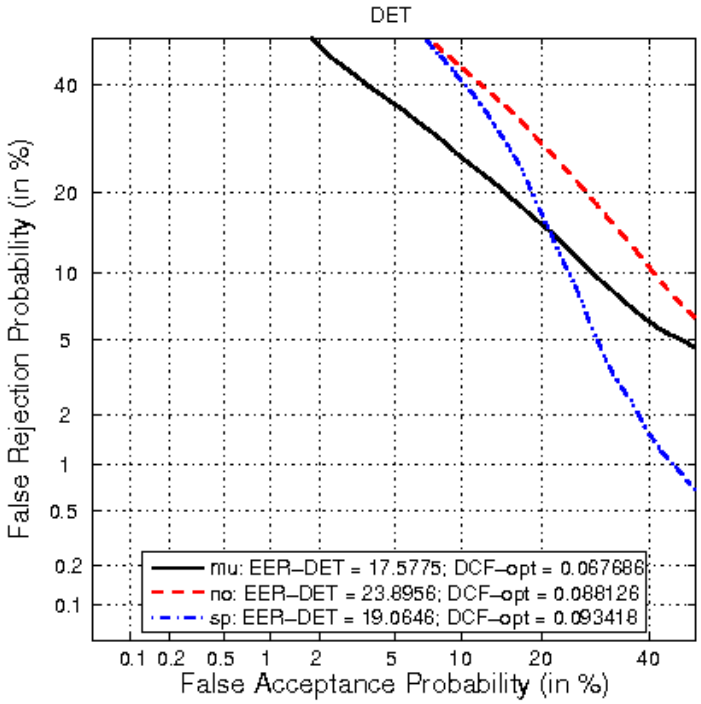


Figura 5.27: Curvas DET obtenidas sobre tracks 16 al 20, fusión características con sistema de referencia.

Los resultados obtenidos son muy similares a nivel de EER con los del sistema basado en agrupación por octavas y subbandas, empeorando ligeramente a nivel de EER la clase música con respecto a éste sistema inicial y mejorando ligeramente para la clase de voz. Sin embargo, de una manera más suavizada, se aprecia cómo se repiten los mismos sucesos que en el caso de la fusión a nivel de características con el sistema basado en estadísticos de la entropía cromática. Esto es, que por lo general los valores de EER obtenidos para cada clase se encuentran comprendidos entre el peor y el mejor valor obtenidos de entre ambos sistemas (el de referencia y el basado en agrupación por octavas y subbandas), por lo que en lugar de optimizar los resultados reduciendo significativamente el EER, el resultado de la fusión a nivel de características ha sido conseguir un rendimiento intermedio con respecto a los otros dos sistemas.

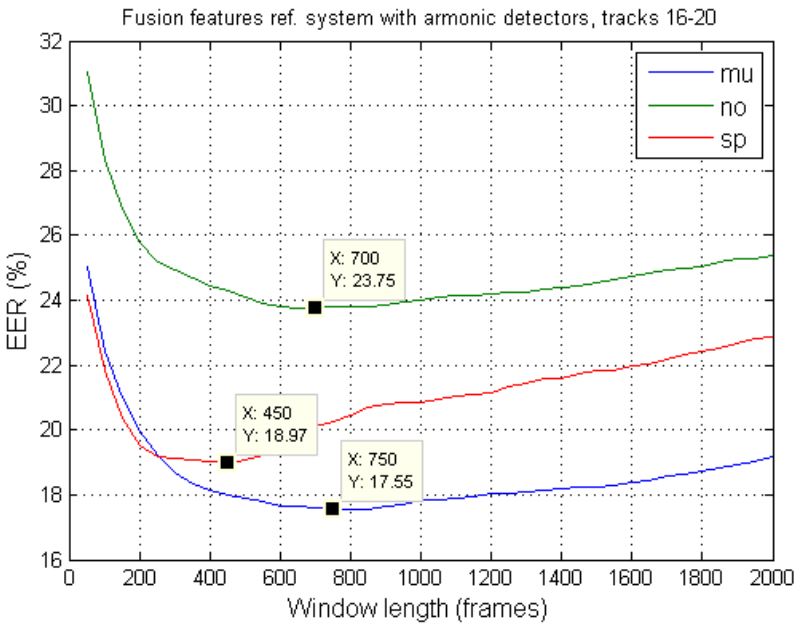


Figura 5.28: Análisis de EER sobre tracks 16 al 20, fusión características con sistema referencia.

5.2.3.1.3. Matrices de confusión

A continuación se evalúa el rendimiento del sistema para los tracks 16 al 20, siguiendo el mismo razonamiento que en los apartados anteriores. Los resultados se exponen a continuación:

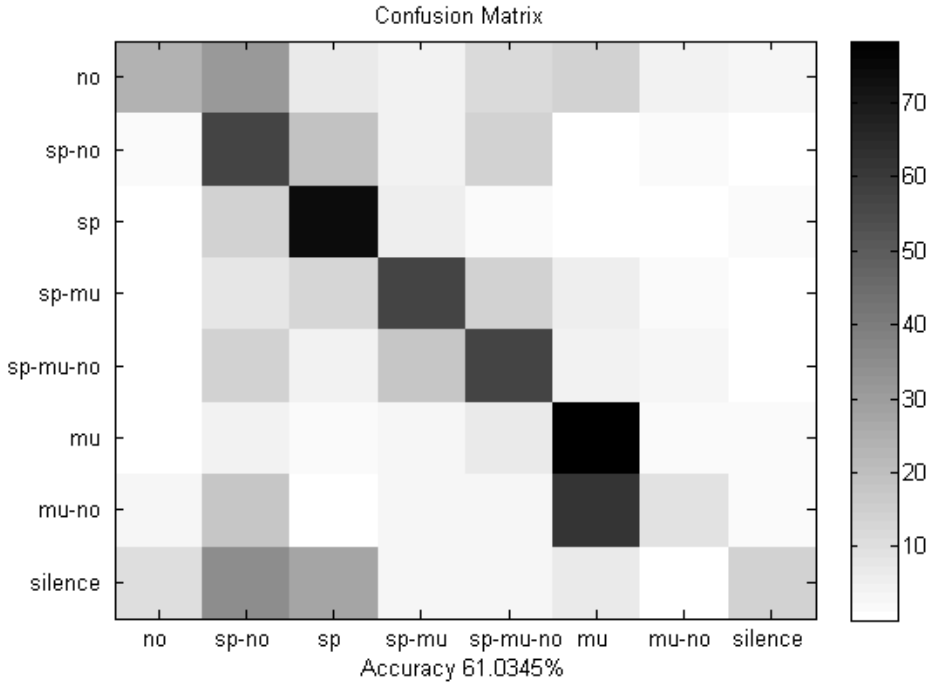


Figura 5.29: Matriz de confusión de la fusión con el sistema de referencia a nivel de características.

Clase	**no**	**sp-no**	**sp**	**sp-mu**	**sp-mu-no**	**mu**	**mu-no**	**silence**
Precisión (%)	24,16	59,56	74,30	56,72	56,76	78,41	9,38	14,08

Tabla 5.13: Valores de precisión por clases de la fusión a nivel de características del sistema de referencia con el sistema de agrupación por octavas y subbandas.

Comparando los resultados obtenidos con cada uno de los sistemas por separado, se puede ver cómo el valor de *accuracy* total ha mejorado, aunque ligeramente, resultando mayor que el obtenido sobre cualquiera de los otros dos sistemas. Por lo tanto, parece que en este caso (figura 5.29 la fusión no ofrece una medida ponderada de rendimiento sino que mejora respecto a valores obtenidos previamente. No obstante, como ya se ha mencionado en otras ocasiones, la medida de *accuracy* global no es suficientemente detallada, por lo que se acude al desglose por clases. En este caso, destaca el valor de precisión de la clase música, que con un 78.41 % (tabla 5.13 es el mejor valor obtenido hasta el momento sobre dicha clase.

5.2.3.2. Resultados de fusionar a nivel de scores con reglas fijas

La fusión a nivel de scores resulta de combinar con la operación suma los scores obtenidos sobre cada uno de los sistemas una vez que ambos han sido calibrados, del mismo modo que se ha operado en casos anteriores. Al igual que en la fusión contemplada a nivel de scores entre el sistema de referencia y el basado en estadísticos de la entropía cromática, se evalúa el error a nivel de matrices de confusión y precisión por clases.

5.2.3.2.1. Matrices de confusión

A continuación se muestran los resultados obtenidos con las matrices de confusión para el conjunto de tracks de evaluación 16 al 20, reflejados en la figura 5.30. Tal y como se ha llevado a cabo en los otros sistemas, se muestra el valor de precisión obtenido para cada clase en la tabla 5.14.

En primer lugar, cabe destacar que el nivel de precisión global obtenido con este sistema es el mayor de todos los anteriores sistemas y combinaciones desarrolladas en el presente proyecto. Con un valor mayor al obtenido para el caso de la fusión a nivel de características (al igual que sucedía con la anterior fusión con estadísticos de la entropía cromática), se podría concluir que la fusión a nivel de características funciona mejor que la agrupación de scores cuando éstas son de naturaleza similar (como es el caso de la fusión del sistema de agrupación por octavas con el sistema de agrupación por subbandas a nivel de características - anexo B - que ha dado lugar a uno de los sistemas desarrollados en este proyecto denominado *agrupación por octavas y subbandas*. En caso contrario, es decir, cuando los vectores de características sean de naturaleza muy diversa, parece que la de fusión a nivel de scores ofrece mejores resultados.

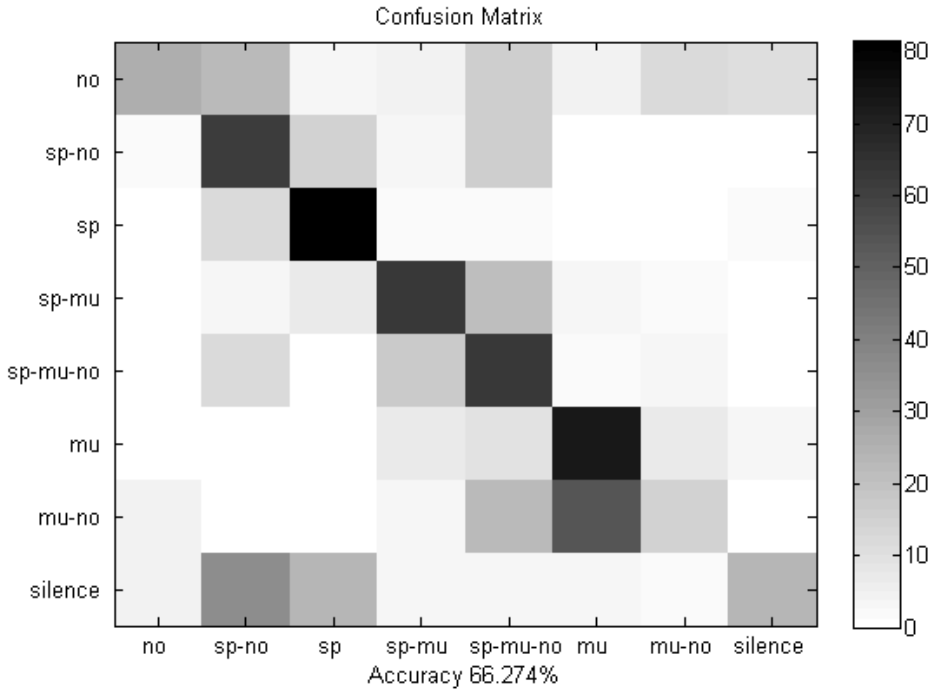


Figura 5.30: Matriz de confusión de la fusión con el sistema de referencia a nivel de scores.

