UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA

"IMPLEMENTACIÓN DE DESCRIPTORES DE MÚSICA PARA SU INDEXACIÓN BAJO LA NORMA MPEG-7"

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN INGENIERÍA

INGENIERÍA ELÉCTRICA– TELECOMUNCIACIONES
P R E S E N T A :

ALEJANDRO MARTÍNEZ CORTÉS

TUTOR:
DR. VÍCTOR GARCÍA GARDUÑO

2011

JURADO ASIGNADO:

Presidente: DR. BORIS ESCALANTE RAMÍREZ

Secretario: DR. JOSÉ MARÍA MATÍAS MARURI

Vocal: DR. VÍCTOR GARCÍA GARDUÑO

1$^{er.}$ Suplente: DR. PSENICKA BOHUMIL

2$^{do.}$ Suplente: DR. MIGUEL MOCTEZUMA FLORES

Lugar donde se realizó la tesis:

FACULTAD DE INGENIERÍA, UNAM.

TUTOR DE TESIS:

_____

DR. VÍCTOR GARCÍA GARDUÑO

DEDICATORIA

Mi más sincero agradecimiento
al Dr. Víctor García Garduño,
por su orientación, dedicación
y apoyo en la realización de este trabajo.

Agradezco a Dios,
por darme la fortaleza interna
para la culminación de este trabajo.

Agradezco a mi familia,
por su confianza y apoyo
que siempre me han brindado.

Alejandro Martínez Cortés

Índice

Antecedentes

La estandarización MPEG ha identificado una necesidad por crear un grupo de herramientas audiovisuales, cuyo propósito es permitir a los usuarios realizar búsquedas, identificar, clasificar, filtrar y buscar contenido multimedia como imágenes, gráficos, modelos tridimensionales, audio, voz así como de archivos multimedia compuestos, etc. a través de su contenido a través de una descripción estructural de su contenido (formas, colores, texturas, movimientos, sonidos), estas herramientas deben ser capaces de soportar una rápida búsqueda y robusta detección de contenido debido al incremento exponencial de los formatos multimedia digitales.

Las aplicaciones de las técnicas de análisis y descripción de contenido serán importantes en sectores de entretenimiento y serán amplias en los sectores de medicina, educación, industria, seguridad o arquitectura, por ejemplo: una biblioteca digital sería un cliente potencial ya que recibe entradas multimedia y esta misma necesita clasificarlas, archivarlas y ponerlas fácilmente a disposición de los usuarios en formatos digitales.

En el caso del audio este estándar permite reconocer si un archivo de audio contiene voz, reconocer la persona o el artista que está hablando, o cuantas voces de personas están en la grabación. También permitirán realizar una clasificación de la música por el tipo o género de esta misma (jazz, clásica, etc), identificar patrones de voz que permitan ejecutar comandos, convertir la voz en texto, o identificar sonidos en particular. En este sector algunos Institutos como empresas han empezado a realizar avances en este tema, tales como:

El Instituto Fraunhofer, creador del MP3, desarrollo el AudioID, un sistema para la identificación automática de una canción en 5 segundos entre más de 30000 archivos; el Insituto de telemática e IBM prueban en Holanda servicios personalizados de video sobre IP (educación y negocios); LG electronics, crea un buscador de información audiovisual para proveedores de contenidos; Ricoh experimenta un software para la recuperación y distribución de video audioclips a través de un interfaz Web, NEC ha desarrollado un equipo de identificación de vídeos en tiempo real.

En el campo de procesamiento de imágenes se han desarrollado herramientas y aplicaciones que nos permiten la recuperación, indexación y clasificación y edición de imágenes; hoy en día existen algunos sistemas como PicSOM, este aplicación nos ayuda a representar las principales características de las imágenes, la cual incluye una base de datos tanto de vistas panorámicas como pinturas y es basado en una estructura o mapeo llamado Self-Organizing Map (SOM).

Existe otro sistema llamado System of Automatic Processing and Indexing of Reports (SAPIR) el cual realiza una búsqueda punto a punto en contenidos audiovisuales y mide la similaridad entre imágenes utilizando descriptores MPEG-7 (scalable color, color, layout, color structure, edge histogram, homogeneous strucutre).

Eptascape (http://www.eptascape.com/) es una empresa la cual ha desarrollado productos para el almacenamiento y procesamiento de datos multimedia, por ejemplo, cuenta con un software llamado EptaAnalyitics y EptaVision, los cuales permiten realizar el monitoreo y detección de eventos desde cámaras analógicas, digitales e IP, cuentan con un equipo EptaScape el cual realiza el análisis de vídeo en tiempo real para la extracción de las características MPEG-7.

UniSay (http://www.unisay.com/) es una empresa que se dedica a dar soluciones para tareas de grabación y edición en los medios audiovisuales.

En el área biomédica se encuentran en estudio dos descriptores MPEG-7: Edge Histogram Descriptor (EHD) y Homogeneous Texture Descriptor (HTD), los cuales han beneficiado en el mejoramiento de la detección de masas en mamografías.

En arquitectura nos permitirá indexar grandes bases de material (imágenes fijas, gráficos, modelos tridimensionales).

