Escola Politècnica Superior de Castelldefels
UNIVERSITAT POLITÈCNICA DE CATALUNYA

TRABAJO FIN DE CARRERA

TÍTULO DEL TFC: Medidas de la calidad del audio en radiodifusión. Descriptores de audio y medidas objetivas.

TITULACIÓN: Ingeniería Técnica de Telecomunicaciones, especialidad Sistemas de Telecomunicaciones

AUTOR: Alejandro Albalá Díaz

DIRECTOR: Francesc Tarrés Ruiz

FECHA: 23 de julio de 2009

Título: Medidas de la calidad del audio en radiodifusión. Descriptores de audio y medidas objetivas

Autor: Alejandro Albalá Díaz

Director: Francesc Tarrés Ruiz

Fecha: 23 de julio de 2006

Resumen

Este proyecto tiene como objetivo principal la evaluación de la calidad subjetiva del audio. Para ello el proyecto se ha dividido en dos grandes bloques: primero conocer qué calidad subjetiva tienen unas ciertas muestras de audio, y posteriormente implementar un programa que pueda medir objetivamente dicha calidad.

Para la primera parte se ha realizado un estudio sobre la calidad del audio en radiodifusión. Este estudio consta de una encuesta a un total de 107 usuarios, donde se le exponen diferentes muestras de audio para que evalúen su calidad. Estas muestras se han obtenido grabando fragmentos de audio en diferentes lugares (como por ejemplo una casa, dentro de un coche…) y con diferente contenido (como por ejemplo un programa de noticias, una canción de pop actual etc.). Se han tomado muestras de varias emisoras para cada tipo de escenario y de programa. El proceso de grabación y adquisición de las muestras ha simulado las condiciones normales de escucha de una persona. Para ello, se ha usado microfonía colocada de una forma para simular la posición y distancia de los oídos humanos.

El mayor objetivo de la segunda parte (que es donde este proyecto se ha centrado más), se centra en la implementación de un software que pueda predecir con qué calidad oirá la audiencia una muestra de audio. Para ello ha sido necesario buscar métodos de extracción y representación de parámetros y características de señales de audio. La parte importante del proyecto consta en encontrar unos parámetros significativos que permitan ser relacionados con la calidad del audio, y posteriormente tratar esos parámetros para conseguir un factor de calidad objetivo que se corresponda con las pruebas subjetivas realizadas en la primera parte.

Merece la pena destacar que sobre este tema existe poca documentación, así que se trata de un primer prototipo que intenta probar si es factible medir una calidad subjetiva de manera objetiva.

Title: Audio broadcasting quality measurement. Audio descriptors and objective measurements

Author: Alejandro Albalá Díaz

Director: Francesc Tarrés Ruiz

Date: June, 23th 2009

Overview

The aim of this project is to evaluate the subjective audio quality of radio signals. The project has been divided in two main blocks: The first one intends to measure the subjective quality of some audio samples, whereas the second part is focused on the implementation of a program that could measure this quality in an objective way.

For the first block a subjective evaluation process has been defined. This process consists of asking a total of 107 people about the subjective quality of different audio samples. The samples consist of audio fragments from different radio stations acquired in different scenarios (home, inside the car…) and with different contents (like news, pop song…). The acquisition set-up tries to simulate
faithfully the way a person will listen to the samples. Thus, microphones have been placed in a manner to simulate the distances and positions of the human ear.

The second part (which is where this project has focused more), consists of the implementation of a software that can predict the quality experience of any audio sample in an objective way. This prediction is based on the extraction and representation of parameters and characteristics of audio signals. The important part of the project is to find significant parameters that can be related to audio quality, and subsequently treat these parameters to achieve an objective quality factor which corresponds to the subjective results obtained in the first part of the project

It should be remarked that there is not too many documentation about this topic in the literature. So this project should be considered as a first prototype that attempts to test whether it is feasible to measure a subjective quality with objective methods.

ÍNDICE

INTRODUCCIÓN

Este proyecto tiene como objetivo principal hacer un estudio sobre la calidad subjetiva del audio y conseguir evaluarla de manera objetiva. Para ello se ha decidido realizar una serie de cuestionarios con fragmentos de audio para que la gente opine sobre su calidad. Después estos resultados serán analizados y se intentará encontrar un patrón que permita medir esa calidad subjetiva de manera objetiva.

Cuando se valora la calidad de un fragmento de audio hay muchos parámetros en los que centrarse. Sí que es cierto que lo más importante en un fragmento de audio es el contenido, pero también es importante la calidad del mismo.

Existen muchos elementos que pueden afectar a la calidad del audio. Y en una cadena de radiodifusión la señal puede verse distorsionada por muchos motivos, como por ejemplo amplificadores, compresores, el propio canal, elementos digitales, tablas de mezclas, filtros etc. Por este motivo el objetivo básico de este proyecto ha sido el estudio y parametrización de cómo es posible medir la calidad de la señal de audio de manera objetiva pero teniendo una relación directa con la opinión de los oyentes.

Para cumplir el objetivo se ha hecho una comparativa entre dos emisoras que comparten características similares. Las emisoras han sido las del grupo CCRTV (Corporación Catalana de Radio y Televisión), y las del grupo Godó. En algún caso se ha utilizado muestras de Radio Nacional de España (RNE).

Se han escogido diferentes tipos de programas y diferentes escenarios donde se cree que la radio puede tener más importancia. Para cada escenario y programa se graban unas muestras para posteriormente generar un cuestionario con la finalidad de obtener unas opiniones de un número significativo de oyentes.

Una vez obtenidos los resultados, comienza la segunda parte: el análisis y la implementación de un programa experto que sea capaz de extraer un factor de calidad de la señal de audio de manera objetiva.

El análisis se ha basado en los descriptores definidos en el estándar MPEG-7 (ver [1]). Una vez extraídos los parámetros y los datos correspondientes se intenta encontrar la forma de encajar los resultados de las encuestas con características objetivas de la señal. Se analiza el comportamiento de cada descriptor en cada escenario y programa y se procesan para adecuar los resultados a las encuestas realizadas. Esto ha permitido tener una serie de referencias y plantillas que se han utilizado para obtener una nota del 0 al 10 (factor de calidad) que indica la calidad del audio. La aplicación se programa bajo este criterio en el entorno MATLAB.

Dada la complejidad del proyecto que se ha descrito se ha dividido en dos proyectos coordinados, que son: El primero contempla la parte de adquisición y tratamiento de muestras (ver [2]). El segundo (este trabajo), contiene la

explicación y el análisis de las muestras y la implementación de un programa experto.

La estructura que se ha seguido en este proyecto es la siguiente:

En el capítulo 1 se resume la primera parte del proyecto: las grabaciones y el método de adquisición de muestras.

En el capítulo 2 se define el estándar MPEG-7 y los descriptores que se han usado.

En el capítulo 3 se especifica el algoritmo y los cálculos aplicados en los descriptores para obtener resultados correlados con los resultados de las encuestas.

En el capítulo 4 se muestra la aplicación experta que se ha creado.

En el capítulo 5 se da una muestra de los resultados obtenidos. Se compara el resultado de la aplicación con las encuestas realizadas.

CAPÍTULO 1. ADQUISICIÓN DE MUESTRAS

1.1 Introducción

Este primer capítulo es un resumen de los conceptos que se detallan en el primer proyecto de los dos coordinados (ver [2]).

Uno de los pasos más importantes en este proyecto ha sido la definición de las condiciones de contorno (parámetros, escenarios, tipo de programas, etc.) de las pruebas realizadas a los usuarios que nos permitan obtener una idea clara de la calidad de la señal de audio.

El primer paso ha sido seleccionar entre bastantes características de la señal de audio: aquellas que se han considerado las más importantes para evaluar la calidad de las muestras.

Estos parámetros son:

- **Ecualización:** Ajuste de las frecuencias de reproducción de un sonido o fragmento de audio.
- **Dinámica:** De forma breve, es la diferencia entre el nivel de volumen más bajo y más alto en una muestra de audio.
- **Volumen:** La potencia de un determinado sonido.
- **Inteligibilidad:** La claridad y la definición de cada instrumento (individualmente) o de cada tertuliano en programas con intervención de personas.

1.2 Escenarios y programas

El entorno propuesto en este proyecto hace referencia a los sistemas donde el canal de comunicación es terrenal, ya sea FM, o internet.

No todos los programas utilizan el canal de comunicación de la misma forma, así pues, hace falta definir una lista de programas tipo que intente cubrir las diversas opciones. De entre muchas posibilidades se han escogido las siguientes:

- **Noticias:** Una voz humana principal.
- **Tertulias:** Varias voces humanas simultáneas.
- **Música moderna**
- **Música clásica**

También se ha decidido acotar las medidas a una serie de escenarios donde se ha creído que es donde más importancia tiene la radio:

- **Casa**
- **Coche**
- **Internet**
- **Auriculares**

En la **Tabla 1.1** se puede ver una relación de las emisoras escogidas y los programas seleccionados.

Tabla 1.1 Relación de emisoras a comparar para cada programa

Emisora	Programa	Emisora
Catalunya ràdio	Tertulias	**“RAC 1”**
iCat FM	Música Moderna	**“RAC 105”**
Catalunya informació	Noticias	**“RAC 1”**
Catalunya Música	Música clásica	**“Radio Nacional de España”**

1.3 Objetivos

El objetivo es grabar una serie de muestras para los diferentes escenarios. Después se elaborará un test donde se permite evaluar los parámetros mencionados en el punto 1.1. Se ha intentado hacer grabaciones de unos 5 a 10 minutos para poder escoger una muestra de 10 segundos que pudiera ser significativa para su análisis. Esta muestra es de 10 segundos para cumplir un tiempo de test límite entre 15 o 20 minutos.

1.4 Adquisición de muestras

1.4.1 Set-up para la adquisición de las muestras

La selección del material utilizado para la adquisición de las muestras es muy importante ya que todos los elementos introducen cierta distorsión o variación de la señal. Por ese motivo, y teniendo en cuenta los recursos disponibles, se han utilizado unos elementos en la cadena de adquisición y reproducción de las muestras cuyas respuestas frecuenciales son lo más planas y uniformes posibles.

El material usado ha sido el siguiente:

- **Micrófonos:** Shure PG81[1].
- **Tarjeta de sonido:** M-Audio MobilePre USB[2].
- **Auriculares:** AKG K240[3]y SonyTec Netsound 280.

Para más información sobre el material ver [2]. A continuación se describen brevemente los elementos en la cadena de adquisición para cada uno de los escenarios.

1.4.1.1 Escenario Coche

Se ha captado con los micrófonos el sonido ambiente dentro del coche mientras se reproducían los programas de radio.

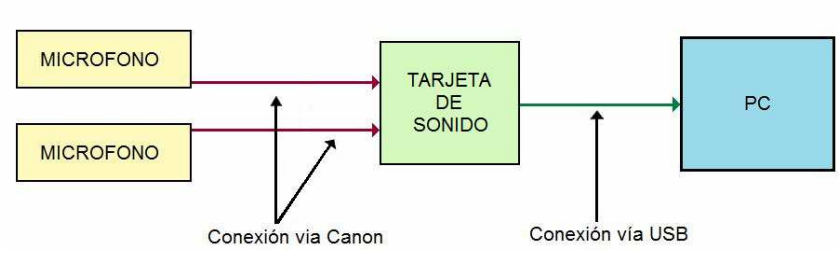


Fig 1.1 Diagrama de bloques del conexionado en la grabación del escenario *Coche***.**

La **Fig 1.1** muestra el conexionado que se ha realizado para este escenario. Los dos micrófonos se han posicionado en la misma dirección y sentido opuesto. Esto se ha conseguido con un sistema binaural como muestra la **Fig 1.2.**

[1] http://akg.com/site/products/powerslave,id,252,pid,252,nodeid,2,_language,EN.html
[2] http://www.m-audio.com/products/en_us/MobilePreUSB.html
[3] http://es.shure.com/ProAudio/Products/WiredMicrophones/uses_pro_PG81-XLR_content

Fig 1.2 Diseño y disposición de los micrófonos

1.4.1.2 Escenario Casa

Para la grabación en casa se utiliza el mismo conexionado y la misma disposición de micrófonos que en el escenario *Coche*. La diferencia es que el audio se reproduce con altavoces de una minicadena estándar.

1.4.1.3 Escenario Auriculares

Para este caso el conexionado cambia: se han sustituido los micrófonos y directamente se adquiere la muestra desde la salida de auriculares de una minicadena (*Kenwood DP-722*). Esto se muestra en la **Fig 1.3**.

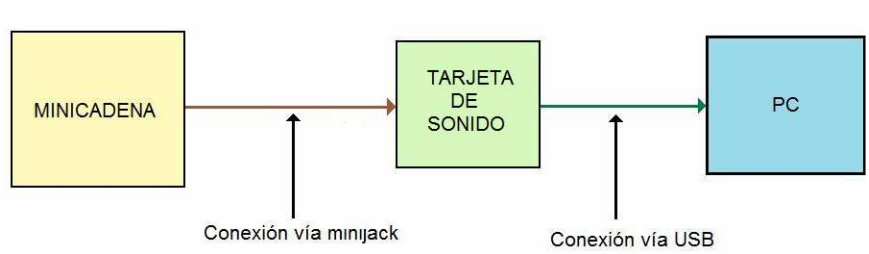


Fig 1.3 Diagrama de bloques del conexionado en la grabación del escenario *Auriculares*

1.4.1.4 Escenario Internet

En el caso del escenario Internet, directamente se cogen las muestras descargadas de la red, mediante *podcast* almacenados en los servidores de cada emisora.

1.4.2 Pre procesado

A pesar de que la longitud original de las muestras adquiridas es de unos 5 minutos, se ha extraído 10 segundos de cada muestra grabada.

Dicho fragmento seleccionado, será lo que denominaremos la muestra original. A parte de la original se dispondrá de 4 ecualizaciones más para filtrar la señal original y preguntar a los oyentes.

Estas ecualizaciones son:

- **Ecualización creciente:** Contrarresta el posible exceso de graves en la transmisión o la falta de agudos.
- **Ecualización decreciente:** Contrarresta el posible exceso de agudos en la transmisión o la falta de graves.
- **Ecualización en U invertida:** Misma respuesta que el oído humano. Amplifica la zona media (220-3500 Hz). Atenúa graves y agudos.
- **Ecualización en U:** Respuesta frecuencial inversa a la del oído humano.

A continuación se muestran las ilustraciones de las curvas.

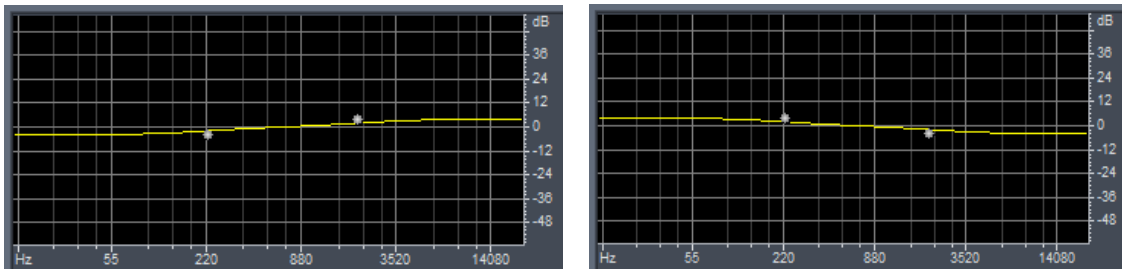


Fig 1.4 Ecualización creciente y ecualización decreciente

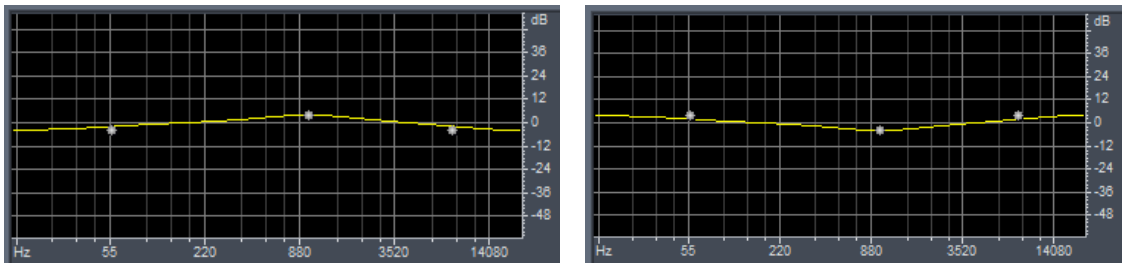


Fig 1.5 Ecualización en 'U' invertida y ecualización en 'U'

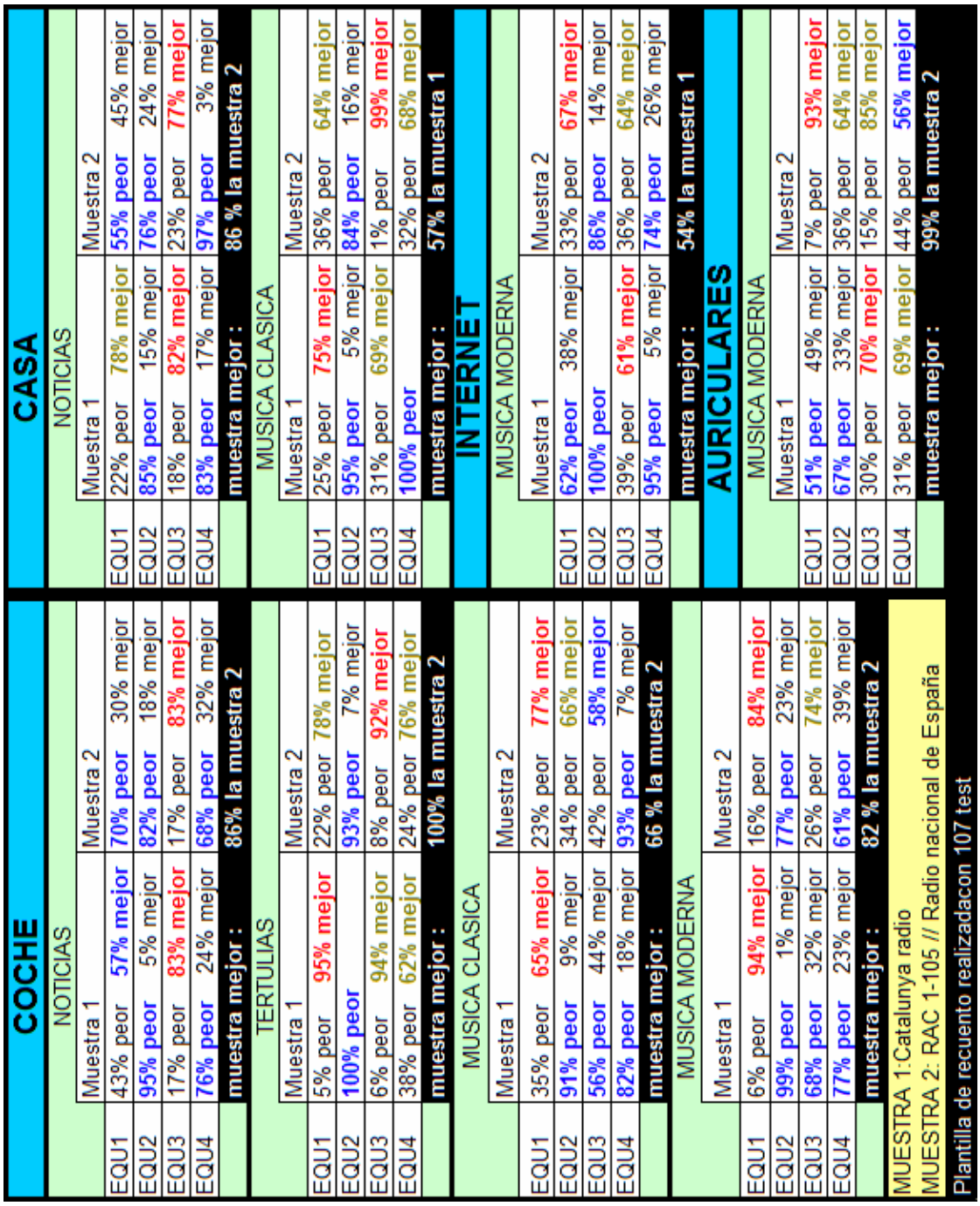
1.4.3 Tests y resultados

De los parámetros definidos en el punto 1.1, se ha elaborado un test para poder evaluarlos. A continuación se explica qué método se ha usado para su evaluación:

- **Volumen:** El oyente elige manualmente un volumen del 0 al 100 que le es más cómodo para oír el fragmento.
- **Dinámica:** Después de explicarle el significado de este parámetro, se le pregunta al oyente si cree que el fragmento tiene dinámica o carece de ella.
- **Inteligibilidad:** El oyente pone una nota del 0 al 10 según la claridad y nitidez del fragmento.
- **Ecualización:** Se le pregunta para cada una de las ecualizaciones mostradas en el punto 1.4.2 si oye mejor el fragmento original o filtrado con esa ecualización.