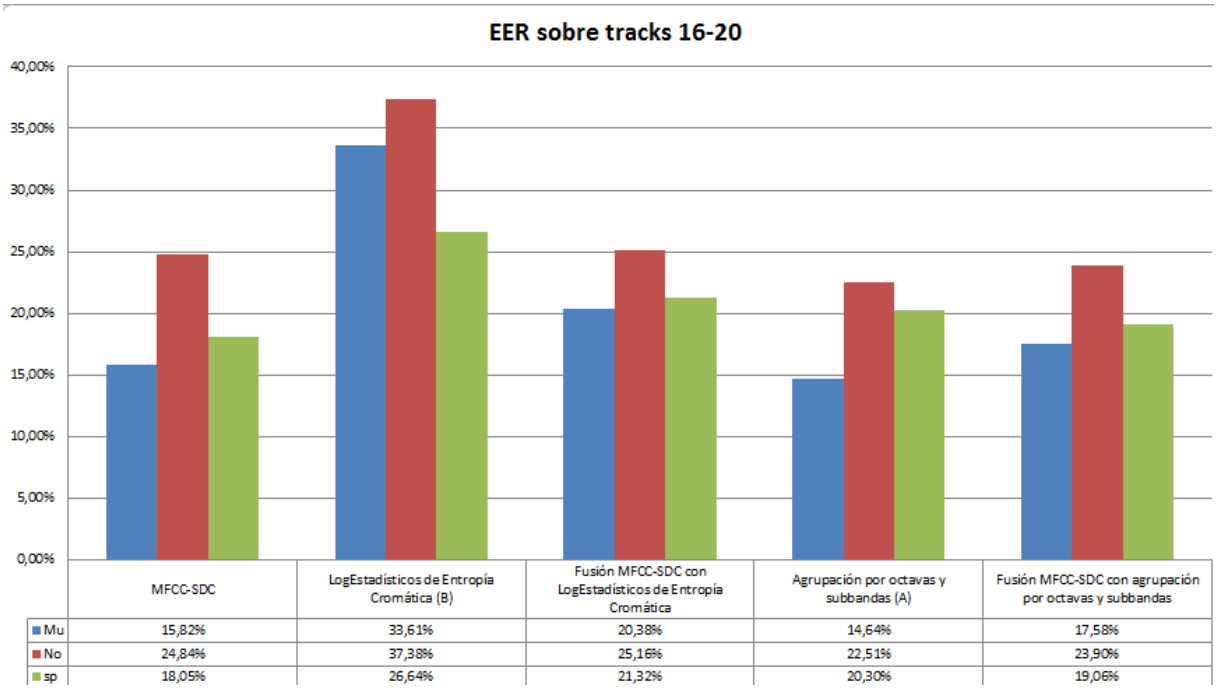
Clase	**no**	**sp-no**	**sp**	**sp-mu**	**sp-mu-no**	**mu**	**mu-no**	**silence**
Precisión (%)	26.53	61.41	81.36	63.38	63.21	72.48	14.99	23.77

Tabla 5.14: Valores de precisión por clases de la fusión a nivel de scores del sistema de referencia con el sistema de agrupación por octavas y subbandas.

5.3. Comparativa de sistemas

Una vez analizado el rendimiento de cada uno de los sistemas que han sido estudiados y desarrollados en el presente proyecto por separado, se procede a realizar una comparativa de resultados del siguiente modo: en primer lugar se contrastan los valores óptimos de ventana escogidos para cada sistema, en segundo lugar se contrastan los valores de EER obtenidos sobre cada uno de los sistemas para el conjunto de evaluación (tracks 16 al 2), y finalmente se compara el rendimiento de cada sistema tanto a nivel de precisión global como local por clases.

5.3.1. Rendimiento de detección



	MFCC-SDC	LogEstadísticos de Entropía Cromática (B)	Fusión MFCC-SDC con LogEstadísticos de Entropía Cromática	Agrupación por octavas y subbandas (A)	Fusión MFCC-SDC con agrupación por octavas y subbandas
Mu	15,82%	33,61%	20,38%	14,64%	17,58%
No	24,84%	37,38%	25,16%	22,51%	23,90%
sp	18,05%	26,64%	21,32%	20,30%	19,06%

Figura 5.31: Comparativa de sistemas en niveles de EER por detector.

Los resultados de la gráfica 5.31 evidencian que el mejor sistema para detección de música es el sistema basado en agrupación por octavas y subbandas, al igual que sucede con la clase ruido. Por el contrario, el sistema que mejor detecta la clase de voz es el sistema de referencia, basado en características MFCC-SDC. Estos resultados resultan coherentes con la lógica que inspira al diseño de cada uno de los sistemas. Por un lado, las características MFCC son comúnmente usadas para reconocer voz (y aunque se esperaba que la agrupación con los derivados SDC resultasen adecuados para música), los MFCC-SDC han resultado el mejor sistema de detección de voz en el proyecto. Por otro lado, la agrupación de la energía por filtros de subbandas imita la naturaleza de los cromagramas, por lo que se buscaban buenos resultados en detección de música especialmente. La agrupación por octavas buscaba por otro lado intentar caracterizar al ruido por no seguir ninguno de los patrones mencionados.

5.3.2. Rendimiento de clasificación

En este apartado se trabaja con 8 clases en lugar de 3 (figura 4.1). Como ya se ha explicado previamente, los tres detectores se han entrenado con clase mixtas, esto es, que se considera voz a cualquier fragmento que al menos contenga voz (ya sea un vector de voz, de voz con música, de voz con ruido o de voz con ruido y música) y lo mismo para las otras dos clases. Si, en cambio, se consideran todas las posibles combinaciones de estas tres clases aisladas surgen las ocho clases con las que trabajan las matrices de confusión.

En este apartado, en primer lugar se muestra un diagrama de barras que agrupa los resultados obtenidos sobre cada una de las clases para cada sistema y seguidamente se muestra el nivel de precisión global y por clases en figuras independientes.

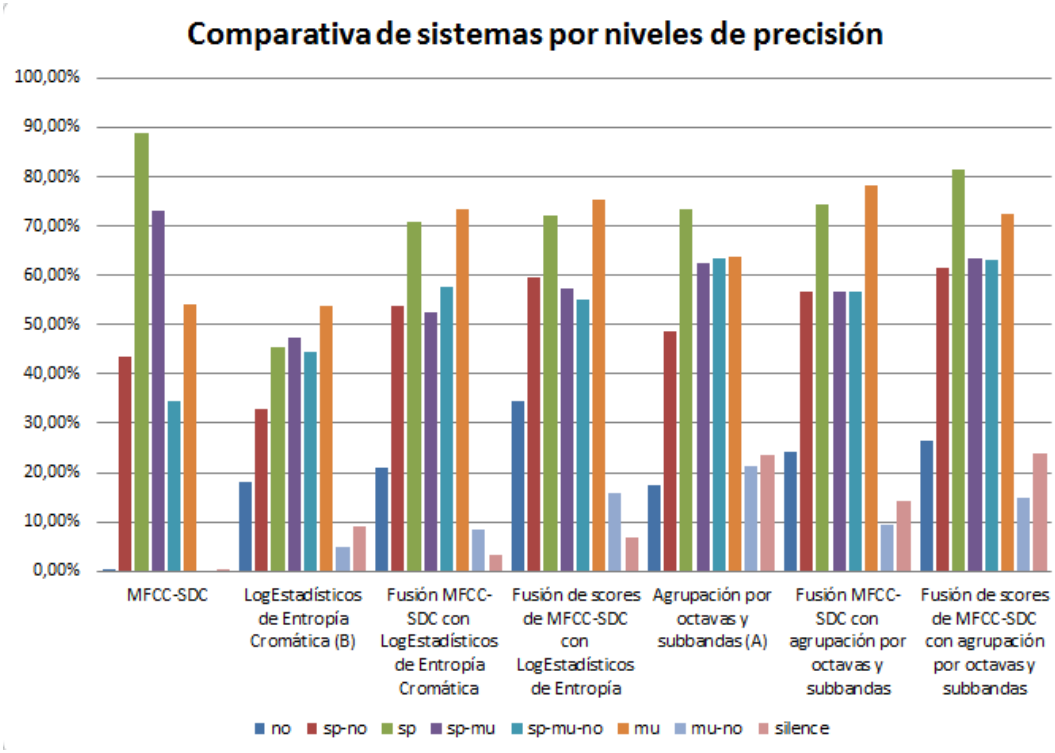


Figura 5.32: Comparativa de sistemas por niveles de precisión por clase.

A partir de la figura 5.32 se puede ver cómo de manera general las clases que mejor se etiquetan son la clase de voz (sp) y música aisladas (mu), seguidas de las clases de voz con ruido (sp-no), voz con música (sp-mu), y voz con música y ruido (sp-mu-no). No obstante, esta tendencia varía en función de cada sistema, motivo por el cual se ha detallado la precisión de cada clase en gráficas independientes para poder comparar a simple vista los resultados por clase.

No obstante, también de manera global a todas las clases se aprecia cómo las clase de ruido (no) y de silencio (silence) son clasificadas con una precisión muy baja, por lo que el sistema ofrece muy poco rendimiento en estas clases. Sin embargo, de nuevo la tendencia es diferente para cada sistema.

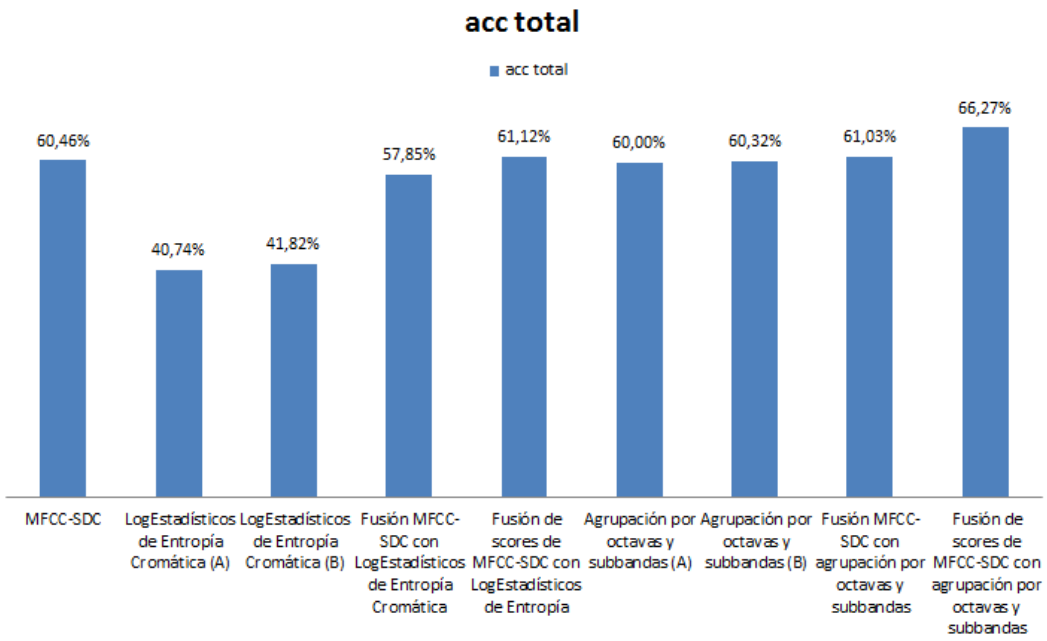


Figura 5.33: Comparativa de sistemas por nivel de precisión global.

La figura 5.33 permite ver a simple vista que el sistema que mejor segmenta el audio a nivel global es el obtenido de fusionar a nivel de scores el sistema de referencia con el sistema basado en agrupación por octavas y subbandas.