Otro campo de futuras aplicaciones de este estándar cubre el ramo de la Inteligencia Artificial, en donde se hace el uso de las anotaciones semánticas, en donde por medio de fonemas se pretende recuperar la información, lo que permite que la "anotación semántica" haga explícito el significado de un audiovisual para un equipo de cómputo, por lo que este tema se ha convertido en un punto clave y de interés.

Entre las herramientas de anotación y recuperación de imágenes se encuentran Caliph & Emir, esta herramienta permite la extracción automática de características de bajo nivel y soporta la anotación semántica. Otras herramientas desarrolladas en este campo son M-Ontomat-Annotizer como proyecto de AceMedia, IBM MPEG-7 Annotation Tool, VIZARD, MARVel y POLYSEMA, entre otras.

Con estas últimas aplicaciones se pretende conseguir que las máquinas entiendan el significado, la semántica, de los textos escritos, audios videos y de las propias páginas Web, ya que actualmente con el crecimiento de la información contenida en Internet resulta imposible que un único usuario realice las tareas de búsquedas en un tiempo aceptable.

Introducción

Este trabajo de investigación esta divido en 5 capítulos de acuerdo a los objetivos que se pretenden, donde:

Capítulo 1. MPEG. En este capítulo se hace mención de los objetivos y características del estándar MPEG y cuáles son sus avances en el área de procesamiento de audio, imagen y video, así como los objetivos de cada estándar que conforman esta familia, con este capítulo se pretende dar un marco teórico con un enfoque introductorio al estándar MPEG-7.

Capítulo 2. MPEG-7. En este capítulo se hace mención propiamente del estándar que es de interés en este trabajo de investigación, se habla de cuáles son sus objetivos, en qué consiste, cual es su arquitectura, normas, elementos y cuáles son sus herramientas de descripción que hacen posible el análisis tanto en tiempo como en frecuencia de un archivo multimedia que nos permitirán hacer un clasificado e indexado adecuado de audio; al término de este capítulo se tendrá un panorama general de cómo está constituido dicho estándar para que pueda ser comprensible y utilizar sus herramientas adecuadamente.

Capítulo 3. Esquemas de Descripción y descriptores de audio. Para este capítulo, ya que se tiene las nociones necesarias de en qué consiste el estándar MPEG-7, cómo se encuentra constituido y cuáles son sus objetivos y posibles aplicaciones de dicho estándar, se hará mención de las herramientas principales de descripción que son de interés para este trabajo de investigación, enfocándonos principalmente a un grupo en especial que consta de un análisis espectral, y que cuyos descriptores nos dan ciertos datos que nos permiten saber por ejemplo: la cantidad de energía contenida y distribuida a lo largo del espectro de frecuencias (ASE), en que banda de frecuencia se encuentra la mayor parte de la energía o centroide (ASC), cuánta energía se distribuye a lo largo del centroide (ASS) y si esta energía corresponde a un tono o a ruido (ASF).

Capitulo 4. Métodos de clasificación. Una vez que se ha realizado el análisis de los descriptores, se hace mención de los métodos de clasificación existentes y más en particular se habla del método SVM que para la bibliografía consultada, se encontró ser el más adecuado y eficaz.

Capítulo 5. Estrategias de reconocimiento. Una vez que se tienen los conocimientos necesarios, se procede en este capítulo a realizar una serie de pruebas, construyendo diferentes vectores de entrenamiento para el SVM con los valores promedio de los descriptores ASC y ASS, también se hace mención de los coeficientes MFCC y se utilizan estos mismos para la clasificación del audio. Finalmente se mencionan los resultados obtenidos, las conclusiones y las áreas donde se puede continuar el estudio de este trabajo de investigación.

Justificación

En los últimos años, se ha dado un gran avance en el desarrollo del software y hardware de las Tecnologías de la Información y las Telecomunicaciones, ya que se han encontrado métodos y equipos que mejoran la eficiencia y el procesamiento de la información, y se han desarrollado mejores sistemas de telecomunicaciones que nos permiten tener una mayor eficiencia y velocidad de transmisión de la información, pero así como han avanzado dichos sistemas de telecomunicaciones, también se ha incrementado la demanda de usuarios que utilizan constantemente diversas aplicaciones y que por consiguiente, los sistemas de comunicaciones tienen que soportar en sus diversos medios ya sean inalámbricos o alámbricos, por lo que una aplicación debe considerar un buen sistema de procesamiento que permita lograr sus objetivos utilizando la menor cantidad de información posible en la etapa de procesamiento y transmisión, considerando varios factores técnicos.