El prototipo de test se puede ver en el **ANEXO A.**

Después de realizar estos tests y hacer un recuento de todos los resultados se obtienen las plantillas de la **Fig 1.6** y la **Fig 1.7**.



COCHE					CASA				
NOTICIAS					NOTICIAS				
	Muestra 1		Muestra 2			Muestra 1		Muestra 2	
EQU1	43% peor	57% mejor	70% peor	30% mejor	EQU1	22% peor	78% mejor	55% peor	45% mejor
EQU2	95% peor	5% mejor	82% peor	18% mejor	EQU2	85% peor	15% mejor	76% peor	24% mejor
EQU3	17% peor	83% mejor	17% peor	83% mejor	EQU3	18% peor	82% mejor	23% peor	77% mejor
EQU4	76% peor	24% mejor	68% peor	32% mejor	EQU4	83% peor	17% mejor	97% peor	3% mejor
	muestra mejor :		86% la muestra 2			muestra mejor :		86 % la muestra 2	
TERTULIAS					MUSICA CLASICA				
	Muestra 1		Muestra 2			Muestra 1		Muestra 2	
EQU1	5% peor	95% mejor	22% peor	78% mejor	EQU1	25% peor	75% mejor	36% peor	64% mejor
EQU2	100% peor		93% peor	7% mejor	EQU2	95% peor	5% mejor	84% peor	16% mejor
EQU3	6% peor	94% mejor	8% peor	92% mejor	EQU3	31% peor	69% mejor	1% peor	99% mejor
EQU4	38% peor	62% mejor	24% peor	76% mejor	EQU4	100% peor		32% peor	68% mejor
	muestra mejor :		100% la muestra 2			muestra mejor :		57% la muestra 1	
MUSICA CLASICA					INTERNET				
	Muestra 1		Muestra 2		MUSICA MODERNA				
EQU1	35% peor	65% mejor	23% peor	77% mejor		Muestra 1		Muestra 2	
EQU2	91% peor	9% mejor	34% peor	66% mejor	EQU1	62% peor	38% mejor	33% peor	67% mejor
EQU3	56% peor	44% mejor	42% peor	58% mejor	EQU2	100% peor		86% peor	14% mejor
EQU4	82% peor	18% mejor	93% peor	7% mejor	EQU3	39% peor	61% mejor	36% peor	64% mejor
	muestra mejor :		66 % la muestra 2		EQU4	95% peor	5% mejor	74% peor	26% mejor
MUSICA MODERNA						muestra mejor :		54% la muestra 1	
	Muestra 1		Muestra 2		AURICULARES				
EQU1	6% peor	94% mejor	16% peor	84% mejor	MUSICA MODERNA				
EQU2	99% peor	1% mejor	77% peor	23% mejor		Muestra 1		Muestra 2	
EQU3	68% peor	32% mejor	26% peor	74% mejor	EQU1	51% peor	49% mejor	7% peor	93% mejor
EQU4	77% peor	23% mejor	61% peor	39% mejor	EQU2	67% peor	33% mejor	36% peor	64% mejor
	muestra mejor :		82 % la muestra 2		EQU3	30% peor	70% mejor	15% peor	85% mejor
					EQU4	31% peor	69% mejor	44% peor	56% mejor
						muestra mejor :		99% la muestra 2	

MUESTRA 1:Catalunya radio
MUESTRA 2: RAC 1-105 // Radio nacional de España
Plantilla de recuento realizadacon 107 test

Fig 1.6 Plantilla de resultados de los parámetros de *ecualización*

En la plantilla se muestra lo siguiente:

- En color negro aparecen los porcentajes de votación obtenidos de preguntar qué muestra (original sin reecualizar) creen que tiene mejor calidad. Esta comparativa se realiza entre las dos emisores que se han grabado para el mismo tipo de escenario y de programa.
- En cada escenario aparece cada muestra con sus cuatro ecualizaciones. Los porcentajes de cada ecualización se ha obtenido preguntando si la ecualización le parece mejor que la original. Por ejemplo, en escenario Coche/Noticias, un 83% de los encuestados prefieren la ecualización 3 a la muestra original (en ambas emisoras).

Los resultados mostrados pertenecen a un muestreo de 107 personas.

Como conclusiones de esta plantilla se puede destacar:

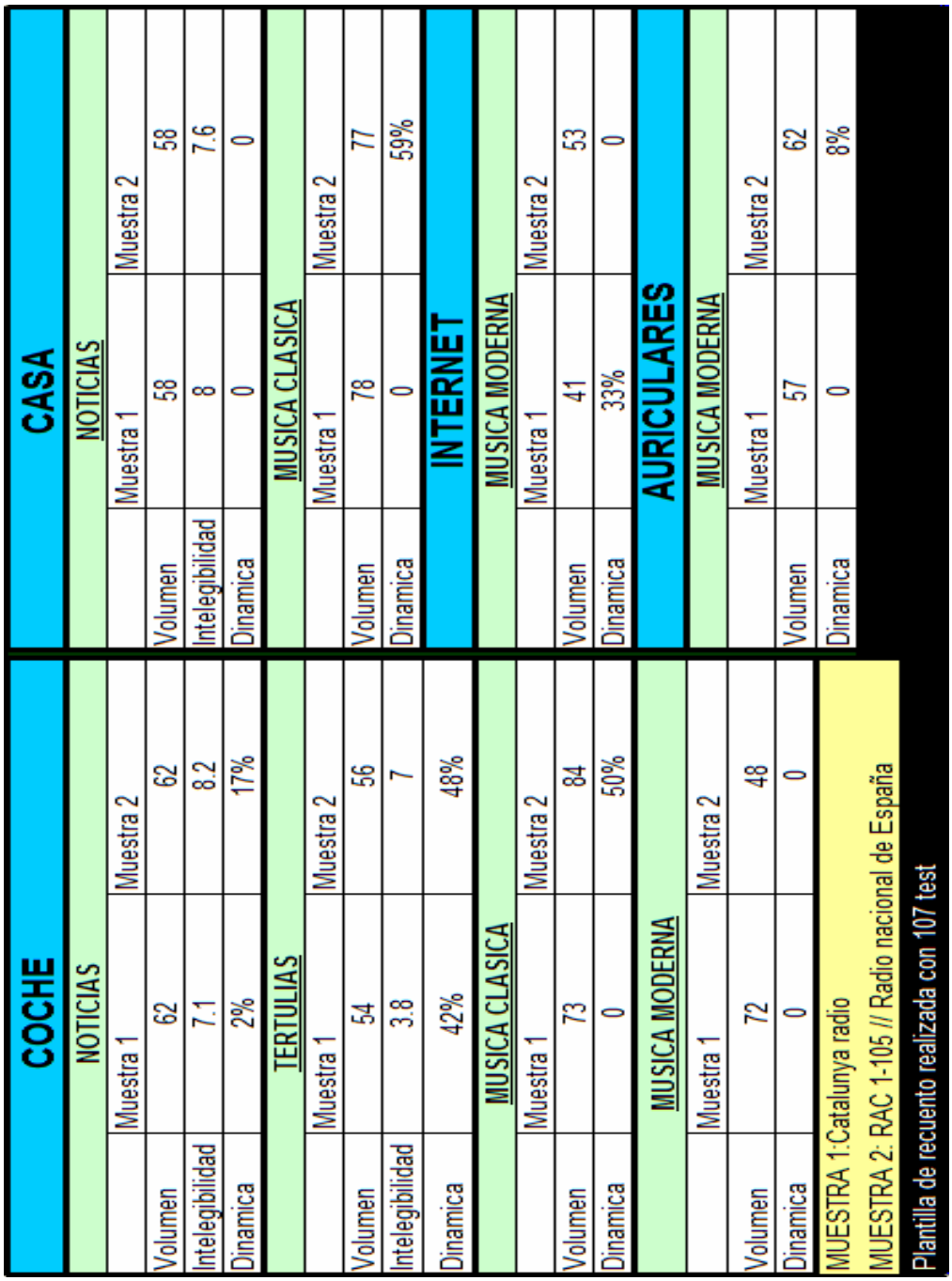
- La emisora correspondiente a la muestra 1 (Catalunya *radio*) tiene subjetivamente peor calidad que la correspondiente a la muestra 2 (*RAC1*, *RAC105*, *Radio Nacional de España*) en todos los casos estudiados, excepto para los escenarios *Casa* (Música clásica) y *Internet* (Música moderna), donde la población estima una calidad similar para ambas emisoras,

- La ecualización 1, *Ecualización creciente,* enfatiza las frecuencias agudas y atenúa las graves, para los oyentes mejora la calidad. Al amplificar los agudos, se amplifica los armónicos de las señales y resulta un sonido que los oyentes lo perciben con mejor calidad.

- La ecualización 2 (ecualización decreciente) reduce la calidad del audio transmitido por estas emisoras según los resultados subjetivos sonsacados del sondeo. Una causa muy probable es que atenúa la banda aguda, que es donde se concentra la calidad del audio.

- La ecualización 3 mejora considerablemente la calidad.

- La ecualización 4 reduce la calidad en la mayoría de escenarios excepto *Coche* (Tertulias) y *Auriculares* (Música moderna). Esta ecualización atenúa las frecuencias medias (alrededor de 1 KHz.), que es donde se concentra la información, por lo tanto , parece ser que las emisoras referidas a las muestras 1 y 2 no enfatizan mucho las frecuencias medias; o lo que se ha creído más probable, enfatizan más los graves y agudos.



COCHE		
NOTICIAS		
	Muestra 1	Muestra 2
Volumen	62	62
Intelegibilidad	7.1	8.2
Dinamica	2%	17%
TERTULIAS		
	Muestra 1	Muestra 2
Volumen	54	56
Intelegibilidad	3.8	7
Dinamica	42%	48%
MUSICA CLASICA		
	Muestra 1	Muestra 2
Volumen	73	84
Dinamica	0	50%
MUSICA MODERNA		
	Muestra 1	Muestra 2
Volumen	72	48
Dinamica	0	0
MUESTRA 1:Catalunya radio		
MUESTRA 2: RAC 1-105 // Radio nacional de España		
Plantilla de recuento realizada con 107 test		

CASA		
NOTICIAS		
	Muestra 1	Muestra 2
Volumen	58	58
Intelegibilidad	8	7.6
Dinamica	0	0
MUSICA CLASICA		
	Muestra 1	Muestra 2
Volumen	78	77
Dinamica	0	59%
INTERNET		
MUSICA MODERNA		
	Muestra 1	Muestra 2
Volumen	41	53
Dinamica	33%	0
AURICULARES		
MUSICA MODERNA		
	Muestra 1	Muestra 2
Volumen	57	62
Dinamica	0	8%

Fig 1.7 Plantilla de resultados de los parámetros de *volumen*, *inteligibilidad* y *dinámica*.

La plantilla de la **Fig 1.7** muestra los resultados de los parámetros: volumen, inteligibilidad y dinámica.

De esta plantilla podemos concluir lo siguiente:

- Hablando de volumen, las emisoras emiten con un volumen más bajo que lo que prefiere la población. Esto no debería ser un problema ya que todos los aparatos de reproducción disponen de regulador de volumen con el cual pueden situarlo a su gusto.

- Los resultados referentes al parámetro de dinámica son poco coherentes, por ejemplo: en Música Clásica, independientemente de la emisora o escenario, la dinámica debería aparecer con un porcentaje alto, a diferencia de la opinión de los oyentes. Otro caso incoherente es que en programas como noticias o tertulias los oyentes opinan que hay algo de dinámica, incluso más que en música clásica, y en principio un composición clásica tiene muchos más niveles de volumen que una tertúlia. Se cree que pueden diferir de la realidad porque la población no ha llegado a entender el concepto en su totalidad.

- En cuanto a inteligibilidad se puede ver que para el programa de noticias, ambos escenarios (*Coche* y *Casa*) tienen una inteligibilidad similar. En cambio, para el programa tertulias (escenario *Coche*) la inteligibilidad de la muestra 1 (perteneciente a la emisora *Catalunya radio*) es mucho peor que la de la muestra 2 (RAC1). La inteligibilidad es un parámetro difícil de medir objetivamente y complicado de testear, así que puede ser que los resultados no sean todo lo fiables que se pretende.

CAPÍTULO 2. DESCRIPTORES

2.1 Introducción

En este capítulo se explicará las herramientas utilizadas para el análisis de las muestras.

En este punto se sabe la opinión de los oyentes de radio sobre unos escenarios y programas concretos. A partir de aquí se puede encontrar tanto lo que más gusta como lo que menos gusta. Con esto se pretende encontrar la forma de relacionar características o parámetros objetivos de la señal con la opinión subjetiva de los oyentes.

Para cumplir este objetivo se ha recurrido a técnicas de extracción de datos sobre fragmentos de audio.

Para ello, se ha estudiado el estándar de descripción de contenido multimedia MPEG-7. Dicho estándar define las herramientas necesarias para describir y almacenar la información asociada a la descripción de señales de audio (ver [3]) o de vídeo. En este estándar existen los denominados descriptores, una herramienta que permite extraer parámetros sobre un fragmento dado. Estos datos pueden ser, por ejemplo, el espectro, el timbre, la potencia etc.

Estos descriptores ya han sido previamente probados por el propio estándar para el reconocimiento, clasificación e indexación de señales de audio. Por ello se ha creído conveniente su uso

2.2 Clasificación

Existen un total de 17 descriptores de audio en el estándar MPEG-7, y están clasificados en 2 grandes grupos dependiendo de su nivel de descripción semántico:

- Bajo nivel (*Low Level Descriptors).* Extraen parámetros y características del fragmento. Como por ejemplo, la potencia, la envolvente del espectro, la forma de onda etc.
- Alto nivel *(High Level Descriptors):* Combinan descriptores de bajo nivel para extraer diferentes características como por ejemplo el reconocimiento de sonidos, el tempo de una pieza musical etc.

2.2.1 Bajo nivel

Se clasifican en 6 subgrupos (ver [4]):

1. *Basic*

- **AudioWaveform (AWF):** Representa la forma de onda con su máximo y su mínimo.
- **AudioPower (AP):** Define la potencia de las muestras de la señal.

2. *BasicSpectral*

- **AudioSpectrumEnvelope (ASE):** Representa la envolvente del espectro.
- **AudioSpectrumFlatness (ASF):** Es un indicativo de cómo es de plano el espectro para cada banda de frecuencias.
- **AudioSpectrumCentroid (ASC):** Indica si el espectro es una representación de una señal con más presencia de altas o bajas frecuencias**.**
- **AudioSpectrumSpread (ASS):** Muestra la desviación del espectro respecto al resultado del descriptor *AudioSpectrumCentroid.*

3. *SignalParameters*

- **AudioHarmonicity (AH):** Indica: HarmonicRatio (ratio de los armónicos del fragmento) y UpperLimitofHarmonicity (punto del espectro a partir del cual no hay armónicos).
- **AudioFundamentalFrequency (AFF):** Indica la frecuencia fundamental del fragmento analizado.

4. *SpectralBasis*

- **AudioSpectrumBasis (ASB):** Basado en el AudioSpectrumEnvelope, transforma el espectro de una señal en una representación más sencilla para un bajo coste computacional.
- **AudioSpectrumProjection (ASP):** Se utiliza generalmente para el reconocimiento de sonidos junto al AudioSpectrumBasis. Transforma un espectro en decibelios a un espectro previamente computado por el AudioSpectrumBasis.

5. *TimbralTemporal*

- **LogAttackTime (LAT):** Caracteriza el ataque de un sonido: Tiempo que tarda el sonido desde que comienza hasta que consigue su amplitud máxima. El descriptor devuelve el logaritmo del tiempo de ataque.
- **TemporalCentroid (TC):** El punto en el tiempo donde se focaliza la máxima energía.

6. *TimbralSpectral*

- **SpectralCentroid (SC):** La media ponderada de la potencia espectral.
- **HarmonicSpectralCentroid (HSC):** El valor medio de las amplitudes de los armónicos.
- **HarmonicSpectralDeviation (HSD):** La desviación de los picos respecto a la envolvente calculada.
- **HarmonicSpectralSpread (HSS):** La desviación típica de la media.
- **HarmonicSpectralVariation (HSV):** La correlación en frecuencia entre un paquete de muestras y el siguiente.

2.2.2 Alto nivel

- **AudioSignature:** Realiza una estadística sobre el descriptor AudioSpectrumFlatness. Analiza si todas las medidas con el citado descriptor se correlan de alguna forma.
- **Timbre:** Contiene LAT, HSC, HSD, HSS, y HSV para la identificación de instrumentos armónicos y SC, TC, y LAT para la identificación de instrumentos percusivos. Se evalúa la calidad del timbre para comprobar si es una característica descriptiva de un sonido.
- **SoundModel:** Consiste en ASB y ASP. Se utiliza para el reconocimiento de sonidos ambientales.

2.3 Descriptores utilizados

Para cumplir el principal objetivo de este proyecto, se deben analizar las muestras en cuanto a dinámica, volumen, inteligibilidad y ecualización (frecuencias y espectro), así que no todos los descriptores mencionados se van a utilizar.

Se han escogido los que pueden dar más información sobre los parámetros que se quieren medir. Para medir volumen y dinámica hacen falta descriptores del dominio temporal, como son *AudioWaveform* y *AudioPower*. Para medir la ecualización hacen falta descriptores en el dominio frecuencial como pueden ser *AudioSpectrumEnvelope*, *AudioSpectrumCentroid*, *AudioSpectrumSpread*, *AudioSpectrumFlatness* y *AudioFundamentalFrequency*.