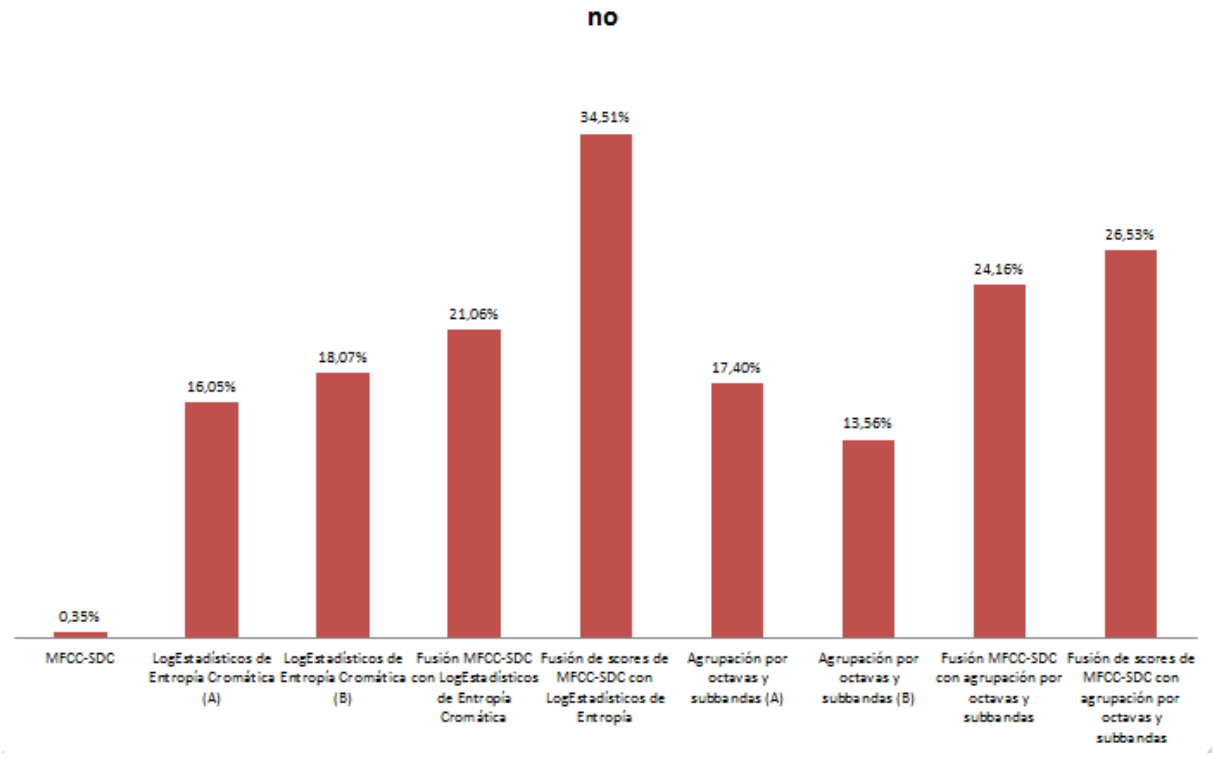


Figura 5.34: Comparativa de sistemas por nivel de precisión de ruido.

A partir de la gráfica 5.34 se puede ver que la clase de ruido se clasifica con muy poca precisión, lo que guarda relación directa con la poca cantidad de datos de entrenamiento de esta naturaleza. Sin embargo, el sistema que mejor clasifica el ruido aislado es el obtenido de fusionar a nivel de scores el sistema de referencia con el sistema basado en estadísticos de la entropía cromática.

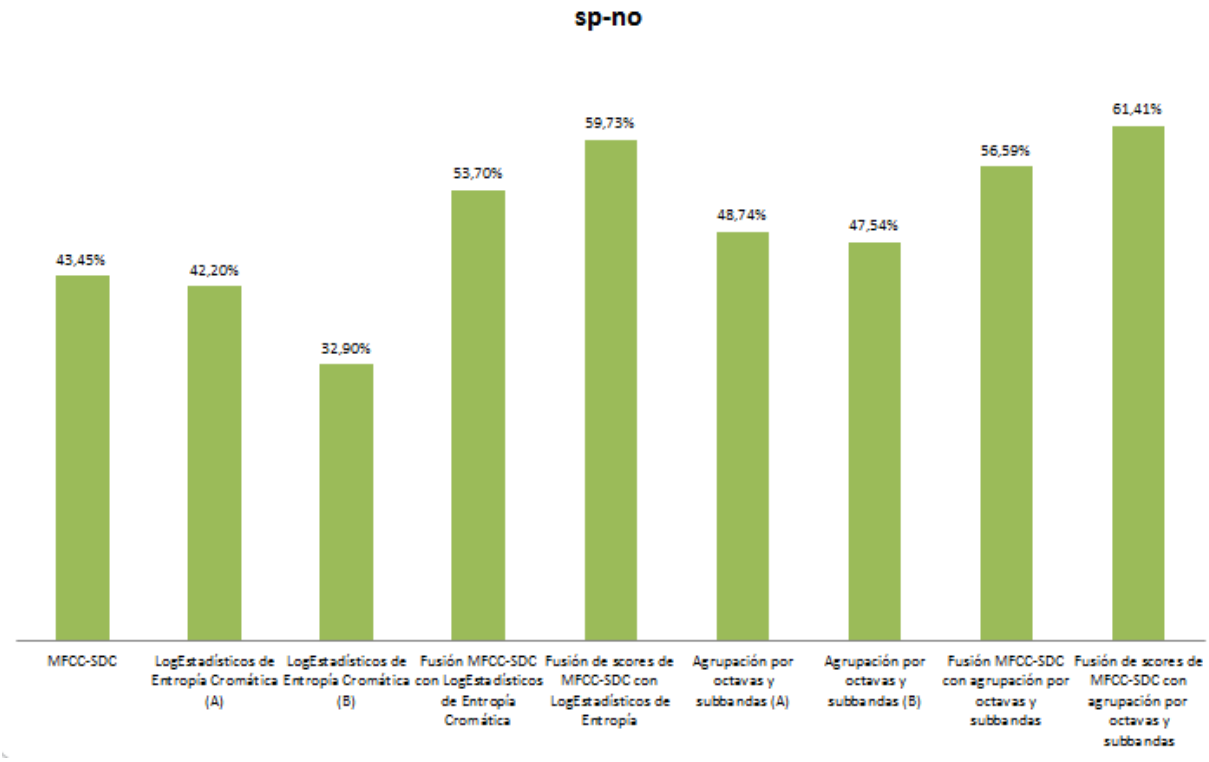


Figura 5.35: Comparativa de sistemas por nivel de precisión de voz con ruido.

La figura 5.35 permite ver que la clase de voz con ruido tiene más presencia en los datos de entrenamiento que se han empleado para generar los modelos. En este caso, el sistema que mejor clasifica esta clase de voz con ruido es la fusión a nivel de scores del sistema de referencia con el sistema de agrupación por octavas y subbandas (al igual que para la precisión global).

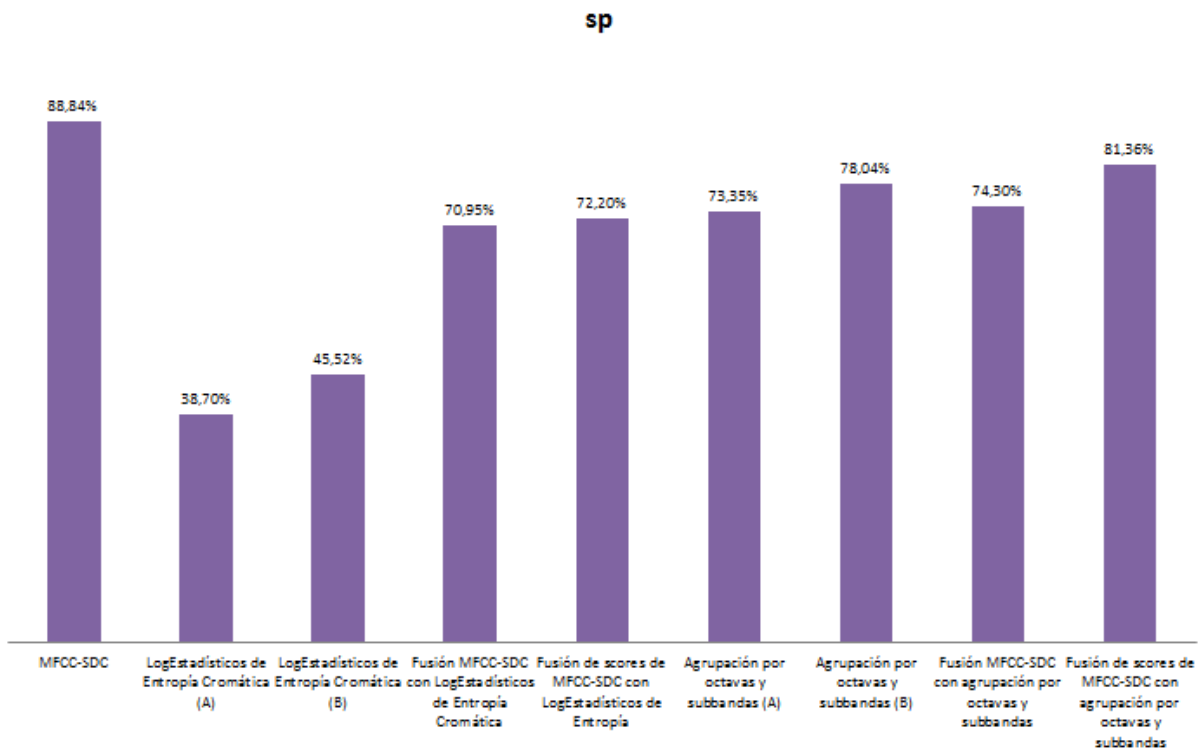


Figura 5.36: Comparativa de sistemas por nivel de precisión de voz.

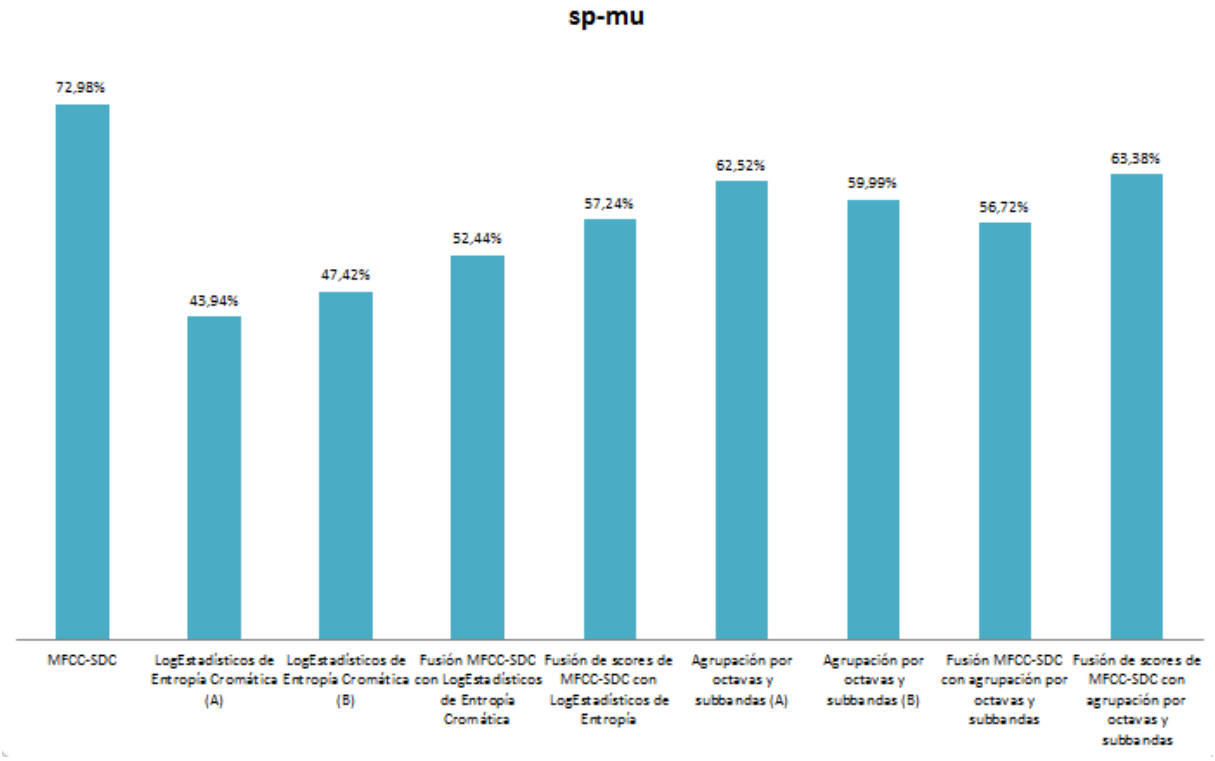


Figura 5.37: Comparativa de sistemas por nivel de precisión de voz con música.

La figura 5.36 permite ver que el sistema que mejor clasifica y segmenta los fragmentos de voz es el sistema de referencia basado en características MFCC-SDC, que además destaca por un valor de precisión significativamente alto, de casi el 90 %. Cabe destacar que a diferencia de los sistemas basados en entropía cromática el nivel de precisión de detección de la clase música no es inferior al 70 %.

Para la clase de voz con música, a partir de la figura 5.37 se aprecia que el sistema que mejor clasifica los fragmentos de esta naturaleza coincide con el resultado anterior, esto es, el sistema de referencia.