Con el auge de las diversas aplicaciones que hoy en día existen y con el aumento de la cantidad de información multimedia que se encuentra disponible en Internet, el hombre ha visto la necesidad de desarrollar nuevos aplicativos y buscar métodos de búsqueda que nos permitan encontrar la información deseada de una manera rápida y eficiente, estos métodos de búsqueda se basan en el análisis del contenido multimedia, ya que con este tipo de análisis, cada archivo multimedia cuenta con cierta información tanto en el dominio del tiempo y espectral o que cumplen ciertos patrones, como en el caso de una imagen, que puede verse como si fuese una huella digital, por ello este tipo de análisis basado en su contenido es de gran interés de estudio, ya que últimamente se han desarrollado aplicaciones que nos hacen posible realizar clasificaciones basadas en su contenido, pero que todavía tiene un gran campo de estudio debido a que se cuenta con una gran cantidad de información que muchas veces puede ser clasificada erróneamente; un ejemplo de estas aplicaciones son: suponiendo que un usuario está buscando las últimas noticias de deportes en los canales de su televisor, una aplicación sería la búsqueda de ciertos patrones en las imágenes de cada canal, como balones de futbol, de basquetbol, de tenis, etc, y/o que sea posible diferenciar entre las playeras de los jugadores y la playera de una persona de una película, otra aplicación sería la clasificación de los archivos de audio en géneros o de acuerdo a los instrumentos utilizados en cada uno de ellos.

Por lo que en este trabajo de investigación nace el estudio de una aplicación que es bastante reciente e innovadora y que tiene por objetivo el indexado y recuperación de música de acuerdo al análisis de su contenido, y que en un futuro no muy lejano será de gran explotación por los usuarios, y de no contar con un buen sistema de procesamiento el cual se limite a transmitir sólo la información necesaria para lograr su objetivo, ocasionará un incremento de tráfico en las redes ocasionando un problema mayor como es la saturación de las redes mismas.
También otro punto de vista que se debe de considerar, es ver hacia donde estará enfocada dicha aplicación y cuáles serán sus limitaciones, por ejemplo, si este aplicativo será utilizado en un

dispositivo móvil, una tableta, una computadora o cualquier dispositivo móvil, por lo que también este aplicativo tendrá que considerar cuales serían sus limitaciones para cumplir sus objetivos, como son los formatos multimedia comúnmente soportados en los dispositivos móviles y el nivel de ruido que estará presente durante la grabación. Estas dos últimas variables son de gran importancia ya que, cada uno de los formatos multimedia tienen diferentes tasas de muestreo, y a la hora de hacer la conversión de un formato (.mp3, .aac, etc) al formato estándar *.wav que es necesario en la etapa de procesamiento utilizando el programa Matlab; ya que durante dicha conversión existen diferentes pérdidas de información y/o adición de ruido, que nos afectarán de manera directa en el procesamiento del audio y por consiguiente en los resultados que deseamos obtener.

Objetivos

En este trabajo de investigación se abarcan dos puntos u objetivos principales, el primero de ellos es investigar las propiedades, parámetros o valores que nos son aportados por los diferentes descriptores del estándar MPEG-7 e identificar que descriptores son necesarios y/o indispensables para poder llevar a cabo una clasificación y recuperación adecuada de música, y como segundo objetivo es, una vez que se han identificado qué descriptores son necesarios para la recuperación adecuada de música, hacer un estudio de cuanta información será requerida en la transmisión a través de un sistema de comunicaciones para hacer uso de esta aplicación, por lo que se hará un análisis de que es más factible, si enviar la información de los descriptores necesarios o enviar un extracto de canción de "n" segundos con un formato determinado desde un dispositivo final, considerando la mejor eficiencia posible, utilizando la menor cantidad de información y con una dependencia del tipo de servicio.

Considerando que la información será enviada a través de un dispositivo móvil, existirán dos variables que nos afectarán en la recuperación de música, la primera de ellas estará dada por las limitaciones de un sistema móvil (teléfono celular) principalmente por los formatos aceptados en dichos dispositivos, debido principalmente a la tasa de muestreo, ya que los archivos multimedia tendrán que ser posteriormente convertidos a archivos con extensión *.wav, y la segunda variable estará condicionada debido al nivel de ruido existente durante la grabación del audio.

Referencias

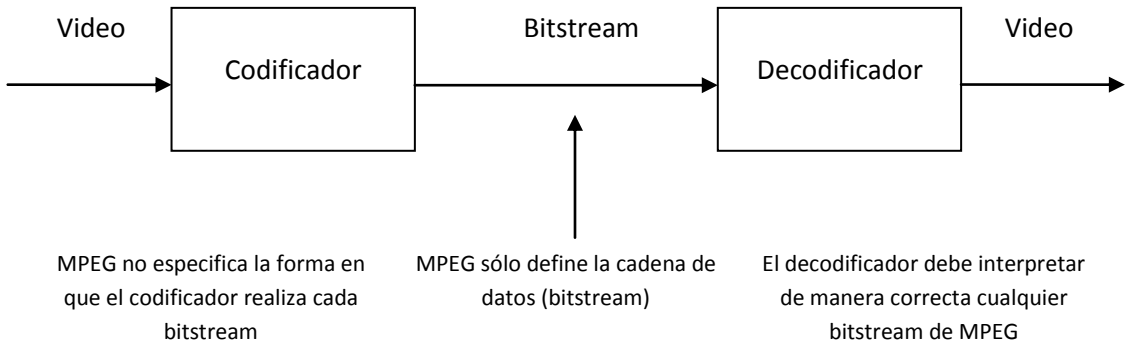
1. **ANGELIDES, Marios C, AGIUS Harry***, The Handbook of MPEG Applications,* Gran Bretaña, Wiley, 2011, p 531.
2. **ADIEGO, Joaquin,** *Anotaciones Semánticas,* edición en Internet, sección Investigación, http://www.infor.uva.es/~sblanco/Tesis/Anotaciones%20Sem%C3%A1nticas.pdf, consultada septiembre de 2011.
3. **PicSOM,** *Content-based self-organizing information retrieval,* http://www.cis.hut.fi/picsom/, consultada septiembre de 2011.
4. **Eptascape**, Compañía dedicada al desarrollo de software y hardware, 2005, http://www.eptascape.com/, consultada noviembre de 2011.
5. **Unisay**, Compañía dedicada al desarrollo de aplicaciones bajo el estándar MPEG-7, febrero de 2003, http://www.unisay.com/, consultada noviembre de 2011.
6. **LUX, Matias¸** Desarrollo de software, http://www.semanticmetadata.net/features/, consultada noviembre de 2011.
7. **ACEMEDIA,** Compañía dedicada al desarrollo de software, http://www.acemedia.org/aceMedia/results/software/m-ontomat-annotizer.html, consultada noviembre de 2011.
8. **IBM**, Compañía dedicada al desarrollo de software, http://www.ibm.com/developerworks/forums/forum.jspa?forumID=631, consultada noviembre de 2011.
9. **POLYSEMA,** Compañía dedicada al desarrollo de software, Junio de 2006, http://polysema.di.uoa.gr/en/objectives.html, consultada noviembre de 2011.