Los descriptores de alto nivel en el estándar MPEG-7 se centran en el reconocimiento y clasificación de sonidos, no de la calidad del audio. Por este motivo no se ha escogido ninguno de alto nivel.

Se podría decir que en este proyecto se ha creado un descriptor de alto nivel propio, ya que se han combinado descriptores de bajo nivel para extraer un factor de calidad.

Todos los descriptores en el estándar MPEG-7 utilizan un parámetro llamado *hopSize*. Este parámetro indica la resolución temporal de los descriptores. Cuando se analiza una muestra de audio se computa por fragmentos de la longitud que especifica el parámetro *hopSize*. Este parámetro se expresa en milisegundos. El valor por defecto de este parámetro corresponde a 10 ms de la señal de audio, que es el valor que se ha utilizado en este proyecto.

2.3.1 AudioWaveform

Una buena forma de representar la forma de onda de una señal es mostrar su máximo y su mínimo en los diferentes fragmentos temporales (no solapados entre sí). Para cada fragmento se guarda el valor máximo y el mínimo.

El descriptor *AudioWaveform* (AWF) representa la serie temporal de cada pareja de valores (el máximo y el mínimo) en cada fragmento temporal.

Este descriptor proporciona una estimación de la forma de onda, y se representa dibujando las dos series de valores:

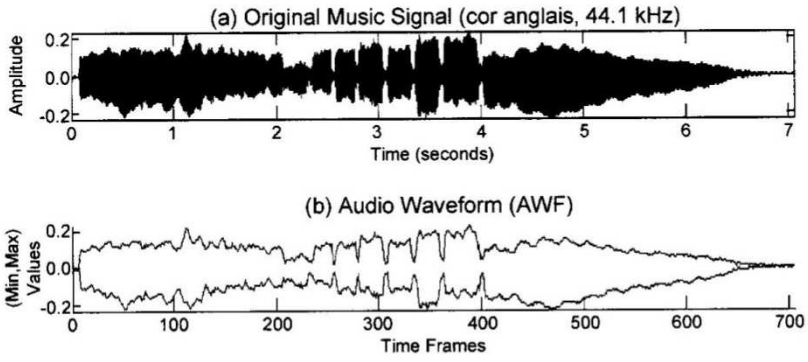


Fig 2.1 Ejemplo del descriptor AudioWaveform

2.3.2 AudioPower

El descriptor *AudioPower* (AP) describe la potencia instantánea de la señal de audio. Es la media al cuadrado de los valores de la señal en los diferentes fragmentos temporales (no solapados entre sí) al igual que en el descriptor *AudioWaveform*. El coeficiente del descriptor AP en el fragmento *'l'* se define como:

$$AP(l) = \frac{1}{N_{hop}} \sum_{n=0}^{N_{hop}-1} \left| s(n + lN_{hop}) \right|^2 \qquad (0 \leq l \leq L-1)\,, \qquad \textbf{(2.1)}$$

,donde L es el total número de los fragmentos temporales y N_{hop} es el número de muestras correspondiente al tiempo definido por el parámetro *hopSize*.

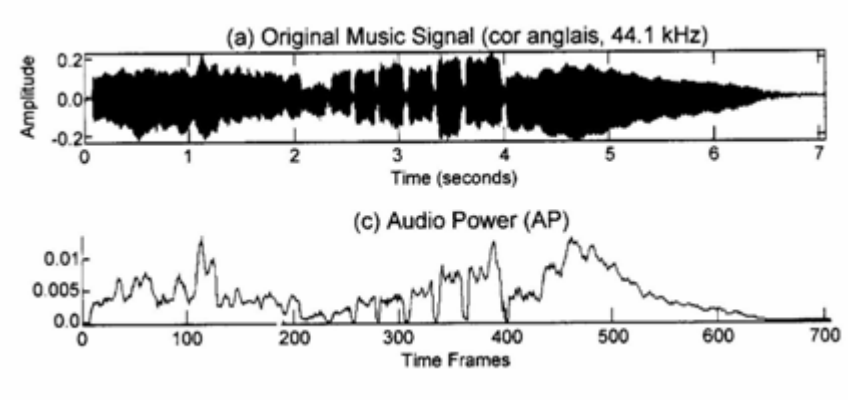


Fig 2.2 Ejemplo del descriptor AudioPower

2.3.3 AudioSpectrumEnvelope

El descriptor *AudioSpectrumEnvelope* (ASE) representa un espectro de potencia sobre una escala logarítmica de frecuencias. Se obtiene sumando la energía del espectro de potencia original dentro de una serie de bandas frecuenciales.

Las bandas están distribuidas logarítmicamente (en base 2) entre dos márgenes frecuenciales definidos por los parámetros *loEdge (margen inferior)* y *hiEdge (margen superior)*. Estos parámetros los especifica el usuario. La resolución espectral *r* de las bandas entre esos dos márgenes puede tomar hasta 8 valores, desde 1/16 de octava, hasta 8 octavas:

$$r = 2^j \ octavas \quad (-4 \leq j \leq 3) \qquad \textbf{(2.2)}$$

Los márgenes (*Edge),* ya sea el inferior o el superior, deben cumplir la siguiente condición:

$$Edge = 2^{rn}\ x\ 1KHz \tag{2.3}$$

donde *r* es la resolución en octavas y *n* es un número entero. El rango de frecuencias por defecto va desde los 62.5 Hz, a los 16000Hz y *r* = 1. Así que el intervalo por defecto corresponde a uno de 8 octavas, logarítmicamente centrado en 1KHz.

El número de bandas que corresponden a *r* es:

$$B_{in} = 8/r. \tag{2.4}$$

y los ejes inferior y superior de cada banda son:

$$loF_b = loEdge\ x\ 2^{(b-1)r}$$
$$hiF_b = loEdge\ x\ 2^{br} \qquad (1 \leq b \leq B_{in}) \tag{2.5}$$

La suma de los coeficientes de potencia en la banda $[loF_b, hiF_b]$ da como resultado el coeficiente *ASE* para esa banda frecuencial. El coeficiente para la banda *b* es:

$$hiK_b = round\left(\frac{hiF_b N_{FT}}{Fs}\right);\ loK_b = round\left(\frac{loF_b N_{FT}}{Fs}\right) \tag{2.6}$$

$$ASE(b) = \sum_{k=loK_b}^{hiK_b} P(k) \quad (1 \leq b \leq B_{in})$$

,donde $P(k)$ son los coeficientes del espectro de potencia definidos en **(2.7)**:

$$P(k) = \frac{1}{N_{FT} E_w} |S(k)|^2 \qquad si\ k = 0, k = \frac{N_{FT}}{2} \tag{2.7}$$
$$P(k) = \frac{1}{N_{FT} E_w} |S(k)|^2 \qquad si\ 0 < k < \frac{N_{FT}}{2}$$

donde $S(k)$ son los coeficientes de la DFT aplicada después de enventanar un fragmento, y donde E_w es la energía del fragmento (sumatorio de todas sus muestras elevadas al cuadrado). N_{FT} es el número de puntos de la DFT calculada.

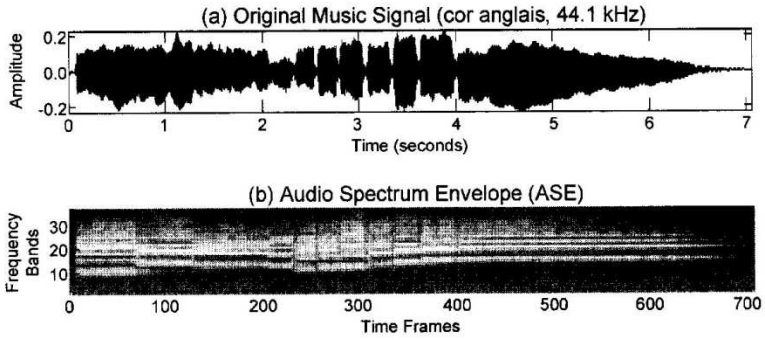


Fig 2.3 Ejemplo del descriptor AudioSpectrumEnvelope

2.3.4 AudioSpectrumCentroid

El *AudioSpectrumCentroid* (ASC) da como resultado el "centro de gravedad" del espectro de potencia (expresado logarítmicamente).

Para evitar componentes de continua (DC) y/o componentes de muy baja frecuencia con demasiado peso en el espectro, no se han tenido en cuenta valores por debajo de 62.5 Hz. Se define K_{low} como dicho límite en el dominio discreto frecuencial.

Esto resulta en un nuevo espectro, en el cual el primer coeficiente contiene la suma de todos los coeficientes originales hasta el límite mencionado anteriormente, y el resto de coeficientes persisten iguales. A este espectro se le denomina $P'(k')$. Y por tanto existen otras frecuencias correspondientes a ese espectro, denominadas $f'(k')$, donde $f'(0) = 31.25\ Hz$; $f'(k') = f(k' + k_{low})$.

Finalmente y teniendo en cuenta lo explicado anteriormente, ASC se define como:

$$ASC = \frac{\sum_{k'=0}^{\left(\frac{N_{FT}}{2}\right)-k_{low}} \log_2\left(\frac{f'(k')}{1000}\right) P'(k')}{\sum_{k'=0}^{\left(\frac{N_{FT}}{2}\right)-k_{low}} P'(k')} \qquad \textbf{(2.8)}$$

Una referencia para interpretar el ASC es que un valor 0 significa que la frecuencia central es cercana a 1 KHz.

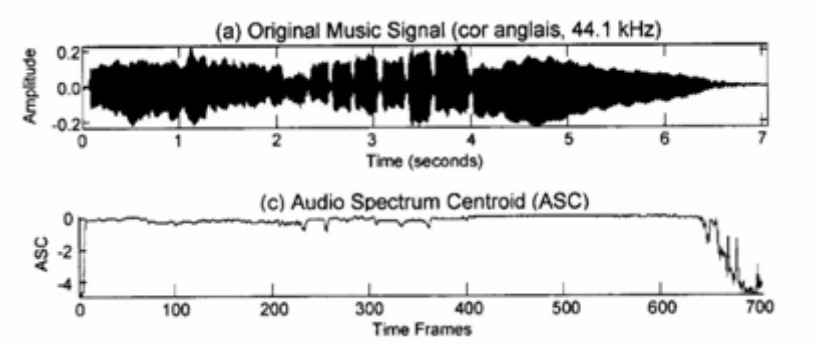


Fig 2.4 Ejemplo del descriptor AudioSpectrumCentroid

2.3.5 AudioSpectrumFlatness

El descriptor *AudioSpectrumFlatness* (ASF) muestra cuánto de plano es el espectro de un fragmento de audio. Dicho de otra forma: indica la desviación dentro de una banda frecuencial respecto una forma espectral plana (ruido blanco).

El primer paso del ASF es el cálculo del espectro de potencia, definido en la ecuación 2.7, por cada uno de los fragmentos que completan toda la señal. En este caso los coeficientes son obtenidos de fragmentos consecutivos (no solapados entre sí).

Dentro del rango definido por el intervalo [*loEdge, hiEdge*] el espectro se divide en diferentes bandas frecuenciales de un cuarto de octava. Por tanto el parámetro *hiEdge* debe ser:

$$loEdge = 2^{n/4}\ x\ 1KHz \qquad \textbf{(2.9)}$$

$$hiEdge = 2^{B/4}\ x\ loEdge \qquad \textbf{(2.10)}$$

donde *B* es el número de bandas, y *n* es un número entero.

Se agranda un 10 % cada banda para cubrir un posible error en la frecuencia de muestreo. Los límites de la banda son:

$$loF_b = 0.95\ x\ loEdge\ x\ 2^{\frac{1}{4}(b-1)}$$

$$hiF_b = 1.05\ x\ loEdge\ x\ 2^{\frac{1}{4}b} \qquad (1 \leq b \leq B) \qquad \textbf{(2.11)}$$

Finalmente:

$$ASF(b) = \frac{\sqrt[hiK_b - loK_b + 1]{\prod_{k=loK_b}^{hiK_b} P_g(k)}}{\frac{1}{hiK_b - loK_b + 1} \sum_{k=loK_b}^{hiK_b} P_g(k)} \qquad (1 \leq b \leq B) \qquad \textbf{(2.12)}$$

Donde hiK_b y loK_b son loF_b y hiF_b en el dominio frecuencial digital obtenido por **(2.6)**.

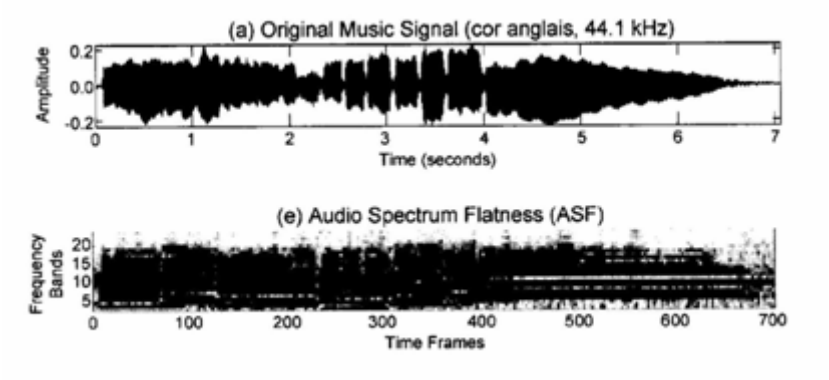


Fig 2.5 Ejemplo del descriptor AudioSpectrumFlatness

2.3.6 AudioSpectrumSpread

El *AudioSpectrumSpread* (ASS) define la desviación del espectro en función del resultado obtenido por el *AudioSpectrumCentroid*: Un valor pequeño indica una gran concentración del espectro en ese punto. Puede ser útil para saber si el espectro es muy concentrado o está muy disperso. En cierta manera se puede encontrar información sobre la inteligibilidad y ecualización.

$$ASS = \sqrt{\frac{\sum_{k'=0}^{\left(\frac{N_{FT}}{2}\right)-k_{low}} \left[\log_2\left(\frac{f'(k')}{1000}\right) - ASC\right]^2 P'(k')}{\sum_{k'=0}^{\left(\frac{N_{FT}}{2}\right)-k_{low}} P'(k')}} \qquad \textbf{(2.13)}$$

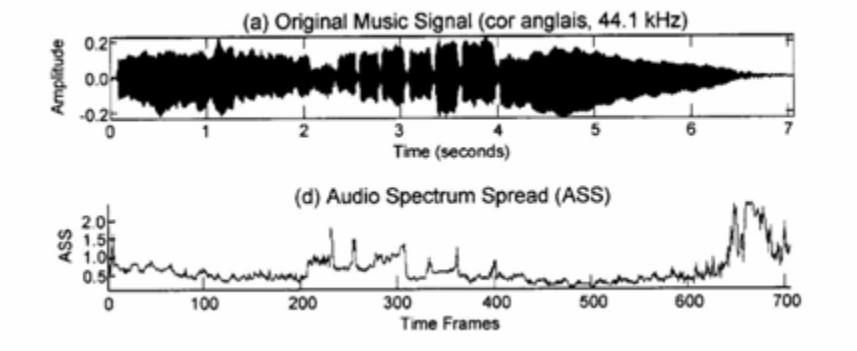


Fig 2.6 Ejemplo del descriptor AudioSpectrumSpread

2.3.7 AudioFundamentalFrequency

El descriptor *AudioFundamentalFrequency* (AFF) estima la frecuencia central en segmentos donde la señal se asume periódica. Puede ser útil para encontrar el *pitch* de un fragmento musical, o hablado.

El estándar no define una forma definida ni una normativa específica para la extracción de este descriptor. Existen varias técnicas diferentes de estimación de *pitch*.

Se permite especificar el rango frecuencial donde se va a buscar esa frecuencia fundamental. Este rango se configura con los parámetros *loLimit* y *hiLimit.*

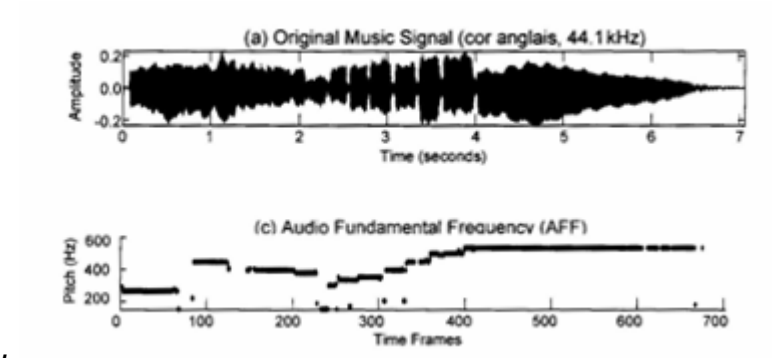


Fig 2.7 Ejemplo del descriptor AudioFundamentalFrequency

2.4 Mel-Frequency Cepstrum Coefficients (MFCC)

A parte de los descriptores definidos anteriormente, existen otras técnicas para caracterizar una señal.

Los *Mel-Frequency Cepstrum Coefficients* se utilizan para el reconocimiento de señales de voz. Se trata de una serie de filtros triangulares, cuyas frecuencias centrales están separadas en base a una escala de Mel. La escala de Mel viene definida por la siguiente ecuación:

$$F_{mel} = 1127.01048 \log_e \left(1 + \frac{f(Hz)}{700}\right) \quad \textbf{(2.14)}$$

$$f(Hz) = 700(e^{F_{mel}/1127.01048} - 1) \quad \textbf{(2.15)}$$

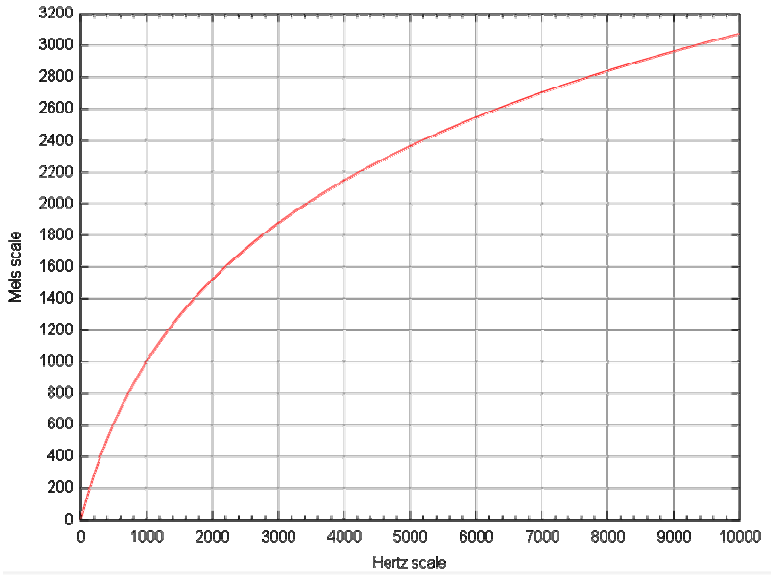


Fig 2.8 Función de transferencia de la escala de Mel

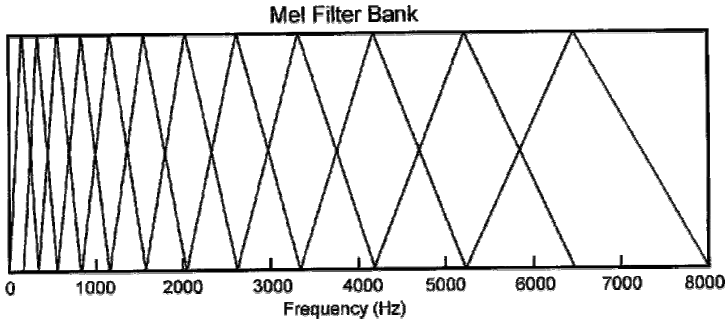


Fig 2.9 Filtros de Mel en el margen frecuencial de la voz

La escala de Mel es una escala de *pitch* equidistante, es decir, que el incremento en frecuencia que percibe el oyente es directamente proporcional al incremento de valor dentro de esta escala. Una referencia entre la escala de Mel y las frecuencias en hercios es 1KHz. 1000Hz, 40dB por encima del umbral de audición del oyente, equivale a un pitch de 1000 Mels.