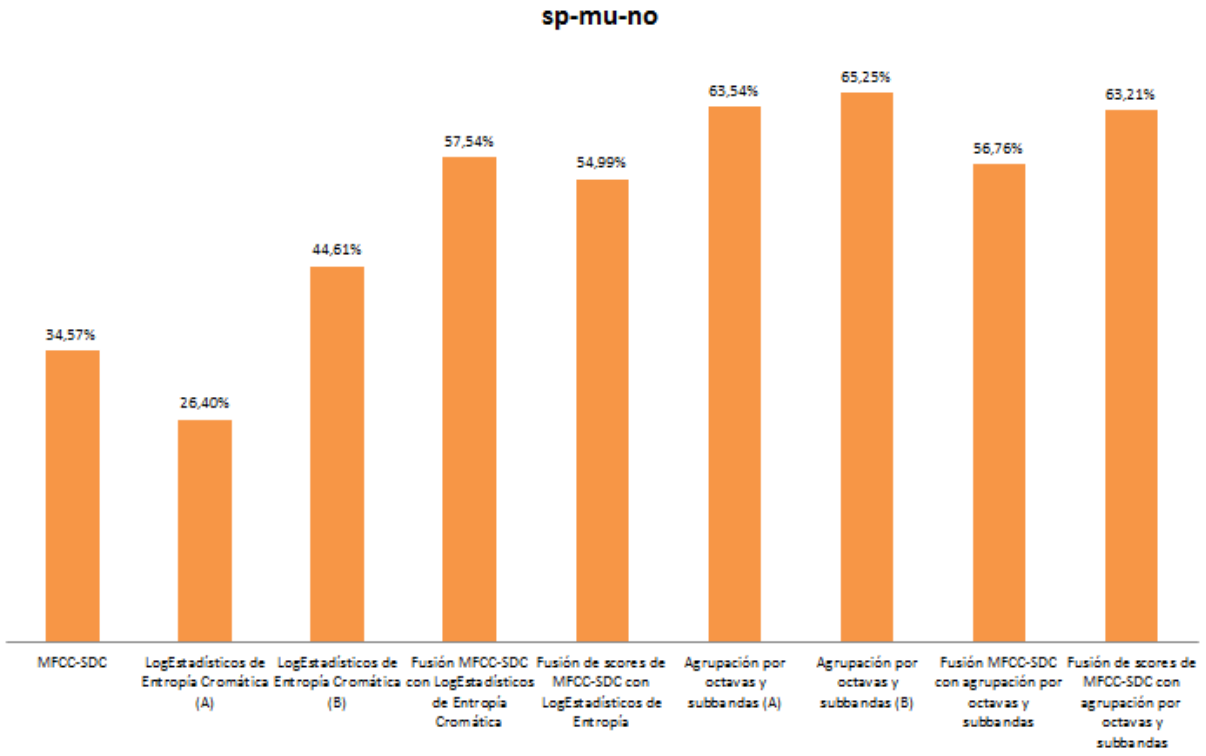


Figura 5.38: Comparativa de sistemas por nivel de precisión de voz con música y ruido.

Los mejores resultados para la clasificación de los fragmentos que contienen voz, música y ruido se encuentran en cualquiera de los sistemas basados en agrupación por octavas y subbandas tal y como evidencia la figura 5.38, y la fusión a nivel de scores de este sistema con el sistema de referencia. Curiosamente, los resultados de clasificación para esta clase en el sistema de referencia son significativamente peores.

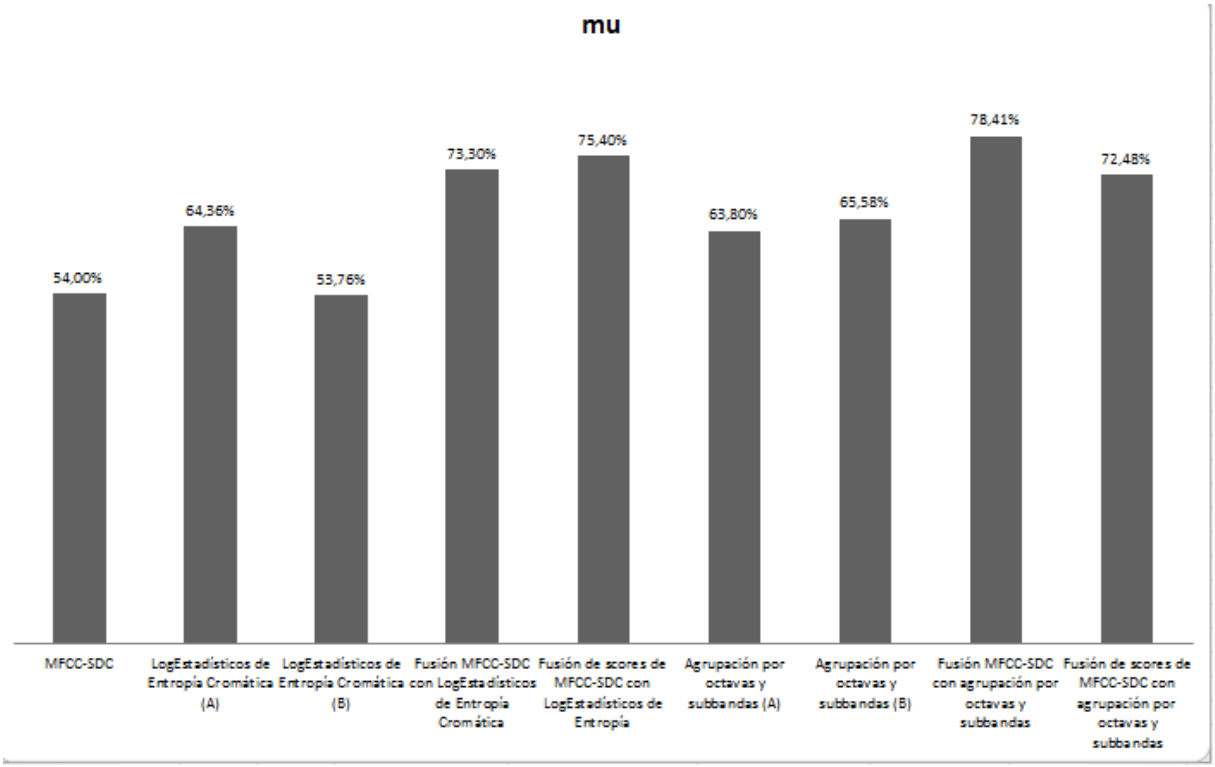


Figura 5.39: Comparativa de sistemas por nivel de precisión de música.

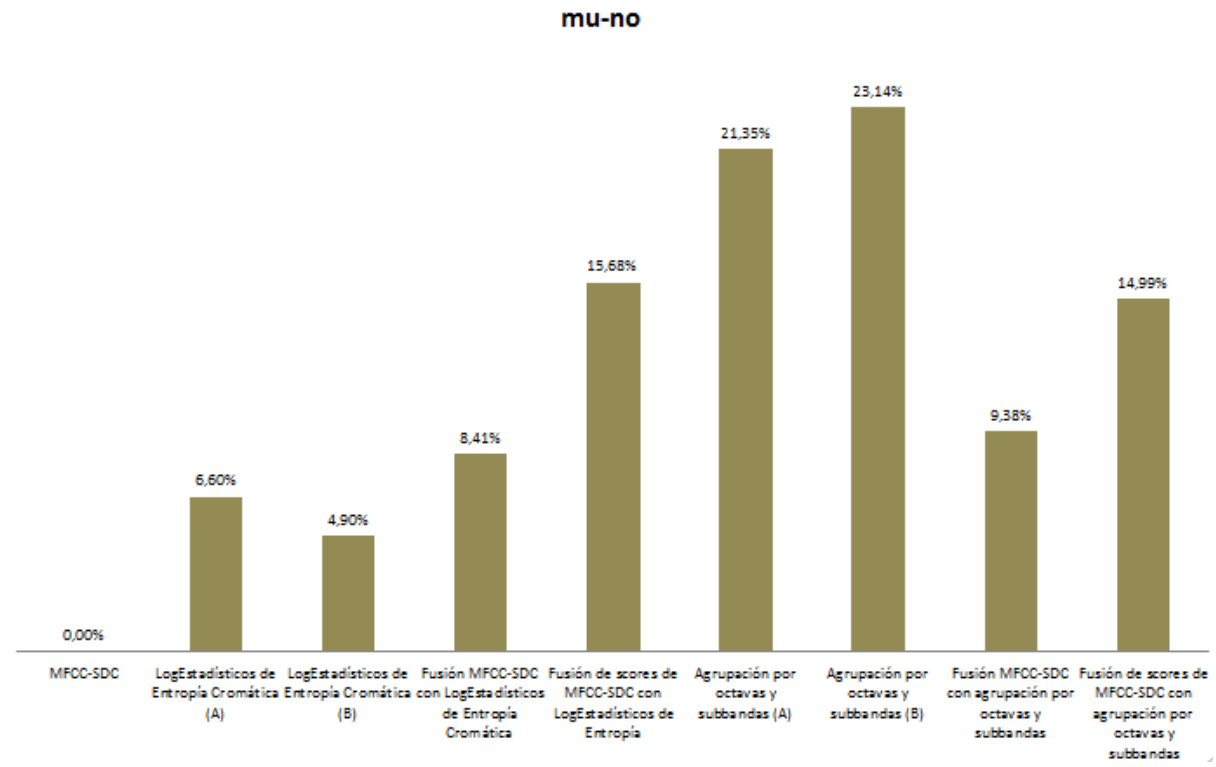


Figura 5.40: Comparativa de sistemas por nivel de precisión de música con ruido.

A partir de los resultados obtenidos en la figura 5.39, cualquiera de los sistemas de fusión del sistema de referencia con características cromáticas ofrecen para la clase música mejores resultados que el resto. Concretamente, sobre el conjunto de datos de desarrollo los mejores resultados se obtienen para la fusión a nivel de características del sistema de referencia con agrupación por octavas y subbandas.

La figura 5.40 muestra que la clase de música y ruido en general se clasifica con muy poca precisión, lo que coincide proporcionalmente con el total de datos de esta naturaleza empleados para generar los modelos.

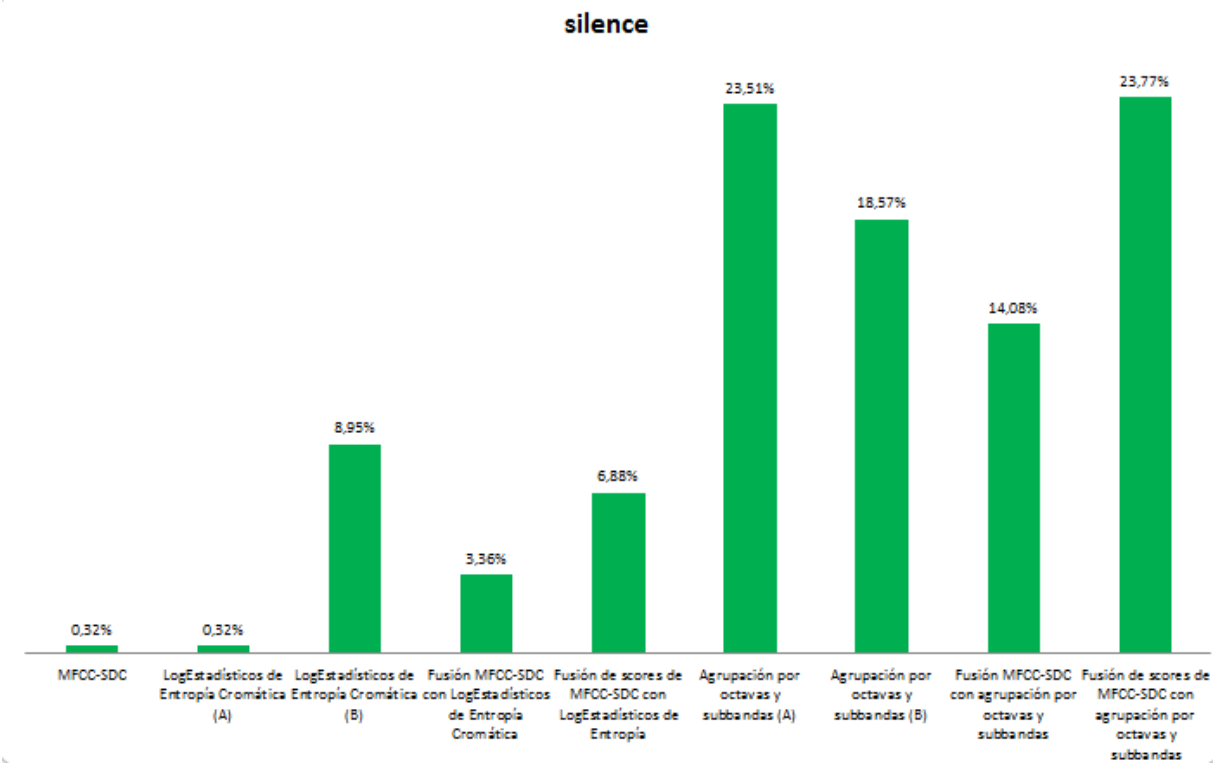


Figura 5.41: Comparativa de sistemas por nivel de precisión de silencio como ausencia de clases.