Capítulo1.MPEG

1.1. Introducción

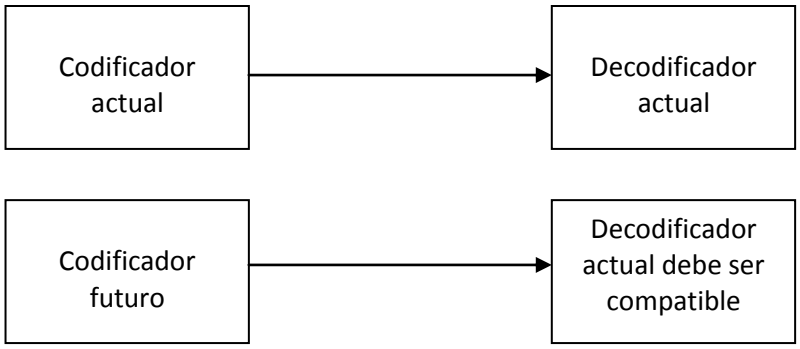
MPEG (Moving Pictures Experts Group) fue creado por la ISO (International Standards Organization) en el cual se definen los siguientes objetivos:

- Estándares de compresión y transmisión de audio y vídeo.
- Brinda información de la estructura y operación del codificador.
- Estandariza la forma y la sintaxis del protocolo que permite multiplexar datos de audio junto con los de video, para producir un audio-visual o programas digitales de TV.
- Define la manera en que operarán los multiplexadores y la forma en que se transportará la información audio-visual (metadatos).
- Define la forma en que los decodificadores procesarán los metadatos para ser demultiplexados correctamente.



Algunas de las características de MPEG son:

- No estandariza el codificador sino la manera en que el decodificador interpretará la cadena de bits (*bitstream*).
- El decodificador debe interpretar cualquier cadena de bits (bitstream) definidas por el estándar MPEG.
- El codificador producirá una estructura de código de acuerdo al estándar conocido por el mismo.
- En la figura se puede observar que el decodificador debe ser capaz de interpretar algoritmos de codificación que han sido mejorados y garantizar que seguirán funcionando con ellos.



La compresión de video en MPEG es usada en muchas aplicaciones, por ejemplo en la conversión de imágenes de baja resolución a alta definición, televisión digital, servicios digitales, decodificadores de alta definición, reproductores de video digital, Internet, ya que existen diversas razones de utilizar la compresión de video, algunas de estas son:

- Hay un ahorro de espacio para el almacenamiento, ya que se reduce la cantidad de información para ser almacenada.
- En los sistemas de transmisión, se logra una reducción en el ancho de banda, reduciendo costos.
- Se puede aumentar la calidad de la señal transmitida al reducir la tasa de bits propagada en el mismo ancho de banda.

La idea básica de la compresión de video es eliminar la redundancia espacial dentro de una trama de video y la redundancia temporal entre las tramas de video.

En la compresión de audio, lo que se realiza es una reducción del rango dinámico del sonido, MPEG reduce la tasa de bits o bit rate, donde existirá un pequeño grado de pérdida en la calidad del sonido, la cual a veces no es perceptible por el oído. Por lo que, existen diversos codificadores que al realizar la compresión de audio o video, presentaran en un mayor o menor grado pérdidas de información.

1.2. MPEG-1, 2, 4 y H.264

MPEG-1 fue básicamente diseñado (1993) para permitir que las imágenes en movimiento y el sonido fueran codificados en una tasa de bits de mediana calidad, ya que se disminuye de manera considerable la resolución de las imágenes y utiliza frecuencias de 25-30 Hz o cuadros por segundo con 352 x 288 pixeles y 352 x 240 pixeles respectivamente, no soporta el entrelazado por lo que la calidad de la imagen es moderada. El estándar MPEG-1 realiza una compresión de audio y video a una tasa de 1.5 Mbps.