Generalmente se usa un banco de entre 25 o 30 filtros en la banda que se quiere analizar. En este proyecto se usa un banco de 28 filtros de 0 a 22100 Hz.

El diagrama de bloques de este procedimiento es el mostrado en la **Fig 2.10**.

En el proceso de enventanado, la ventana que se ha usado ha sido una ventana *hamming.* Es una de las más usadas junto con la *hanning*.

Después de filtrar con los filtros de Mel, se obtiene una serie de coeficientes que indican la energía de cada banda a la salida del filtro, la etapa Log realiza el logaritmo de cada uno de estos coeficientes.

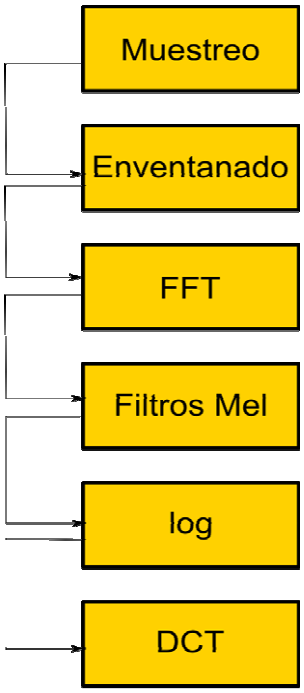


Fig 2.10 Etapas de los MFCCs

Dado su uso en el reconocimiento de patrones de voz y reconocimiento de locutor, se ha creído conveniente el probar dicho parámetro también para la medida objetiva de la calidad de señales de audio.

CAPÍTULO 3. SISTEMA EXPERTO PARA MEDIR LA CALIDAD DE LA SEÑAL DE AUDIO

Uno de los objetivos principales de este proyecto es crear un método capaz de medir la calidad de la señal de audio radiofónica de manera objetiva. Dicho de otra manera, un sistema que puntúe la calidad de un fragmento de audio de tal manera que el resultado esté correlado con los resultados de los cuestionarios explicados en el **Capítulo 1**.

En este capítulo se explicará cómo se han seleccionado los diferentes descriptores que tendrán un peso importante en la puntuación final del sistema. También se detallará las contribuciones y cambios realizados para poder resaltar aún más los cambios de calidad entre diferentes muestras.

Se podría haber optado por diversos métodos, por ejemplo, métodos estadísticos avanzados que intenten aprender a partir de las muestras de entrenamiento (cuestionarios). El principal problema de estos métodos es que se pueden adecuar demasiado al conjunto de muestras de entrenamiento con el peligro que estas no sean suficientemente representativas. Por eso, se ha optado más por el siguiente procedimiento:

1- Selección de la muestra de referencia.
2- Selección de los descriptores.
3- Combinación de descriptores. Asignación de pesos.

3.1 Selección de la muestra de referencia

Para lo que se explicará en este capítulo, es necesario tener una plantilla de referencia, es decir, una muestra o conjunto de muestras que representen la máxima calidad o calidad óptima. Esta plantilla se obtiene mediante la combinación de las muestras que los oyentes han considerado mejores.

COCHE				
	NOTICIAS			
	Muestra 1		Muestra 2	
EQU1	43% peor	57% mejor	70% peor	30% mejor
EQU2	95% peor	5% mejor	82% peor	18% mejor
EQU3	17% peor	83% mejor	17% peor	83% mejor
EQU4	76% peor	24% mejor	68% peor	32% mejor
	muestra mejor :		86% la muestra 2	
	TERTULIAS			
	Muestra 1		Muestra 2	
EQU1	5% peor	95% mejor	22% peor	78% mejor
EQU2	100% peor		93% peor	7% mejor
EQU3	6% peor	94% mejor	8% peor	92% mejor
EQU4	38% peor	62% mejor	24% peor	76% mejor
	muestra mejor :		100% la muestra 2	
	MUSICA CLASICA			

CASA				
	NOTICIAS			
	Muestra 1		Muestra 2	
EQU1	22% peor	78% mejor	55% peor	45% mejor
EQU2	85% peor	15% mejor	76% peor	24% mejor
EQU3	18% peor	82% mejor	23% peor	77% mejor
EQU4	83% peor	17% mejor	97% peor	3% mejor
	muestra mejor :		86 % la muestra 2	
	MUSICA CLASICA			
	Muestra 1		Muestra 2	
EQU1	25% peor	75% mejor	36% peor	64% mejor
EQU2	95% peor	5% mejor	84% peor	16% mejor
EQU3	31% peor	69% mejor	1% peor	99% mejor
EQU4	100% peor		32% peor	68% mejor
	muestra mejor :		57% la muestra 1	
	INTERNET			

Fig 3.1 Excel con el resultado de las encuestas

Como se ha comentado en el punto 1.4.3, aquí se muestran los resultados de preguntar qué muestra original se cree que tiene más calidad, y qué muestra reecualizada es mejor en calidad que su muestra original.

Los fragmentos más votados son elegidos para formar la plantilla. Esta plantilla es utilizada como la máxima calidad, la muestra con la que hay que comparar. Es importante destacar que la plantilla es un fragmento o varios fragmentos combinados con ponderaciones:

Existen dos niveles de ponderación:

- **Ponderación de muestras originales:** Pondera la opinión de los oyentes sobre las muestras originales (sin reecualizar). En el caso que la más votada no supere el 75% de los votos, se aplica esta ponderación. Las muestras a ponderar son las mejores ecualizaciones de cada emisora.
- **Ponderación de muestras ecualizadas:** Si (y sólo si) no ha sido necesaria la primera ponderación, dentro de la muestra más votada se ponderan las dos mejores ecualizaciones. Si hay una ecualización que destaca muy por encima de las demás solo se elige ésta como plantilla final. Solamente en este caso no hace falta ninguna ponderación.

Por ejemplo, en la **Fig 3.2** se muestra los porcentajes elegidos por los oyentes para dos fragmentos de Música moderna a través del canal Coche. La muestra 2 ha sido elegida por un 84% de los oyentes, así que la ponderación entre muestras originales no se hace. Las ecualizaciones 1 y 3 de la muestra 2 han sido las más elegidas, con un 84 y 74% respectivamente.

MUSICA MODERNA				
	Muestra 1		Muestra 2	
EQU1	6% peor	94% mejor	16% peor	84% mejor
EQU2	99% peor	1% mejor	77% peor	23% mejor
EQU3	68% peor	32% mejor	26% peor	74% mejor
EQU4	77% peor	23% mejor	61% peor	39% mejor
	muestra mejor :		**82 % la muestra 2**	

Fig 3.2 Resultados de las encuestas para el escenario Coche/Música moderna

Los factores que dan el peso adecuado a cada muestra se calculan de la siguiente forma:

$$P_{EQU1} = \frac{0.84}{0.84+0.74} = 53.16\% \quad ; \quad P_{EQU3} = \frac{0.74}{0.84+0.74} = 46.84\% \qquad \textbf{(3.1)}$$

Otro ejemplo es el mostrado en la **Fig 3.3**. En este caso la muestra 1 ha sido elegida por un 54% de los oyentes. En este caso sí que se pondera por muestras originales.

INTERNET				
	MUSICA MODERNA			
	Muestra 1		Muestra 2	
EQU1	62% peor	38% mejor	33% peor	67% mejor
EQU2	100% peor		86% peor	14% mejor
EQU3	39% peor	61% mejor	36% peor	64% mejor
EQU4	95% peor	5% mejor	74% peor	26% mejor
	muestra mejor :		54% la muestra 1	

Fig 3.3 Resultados de las encuestas para el escenario Internet/Música moderna

Se escoge la mejor ecualización de cada muestra. En este caso:

- Muestra 1 Ecualización 3
- Muestra 2 Ecualización 1

Y se pondera de la siguiente manera:

$$P_{M1} = \frac{0.54x0.61}{0.54x0.61+0.46x0.67} = 51.66\% \quad P_{M2} = \frac{0.46x0.67}{0.54x0.61+0.46x0.67} = 48.34\% \qquad \textbf{(3.2)}$$

Es la operación equivalente a **(3.1)** ponderando con los porcentajes de las muestras originales.

3.2 Selección de los descriptores.

3.2.1 Selección de la mejor y peor muestra para cada escenario

Una vez seleccionada la muestra o muestras de referencia, es necesario decidir qué descriptores son importantes para medir la calidad de la señal de audio. Para ello, se define un proceso por el cual cada descriptor se evalúa por separado: se elige la mejor muestra y la peor a partir de las plantillas y se define un factor de correlación entre ellas dos para cada descriptor.

AURICULARES				
MUSICA MODERNA				
	Muestra 1		Muestra 2	
EQU1	51% peor	49% mejor	7% peor	93% mejor
EQU2	67% peor	33% mejor	36% peor	64% mejor
EQU3	30% peor	70% mejor	15% peor	85% mejor
EQU4	31% peor	69% mejor	44% peor	56% mejor

Fig 3.4 Resultados para el escenario Auriculares/Música moderna

Para ello comparamos la peor muestra con la plantilla explicada en el punto 3.1.

El procedimiento es comparar la opinión de la gente y también comparar la "opinión" de los descriptores. Analizando el resultado de las dos comparaciones se puede decidir qué descriptores pueden ser más importantes.

3.2.2 Método de comparación. Factor de correlación.

Una forma de comparar es la correlación **(3.3)**. Se trata de una correlación que mide la proporcionalidad de una señal con otra, es decir, si una señal crece con una pendiente 3, y la otra decrece con pendiente 3 (es decir, pendiente 3 negativa), el resultado de la correlación es -1. Si crecen por igual el resultado es 1. La **Fig 3.6** es un ejemplo de una correlación con resultado -1.

$$\cos(\alpha) = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}} \qquad \textbf{(3.3)}$$

En **(3.3)** se muestra la correlación de dos vectores Donde x_i y y_i son los valores de los vectores a comparar y $\bar{x}$ y $\bar{y}$ son los valores promedio de cada vector.

Sabiendo que $c = \cos(\alpha)$:

- Si c = 1, α = 0°, ambos vectores son paralelos.
- Si c = 0, α = 90°, ambos vectores son ortogonales.
- Si c =-1, α = 180°, ambos vectores son paralelos con sentido opuesto.

3.2.3 Lectura y cálculo de los descriptores

Los descriptores de un fragmento de audio están guardados en disco sobre un formato de fichero XML[4]. Se han conseguido extraer esos datos mediante la librería *Mpeg7AudioEnc* (ver [5]). En el **ANEXO B** se pueden ver detalles más concretos sobre el proceso de cálculo y extracción de los descriptores y su posterior escritura en el fichero XML en cuestión.

Con el procedimiento de correlación se obtienen las correlaciones mostradas en la **Tabla 3.1**.

Tabla 3.1 Correlaciones obtenidas para cada descriptor y escenario

Escenario	Awf	Ap	Aff	Ase	Ass	Asc	Asf
NotCoche	6.75%	5.22%	18.64%	65.15%	1.59%	-9.55%	93.23%
TertCoche	-9.01%	-4.11%	1.89%	62.05%	-6.13%	-5.39%	95.34%
McCoche	-0.6%	2.34%	5.61%	77.19%	10.73%	5.72%	82.56%
MmCoche	6.57%	3.01%	-11.77%	78.83%	3.19%	10.22%	90.39%
NotCasa1	-5.08%	-6.89%	9.07%	80.78%	-1.85%	-5.32%	90.02%
NotCasa2	-6.19%	-7.53%	12.29%	77.02%	-0.24%	-6.68%	90.12%
McCasa	2.05%	0.06%	-1.92%	81.44%	12.02%	-22.61%	84.75%
MmInternet	-4.73%	-2.42%	3.61%	28.44%	1.80%	-7.29%	60.71%
MmAuriculares1	3.04%	1.5%	5.24%	19.31%	-1.72%	-4.12%	88.63%
MmAuriculares2	4.05%	2.12%	6.51%	76.69%	-5.1%	-7.67%	87.51%

La tabla ha sido extraída con la correlación entre las mejores y las peores muestras es decir, entre dos muestras muy diferentes (en términos de calidad subjetiva). Para concordar con este resultado, los descriptores comparados también deberían ser muy diferentes (correlaciones cercanas a 0), así que aquellos descriptores que hayan sido menos correlados son los que más concuerdan con la opinión de los oyentes.

Un dato a destacar es que los descriptores *AudioSpectrumFlatness (Asf)* y *AudioSpectrumEnvelope (Ase)* han resultado muy correlados.

El *Asf*, como se ha explicado en el **Capítulo 2**, indica cómo de plano es el espectro. En este caso es normal que de un valor alto de correlación, ya que aunque sean fragmentos de diferente calidad, hay muchas bandas frecuenciales en común que son planas, y también otras bandas (como las frecuencias bajas) que suelen variar en todas las muestras. Si son las mismas bandas las que son planas, o muy variantes, el descriptor se correla mucho. Por este motivo se ha descartado este descriptor.

[4] *Extensible Markup Language*. Lenguaje de etiquetas jerárquico que permite intercambiar información organizadamente.

En el caso del *Ase* también se podría descartar, pero al ser la envolvente del espectro, se ha intentado extraer toda la información posible, así que no se ha descartado y se ha optado por aplicarle algunas operaciones para poder corregir esa alta correlación y conseguir que concuerde con los resultados de las encuestas.

Tal y como sucede con las muestras de mejor calidad, en algunos escenarios como es Música moderna Auriculares (ver **Fig 3.5**) no hay una muestra claramente peor, sino que hay dos. En este caso esas dos muestras son:

- Muestra 1 ecualización1
- Muestra 1 ecualización 2

Primero se correla primero una y luego otra. Por esto motivo aparecen dos filas de correlaciones en este escenario (MmAuriculares1 y MmAuriculares2).

Existen más casos donde sucede esto mismo, pero no sólo con dos muestras, sino con más. En estos casos no se ha aplicado ningún tipo de ponderación debido a que la aplicación alcanza un grado de complejidad muy alto.

AURICULARES				
	MUSICA MODERNA			
	Muestra 1		Muestra 2	
EQU1	51% peor	49% mejor	7% peor	93% mejor
EQU2	67% peor	33% mejor	36% peor	64% mejor
EQU3	30% peor	70% mejor	15% peor	85% mejor
EQU4	31% peor	69% mejor	44% peor	56% mejor
	muestra mejor :		**99% la muestra 2**	

Fig 3.5 Resultados para el escenario Auriculares/Música Moderna

3.2.4 Enfatización de las diferencias

Interesa enfatizar lo máximo posible los pequeños cambios en la señal que contengan interés para la evaluación de la calidad de la señal. Por este motivo a cada descriptor se le aplican unas operaciones para enfatizar las características que pueden ser distintivas a la hora de evaluar la calidad.

3.2.4.1 AudioSpectrumCentroid

Después de analizar los resultados se ha decidido prescindir de este descriptor por el siguiente motivo:

De dos fragmentos de audio se extrae el descriptor *AudioSpectrumCentroid* de cada uno de ellos. Se obtienen las gráficas de la **Fig 3.6**.

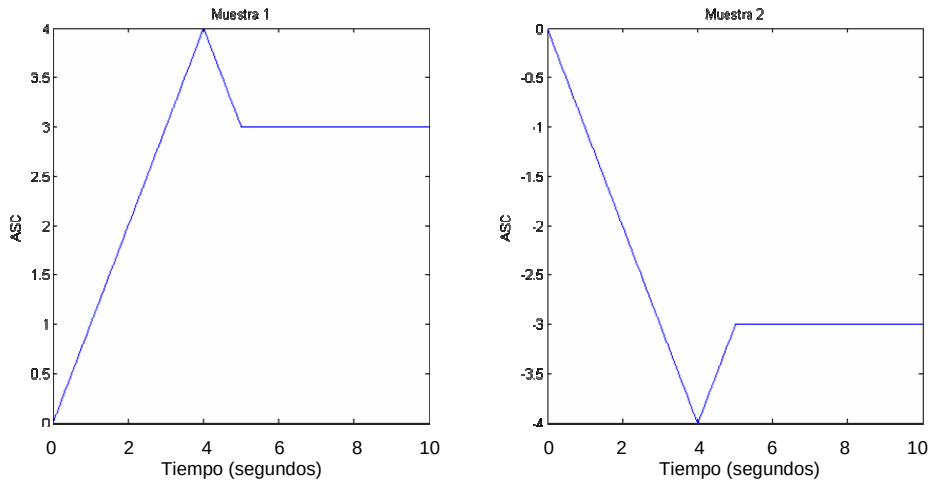


Fig 3.6 Comparativa de los AudioSpectrumCentroid de dos muestras diferentes

Si estos dos resultados se correlan, se obtiene un -1 (-100%). Siguiendo el significado de correlación, quiere decir que estos dos elementos son casi iguales, multiplicando uno por '-1' se obtiene el otro. Pero realmente, la información que da cada uno no es igual, sino todo lo contrario. Cada uno nos da un centro de gravedad del espectro (en cada instante de tiempo) muy diferente al otro.

Debido a esto el resultado de la correlación no es lo que se pretende obtener. Por este motivo se ha descartado este descriptor. No obstante, cabe decir que la información que proporciona este descriptor es espectral y existen más descriptores para este tipo de información.

3.2.4.2 AudioSpectrumEnvelope

A pesar de que en la **Tabla 3.1**, el descriptor *AudioSpectrumEnvelope*, en todas las muestras da un alto valor de correlación, este descriptor proporciona una gran cantidad de información, ya que es la envolvente del espectro, así que las diferencias que notan los oyentes, seguramente se puedan ver reflejadas aquí también. Es necesario poder encontrar los elementos que permitan distinguir estas diferencias.

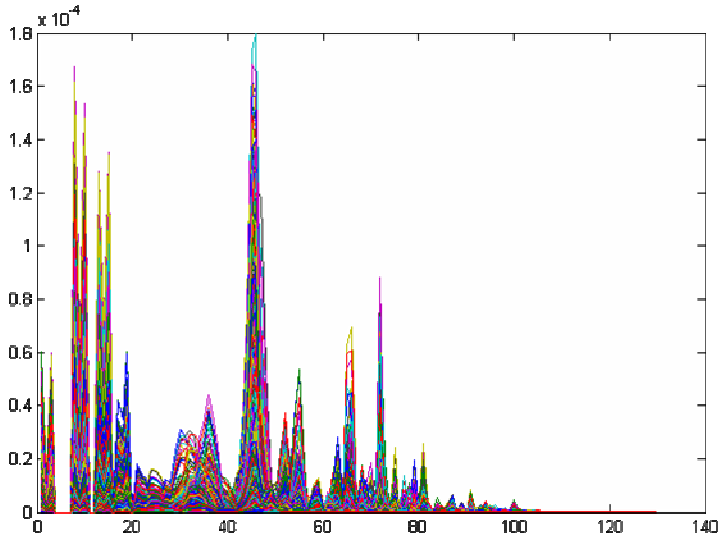


Fig 3.7 AudioSpectrumEnvelope de un fragmento

Viendo la **Fig 3.7**, se observa que un gran porcentaje de la energía se concentra en las frecuencias bajas. Esta parte será la que tenga todo el peso en el resultado de la correlación.