La figura 5.41 permite observar que al igual que sucede con las clases de ruido y de música con ruido, los fragmentos de silencio son escasos en la base de datos. Adicionalmente, no existe un modelo de silencio, sino que se segmentan con esta clase los segmentos que no se detectan como pertenecientes a ninguna de las clases. Aún así, llama la atención que si bien el sistema de referencia apenas es capaz de clasificar correctamente un 1 % de los fragmentos de silencio, la fusión a nivel de scores del sistema de referencia con la agrupación por octavas y subbandas clasifica correctamente más de un 20 % de los fragmentos.

6

Conclusiones y trabajo futuro.

6.1. Conclusiones

Las conclusiones que se pueden obtener a raíz de la investigación llevada a cabo en este proyecto son variadas e interesantes de analizar, si bien es preciso hacer una distinción entre conclusiones generales (aplicadas a cualquier sistema de segmentación de audio) y más particulares (las cuales resumen los mejores resultados de la investigación).

En primer lugar se enumeran y detallan cada una de las conclusiones generales obtenidas a raíz de desarrollar y/o evaluar cada uno de los seis sistemas de segmentación de audio que se han propuesto en este proyecto.

- Los resultados finales obtenidos de entrenar el sistema con agrupación por octavas y subbandas evidencian que la extracción de características cromáticas para entrenar sistemas de segmentación de audio resulta una línea de investigación de gran interés con resultados prometedores.

- En segundo lugar, cabe destacar la relevancia que ha supuesto para el desarrollo de este proyecto participar a comienzos del mismo en una evaluación de segmentación de audio competitiva como es la evaluación *Albayzin* en segmentación de audio. Presentarse a la competición no sólo ha permitido poder conocer de primera mano el estado del arte en sistemas de segmentación de audio sino que también ha supuesto una gran motivación por la investigación en esta línea con objeto de obtener resultados novedosos y satisfactorios que mejoren los sistemas existentes hasta el momento.

- A raíz de la mención de la evaluación, cabe destacar el reto que supone desarrollar sistemas de segmentación de audio cuando el acceso a bases de datos de calidad es muy limitado. A la hora de desarrollar un sistema con éxito no solamente influye la potencia algorítmica del sistema sino que es muy influyente que la base de datos con la que se entrena sea representativa del audio que se puede encontrar en cada uno de los diferentes escenarios de la vida real. En esta línea, la participación en la evaluación ha sido un éxito, al obtener

además una base de datos muy completa y de un número elevado de horas, que si bien nunca es suficiente, es representativa.

- Otro dato observable tras finalizar este proyecto, es la gran cantidad de parámetros que definen un modelo y deben ser ajustados. Generalmente, la generación de un modelo implica seleccionar y adecuar una serie de parámetros que influyen significativamente en los resultados. En este caso, el número de mezclas de cada sistema GMM-UBM y de iteraciones de entrenamiento del UBM y adaptación de los modelos por clase son infinitos si se decide hacer un barrido de valores de entrenamiento que optimice los resultados. El principal objetivo de este proyecto ha sido probar y testear diferentes vectores de características para desarrollar un modelo de segmentación de audio basado en detectores de música, voz y ruido. En este sentido, ha resultado prudente pre-seleccionar ciertos valores del estado del arte tomados como base para entrenar los modelos, si bien se ha realizado cierto *tunning* al probar los sistemas con los diferentes valores pre-seleccionados.

- Como última conclusión aplicable a cualquier sistema de segmentación de audio, cabe resaltar la importancia de elegir una medida de error adecuada y representativa. En este proyecto se comenzó trabajando con el SER como comparativa de rendimiento, para después descubrir que medidas de *accuracy* obtenidas a raíz de las matrices de confusión ofrecían resultados más detallados.

Una vez reflejadas las conclusiones generales, se resumen los mejores resultados de segmentación de audio sobre cada detector.

- Si se realiza una clasificación de sistemas en base al valor de EER obtenido, cabría decir que **el sistema que mejor detecta la música es el sistema basado en agrupación por octavas y subbandas**. En contraste, el sistema que peores resultados ha ofrecido para detectar música ha sido con diferencia el basado en log-estadísticos de la entropía cromática. En este punto, es interesante destacar la importancia de elegir un buen conjunto de características dentro de un mismo ámbito, ya que si bien en general las características cromáticas parecen adecuadas para detectar especialmente música, igual de importante es elegir cómo agrupar las energías de cada muestra y el número de características que se considere oportuno calcular en cada caso. Por otro lado, si bien por lo general el **ruido** es la clase que peor se detecta, **el sistema que lo hace con más acierto es de nuevo el sistema basado en agrupación por octavas y subbandas**. Por contraste, el peor de los sistemas para detectar ruido vuelve a ser el basado en log-estadísticos de la entropía cromática. En cuanto a la **clase voz**, una clase más estudiada quizás en el estado del arte debido a la fuerte línea de investigación que existe en el reconocimiento de voz, idioma y locutor, **el sistema que ofrece mejores resultados es el sistema basado en características MFCC-SDC**, pertenecientes al estado del arte desde hace ya varios años. De este modo, el mejor sistema de segmentación de audio utilizaría detectores basados en agrupación por octavas y subbandas para música y ruido, y en MFCC-SDC para voz.

- Ahora bien, los resultados obtenidos a partir de las matrices de confusión permiten evaluar el rendimiento de clasificación, es decir, cómo se comporta el sistema de forma global, ya que para definir cada una de las clases, se han empleado todos los detectores individuales, de modo que al combinar sus decisiones se han obtenido las nuevas etiquetas por clases homogéneas. Así pues, los mejores resultados obtenidos para cada una de las ocho clases se detallan en el siguiente listado:

- Para la clase ruido (no), la cual en general se clasifica bastante mal, el sistema que ofrece mejores resultados es el de la fusión de scores de los sistemas entrenados con características MFCC-SDC con log-estadísticos de la entropía cromática. Resalta sin embargo que el sistema que peor clasifica el ruido aislado es el entrenado con características MFCC-SDC, sin llegar al 1 % de precisión, por lo que es nulo sobre esta clase.

- Sobre la clase sp-no (voz con ruido), los resultados están más igualados y son en general mejores, destacando con más de un 60 % de precisión en este el caso también un sistema de fusión de scores de MFCC-SDC pero con agrupación por octavas y subbandas, seguido del otro sistema estudiado de fusión de scores.

- La clase de voz aislada es en general la que mejores valores de precisión obtiene alcanzando casi un 90 % de precisión, destacando como mejor sistema el de referencia (al igual que sucedía al evaluar la clase de voz en niveles de EER).

- Al igual que en el caso anterior, el sistema que mejor clasifica la clase de sp-mu (voz con música) es el de referencia alcanzando en este caso más de un 70 % de precisión y seguido de nuevo por el sistema último de fusión de scores con el sistema basado en agrupación por octavas y subbandas.

- El alcance de los resultados de la clase de sp-mu-no (voz con música y ruido) es similar al caso anterior. No obstante, en este caso es otro de los seis sistemas implementados el que obtiene mejores resultados, el sistema obtenido con agrupación por octavas y subbandas del audio.

- En cuanto a la clase de música aislada, cabe destacar que los sistemas con mejores resultados son obtenidos de fusionar tanto a nivel de características como a nivel de scores, si bien el que mayor nivel de precisión ha obtenido es el que fusiona a nivel de características MFCC-SDC con agrupación de octavas y subbandas.

- Al igual que sucedía con cualquiera de las otras clases que contenían ruido, la clase mu-no (música con ruido) obtiene unos resultados de precisión muy bajos, lo que en realidad es coherente con el bajo nivel de datos de entrenamiento sobre esta clase en comparación con el resto. Sin embargo, es el sistema basado en agrupación por octavas y subbandas el que con diferencia ofrece mejores resultados, aunque éstos apenan superan el 20 %.

- En último lugar, se ha querido medir la precisón de la clase silencio, como ausencia de cualquiera de las otras, y si bien los resultados siguen siendo bastante pobres, el sistema que mejor clasifica la ausencia de sonidos es el basado en agrupación por octavas y subbandas.

Finalmente, se puede decir que la fusión de scores en general funciona mejor que la fusión de características, por lo que resulta más complentario fusionar puntuaciones de sistemas independientes que intentar agrupar característidcas de dos naturalezas diferentes bajo un mismo vector. No obstante, generalmente la fusión de dos sistemas supera los resultados de ambos por separado, siempre y cuando el rendimiento de éstos por separado sea de algún modo similar. Si se combina un sistema que segmenta con alta precisión con uno de significativamente menor precisión (tal es el caso del sistema de referencia y el entrenado con log-estadísticos de la entropía cromática) el resultado de la fusión tenderá a ser un resultado intermedio de ambos sistemas.

6.2. Trabajo futuro

El desarrollo de este proyecto abre nuevas líneas de investigación en el área de segmentación de audio. Las más interesantes se detallan a continuación:

- Dado que cada sistema de segmentación está basado en detectores individuales, se contempla como línea de futuro entrenar un sistema de segmentación de audio con detectores para música y ruido basados en agrupación por octavas y subbandas y un detector de voz basado en coeficientes MFCC-SDC, siendo cada detector un sistema GMM-UBM
- Implementar un sistema de back-end que permita tomar decisiones de detección en base a las puntuaciones de una trama para todos los detectores. Es decir, que para tomar la decisión de pertenencia de una trama a una clase, por ejemplo la música, el sistema tenga en cuenta no sólo el score del detector de música sino también los scores de los detectores de voz y ruido.
- Entrenar un sistema con clases aisladas utilizando la base de datos de la que se ha dispuesto en este proyecto. Esto es, a partir de las etiquetas proporcionadas para entrenar, desarrollar y testear el sistema de segmentación (las cuales se basan en un sistema de etiquetado mixto, figura 2.4) generar nuevas etiquetas basadas en un etiquetado homogéneo (figura 2.4, lo que permita trabajar con las ocho clases contempladas en las matrices de confusión, y, con este nuevo modo de entender los datos, entrenar los 8 detectores GMM-UBM que constituirían el sistema de segmentación.

Bibliografía

[1] Javier Ortega. *Tratamiento de señales de vídeo y audio (UAM-EPS)*, 2014.

[2] Iván Gómez Piris. Extracción de información en señales de voz para el agrupamiento por locutores de locuciones anónimas. Technical report, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 2014.

[3] Pedro A. Torres-Carrasquillo, Elliot Singer, Mary A. Kohler, Richard J. Greene, Douglas A. Reynolds, and John R. Deller Jr. Approaches to language identification using gaussian mixture models and shifted delta cepstral features. In *7th International Conference on Spoken Language Processing*, 2002.

[4] Nikos Fakotakis Theodoros Theodorou, Iosif Mporas. An overview of automatic audio segmentation. *International Journal of Information Technology and Computer Science*, 6(11):1–9, October 2014.

[5] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. In *Digital Signal Processing*, page 2000, 2000.

[6] Daniel Ramos Javier Franco-Pedroso, Elena Gomez Rincon and Joaquin Gonzalez-Rodriguez. Atvs-uam system description for the albayzin 2014 audio segmentation evaluation. *Iberspeech 2014: VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop*, 2014.

[7] D. Reynolds and R.Rose. Robust text-independent speaker identification using gaussian mixture speaker models. In *IEEE Trans. Speech Audio Process. 3*, pages 72–83, 1995.

[8] J. Mariani. Recent advances in speech processing. In *Proc. IEEE ASSP89, Vol S1*, pages 429–440, 1989.

[9] Alfonso Ortega Diego Castan Antonio Miguel and Eduardo Lleida. The albayzin 2014 audio segmentation evaluation. *VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop*, 2014.