La codificación del audio en MPEG-1, ésta basada en esquemas de codificación, llamados Capa-1, 2 y 3, donde el codificador aumenta su complejidad y rendimiento de la Capa-1 a la Capa-3, en la Capa-1 se usan 384 kbps, en Capa-2 192 kbps y en Capa-3 32 kbps, aunque se puede codificar uno o dos canales a velocidades de hasta 448 kbps.

Con MPEG-2 es posible codificar y combinar señales de audio y video para producir programas de televisión y multiplexar estos mismos, MPEG-2 fue diseñado para la emisión de televisión y aplicaciones que utilizan imágenes en movimiento (1994), por lo que se realizaron mejoras en el video, una de ellas fue soportar entrelazado en cada cuadro y con esto se logra una resolución de Alta Definición (HD), aunque utiliza una mayor tasa de bits. Algunos ejemplos de aplicación de este estándar se encuentran en las tecnologías utilizadas en los sistemas DVB, ATSC, VCD y DVD.

Al tener varias aplicaciones, MPEG-2 es subdividido en Perfiles y Niveles, lo que permite el uso de diferentes formatos o tamaños de imágenes, a la vez que utiliza diferentes tasas de bits en la codificación, un perfil describe el grado de complejidad de la codificación, mientras que un nivel describe el tamaño de la imagen o la resolución de la misma que utiliza cada perfil. Esta subdivisión permite una inter-operabilidad y una flexibilidad entre los dispositivos y receptores que utilizan monitores de diferentes formatos.

La codificación de audio en MPEG-2, se incrementó de 2 a 5 canales con efecto envolvente, teniendo compatibilidad con la codificación de audio en MPEG-1 (2 canales). También en MPEG-2 se introdujo un esquema de codificación de audio más eficiente conocido como MPEG-2 AAC (Advanced Audio Coding), el cual no es compatible con los esquemas anteriores de codificación de MPEG. MPEG-2 también utiliza los esquemas de codificación de MPEG-1, en el caso de la Capa-3, la tasa de bits llega a los 8 kbps.

El Instituto Fraunhofer desarrolló una extensión del estándar MPEG-2 llamado MPEG-2.5, el cual mejora el rendimiento de MPEG-2 Capa-3 a bajas tasas de bits, donde se muestrea a frecuencias de 8, 11.025 y 24 kHz. Un ejemplo de aplicación es en los archivos de audio con formato MP3, que es la combinación de MPEG-1/2 Capa-3 y MPEG-2.5, donde para bajas tasas (menores a 24 kbps) utiliza MPEG-2.5 y para tasas superiores a los 24 kbps utiliza la codificación de MPEG-2 Capa-3.

MPEG-4 (1999) está enfocado hacia aplicaciones gráficas, por lo que es muy útil e importante en aplicaciones como Internet, redes, sistemas de telecomunicaciones e inalámbricos, contiene herramientas de codificación y compresión de mayor complejidad, con lo que se consiguen factores de compresión más altos (baja tasa de bits) que en MPEG-2.

Por ejemplo, una de las herramientas de video de MPEG-4, se basa en la mejora de la compensación de movimiento. En MPEG-1 y 2, la compensación de movimiento, se basa en el procesamiento de áreas conocidas como *macroblocks*, originando una mayor tasa de bits, mientras que en MPEG-4 como está orientada hacia objetos, los objetos en movimiento pueden ser codificados en formas arbitrarias, lo que permite disminuir dicha tasa de bits.

Algunas aplicaciones de MPEG-4,utilizan la técnica mesh, la cual permite determinar si un objeto ha cambiado su posición y cuanto se ha desplazado, también MPEG-4 permite la animación de rostros, cuerpos, escenas, sistemas interactivos, simuladores y video juegos.

Su gran importancia ha permitido la creación del estándar H.264 que es una extensión de este estándar, el cual define aspectos de codificación de audio de MPEG-4 y el cual cuenta con las mismas funciones de MPEG-2.

MPEG-4 extiende las capacidades de codificación de audio (MPEG-2 ACC), el cual introduce un procesamiento de audio estructurado que es basado en objetos, donde el audio se ve como un objeto que puede ser codificado de manera independiente, lo que permite que este audio sea decodificado y mezclado con otros, permitiendo tener sistemas interactivos en "tiempo real",

Otras herramientas agregadas a MPEG-4 son: Perceptual Noise Substitution (PNS) y vector de cuantización. Estas dos herramientas aprovechan la ventaja que el oído presenta, al no poder distinguir entre el ruido que tiene la señal de audio y el ruido generado en el decodificador. Por lo que si en un cierto rango de frecuencias se detecta que no existe un tono dominante y/o la forma de onda permanece constante, la codificación o los coeficientes que representan a las frecuencias de la forma de onda serán reemplazadas por una bandera PNS. En el decodificador estos coeficientes serán obtenidos por un vector que recreará un ruido aleatorio.

Entre las herramientas principales de codificación de voz son: HVXC (Harmonic Vector eXcitation Coding) o CELP (CodeExcited Linear Predicition). En este estándar se define el formato de audio MP4.