Cualquier señal de audio se puede descomponer en una frecuencia fundamental y otras frecuencias múltiples de ésta llamadas armónicos. Todas estas frecuencias conforman lo que se llama el timbre de un sonido. Si los armónicos de alta frecuencia se atenúan por cualquier motivo, resulta un sonido de igual timbre (o prácticamente igual) pero un poco más pobre (o menos brillante) que el original. Por este motivo se puede definir que la calidad del audio se concentra en las frecuencias altas.

Con el fin de darle mayor importancia a esta calidad se ha utilizado una curva de ponderación como la representada en la **Fig 3.8**.

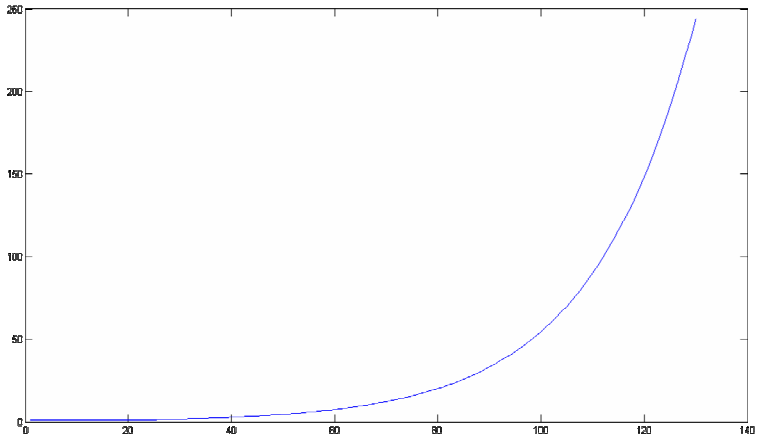


Fig 3.8 Curva de corrección del AudioSpectrumEnvelope

Además de la curva de ponderación, para que la varianza de la señal crezca (y así sea más sensible a la correlación) se eleva al cuadrado la señal resultante.

En la **Fig 3.9** se muestra el resultado de la operación. Se ve que el espectro corregido es más equilibrado que el original. Se han reducido notablemente los graves y por el contrario se realzan los agudos.

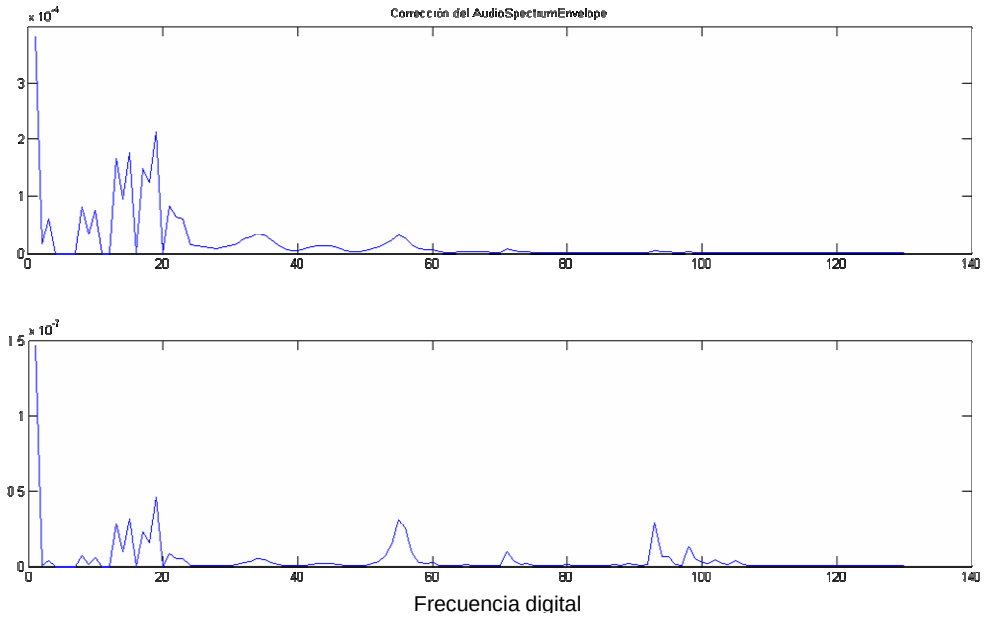


Fig 3.9 Corrección del AudioSpectrumEnvelope

Tabla 3.2 Correlaciones con el Ase modificado

Escenario	Awf	Ap	Aff	Ase	Ass
NotCoche	6.75%	5.22%	18.64%	**4.71%**	1.59%
TertCoche	-9.01%	-4.11%	1.89%	**4.96%**	-6.13%
McCoche	-0.6%	2.34%	5.61%	**2.40%**	10.73%
MmCoche	6.57%	3.01%	-11.77%	**1.57%**	3.19%
NotCasa1	-5.08%	-6.89%	9.07%	**85.9%**	-1.85%
NotCasa2	-6.19%	-7.53%	12.29%	**70.2%**	-0.24%
McCasa	2.05%	0.06%	-1.92%	**46.8%**	12.02%
MmInternet	-4.73%	-2.42%	3.61%	**46.6%**	1.80%
MmAuriculares1	3.04%	1.5%	5.24%	**-0.408%**	-1.72%
MmAuriculares2	4.05%	2.12%	6.51%	**8.35%**	-5.1%

Se ha mejorado bastante ya que las correlaciones de este descriptor han bajado notablemente después de la corrección aplicada. Aún así en escenarios como Notícias/Casa o incluso Música Clásica/Casa o Música Moderna/Internet no resulta un valor lo suficientemente bajo.

3.2.4.3 AudioFundamentalFrequency y AudioSpectrumSpread

Como se muestra en la **Fig 3.10**, y como ya se ha comentado anteriormente, gran parte de la energía se concentra en las bajas frecuencias.

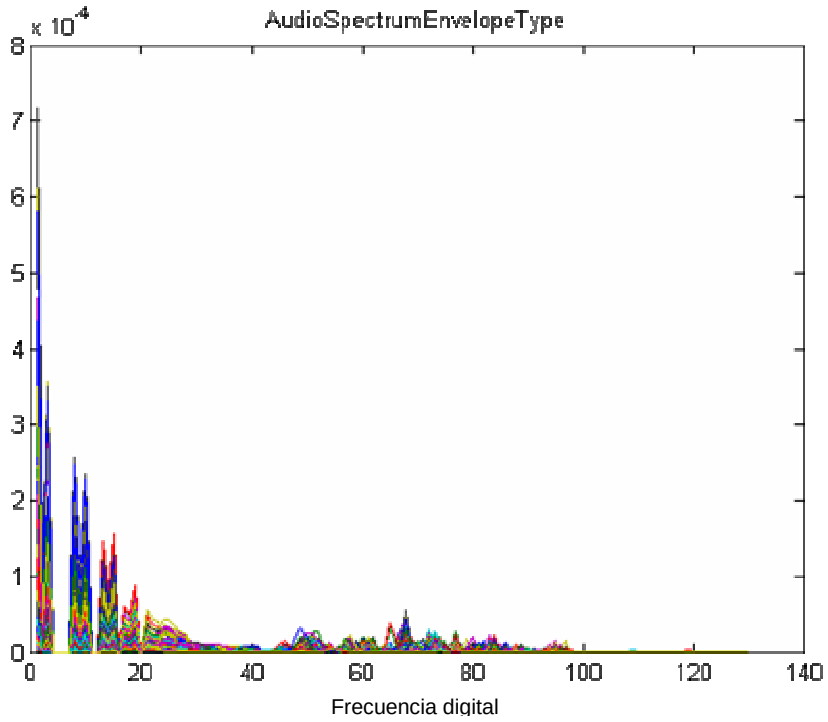


Fig 3.10 AudioSpectrumEnvelope de una muestra

Debido a esto la gran parte de la información de los descriptores *AudioFundamentalFrequency* y *AudioSpectrumSpread*, estarán en un rango de frecuencias bajas. La combinación de estos descriptores puede dar información sobre qué banda de frecuencias tiene más energía y si ésta tiene un nivel de concentración alto. Esta misma información se puede obtener con los MFCCs (Mel Frequency Cepstrum Coefficients) explicados en el **capítulo 2**. Además es una ventaja porque comparar los MFCCs es más sencillo e intuitivo que combinar los dos descriptores comentados y compararlos.

Para la correlación de los ya citados coeficientes, se le aplican las mismas operaciones que para el descriptor *AudioSpectrumEnvelope,* ya que la información que muestran los dos tienen que ver con el espectro de potencia. A parte, los primeros coeficientes no se tienen en cuenta, ya que son coeficientes de muy baja frecuencia que no aportan información en cuanto a calidad.

3.2.4.4 AudioWaveform y AudioPower

Estos dos descriptores se han elegido para medir la dinámica de las señales, es decir, la variación de nivel de volumen dentro de una misma señal.

La **Fig 3.11** muestra un *AudioWaveform* extraído de uno de los fragmentos analizados:

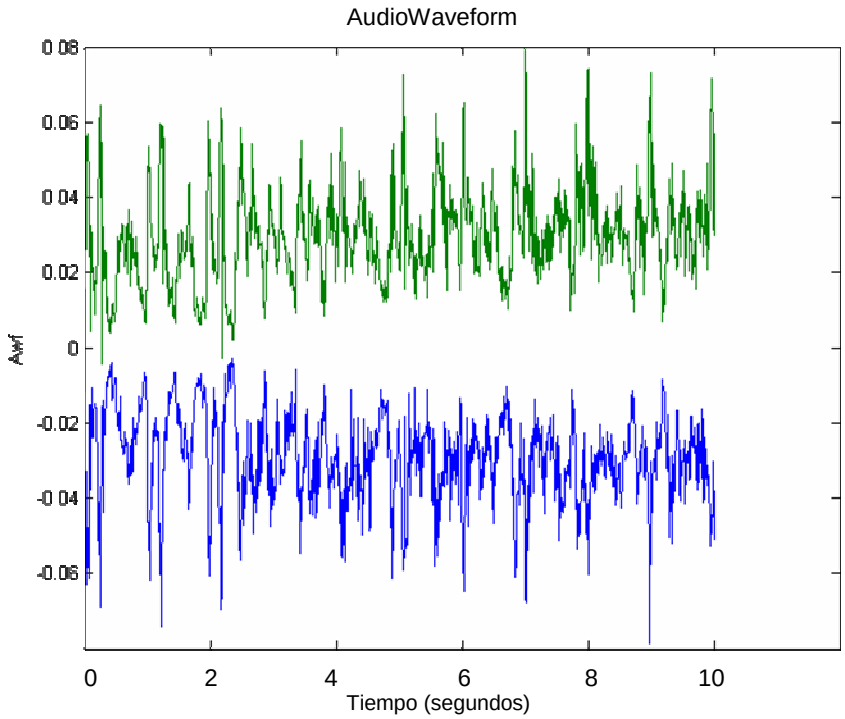


Fig 3.11 AudioWaveform de una muestra

De la **Fig 3.11** es difícil extraer matemáticamente una conclusión en cuanto a dinámica. Esto se debe a la exagerada variación de la forma de onda y a la aparición de múltiples picos. Se ha decidido aplicar un filtro de mediana para este descriptor.

Un filtro de mediana consiste en ordenar de menor a mayor (o viceversa) las muestras, y escoger la muestra que está posicionada en el centro. Es decir:

1. Considerando estas muestras: **[34 2 8 90 57 63 12]**,
2. Se ordenan de menor a mayor: **[2 8 12 34 57 63 90]**
3. El filtro de mediana para este fragmento sería el **34**.

Con este filtro se consiguen eliminar las muestras con una amplitud muy distante a las de sus vecinas. El resultado se puede ver en la **Fig 3.12**.

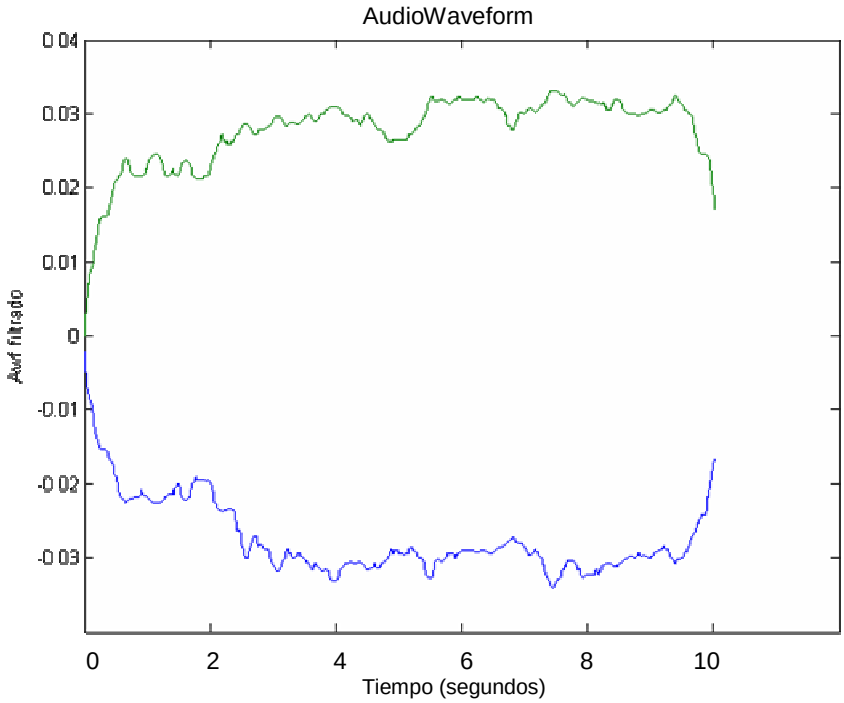


Fig 3.12 AudioWaveform después de un filtro de mediana con ventana 128 muestras

Después del filtrado ya no aparecen los picos indeseados. En el momento de correlar, las señales se elevan al cuadrado para hacer más sensible el resultado de la operación Si se eleva al cuadrado se obliga a que sean muy parecidos los dos vectores para que la correlación de próxima a 0.

Con el descriptor *AudioPower* se ha aplicado la misma operación que en el *AudioWaveform*, ya que dan una información que está muy relacionada.

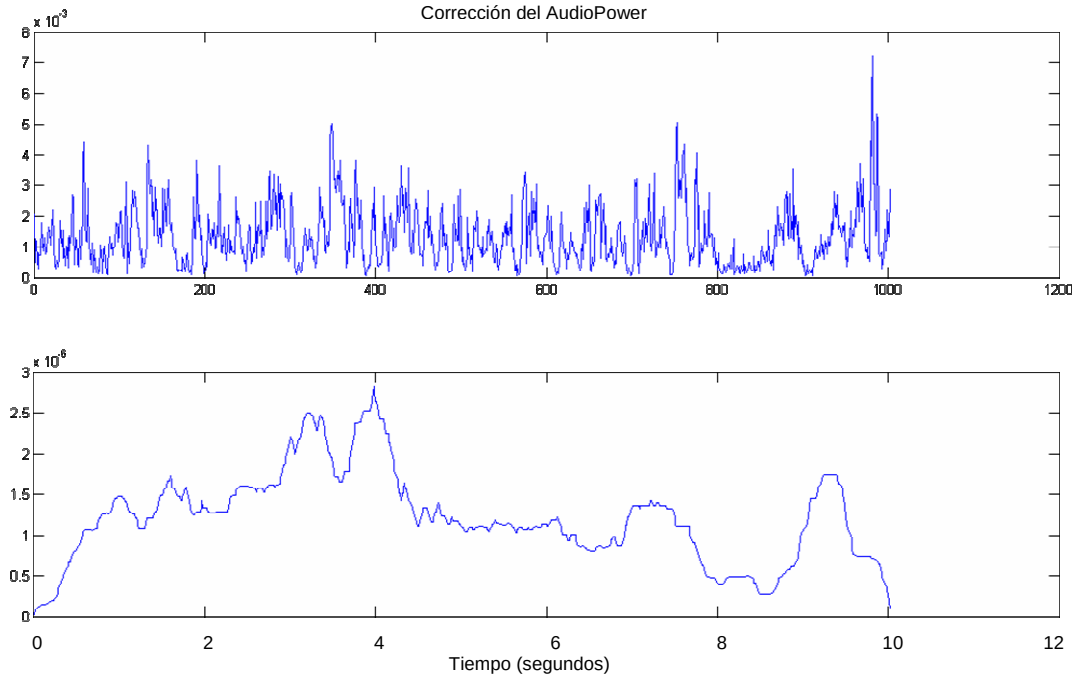


Fig 3.13 Corrección del descriptor AudioPower

Tabla 3.3 Correlaciones con los descriptores Awf y Ap modificados

Escenario	Awf	Ap	Ass
NotCoche	**52.9%**	**37.8%**	1.59%
TertCoche	**-21.4%**	**-30.4%**	-6.13%
McCoche	**32.1%**	**17.1%**	10.73%
MmCoche	**50.0%**	**35.1%**	3.19%
NotCasa1	**16.9%**	**-29.2%**	-1.85%
NotCasa2	**13.4%**	**-31.8%**	-0.24%
McCasa	**70.7%**	**81.7%**	12.02%
MmInternet	**79.8%**	**59.5%**	1.80%
MmAuriculares1	**59.0%**	**37.4%**	-1.72%
MmAuriculares2	**42.8%**	**32.4%**	-5.1%

Si se compara con la tabla original (**Tabla 3.1**) a simple vista se ha empeorado ya que los valores actuales son más altos. No obstante si no se aplica la corrección siempre salen valores muy bajos de correlación con cualquier muestra debido a que estos descriptores varían muy bruscamente su forma. Además los resultados obtenidos son más lógicos ya que siendo el mismo escenario, en volumen y dinámica es difícil encontrar diferencias que destaquen.

3.2.5 Resultados finales

Después de aplicar cada operación a cada descriptor y de añadir los MFCCs con su respectiva modificación, las nuevas correlaciones están definidas en la **Tabla 3.4**

Tabla 3.4 Correlaciones después de aplicar las correcciones

Escenario	Awf	Ap	Ase	MFCCs
NotCoche	52.9%	37.8%	4.71%	27.07%
TertCoche	-21.4%	-30.4%	4.96%	41.1%
McCoche	32.1%	17.1%	2.40%	49.4%
MmCoche	50.0%	35.1%	1.57%	15.4%
NotCasa1	16.9%	-29.2%	85.9%	40.9%
NotCasa2	13.4%	-31.8%	70.2%	41.5%
McCasa	70.7%	81.7%	46.8%	61.9%
MmInternet	79.8%	59.5%	46.6%	74.3%
MmAuriculares1	59.0%	37.4%	-0.408%	56.9%
MmAuriculares2	42.8%	32.4%	8.35%	60.4%

En los escenarios con muestras duplicadas ahora se hace la media aritmética de los dos (para Noticias en casa y música moderna con auriculares).