[10] Alvin F. Martin, George R. Doddington, Terri Kamm, Mark Ordowski, and Mark A. Przybocki. The DET curve in assessment of detection task performance. In *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*, 1997.

[11] Javier Franco-Pedroso Ignacio Lopez-Moreno Doroteo T. Toledano and Joaquin Gonzalez-Rodriguez. Atvs-uam system description for the audio segmentation and speaker diarization albayzin 2010 evaluation. *VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop*, pages 415–418, 2010.

[12] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, 1980.

[13] Xuedong Huang and Li Deng. An overview of modern speech recognition. In *Handbook of Natural Language Processing, Second Edition.*, pages 339–366. 2010.

[14] Tomi Kinnunen and Haizhou Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1):12–40, 2010.

[15] Hsiao-Wuen Hon Xuedong Huang, Alex Acero. *Spoken language processing: a guide to theory, algorithm, and system development.* Prentice Hall, 2001.

[16] F.K. Soong and A.E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 36, pages 871–879, 1988.

[17] M.A. Kohler and M. Kennedy. Language identification using shifted delta cepstra. In *Circuits and Systems, 2002. MWSCAS-2002. The 2002 45th Midwest Symposium on*, volume 3, pages III–69–72 vol.3, Aug 2002.

[18] B S Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. In *The Journal of the Acoustical Society of America*, volume 56. 7 1974.

[19] Furui. Cepstral analysis technique for automatic speaker verification. In *IEEE Trans. Acoustics, Speech Signal Process*, volume 29. 1981.

[20] H. Hermansky and N. Morgan. Rasta processing of speech. *Speech and Audio Processing, IEEE Transactions on*, 2(4):578–589, Oct 1994.

[21] Jason W. Pelecanos and Sridha Sridharan. Feature warping for robust speaker verification. In *2001: A Speaker Odyssey - The Speaker Recognition Workshop, Crete, Greece, June 18-22, 2001*, pages 213–218, 2001.

[22] Francois Pachet Jean-Julien Aucouturier. Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 2004.

[23] Ricardo Landriz Lara. Evaluación de características musicales para detección de tipos de audio. Technical report, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 2014.

[24] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[25] P. Gómez, A. Pikrakis, J. Mora, Jose Miguel Díaz-Báñez, Emilia Gómez, F. Escobar, S. Oramas, and J. Salamon. Automatic detection of melodic patterns in flamenco singing by analyzing polyphonic music recordings. In *III Interdisciplinary Conference on Flamenco Research (INFLA) and II International Workshop of Folk Music Analysis (FMA)*, Seville, Spain, 19/04/2012 2012.

[26] C. Gao Z. Shi H. Xue, H. Li. Computationally efficient audio segmentation through a multi-stage bic approach. *3rd International Congress on Image ad Signal Processing CISP*, 8:3774–3777, 2010.

[27] J. H.L. Hansen R. Huang. Advances in unsupervised audio classification and segmentation for broadcast news and ngsw corpora. *Audio, Speech and Language Processing, IEEE Transactions on*, 14(3):907–919, May 2006.

[28] David Tavarez Eva Navas Agustin Alonso Daniel Erro Ibon Saratxaga Inma Hernaez. Aholab audio segmentation system for albayzin 2014 evaluation campaign. In *VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop*.

[29] LAWRENCE R. RABINER. A tutorial on hidden markov models and selected applications in speech recognition. In *PROCEEDINGS OF THE IEEE*.

[30] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[31] Buzo A. Gray Linde, Y. An algorithm for vector quantizer design. *IEEE Trans. Comm.*, 28(84-95), 1980.

[32] G.; Ouellet P.; Dumouchel P. Kenny, P.; Boulianne. Speaker and session variability in gmm-based speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[33] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[34] Carmen Garcia-Mateo Paula Lopez-Otero, Laura Docio-Fernandez. Gtm-uvigo system for albayzin 2014 audio segmentation evaluation. In *VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop*.

[35] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 20:210–229, 2006.

[36] Alex Solomonoff, W.M. Campbell, and I. Boardman. Advances in channel compensation for svm speaker recognition. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 1, pages 629–632, March 2005.

[37] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, 13:308–311, 2006.

[38] David A. van Leeuwen and Niko Brümmer. An introduction to application-independent evaluation of speaker recognition systems. In *Speaker Classification I: Fundamentals, Features, and Methods*, pages 330–353, 2007.

[39] Stéphane Pigeon, Pascal Druyts, and Patrick Verlinde. Applying logistic regression to the fusion of the nist'99 1-speaker submissions. *Digital Signal Processing*, 10(1-3):237–248, 2000.

A

Tabla frecuencias

Para el cálculo de las frecuencias centrales del banco de filtros empleados en el cálculo de parámetros basados en estadísticos de la entropía cromática, se ha tomado como frecuencia de referencia (f_0) el Do central (Do 1 en notación Franco Belga), y a partir de éste se han calculado las frecuencias restantes atendiendo a la fórmula:

$$f(Hz) = f_0 * 2^{k/12}, //k = 0, 1, ..., L-1 \tag{A.1}$$

Para el cálculo de estos parámetros se ha trabajado sobre 7 octavas, que comprenden desde el Do1 hasta el Si7, cuya equivalencia de frecuencias se detalla en las tablas A.1 y A.2.

Por otro lado, y partiendo de la misma fórmula expuesta recientemente, para el cálculo de parámetros basados en la armonía de la música se ha decidido ampliar a un rango de frecuencias más amplio, concretamente desde el Do -2 (refiriéndonos siempre, a la notación Franco-Belga) hasta el Si7, abarcando un total de 10 octavas.

Nota	**Do -2**	**Do# -2**	**Re -2**	**Re# -2**	**Mi -2**	**Fa -2**
Frec (Hz)	8.1757	8.6618524	9.176913	9.7226006	10.300737	10.91325
Nota	**Do -1**	**Do# -1**	**Re -1**	**Re# -1**	**Mi -1**	**Fa -1**
Frec (Hz)	16.35159	17.323906	18.354039	19.445427	20.601712	21.826754
Nota	**Do**	**Do#**	**Re**	**Re#**	**Mi**	**Fa**
Frec (Hz)	32.7031	34.647728	36.707989	38.890759	41.203324	43.653401
Nota	**Do 1**	**Do# 1**	**Re 1**	**Re# 1**	**Mi 1**	**Fa 1**
Frec (Hz)	65.40639	69.295656	73.41619	77.781744	82.406888	87.307056
Nota	**Do 2**	**Do# 2**	**Re 2**	**Re# 2**	**Mi 2**	**Fa 2**
Frec (Hz)	130.8127	138.59123	146.83229	155.56339	164.81367	174.61401
Nota	**Do 3**	**Do# 3**	**Re 3**	**Re# 3**	**Mi 3**	**Fa 3**
Frec (Hz)	261.62556	277.18263	293.66476	311.12698	329.62755	349.22822
Nota	**Do 4**	**Do# 4**	**Re 4**	**Re# 4**	**Mi 4**	**Fa 4**
Frec (Hz)	523.2511	554.36523	587.3295	622.25393	659.25508	698.45642
Nota	**Do 5**	**Do# 5**	**Re 5**	**Re# 5**	**Mi 5**	**Fa 5**
Frec (Hz)	1046.5022	1108.7305	1174.659	1244.5079	1318.5102	1396.9128
Nota	**Do 6**	**Do# 6**	**Re 6**	**Re# 6**	**Mi 6**	**Fa 6**
Frec (Hz)	2093.0045	2217.461	2349.3181	2489.0158	2637.0204	2793.8258
Nota	**Do 7**	**Do# 7**	**Re 7**	**Re# 7**	**Mi 7**	**Fa 7**
Frec (Hz)	4186	4434.9125	4698.6261	4978.021	5274.0295	5587.6396

Tabla A.1: Detalle de frecuencias asociadas a cada nota musical, empleadas para el banco de filtros (I)

Nota	**Fa# -2**	**Sol -2**	**Sol# -2**	**La -2**	**La# -2**	**Si -2**
Frec (Hz)	11.562186	12.249709	12.978115	13.749834	14.567441	15.433666
Nota	**Fa# -1**	**Sol -1**	**Sol# -1**	**La -1**	**La# -1**	**Si -1**
Frec (Hz)	23.12464	24.499703	25.956531	27.499987	29.135221	30.867692
Nota	**Fa#**	**Sol**	**Sol#**	**La**	**La#**	**Si**
Frec (Hz)	46.249168	48.999286	51.912935	54.999839	58.2703	61.735232
Nota	**Fa# 1**	**Sol 1**	**Sol# 1**	**La 1**	**La# 1**	**Si 1**
Frec (Hz)	92.498604	97.998857	103.82617	110	116.54094	123.47082
Nota	**Fa# 2**	**Sol 2**	**Sol# 2**	**La 2**	**La# 2**	**Si 2**
Frec (Hz)	184.99709	195.99759	207.65222	219.99986	233.08173	246.94149
Nota	**Fa# 3**	**Sol 3**	**Sol# 3**	**La 3**	**La# 3**	**Si 3**
Frec (Hz)	369.99442	391.99543	415.30469	439.99999	466.16375	493.88329
Nota	**Fa# 4**	**Sol 4**	**Sol# 4**	**La 4**	**La# 4**	**Si 4**
Frec (Hz)	739.9888	783.99083	830.60935	879.99995	932.32747	987.76654
Nota	**Fa# 5**	**Sol 5**	**Sol# 5**	**La 5**	**La# 5**	**Si 5**
Frec (Hz)	1479.9776	1567.9817	1661.2187	1759.9999	1864.6549	1975.5331
Nota	**Fa# 6**	**Sol 6**	**Sol# 6**	**La 6**	**La# 6**	**Si 6**
Frec (Hz)	2959.9554	3135.9635	3322.4375	3520	3729.3101	3951.0664
Nota	**Fa# 7**	**Sol 7**	**Sol# 7**	**La 7**	**La# 7**	**Si 7**
Frec (Hz)	5919.898	6271.9134	6644.8608	7039.9848	7458.6041	7902.1157

Tabla A.2: Detalle de frecuencias asociadas a cada nota musical, empleadas para el banco de filtros (II)

B

Análisis previo al desarrollo de nuevas características cromáticas

El sistema presentado en este proyecto de nuevas características cromáticas ha sido resultado de un estudio previo de dos sistemas con diferentes sistemas de extracción de características cromáticas.

En primer lugar, se consideró un extractor de características por subbandas, esto es, el resultado de agrupar la energía de los 12 semitonos que forman una escala musical tal y como lo presenta un cromagrama. Por otro lado, se consideró un extractor de características por octavas, esto es, agrupar la energía resultante por cada escala considerando un total de 10 octavas. Los rangos frecuenciales para ambos casos coinciden, y se han detallado en el anexo anterior A.

De este modo, tras entrenar dos sistemas de segmentación de audio cada uno basado en una de estas nuevas características cromáticas, se decidió estudiar la fusión a nivel de scores y a nivel de características, obteniendo unos resultados óptimos para el caso de la fusión a nivel de características de ambos sistemas. A continuación se detallan los resultados obtenidos con matrices de confusión y nivel de precisión por clases tanto por separado como para la fusión, lo que permite apreciar la mejora aportada por la fusión a nivel de características.

B.1. Sistema basado en agrupación por subbandas

A continuación se presentan los resultados obtenidos de entrenar un GMM-UBM de 1024 mezclas, con un extractor de características basado en agrupación por subbandas, para una banda de trabajo desde los 8 Hz a los 8 KHz aproximadamente.