1.3. Capas de Audio MPEG

Entre los formatos de compresión de audio que dieron origen al procesamiento de Audio en MPEG y de los cuales se toman algunos atributos son:

- ASPEC (Adaptative Spectral Perceptual Entropy Coding)
- MUSICAM (Masking pattern adapted Universal Sub-band Integrated Coding And Multiplexing)

El primero de ellos fue designado para sistemas con un alto nivel de compresión que permitieran la transmisión en ISDN el cual fue desarrollado por AT&T Bell Labs, el segundo fue designado para el uso en los sistemas DAB, el cual fue desarrollado en conjunto por CCETT en Francia, IRT en Alemania y Philips en Irlanda.

El estándar de Audio MPEG cuenta con tres capas de complejidad y rendimiento, las cuales definen el tipo de codificación empleado para diferentes calidades y tasas de compresión:

- MPEG Capa 1 es una versión simplificada de MUSICAM el cual es apropiado para aplicaciones de mediano factor de compresión y que sean de bajo costo.

- MPEG Capa 2 es idéntico a MUSICAM y es usado en DAB y para contenido de audio de transmisoras de televisión digital DVB.
- MPEG Capa 3 contiene las mejores características tanto de ASPEC y MUSICAM y se aplica principalmente en las telecomunicaciones donde se requieren altos factores de compresión.

En general, la codificación de Audio en MPEG cuenta con las siguientes características:

- Permite velocidades de muestreo en la entrada de 32, 44.1 y 48 kHz.
- Soporta velocidades de salida de 32, 48, 56, 64, 96, 112, 128, 192, 256 y 384 kbps.
- La transmisión puede ser mono, dual o estéreo.
- Una aplicación es en el uso de los canales estéreo, donde a una cierta frecuencia estos canales empiezan a funcionar como un canal monoaural, lo que permite disminuir la tasa de bits.

Referencias

1. **WATKINSON, John***, The MPEG Handbook,* Focal Press. 2da edición, Gran Bretaña, 2004, p 435.

2. **KOSCH, Harald,** *Distributed Multimedia Database Technologies, supported MPEG-7 and by MPEG-21,* CRC Press. Estados Unidos, 2004, p 260.
3. **MOREAU, Nicolas, SIKORA, Thomas,** *MPEG-7 Audio and beyond,* Wiley, Alemania, 2005, p 281.

Capítulo 2. MPEG-7

2.1. Introducción

Las aplicaciones y servicios audiovisuales son posibles debido a que existen técnicas que se basan en el análisis y descripción de contenido de audio, de forma similar como los motores de búsqueda o filtros que permiten buscar información sobre algún tema determinado en una base de datos que contiene texto.

Las aplicaciones de las técnicas de análisis y descripción de contenido de audio son muy extensas ya que permiten y pueden conocer si un archivo de audio contiene voz, reconocer la persona o el artista que está hablando, o cuantas voces de personas están en la grabación. También permiten realizar una clasificación de la música por el tipo o género de esta misma (jazz, clásica, etc), identificar patrones de voz que permitan ejecutar comandos, convertir la voz en texto, o identificar sonidos en particular.

Entre los archivos que son soportados por el estándar MPEG-7 (aparte del audio y voz), de los cuales se pueden realizar descripciones son: imágenes, gráficos, modelos tridimensionales, así como de archivos multimedia compuestos.

2.2. Objetivos

Los propósitos del estándar MPEG-7 son buscar información descriptiva y particular a través de descriptores de audio o herramientas con diferentes niveles de discriminación, que permitan comparar y clasificar un archivo de manera eficiente, almacenando su descripción de contenido (metadatos) en una base de datos, pudiéndose apoyar de las herramientas que son establecidas en el estándar MPEG-4 [2, "*Distributed Multimedia Database Technologies, supported MPEG-7 and by MPEG-21",* pág. 26].

Este estándar es de gran utilidad en:

- Almacenamiento de audio y video (base de datos),
- Mejora la capacidad y el uso de contenidos audiovisuales
- Aplicaciones de recuperación basada en contenido (movimientos, color, cuerpos, caras, etc).
- Clasificación, reconocimiento, búsqueda y/o filtrado de datos.
- Producción y estandarización en el intercambio de contenido audiovisual entre las herramientas avanzadas de análisis de contenido audiovisual y las herramientas de búsqueda o semánticas.

2.3. Descripción

MPEG-7 define descriptores estandarizados, cuyo propósito es permitir a los usuarios, realizar una similitud, identificar, clasificar, filtrar y buscar contenido audiovisual. Por ejemplo, basándose en la figura 2.1, un usuario graba una fracción de una canción, quién quiere conocer el título y el cantante, el usuario ingresa al portal o aplicativo según sea el tipo de dispositivo móvil y sube la fracción de canción a la base de datos multimedia, ahí se extrae la información requerida mediante la herramienta Audio Signature DS, cuya información obtenida es comparada con las demás descripciones que se encuentran en el servidor de almacenamiento, descartando todas aquellas que no contengan características similares. Finalmente se entrega la información solicitada al usuario final.