Tabla 3.5 Correlaciones finales

Escenario	Awf	Ap	Ase	MFCCs
NotCoche	52.9%	37.8%	4.71%	27.07%
TertCoche	21.4%	30.4%	4.96%	41.1%
McCoche	32.1%	17.1%	2.40%	49.4%
MmCoche	50.0%	35.1%	1.57%	15.4%
NotCasa	15.15%	30.5%	78.05%	41.2%
McCasa	70.7%	81.7%	46.8%	61.9%
MmInternet	79.8%	59.5%	46.6%	74.3%
MmAuriculares	50.9%	34.9%	4.175%	58.65%

Como se puede ver en la **Tabla 3.5** los descriptores *AudioWaveform* y *AudioPower* no dan unos valores muy bajos. Esto sucede porque las diferentes emisoras utilizan volúmenes parecidos, y en escenarios iguales las variaciones de volumen y de potencia de la señal son semejantes. Se ve que en la mayoría de escenarios el descriptor *AudioSpectrumEnvelope* se correla poco, que es lo que se busca en este apartado, sin embargo en el escenario Noticias/Casa ("NotCasa") da una alta correlación. Un poco menor es en Música Clásica/Casa y Música moderna/internet. Los MFCCs tienen un comportamiento peor que el *AudioSpectrumEnvelope*. En un primer momento se pensó que podría aportar información relevante, pero eso sólo ocurre en un escenario: Noticias/Casa.

3.3 Combinación de múltiples descriptores en la estimación de la calidad de la señal de audio. Asignación de pesos

Hasta ahora se han estudiado los diferentes descriptores por separado. El siguiente paso es darles un peso relativo a cada uno de los 4 descriptores seleccionados (ver **Tabla 3.5**) en la estimación final de la calidad de la señal de audio.

Cuando se tiene clara la plantilla a usar en cada escenario, se calcula la nota final mediante los descriptores citados anteriormente. Éstos descriptores, como ya se ha calculado en la **Tabla 3.5**, no afectan igual en cada escenario, así que de esa misma tabla sacamos la **Tabla 3.6** con los pesos de cada descriptor:

Tabla 3.6 Pesos de cada descriptor para la nota final de calidad

Escenario	Awf	Ap	Ase	MFCCs
NotCoche	6.42%	8.98%	72.06%	12.54%
TertCoche	9.64%	7.11%	77.75%	5.50%
McCoche	5.92%	11.11%	79.13%	3.84%
MmCoche	2.67%	3.80%	84.88%	8.65%
NotCasa	48.57%	24.13%	9.43%	17.86%
McCasa	22.13%	19.15%	33.44%	25.28%
MmInternet	19.50%	26.16%	33.40%	20.94%
MmAuriculares	6.44%	9.40%	78.56%	5.59%

Los pesos se han calculado de la siguiente manera:

Primero se obtienen los porcentajes relativos de cada descriptor (ver **Tabla 3.7**):

$$P_{AWF} = \frac{Awf}{Awf+Ap+Ase+MFCCs} ; P_{AP} = \frac{Ap}{Awf+Ap+Ase+MFCCs} \quad \{\ldots\} \qquad \textbf{(3.4)}$$

Por ejemplo para el escenario noticias coche:

$$P_{AWF} = \frac{0.629}{0.629+0.378+0.0471+0.2707} = 43.19\%; \quad \{\ldots\} \qquad \textbf{(3.5)}$$

Tabla 3.7 Porcentajes relativos de la **Tabla 3.5**.

Escenario	Awf	Ap	Ase	MFCCs
NotCoche	43.19%	30.86%	3.85 %	22.10%
TertCoche	21.86%	31.06%	5.08 %	42.00%
McCoche	31.78%	16.93%	2.38%	48.91%
MmCoche	48.99%	34.39%	1.54%	15.09%
NotCasa	9.19%	18.5%	47.33%	24.98%
McCasa	27.08%	31.29%	17.92%	23.71%
MmInternet	30.67%	22.87%	17.91%	28.55%
MmAuriculares	34.25%	23.48%	2.81%	39.46%

El objetivo es dar mayor peso al que menos correlado esté, es decir, el que menos porcentaje haya obtenido (se trata cada escenario individualmente). Para ello se calcula la inversa de estos porcentajes y se vuelven a calcular los porcentajes relativos, y así se obtiene la **Tabla 3.6**.

3.4 Factor de calidad

En el momento de extraer un factor de calidad, se usa la operación correlación explicada en el punto 3.2. Se correla el descriptor del fragmento a evaluar con el mismo descriptor del fragmento o fragmentos que hemos escogido como plantilla.

Como se ha explicado en el punto 3.1 puede darse el caso en que la plantilla sea una sola muestra, o sean varias muestras con unas ponderaciones.

Si se compara con una plantilla de sólo una muestra, el factor de calidad se calcula de la siguiente manera: se pondera la correlación de cada descriptor con el mismo descriptor de la plantilla. En **(3.6)** los descriptores de la plantilla se definen como $Nom\_Descriptor_{PL}$ y los pesos de cada descriptor se definen como $P_{Nom\_Descriptor}$. (Nom_Descriptor es el nombre del descriptor). La función $corr(x, y)$ es la correlación explicada en el punto 3.2.2.

$$QF_i = (P_{ASE}\ x\ corr(ASE, ASE_{PL}) + [\dots] + P_{MFCCs}\ x\ corr(MFCC, MFCC_{PL})) \quad \textbf{(3.6)}$$

Si la plantilla escogida se compone de más de una muestra, se calcula para cada muestra que compone dicha plantilla su factor de calidad de forma simple como en **(3.7)**, y luego se pondera con los pesos de cada muestra:

$$Q_T = P_1\ x\ QF_1 + P_2\ x\ QF_2 \quad \textbf{(3.7)}$$

CAPÍTULO 4. IMPLEMENTACIÓN

4.1 Representación de los datos

Se ha programado un script en MATLAB (ver **ANEXO C**) para poder extraer todos los datos de los archivos XML y poder representarlos.

En el **ANEXO D** se pueden ver los ejemplos de todos los descriptores extraídos con MATLAB.

4.2 Aplicaciones

Se han implementado dos tipos de aplicaciones bajo el entorno MATLAB. La primera sirve para representar tres descriptores diferentes sobre tres fragmentos de audio diferentes, simultáneamente. La segunda es una aplicación experta que evalúa la calidad subjetiva de un fragmento de audio.

4.2.1 Visor de descriptores

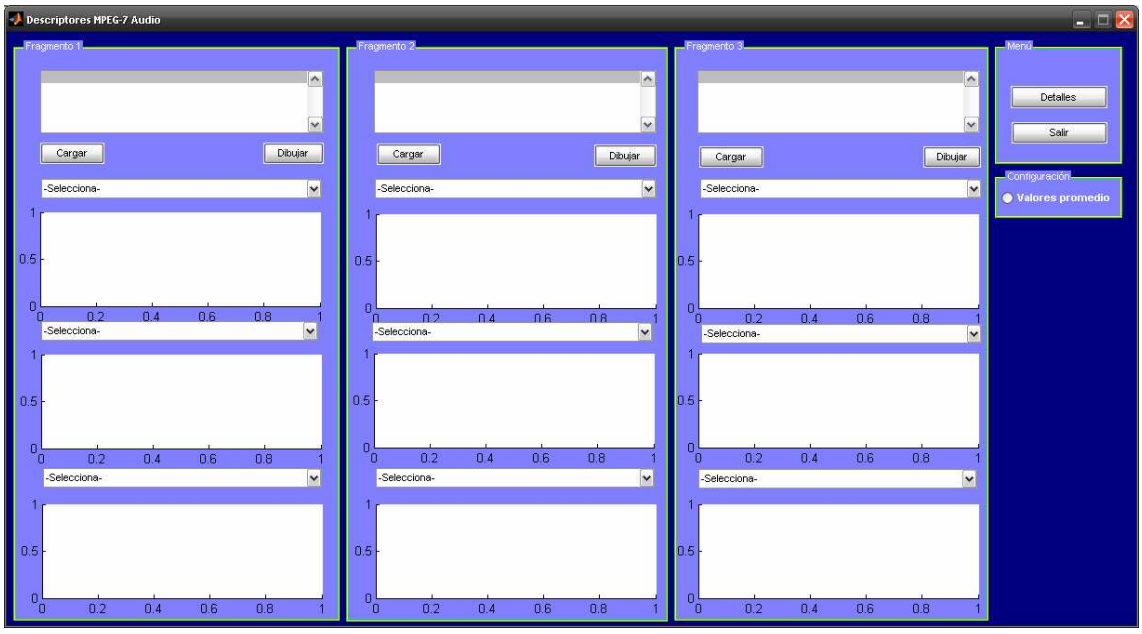


Fig 4.1 Diseño de la aplicación

Esta aplicación permite la representación de cualquiera de los descriptores explicados en el apartado 1.3.

Además de todos esos descriptores, permite mostrar los llamados *Mel-Frequency Cepstrum Coefficients* (MFCC) explicados en el punto 1.3.8.

La aplicación consta de tres grandes bloques. En cada bloque se puede cargar un fichero XML con los datos MPEG-7 de un fragmento de audio previamente procesado.

Por cada fichero cargado, se dispone de 3 ejes donde dibujar los descriptores. En cada eje se puede dibujar el descriptor que el usuario escoja.

En la **Fig 4.2**, se pueden ver representados tres descriptores: *AudioPower*, *AudioWaveform*, *AudioSpectrumEnvelope.*

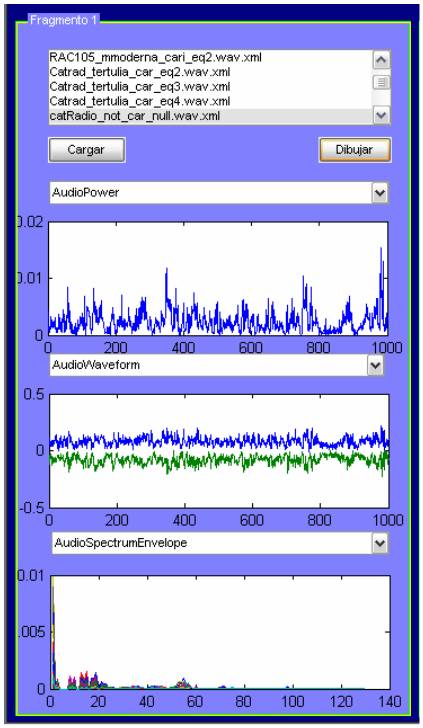


Fig 4.2 Ejemplo con un fragmento

AudioWaveform y *AudioPower* son descriptores con un resultado escalar, es decir, un único número (o en el caso del *AudioWaveform* una pareja de números) para cada paquete de señal. *AudioSpectrumEnvelope*, en cambio, es un descriptor del tipo vector, es decir, para cada paquete de la señal tenemos un vector (que en este caso representa el espectro).

Debido a esto, se representan simultáneamente tantos espectros como fragmentos temporales en los que la señal ha sido dividida.

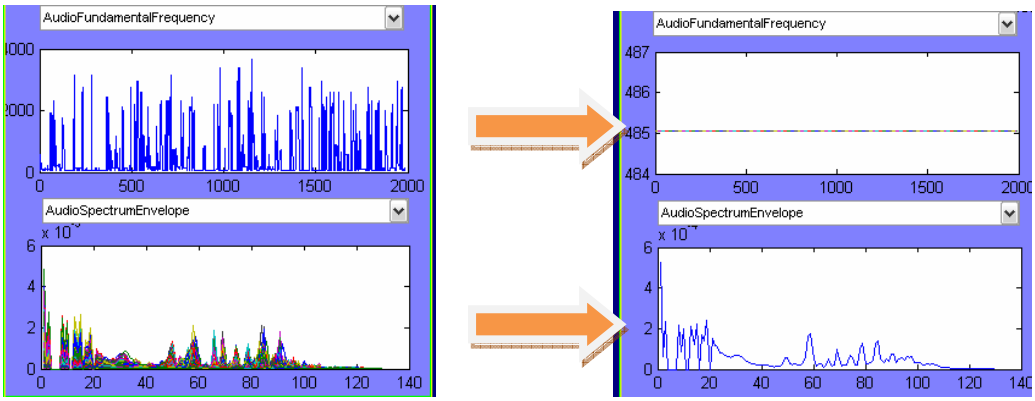


Fig 4.3 Gráfica de la media de los descriptores que lo permiten

Como muestra la **Fig 4.3**, en la aplicación se permite mostrar únicamente el promedio de todos estos vectores. En este caso queda un espectro promedio, que puede resultar más interesante en el momento de extraer conclusiones, o procesar los datos.

4.2.2 Programa experto.

El objetivo final del proyecto es tener una aplicación que pueda predecir si un fragmento de audio, respecto a la calidad del mismo, es bueno o malo para los oyentes.

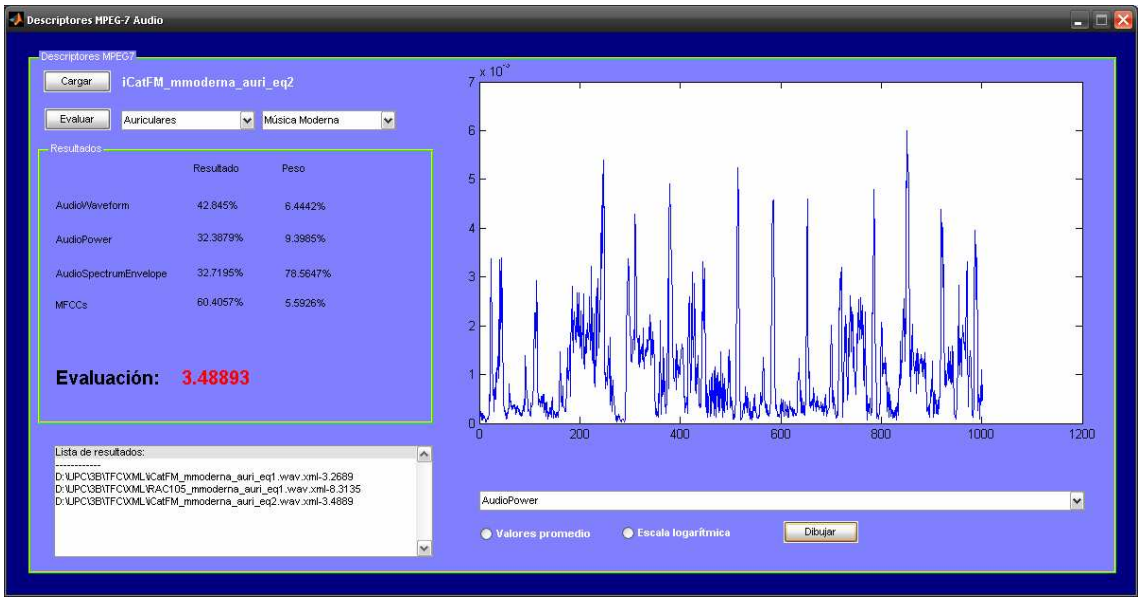


Fig 4.4 Ejemplo con una muestra del programa experto

La **Fig 4.4** muestra una captura de pantalla de la aplicación, que permite visualizar detalladamente los descriptores de un fragmento que el usuario haya cargado.

En la **Fig 4.5** se ha ampliado la sección donde se muestra la nota de calidad. Muestra una nota del 0 al 10 que representa la calidad de ese fragmento.

Fig 4.5 Resultados para el análisis de un fragmento

Los detalles de los cálculos que realiza la aplicación, y de su metodología están en el **Capítulo 3**.

CAPÍTULO 5. RESULTADOS OBTENIDOS

Los resultados obtenidos se muestran a continuación:

Se evaluará el escenario Coche, Tertulias:

	TERTULIAS	
	Muestra 1	Muestra 2
EQU1	5% peor 95% mejor	22% peor 78% mejor
EQU2	100% peor	93% peor 7% mejor
EQU3	6% peor 94% mejor	8% peor 92% mejor
EQU4	38% peor 62% mejor	24% peor 76% mejor
	muestra mejor :	**100% la muestra 2**

Fig 5.1 Resultados para escenario Coche/Tertulias

Comparamos con la aplicación las muestras originales:

Fig 5.2 Comparativa de las muestras originales

Con este procedimiento se analizan todas las muestras del escenario.

En la **Tabla 5.1**, se muestran los resultados obtenidos en las encuestas y los factores de calidad obtenidos con el programa. En rojo se han seleccionado las muestras en las que el factor de calidad y la opinión de los oyentes no coinciden.

De la misma forma se han analizado todos los escenarios:

Tabla 5.1 Comparativa escenario Coche/Tertulias

Muestras	Ecualizaciones	Resultado encuestas	Factor de calidad
Muestra 1	**Original**		1.85
	EQU1	95% mejor	4.46
(100% peor)	**EQU2**	100% peor	1.31
	EQU3	94% mejor	6.66
	EQU4	62% mejor	1.19
Muestra 2	**Original**		9.29
	EQU1	78% mejor	9.96
(100% mejor)	**EQU2**	93% peor	5.48
	EQU3	92% mejor	9.97
	EQU4	76% mejor	4.04

Tabla 5.2 Comparativa escenario Coche/Noticias

Muestras	Ecualizaciones	Resultado encuestas	Factor de calidad
Muestra 1	**Original**		2.61
	EQU1	57% mejor	2.48
(86% peor)	**EQU2**	95% peor	1.36
	EQU3	83% mejor	6.26
	EQU4	76% peor	0.99
Muestra 2	**Original**		8.54
	EQU1	70% peor	9.87
(86% mejor)	**EQU2**	82% peor	5.22
	EQU3	83% mejor	10
	EQU4	68% peor	4.68

Tabla 5.3 Comparativa escenario Coche, Música Clásica

Muestras	Ecualizaciones	Resultado encuestas	Factor de calidad
Muestra 1	**Original**		2.12
	EQU1	65% mejor	2.94
(66% peor)	**EQU2**	91% peor	0.76
	EQU3	56% peor	1.65
	EQU4	82% peor	1.10
Muestra 2	**Original**		6.77
	EQU1	77% mejor	6.93
(66% mejor)	**EQU2**	66% mejor	2.88
	EQU3	58% mejor	6.84
	EQU4	93% peor	2.71

Tabla 5.4 Comparativa escenario Coche, Música Moderna

Muestras	Ecualizaciones	Resultado encuestas	Factor de calidad
Muestra 1	**Original**		0.58
	EQU1	94% mejor	0.75
(82% peor)	**EQU2**	99% peor	0.53
	EQU3	68% peor	0.99
	EQU4	77% peor	0.42
Muestra 2	**Original**		9.18
	EQU1	84% mejor	9.89
(82% mejor)	**EQU2**	77% peor	4.00
	EQU3	74% mejor	9.88
	EQU4	61% peor	2.25

Tabla 5.5 Comparativa escenario Casa, Noticias

Muestras	Ecualizaciones	Resultado encuestas	Factor de calidad
Muestra 1	**Original**		3.10
	EQU1	78% mejor	2.60
(86% peor)	**EQU2**	85% peor	3.06
	EQU3	82% mejor	3.25
	EQU4	83% peor	2.82
Muestra 2	**Original**		7.92
	EQU1	55% peor	7.90
(86% mejor)	**EQU2**	76% peor	7.19
	EQU3	77% mejor	10
	EQU4	97% peor	7.16

Tabla 5.6 Comparativa escenario Casa, Música Clásica

Muestras	Ecualizaciones	Resultado encuestas	Factor de calidad
Muestra 1	**Original**		5.90
	EQU1	75% mejor	6.20
(57% mejor)	**EQU2**	95% peor	5.49
	EQU3	69% mejor	6.15
	EQU4	100% peor	5.48
Muestra 2	**Original**		6.23
	EQU1	64% mejor	6.12
(57% peor)	**EQU2**	84% peor	6.26
	EQU3	99% mejor	6.18
	EQU4	68% mejor	6.22

Tabla 5.7 Comparativa escenario Internet, Música Moderna

Muestras	Ecualizaciones	Resultado encuestas	Factor de calidad
Muestra 1	**Original**		7.07
	EQU1	62% peor	7.02
(54% mejor)	**EQU2**	100% peor	6.77
	EQU3	61% mejor	7.36
	EQU4	95% peor	6.42
Muestra 2	**Original**		7.03
	EQU1	67% mejor	7.17
(54% peor)	**EQU2**	86% peor	6.22
	EQU3	64% mejor	6.79
	EQU4	74% peor	6.44

Tabla 5.8 Comparativa escenario Auriculares, Música Moderna

Muestras	Ecualizaciones	Resultado encuestas	Factor de calidad
Muestra 1	**Original**		1.60
	EQU1	51% peor	1.08
(99% peor)	**EQU2**	67% peor	1.57
	EQU3	70% mejor	1.64
	EQU4	69% mejor	1.55
Muestra 2	**Original**		7.26
	EQU1	93% mejor	8.59
(99% mejor)	**EQU2**	64% mejor	2.87
	EQU3	85% mejor	8.45
	EQU4	56% mejor	4.36

En los casos más acentuados (donde la opinión de los oyentes ha coincido en un alto porcentaje) el programa responde correctamente a las expectativas. En cambio, en los casos donde los resultados no difieren mucho, y sobre todo los relacionados con la ecualización 4, el resultado que ofrece el programa no concuerda con el obtenido mediante las encuestas. Esto es normal ya que el programa se ha adecuado demasiado a lo que se ha denominado plantilla. Los pesos de los descriptores han sido calculados mediante la comparación de la mejor y la peor muestra, cuando se intenta averiguar la calidad de una muestra que no ha sido ni la mejor ni la peor, hay un cierto nivel de incertidumbre.