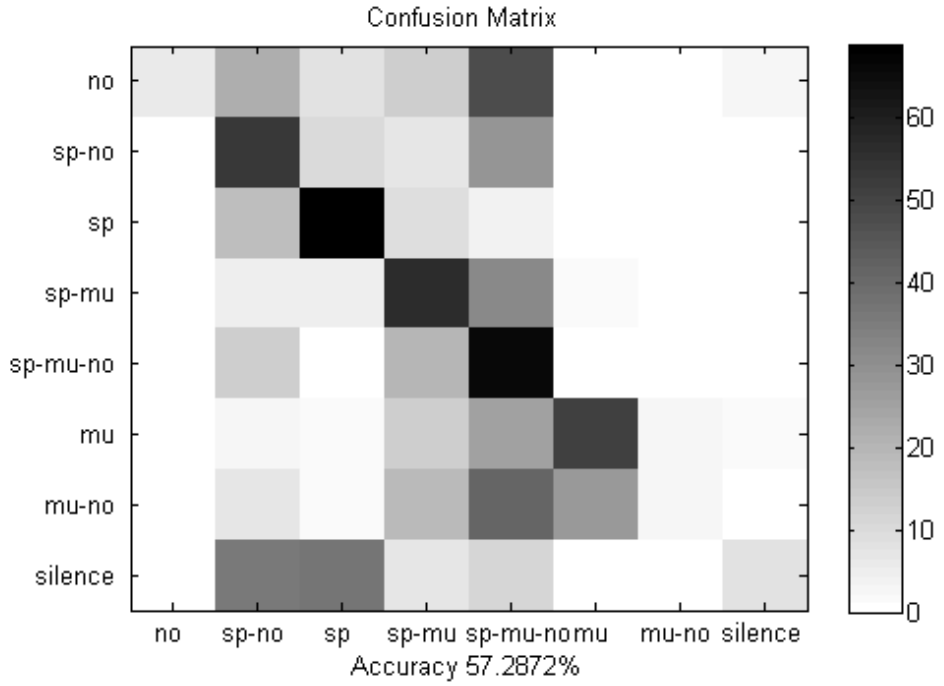


Figura B.1: Matriz de confusión de agrupación por octavas sobre tracks 16 al 20

Clase	**no**	**sp-no**	**sp**	**sp-mu**	**sp-mu-no**	**mu**	**mu-no**	**silence**
Precisión (%)	6.27	52.91	68.79	56.89	65.66	51.56	2.15	7.94

Tabla B.1: Valores de precisión por octavas

B.2. Sistema basado en agrupación por octavas

En este caso se presentan los resultados obtenidos de entrenar un GMM-UBM de 1024 mezclas, con un extractor de características basado en agrupación por pctavas, para una banda de trabajo desde los 8 Hz a los 8 KHz aproximadamente.

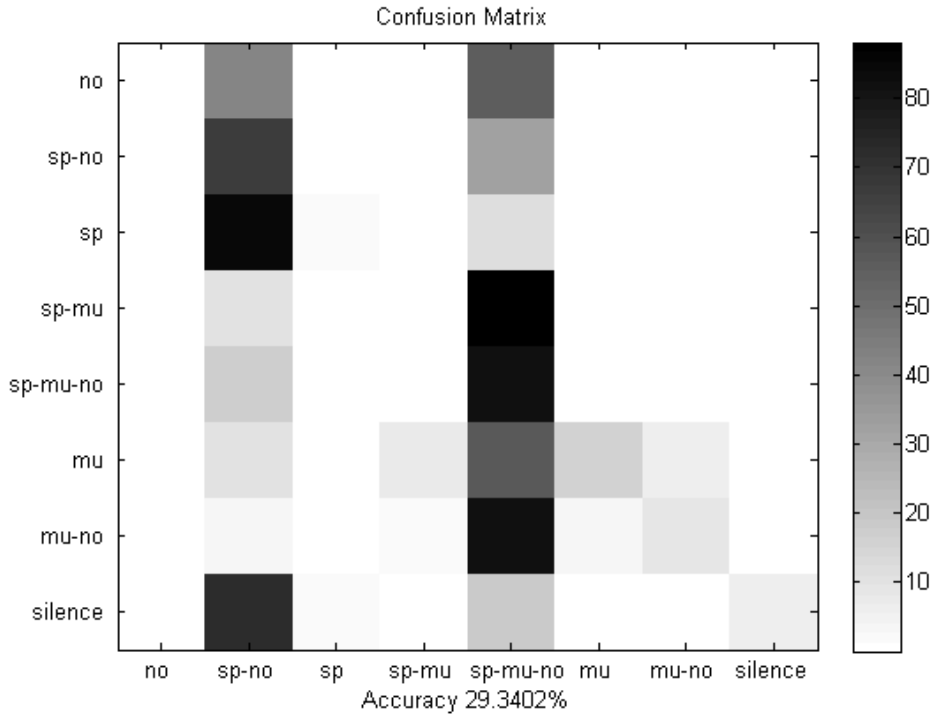


Figura B.2: Matriz de confusión de agrupación por subbandas sobre tracks 16 al 20

Clase	**no**	**sp-no**	**sp**	**sp-mu**	**sp-mu-no**	**mu**	**mu-no**	**silence**
Precisión (%)	0	66.99	1.89	0.82	82	16.14	9.19	5.8

Tabla B.2: Valores de precisión por subbandas

B.3. Fusión de sistemas a nivel de características

Una vez evaluados los sistemas se presenta el resultado de fusionar ambos a nivel de características, consiguiendo un vector que concatena la agrupación por subbandas y por octavas, sumando un total de 22 características por cada vector.

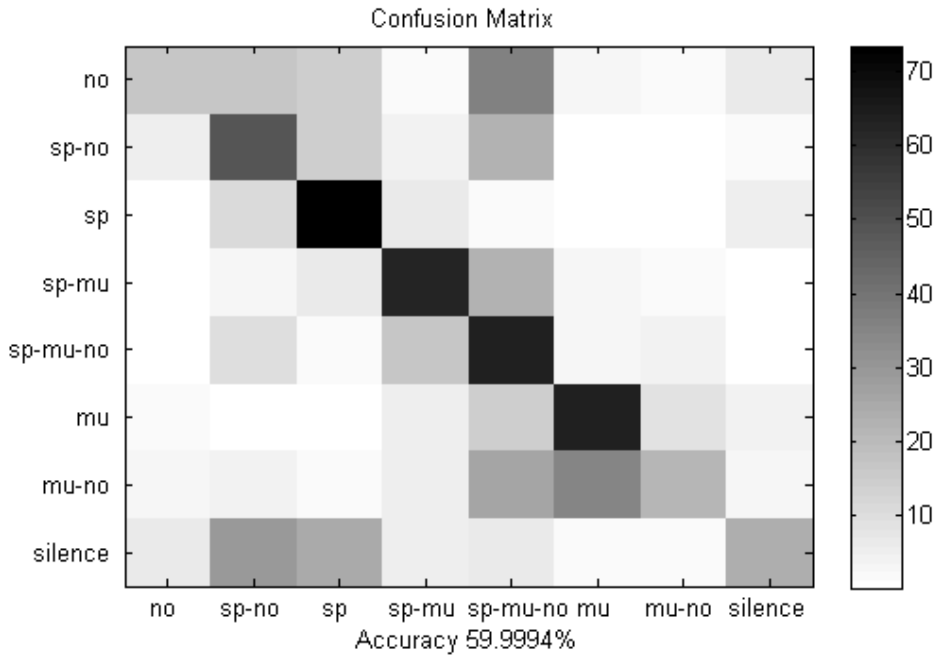


Figura B.3: Matriz de confusión de la fusión a nivel de características

Clase	**no**	**sp-no**	**sp**	**sp-mu**	**sp-mu-no**	**mu**	**mu-no**	**silence**
Precisión (%)	17.4	48.74	73.35	62.52	63.54	63.8	21.35	23.51

Tabla B.3: Valores de precisión por clases de la fusión a nivel de características

B.4. Fusión de sistemas a nivel de scores

Por otro lado se presentan los resultados obtenidos de fusionar ambos sistemas a nivel de scores mediante la función suma.

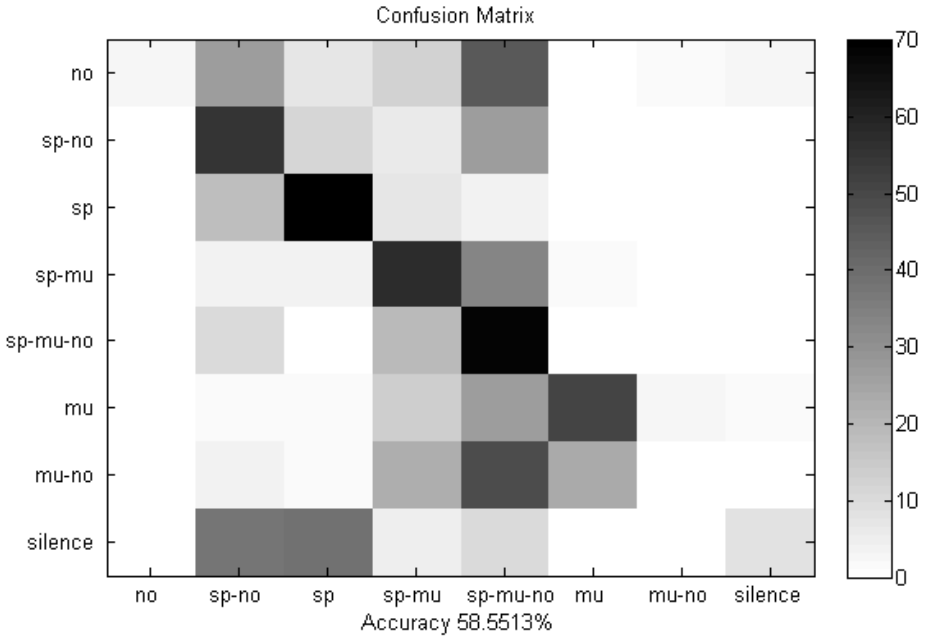


Figura B.4: Matriz de confusión de la fusión a nivel de scores

Clase	**no**	**sp-no**	**sp**	**sp-mu**	**sp-mu-no**	**mu**	**mu-no**	**silence**
Precisión (%)	3.12	55.06	70.14	57.31	68.67	50.79	0	8.42

Tabla B.4: Valores de precisión por clases de la fusión a nivel de scores

C

Artículo publicado

Publicación en IberSpeech 2014 Online Proceedings

A continuación se anexa el artículo que describe el sistema propuesto para la evaluación Albayzin 2014 de Segmentación e Audio, y que forma parte de las actas del congreso *Iberspeech 2014*.

ATVS-UAM System Description for the Albayzin 2014 Audio Segmentation Evaluation

Javier Franco-Pedroso, Elena Gomez Rincon, Daniel Ramos and Joaquin Gonzalez-Rodriguez

ATVS - Biometric Recognition Group
Universidad Autonoma de Madrid (UAM). Spain
http://atvs.ii.uam.es
javier.franco@uam.es,elena.gomezr@estudiante.uam.es,daniel.ramos@uam.es,joaquin.gonzalez@uam.es

Abstract. This document describes the audio segmentation system developed by the ATVS – Biometric Recognition Group, at Universidad Autonoma de Madrid (UAM), for the Albayzin 2014 Audio Segmentation Evaluation (ASE). This system is based on three independent GMM-UBM acoustic-class detectors based on MFCC-SDC features. Each acoustic-class detector ('mu', 'no', 'sp') evaluates test recordings in a frame-by-frame manner, and the score-streams are filtered and calibrated previous to the detect-decision stage. Although the performance of the independent acoustic-class detectors is far from being perfect in terms of EER, the resulting audio segmentation systems achieves low miss (7.9%), false alarm (10.6%) and class error (3.0%) rates, given a final 21.43% SER on our development subset.

Keywords: audio segmentation, MFCC-SDC, GMM-UBM, calibration

1 Introduction

In contrast to our previous participation in Albayzin ASE campaigns (the 2010 edition [2]), this year we present a lighter but more robust system that avoids the overfitting introduced by Maximum Mutual Information discriminative training when the available data is scarce. Moreover, the system developed fits better the approach followed in this campaign by the organizers to the problem of evaluating automatic segmentation systems [3]: instead of labeling non-overlapping segments of (maybe overlapped) different acoustic classes, the presence of each acoustic class should be independently annotated in different segments (maybe overlapped with other acoustic classes). Although the problem can be solved from both perspectives (training different models for each possible acoustic-classes combination as we did in 2010 campaign), considering one independent detector for each acoustic class provides a more scalable solution and avoids the constraints regarding the available data for training the acoustic models.

The system developed consists in three independent acoustic-class detectors (speech –'sp'-, music -'mu'-, and noise –'no'-) based on the classical GMM-UBM

adfa, p. 1, 2011.

framework [4]. Each detector performs a frame-by-frame scoring of the test recordings, obtaining one log-likelihood stream per acoustic class. These score-streams are smoothed through a mean filter over a sliding window in order to deal with the high variability of frame-scores. Finally, smoothed frame-scores are independently calibrated by means of a linear logistic regression trained on a subset of the development dataset.