Un descriptor de datos también es llamado vector característico o “huella digital” y el proceso de extracción de información se llama: extracción característica de audio o “audio finger printing”.

La eficiencia de una “huella digital” en particular que es utilizada para la comparación y clasificación, depende en gran medida de la aplicación, el proceso de extracción y la cantidad de información que esta contenga.

Los sistemas de recuperación basados en contenido, permiten que en el análisis del audio, sea posible describir a través de su distribución de energía espectral sus características particulares, por ejemplo: el radio armónico o su frecuencia fundamental con lo que se logra realizar una comparación entre diferentes archivos audiovisuales y así poder hacer una clasificación del sonido en categorías o clases.

Existen dos maneras de asociar los datos de las descripciones MPEG-7 con el material audiovisual, la primera de ellas es: asociar la información con el material audiovisual dentro de la misma base de datos, la segunda es utilizar un mecanismo el cual nos permita enlazar la información de las descripciones contenida en una base de datos con el material audiovisual contenida en otra base de datos, como se observa en la siguiente figura:

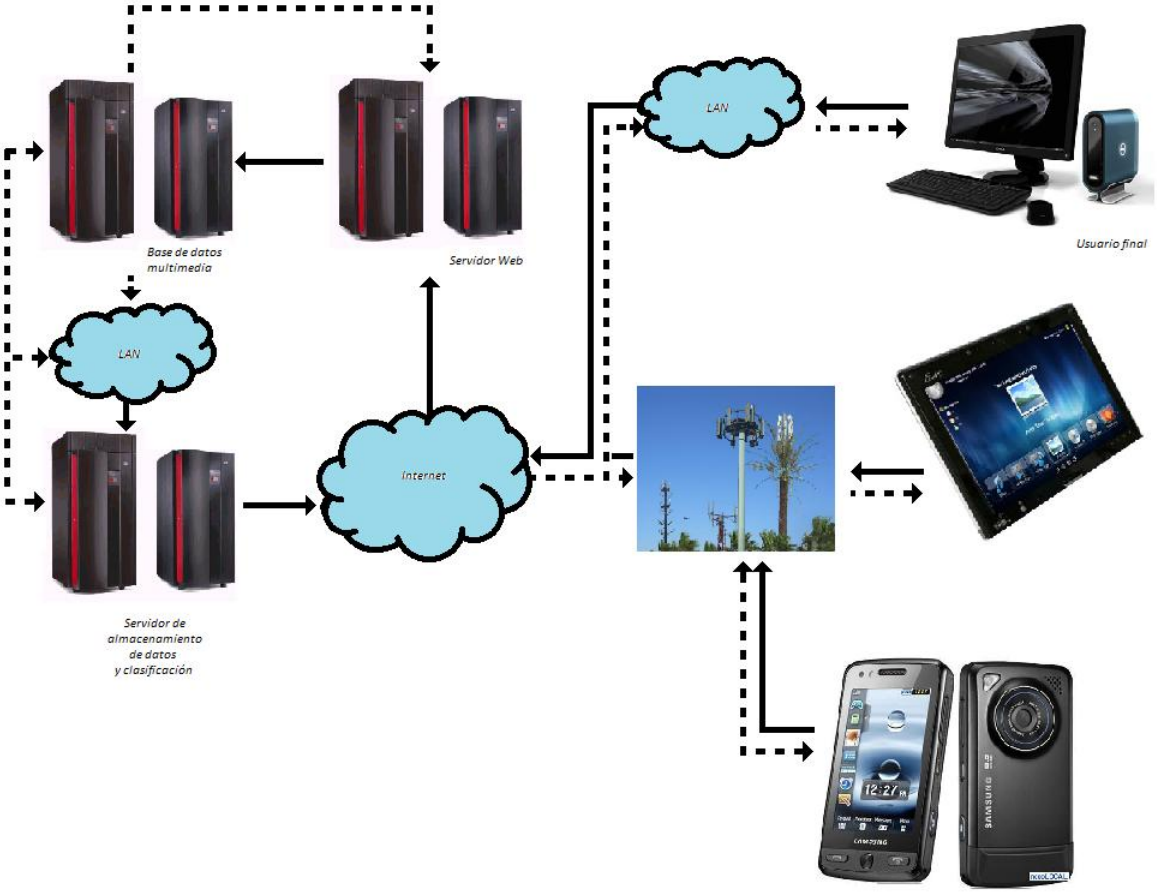


Figura 2.1. Arquitectura, componentes típicos y flujo de información de un sistema multimedia.

Como se observa en la figura anterior existe una intercomunicación bidireccional entre el usuario final (quien solicita la información) y el sistema multimedia, donde en el Servidor Web se realiza una autenticación del usuario a través de un portal, el cual permite la interacción entre el usuario, las herramientas y la base de datos. Al haber una solicitud de búsqueda por parte del usuario, esta se registra, se hace el procesamiento correspondiente y finalmente el servidor de almacenamiento realiza la entrega de la información.

Se puede observar que existe una comunicación direccional entre la base de datos multimedia y el servidor de almacenamiento, de tal manera que al haber un cambio en la base de datos multimedia (contenido audiovisual), se realizará una actualización en el servidor de almacenamiento (datos obtenidos de los descriptores).