La ecualización 4 (ecualización en 'U') atenúa 4 dB a 1 KHz. Lo que produce un cambio importante en la envolvente del espectro. Así pues el descriptor *AudioSpectrumEnvelope* se ve modificado. Objetivamente esto es un cambio importante, no obstante subjetivamente (basándonos en las encuestas) no es así. Por eso da un resultado equivocado.

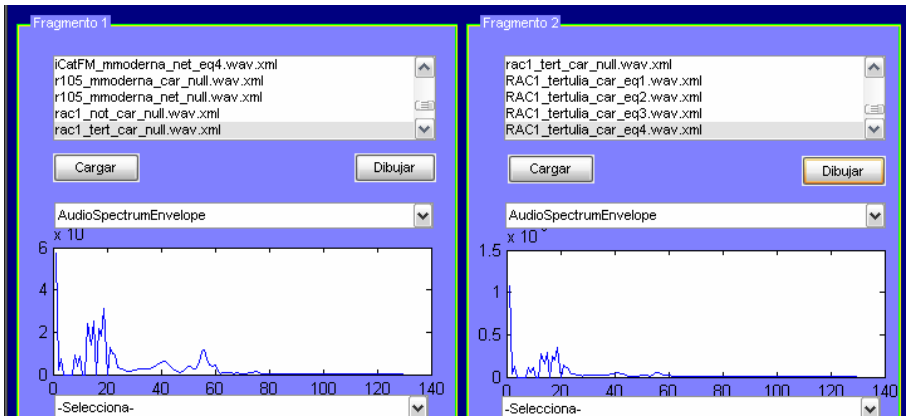


Fig 5.3 Efectos de la ecualización 4 (U)

Dependiendo del escenario, la ecualización 4 sí que reduce la calidad del audio según las encuestas. En estos casos el resultado objetivo concuerda con el subjetivo.

Con la ecualización 2 ocurre algo parecido. Esta ecualización es la decreciente y atenúa los agudos. Esta banda, como ya se ha explicado, se enfatiza mucho en el momento de comparar las muestras. Esto provoca que el descriptor se correle poco. Así independientemente de la opinión de los oyentes siempre se obtiene un factor de calidad menor que el original.

En cuanto a las ecualizaciones 1 y 3 por lo general se adaptan bien a la opinión de los oyentes.

Cabe destacar que el escenario Casa, Música Clásica (ver **Tabla 5.6**) es un escenario en donde las opiniones no difieren mucho, y es por esto que el programa no ha dado los factores de calidad adecuados.

CAPÍTULO 6. CONCLUSIONES Y LÍNEAS FUTURAS

En este proyecto siempre se ha trabajado teniendo presente que se trata de un prototipo. Aún así ha sido complejo debido a que no hay documentación sobre esta cuestión. Encontrar el método correcto de medir objetivamente una cuestión subjetiva como la aquí tratada es difícil. Requiere tiempo para tomar muestras, grabar todos los escenarios y programas y también realizar un número suficiente de encuestas. La contribución del proyecto en este ámbito es alta ya que es un primer paso en la evaluación de la calidad subjetiva del audio. Además, puede ser utilizado como referencia para seguir investigando.

Después de analizar los resultados obtenidos para todos los casos, se observa que el programa responde, en general, adecuadamente para las mismas muestras usadas en el cuestionario. Se puede decir que el principal objetivo del proyecto se ha cumplido.

Aún así hay casos que no concuerdan (aquellas muestras donde los propios oyentes diferían en sus opiniones). Después de este prototipo se ha observado que en el procesado de un fragmento de audio no se distingue entre información y calidad, es decir, se toma todo el espectro como variable para ser comparada, cuando realmente sólo interesa la parte de calidad. Si se consigue realizar esta distinción, los resultados podrían ser mucho más fiables.

A parte de esto, es necesario cambiar el método de comparación. El método actual implica descartar varios descriptores y así perder parámetros que podrían ser significativos en el cálculo del factor de calidad.

Es probable que con el método utilizado se haya adecuado demasiado a las muestras de las encuestas. Se deben aplicar métodos estadísticos avanzados para poder evitar este efecto. Un ejemplo sería la lógica difusa, que permitiría cuantificar cualquier muestra dentro de un rango comprendido entre la máxima calidad y la mínima.

Otro factor a tener en cuenta es la adquisición de la muestra. Lo ideal sería grabar el mismo contenido en emisoras diferentes para que las únicas diferencias que pueda haber sean de la calidad.

Los resultados del test actual en cuanto a los parámetros dinámica e inteligibilidad han sido confusos y en algunos casos ilógicos. Es necesario rehacer esa parte.

En cuanto impacto medioambiental no se ha encontrado ningún elemento que pueda alterar el propio impacto que ya tienen de por si las emisoras de radio. Si en un futuro es implementado en un sistema empotrado afectaría el diseño de placas y elementos relacionados.

BIBLIOGRAFÍA

[1] http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm (14 de julio de 2009)

[2] Martínez Borrell,J. “Medidas de la calidad del audio en radiodifusión, Estudio de la calidad subjetiva”. EPSC, (2009).

[3] H.-G. Kim, N. Moreau, T.Sikora,”MPEG7 Audio and Beyond: Audio Content indexing and Retrieval”,John Willey & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England (2005)

[4] Dalibor Mitrovic, Matthias Zeppelzauer, Horst Eidenberger. “Analysis of the Data Quality of Audio Descriptions of Environmental Sounds”, Vienna University of Technology.

[5] http://sourceforge.net/projects/mpeg7audioenc/ (25 de Abril de 2009)

ANEXO A

1.COTXE

· Fragment de “Noticies 1”

(Ara escoltarà un fragment de noticies com si estigués escoltant-ho en un cotxe)

A quin volum escoltaries aquest fragment?
Com escoltes millor aquest fragment? (pre-programats)
Creus que aquest fragment te molta variació de volums?
Posi una nota del 0-10 sobre la claritat i la Intel·ligebilitat, en resum, s’entén bé?
Podria identificar quina emissora es?

Volum	
Equalització	1: 4: 2: 3:
Dinàmica	
Intel·ligibilitat	

· Fragment de “Noticies 2”

(Ara escoltarà un fragment de noticies com si estigués escoltant-ho en un cotxe)

A quin volum escoltaries aquest fragment?
Com escoltes millor aquest fragment? (pre-programats)
Creus que aquest fragment te molta variacio de volums?
Posi una nota del 0-10 sobre la claritat i la Intel·ligebilitat, en resum, s’entén bé?
Podria identificar quina emissora es?

Volum	
Equalització	1: 4: 2: 3:
Dinàmica	
Intel·ligibilitat	

QUIN DELS DOS FRAGMENTS TROBES QUE ES MILLOR EN QUAN A QUALITAT D’ÀUDIO?

[...]

ANEXO B

Los descriptores y sus parámetros se especifican en un XML en forma de plantilla. Un ejemplo es el siguiente:

```
- <Module mp7ae:enable="true" xsi:type="AudioFundamentalFrequency">
    <lolimit>50.0</lolimit>
    <hilimit>16000.0</hilimit>
  </Module>
  <Module mp7ae:enable="true" xsi:type="HarmonicSpectralCentroid" />
- <Module mp7ae:enable="true" xsi:type="LogAttackTime">
    <threshold>0.02</threshold>
  </Module>
- <Module mp7ae:enable="true" xsi:type="AudioSpectrumBasisProjection">
    <frames>0</frames>
    <numic>8</numic>
  </Module>
```

Fig B.1 Plantilla XML

Todos los parámetros de entrada, tanto su nombre como su tipo, están especificados en un XMLSchema de la siguiente manera:

```
<complexType name="ModuleType" abstract="true" />

<complexType name="OutputModuleType" abstract="true">
  <complexContent>
    <extension base="mp7ae:ModuleType">
      <attribute name="enable" type="boolean" use="required"/>
  </extension>
  </complexContent>
</complexType>

<complexType name="Resizer">
  <complexContent>
    <extension base="mp7ae:ModuleType">
      <sequence minOccurs="0">
        <element name="HopSize" type="decimal" default="10"/>
      </sequence>
    </extension>
  </complexContent>
</complexType>
```

Fig B.2 XMLSchema con la descripción de los parámetros de entrada

Una vez la plantilla está preparada se pude proceder a los cálculos. Para completar esta parte se ha programado una pequeña aplicación en C# que utiliza la librería *mpeg7AudioEnc* (ver [5]) y calcula los descriptores para cada archivo de audio dentro de un directorio. Esta aplicación también se encarga de

guardar los datos resultantes en un XML. El código de esta aplicación está en el **ANEXO E.**

El resultado está representado de la siguiente forma:

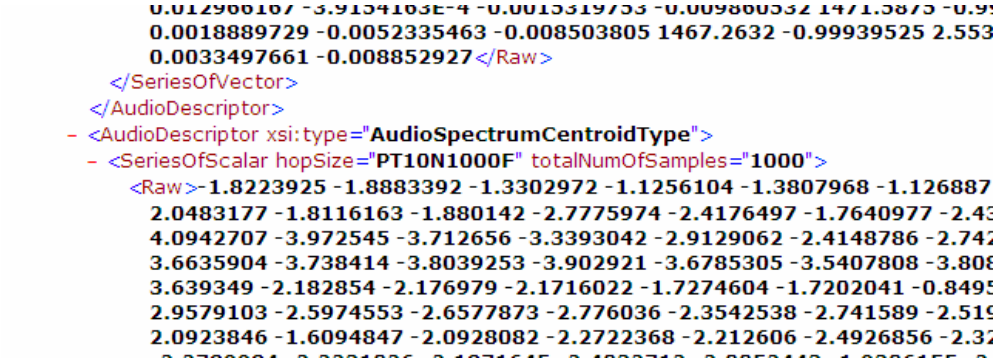


Fig B.3 Fichero XML resultante

ANEXO C

C.1 Descriptores tipo escalar

```
function Values = ReadAudioPowerDescriptor(path)

[file,error] = fopen(path,'r');

if (file > -1) % It's OK

    File = fscanf(file,'%c');
    %findstr or strfind
    DeclarationPosition = findstr(File,'AudioPowerType');

    InitialPositions = findstr(File(DeclarationPosition:end),'<Raw>')
+ 5;
    InitialPosition = InitialPositions(1) + DeclarationPosition - 1;

    FinalPositions = findstr(File(InitialPosition:end),'</Raw>') - 1;
    FinalPosition = FinalPositions(1) + InitialPosition - 1;

    Values = strread(File(InitialPosition:FinalPosition), '%f',
'delimiter', sprintf(' '));

else
    disp('File corrupted' + error);
end
```

C.2 Descriptores tipo vector

```
function Values = ReadAudioSpectrumEnvelopeDescriptor(path)

[file,error] = fopen(path,'r');

if (file > -1) % It's OK

    File = fscanf(file,'%c');
    %findstr or strfind
    DeclarationPosition = findstr(File,'AudioSpectrumEnvelopeType');

    InitialPositions =
findstr(File(DeclarationPosition:end),'mpeg7:dim="') + 11;
    InitialPosition = InitialPositions(1) + DeclarationPosition - 1;

    temp = findstr(File(InitialPosition:end),'">') - 1;

    FinalPositions = findstr(File((temp + 2 +
InitialPosition):end),'</Raw>') - 1;
    FinalPosition = FinalPositions(1) + InitialPosition - 1 + temp +
2;
```

```
    values = File((temp + 2 + InitialPosition):FinalPosition);
    %disp(values);
    InitialPosition = InitialPosition + temp + 2;

    TValues = strread(File(InitialPosition:FinalPosition),
'%s','delimiter',sprintf('\n'));
    for i=1:length(TValues)
        pepe = TValues{i,:};
        values = strread(pepe(1:end-6),'%f','delimiter',sprintf(' '));
        Values(i,:) = values;
    end

    Values = transpose(Values);
    %[pedro] = strread(TValues);
    %Values = textscan(TValues, '%f', 'delimiter', sprintf('&#13;'));
else
    disp('File corrupted' + error);
end
```