The remainder of this paper is organized as follows. Section 2 describes the feature extraction process. Sections 3 and 4 describe, respectively, the acoustic-class modeling and the acoustic-class detection stage. Section 5 explains the experimental protocol followed, and shows the results obtained in our development subset. Finally, Section 6 summarizes the key points of our submission, exposes the computational requirements and draws some conclusions.

2 Feature Extraction

Shifted Delta Coefficients (SDC) [5] have been widely used in Language Recognition due to the fact that they capture the time dependency structure of the language better than the speed or acceleration coefficients (also known as delta and delta-delta). Similarly, SDC features are expected to capture the different time dependency of the music over the speech or noise. In fact, experiments carried out over a subset of the development tracks revealed that GMM-UBM detectors build from MFCC-SDC features outperform those trained on MFCC plus delta coefficients.

For both development and evaluation tracks, one feature vector was extracted every 10 ms by means of a 20 ms Hamming sliding window (50% overlap). For each window, 7 MFCC features (including C0) were computed from 25 Mel-spaced magnitude filters over the whole available spectrum (0-8000 Hz). These features have been mean-normalized, RASTA filtered and Gaussianized through a 3-second window. Finally, their SDC were computed on a 7-1-3-7 (N-D-P-K) configuration and concatenated with them in a 56-coefficient feature vector.

3 Acoustic-Class Modeling

Acoustic classes have been modeled adopting the classical GMM-UBM framework [4] widely used for speaker recognition. First, a 1024-component UBM was trained by means of a 1-iteration k-means initialization followed by a 5-iteration EM stage. For this purpose, one half of the development dataset provided was used (tracks 01-10). Secondly, acoustic-class models were MAP-adapted [4] from this UBM through 1 single iteration and using a relevance factor r=16. Again, tracks 01-10 were used also for this step.

For each acoustic class, training data were extracted from segments belonging to the same acoustic-class as appeared in the provided development labels. This means that, for instance speech segments may contain not only isolated speech but also any of the other acoustic classes overlapped with it. As we are aiming to develop an acoustic-class detector, our assumption is that the acoustic-class models should collect

their own acoustic class in any possible condition it may appear. On the other hand, segments where each class can be found isolated are very scarce in the database provided, so robust acoustic-class models cannot be trained from such small amount of data, as we found out in our preliminary experiments.

4 Acoustic-Class Detection Stage

Acoustic-class detection stage is based on a frame-by-frame scoring of the test track against every acoustic-class model. Frame-by-frame log-likelihoods are highly variable over time, as it can be seen on Figure 1. For a segment with an isolated acoustic-class, it is expected that the mean log-likelihood will converge to a stable value as long as more frames are incorporated, as it has been shown for the speaker recognition task in [6]. For this reason, these score-streams were smoothed through a mean filter over a sliding window in order to have a more stable frame-score that approaches the "true" score of the acoustic class present in the surrounding frames. Figure 2 shows the result of applying this mean filtering stage for a 700-frame sliding window. The window length was independently optimized for each acoustic-class detector, looking for the length that provides the best detection performance in terms of EER. Results are shown in Figure 3 for our development subset (tracks 11-15).

Finally, the frame-by-frame log-likelihoods were calibrated by means of a linear logistic regression implemented in FoCal toolkit [1]. One different logistic regression is used for each acoustic-class detector, all of them trained on the same development subset used for the window length optimization (tracks 11-15).

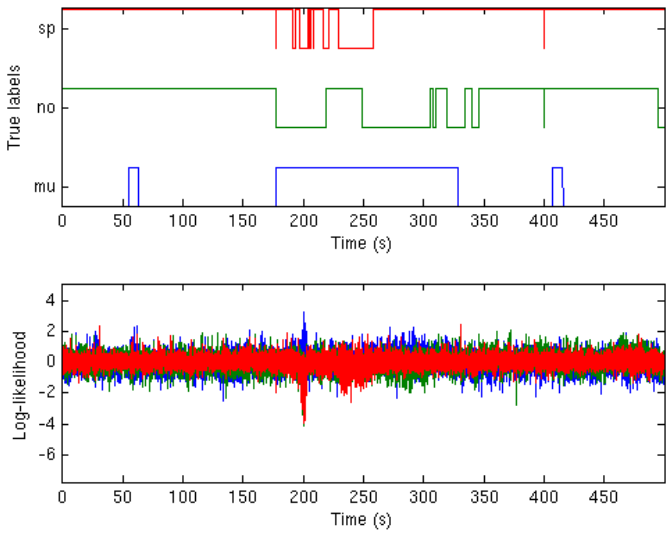


Fig. 1. Detail of the frame log-likelihoods for a 500-second segment of track11.

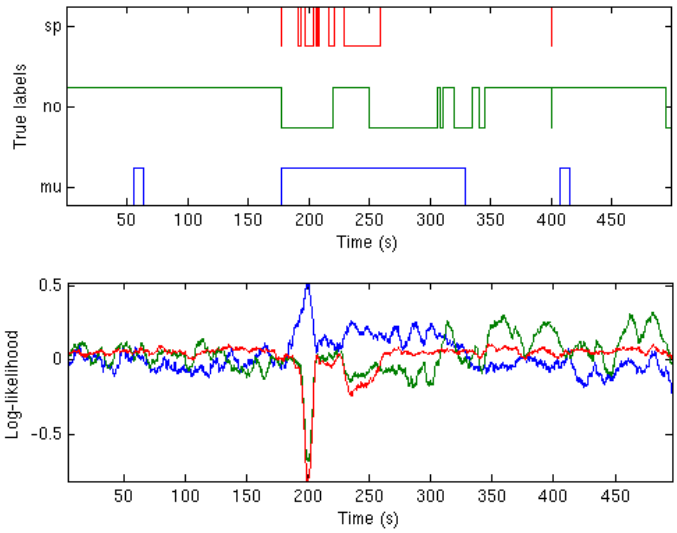


Fig. 2. Detail of the frame log-likelihoods for a 500-second segment of track11 after the mean filtering stage.

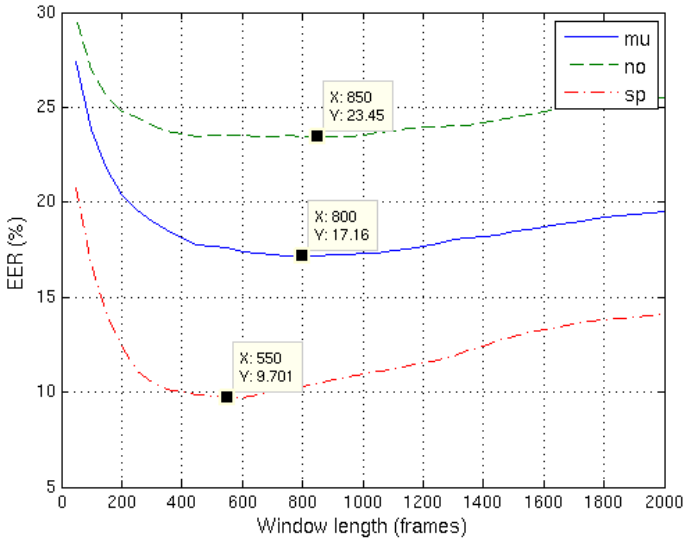


Fig. 3. EER as a function of the mean-filtering window-length, obtained for our development subset (tracks11-15). Best results are highlighted (X: window length, Y: EER).

5 Experimental setup and development results

Table 1 shows how the development data have been partitioned in order to be used for different purposes. One half of the development dataset has been devoted to train the acoustic models. From the remaining subset, one half has been used to find the optimum window length for the frame-scores mean-filtering, and the resulting frame-scores used to train the calibration rule; the final 5-track subset has been left apart in order to test the developed system.

Table 1. Dataset partitioning for system development.

Purpose	Track numbers
UBM training	01-10
Acoustic-class modeling	01-10
Window length optimization	11-15
Calibration training	11-15
Audio segmentation testing	16-20

Segmentation results obtained for our test subset (tracks 16-20) are shown in Table 2. As it can be seen, in spite of having acoustic-class detectors of relatively low detection performance (9.7% EER for 'sp', 17.2% EER for 'mu' and 23.4% EER for 'no'), the whole audio segmentation system achieves good performance compared with results shown in previous Albayzin ASE campaigns.

Table 2. Performance of the audio segmentation system: missed class time, false alarm class time, class error time and overall segmentation error, in seconds and percentages.

Error	Time (s)	% scored class time
Missed Class	2262.51	7.9
False Alarm Class	3057.21	10.6
Class error	853.85	3.0
Overall Segmentation Error	**21.43 %**	

6 Summary and conclusions

ATVS – Biometric Recognition Group has developed an efficient and light audio segmentation system. This system is based on three independent GMM-UBM acoustic-class detectors that can be developed and tuned independently. For instance, detectors in submitted systems make use of a different mean-filtering window-length and independent score-calibration rules, but they could be based in different features as well. Moreover, the adopted approach of modeling broad acoustic classes ('mu', 'no', 'sp') instead of the specific sub-classes given by all the possible combinations ('mu+no', 'sp+no', etc.) allows to develop a more robust system and avoids overfitting when the available training data is scarce. Finally, it can be seen in Table 3 that the computational requirements in terms of CPU time are very low, allowing the

testing to be run in 0.225xRT for each track. Experiments were carried out in a machine equipped with two Xeon Quad Core E5335 microprocessors at 2.0GHz (allowing 8 simultaneous threads) and 16GB of RAM.

Table 3. Testing time per track (~60 min) for the different stages and total time as a real-time (xRT) factor.

Stage	Time
Feature extraction	19 secs
Frame-by-frame scoring	13 min
Scores filtering and calibration	5 sec
Total (xRT)	~0.225

Acknowledgement

This work has been supported by the Spanish Ministry of Economy and Competitiveness (project CMC-V2: Caracterizacion, Modelado y Compensacion de Variabilidad en la Señal de Voz, TEC2012-37585-C02-01).

References

1. Niko Brummer, FoCal: toolkit for evaluation, fusion and calibration of statistical pattern recognizers (2008). Online: http://sites.google.com/site/nikobrummer/focal
2. J. Franco-Pedroso, I. Lopez-Moreno, D. T. Toledano, and J. Gonzalez-Rodriguez, ATVS-UAM System Description for the Audio Segmentation and Speaker Diarization Albayzin 2010 Evaluation. In Proceedings of FALA: VI Jornadas en Tecnologa del Habla and II Iberian SLTech Workshop, 2010, pp. 415–418.
3. Alfonso Ortega, Diego Castan, Antonio Miguel, Eduardo Lleida. The Albayzin 2014 Audio Segmentation Evaluation. Online: http://iberspeech2014.ulpgc.es/images/segm_eval.pdf
4. Reynolds, D., Quatier, T., Dunn, R., Speaker Verification Using Adapted Gaussian Mixture Models Digital Signal Processing, vol. 10, 19–41 (2000).
5. P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller, Jr., Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. Proc. ICSLP 2002, Sept. 2002, pp. 89-92.
6. Robbie Vogt and Sridha Sridharan, Minimising Speaker Verification Utterance Length through Confidence Based Early Verification Decisions. Lecture Notes in Computer Science Volume 5558, 2009, pp 454-463.

D

Presupuesto

1) Ejecución Material

- Compra de ordenador personal (Software incluido) — 2.000 €
- Material de oficina — 150 €
- Total de ejecución material — 2.150 €

2) Gastos generales

- 16 % sobre Ejecución Material — 344 €

3) Beneficio Industrial

- 6 % sobre Ejecución Material — 86 €

4) Honorarios Proyecto

- 1400 horas a 15 €/ hora — 21.000 €

5) Material fungible

- Gastos de impresión — 120 €
- Encuadernación — 180 €

6) Subtotal del presupuesto

- Subtotal Presupuesto — 25.700 €

7) I.V.A. aplicable

- 21 % Subtotal Presupuesto — 5.397 €

8) Total presupuesto

- Total Presupuesto — 31.097 €

Madrid, Junio 2015

El Ingeniero Jefe de Proyecto

Fdo.: Elena Gómez Rincón

Ingeniero de Telecomunicación

E

Pliego de condiciones

Pliego de condiciones

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un *Segmentación de audio mediante características cromáticas en ficheros de noticias*. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales.

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4 % del presupuesto y la provisional del 2 %.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad "Presupuesto de Ejecución de Contrata" y anteriormente llamado "Presupuesto de Ejecución Material" que hoy designa otro concepto.

Condiciones particulares.

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.