ANEXO D

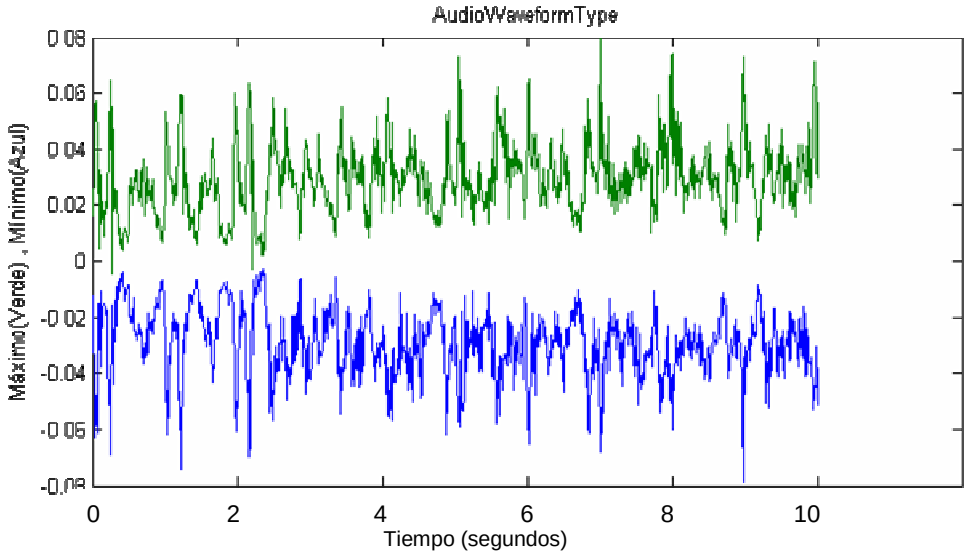


Fig D.1 Ejemplo AudioWaveform

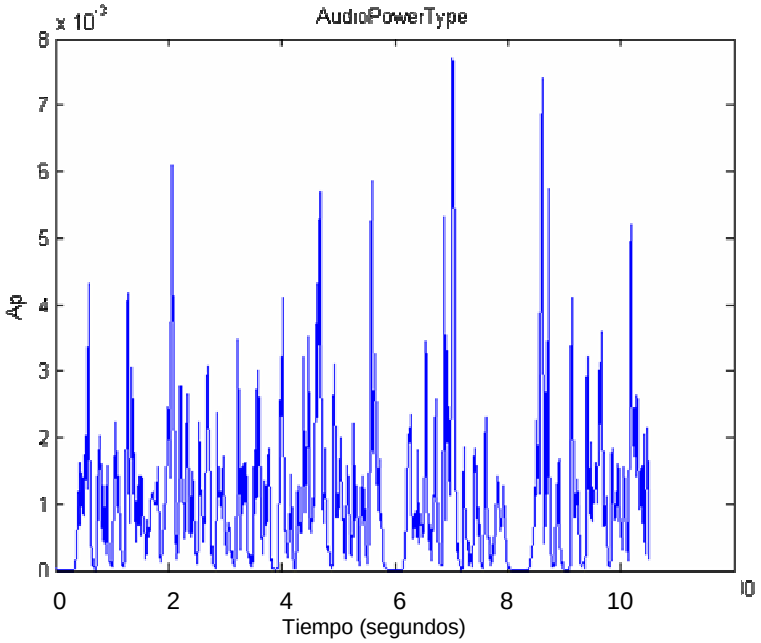


Fig D.2 Ejemplo AudioPower

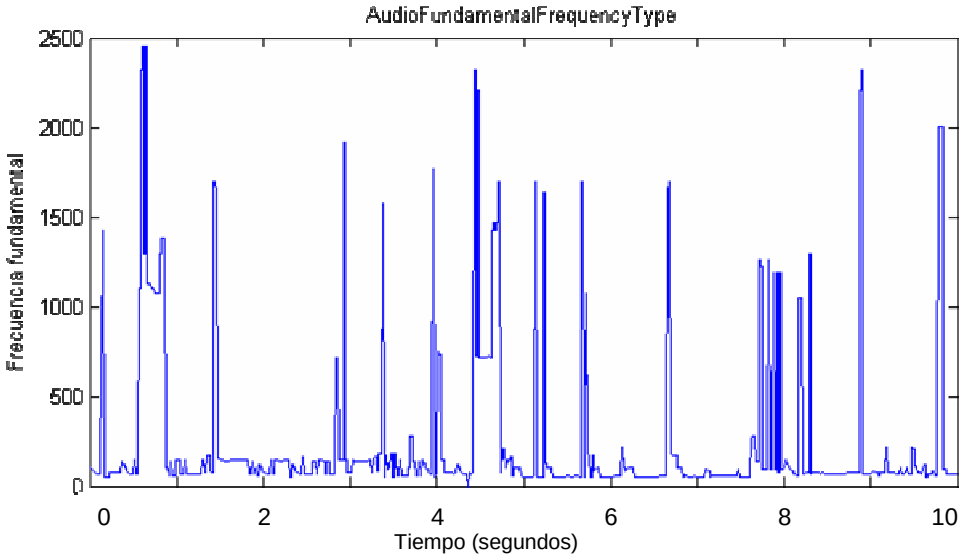


Fig D.3 Ejemplo AudioFundamentalFrequency

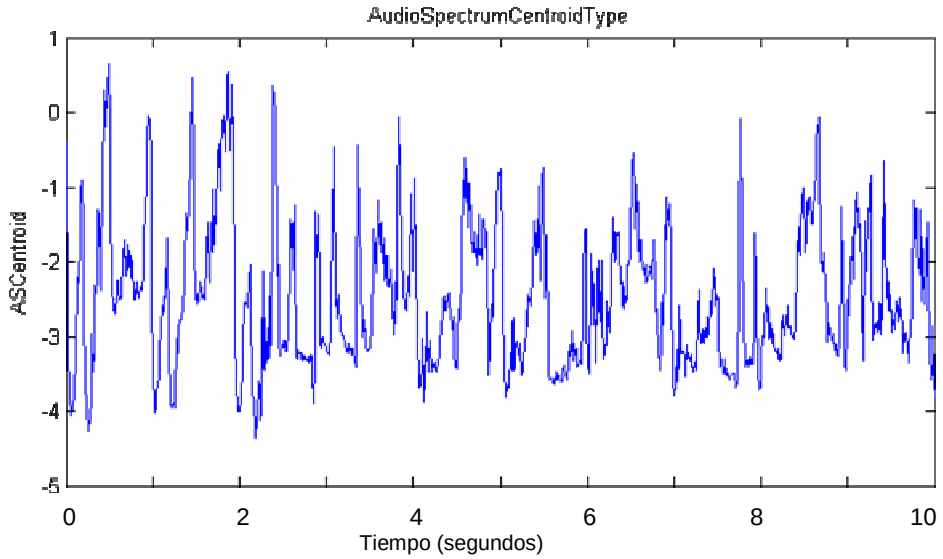


Fig D.4 Ejemplo AudioSpectrumCentroid

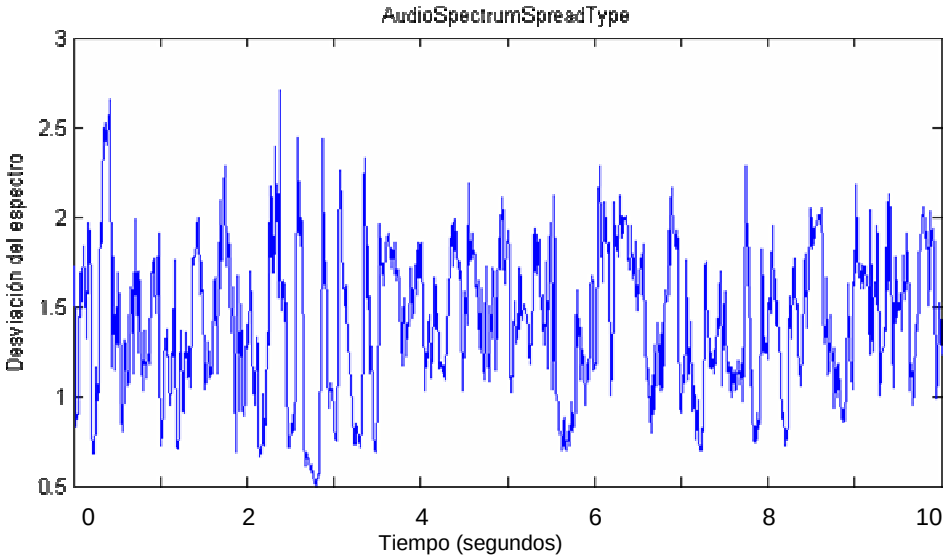


Fig D.5 Ejemplo AudioSpectrumSpread

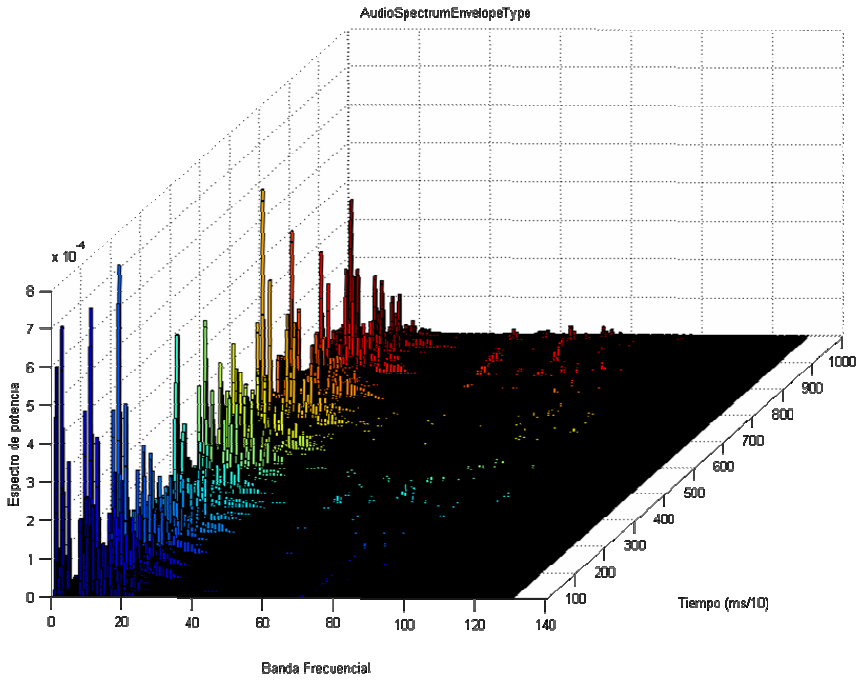


Fig D.6 Ejemplo AudioSpectrumEnvelope a lo largo del tiempo

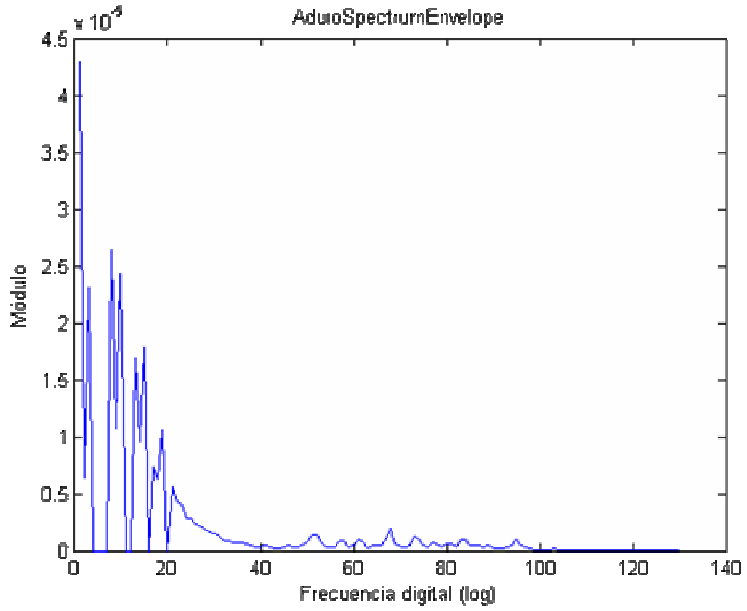


Fig D.7 Ejemplo de la media del descriptor AudioSpectrumEnvelope

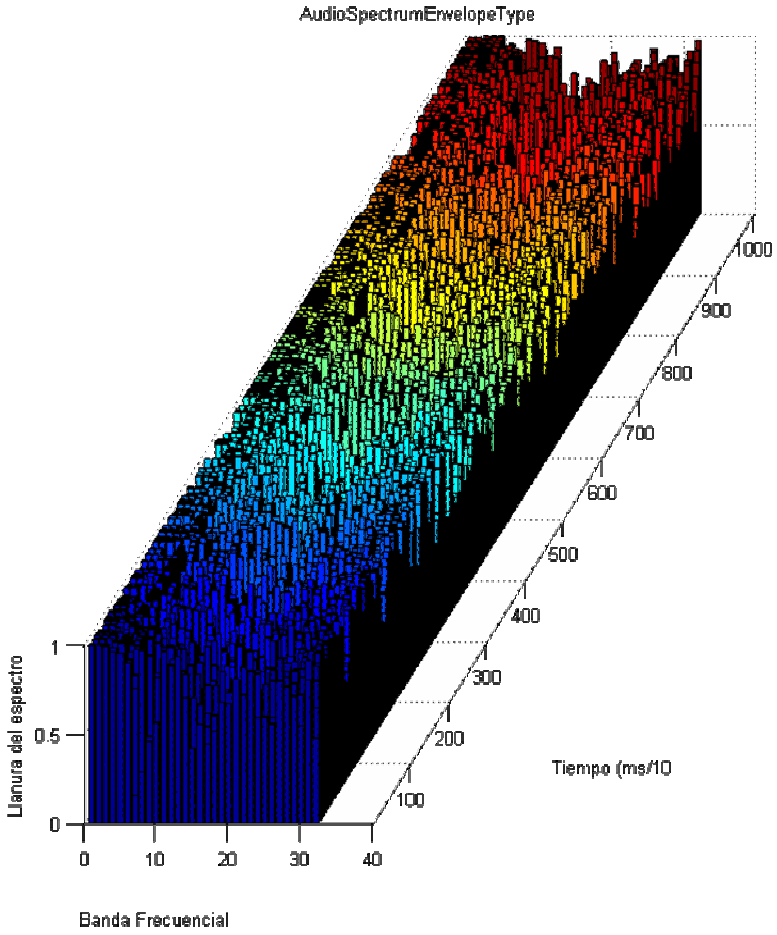


Fig D.8 Ejemplo AudioSpectrumFlatness a lo largo del tiempo

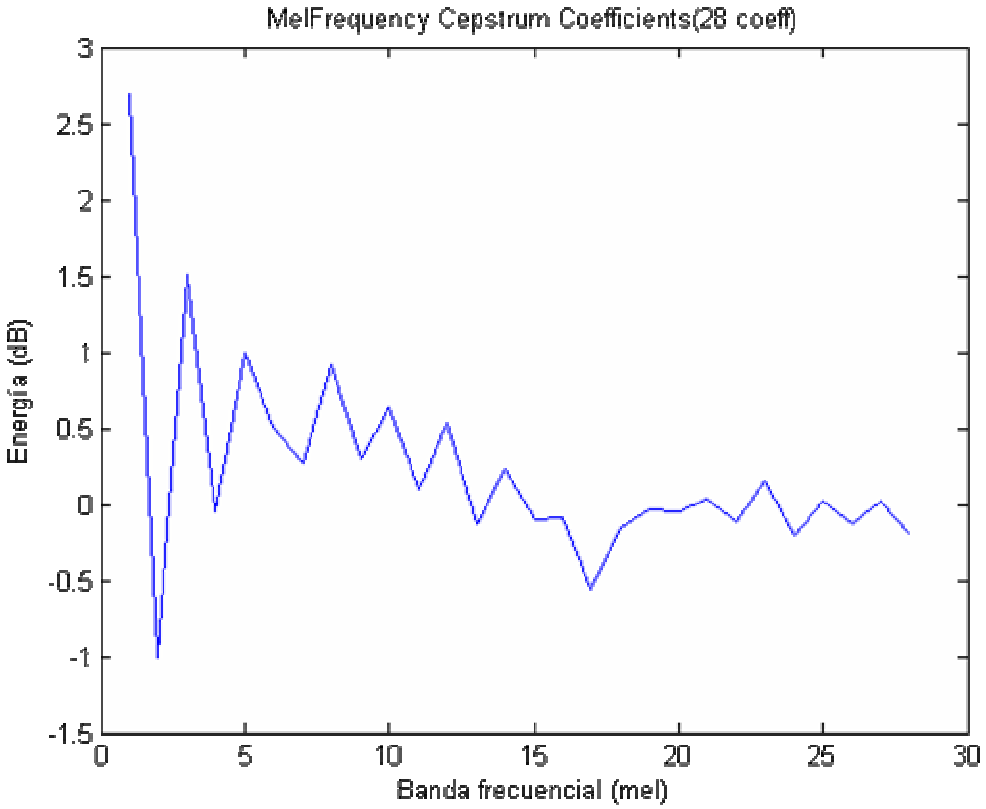


Fig D.9 Eejemplo MelFrequency Cepstrum Coefficients

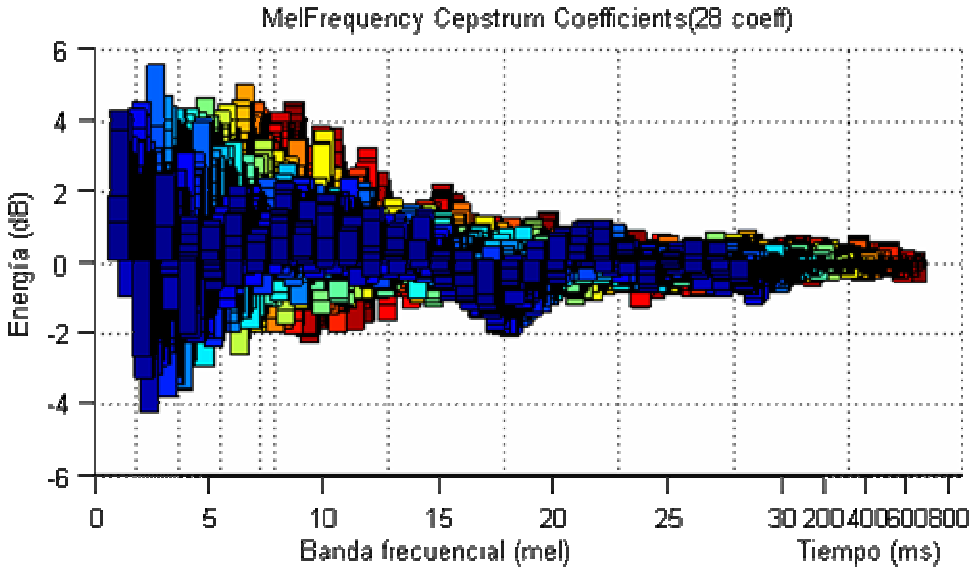


Fig D.10 Eejemplo MelFrequency Cepstrum Coefficients a lo largo del tiempo

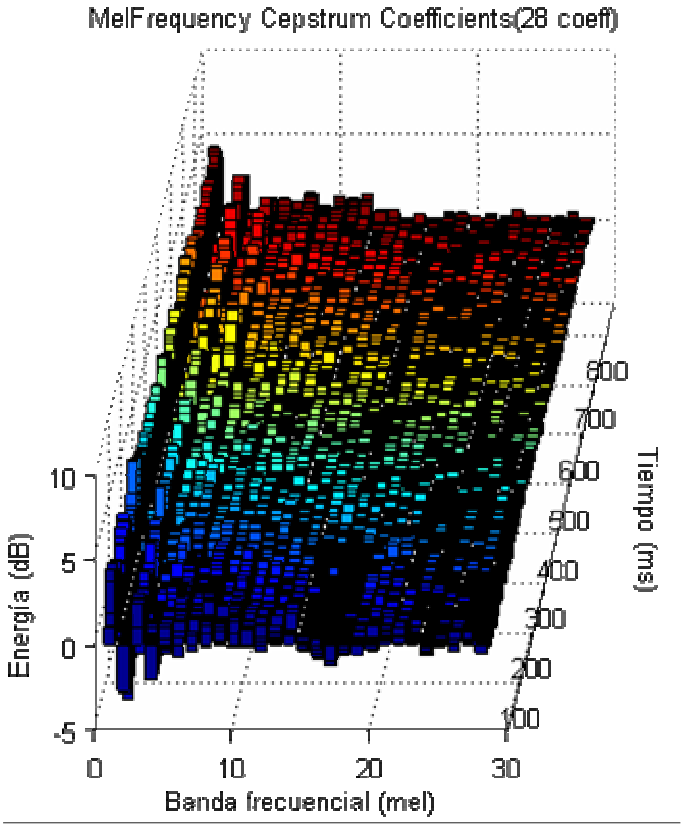


Fig D.11 Eejemplo MelFrequency Cepstrum Coefficients a lo largo del tiempo

ANEXO E

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.IO;
using System.Media;
using System.Net;

namespace DescriptorReader
{
    class Program
    {
        static string WorkingPath =
@"D:\UPC\3B\TFC\XML\MUESTRAS_10s_CON_EQUALIZACIONES";
        static void Main(string[] args)
        {
            ReadFolder(WorkingPath);
            Console.ForegroundColor = ConsoleColor.Green;
            Console.WriteLine("Completed!");
            Console.ReadLine();
        }

        static void ReadFolder(string path)
        {
            int cont = 0;
            int number = 0;
            DirectoryInfo di = new DirectoryInfo(path);

            foreach (DirectoryInfo _di in di.GetDirectories())
            {
                int temp = 0;
                foreach (FileInfo fi in _di.GetFiles())
                {
                    if (fi.Extension == ".wav")
                    {
                        temp++;
                    }
                }
                number += temp * (_di.GetDirectories().Length + 1);
            }


            foreach (DirectoryInfo _di in di.GetDirectories())
            {
                foreach (FileInfo fi in _di.GetFiles())
                {
                    if (fi.Extension == ".wav")
                    {
                        cont++;
                        ReadFile(fi.FullName, fi.Name);
                        Console.Title = (cont * 100 /
number).ToString() + "%";
                    }
                }

                foreach (DirectoryInfo __di in _di.GetDirectories())
                {
```

```
                    foreach (FileInfo fi in __di.GetFiles())
                    {
                        if (fi.Extension == ".wav")
                        {
                            cont++;
                            ReadFile(fi.FullName, fi.Name);
                            Console.Title = (cont * 100 /
number).ToString() + "%";
                        }
                    }
                }
            }
        }

        static void ReadFile(string FullPath, string FileName)
        {
            System.Diagnostics.Process p = new
System.Diagnostics.Process();
            p.StartInfo.FileName = @"C:\Archivos de
programa\Java\jre6\bin\Java.exe";
            p.StartInfo.WorkingDirectory = @"C:\";
            p.StartInfo.Arguments = "-jar " + WorkingPath +
"\\MPEG7AudioEnc-0.4-rc2.jar " + FullPath + " " + WorkingPath +
"\\mpeg7audioenc.xml";
            p.StartInfo.UseShellExecute = false;
            p.StartInfo.RedirectStandardOutput = true;
            p.StartInfo.RedirectStandardError = true;
            p.Start();
            string sr = p.StandardOutput.ReadToEnd();
            string err = p.StandardError.ReadToEnd();

            //Console.WriteLine(sr);
            Console.ForegroundColor = ConsoleColor.Yellow;
            Console.WriteLine(err);
            Console.ForegroundColor = ConsoleColor.Gray;

            p.WaitForExit();
            string SavePath = @"D:\UPC\3B\TFC\XML\"+
FullPath.Substring(FullPath.LastIndexOf('\\')+1) + ".xml";
            StreamWriter sw = new StreamWriter(SavePath);

            Console.WriteLine(FullPath + " audio descriptors saved on
" + SavePath);


            sw.Write(sr);
            sw.Close();

            //Console.ReadLine();
        }
    }
}
```