En la base de datos multimedia se utilizan técnicas de indexado y recuperación de audiovisuales, las cuales son soportadas por SQL (SQL/MM, mejoramiento multimedia SQL). [2, "*Distributed Multimedia Database Technologies, supported MPEG-7 and by MPEG-21",* págs. 3,4]

2.4. Arquitectura

Una de las ventajas de MPEG-7 es que provee una flexibilidad y una estructura para representar la información obtenida de los descriptores. En este estándar se han definido 8 herramientas bajo la norma ISO-IEC 15938, y se representan en el siguiente diagrama [2, "*Distributed Multimedia Database Technologies, supported MPEG-7 and by MPEG-21",* pág. 26]:

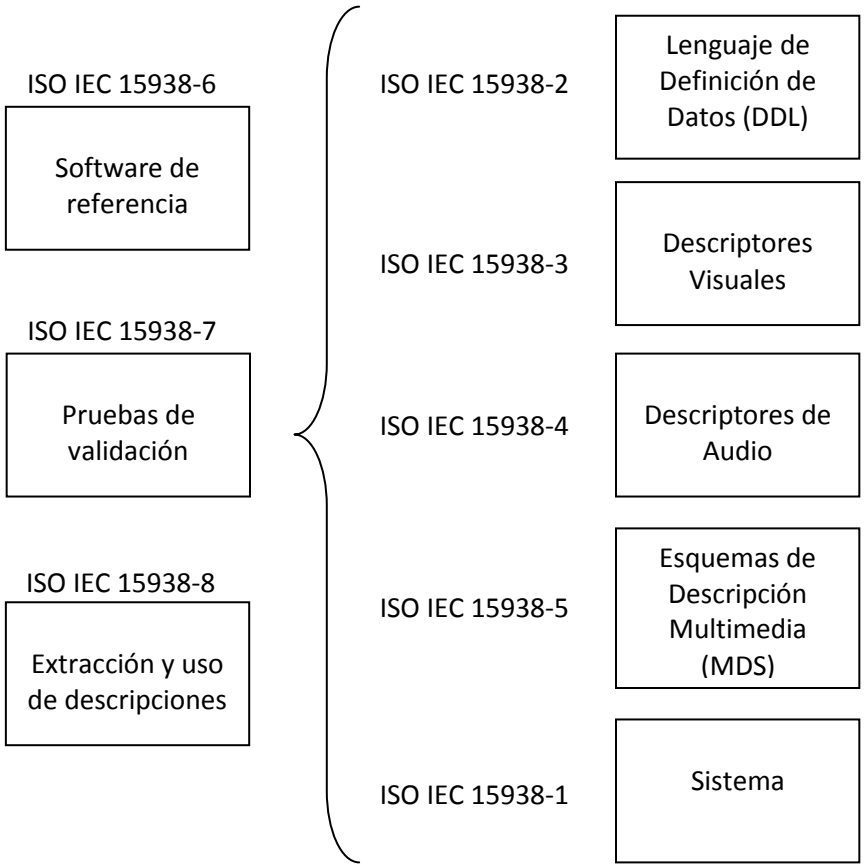


Figura 2.2. Organización general de MPEG-7 [2, "*Distributed Multimedia Database Technologies, supported MPEG-7 and by MPEG-21",* pág. 26].

Dónde:

- Parte 1, provee el formato binario para codificar las descripciones MPEG-7 y poder ser enviados.
- Parte 2, define la sintaxis de las herramientas de descripción de MPEG-7, a través del lenguaje de definición de descripción (DDL).
- Parte 3, contiene las herramientas de descripción para archivos visuales.
- Parte 4, contiene las herramientas de descripción para archivos de audio.
- Parte 5, contiene herramientas de características generales y descripciones multimedia.

- Parte 6, provee las normas para la implementación de las demás partes del estándar.
- Parte 7, Como su nombre lo menciona, en esta parte se definen los procedimientos para hacer las pruebas necesarias para validar el correcto funcionamiento de una nueva herramienta e implementarla.
- Parte 8, provee material de información acerca de la extracción y uso de algunos descriptores.

Las primeras 6 partes fueron publicadas en 2002, mientras que las últimas dos lo fueron en 2003 [2, "*Distributed Multimedia Database Technologies, supported MPEG-7 and by MPEG-21",* pág. 27].

Los elementos principales del estándar MPEG-7 son (ver figura 2.3):

- Descriptores de bajo nivel (LLD), que definen la sintaxis y la semántica de cada vector característico de audio y sus elementos.
- Esquemas de descripción (DSs), los cuales especifican la estructura y la semántica de la relación entre los descriptores y los propios esquemas DSs.
- El lenguaje de definición de descripción (DDL), el cual permite la extensión, modificación o creación de esquemas de descripción y de descriptores, define las reglas para expresar y combinar los esquemas de descripción y la forma en que se escriben los documentos MPEG-7.
- Y las herramientas de sistema, que permiten la representación de los descriptores y esquemas de descripción en formato binario, para el almacenamiento eficiente, transmisión, multiplexado y sincronización de los descriptores de contenido.

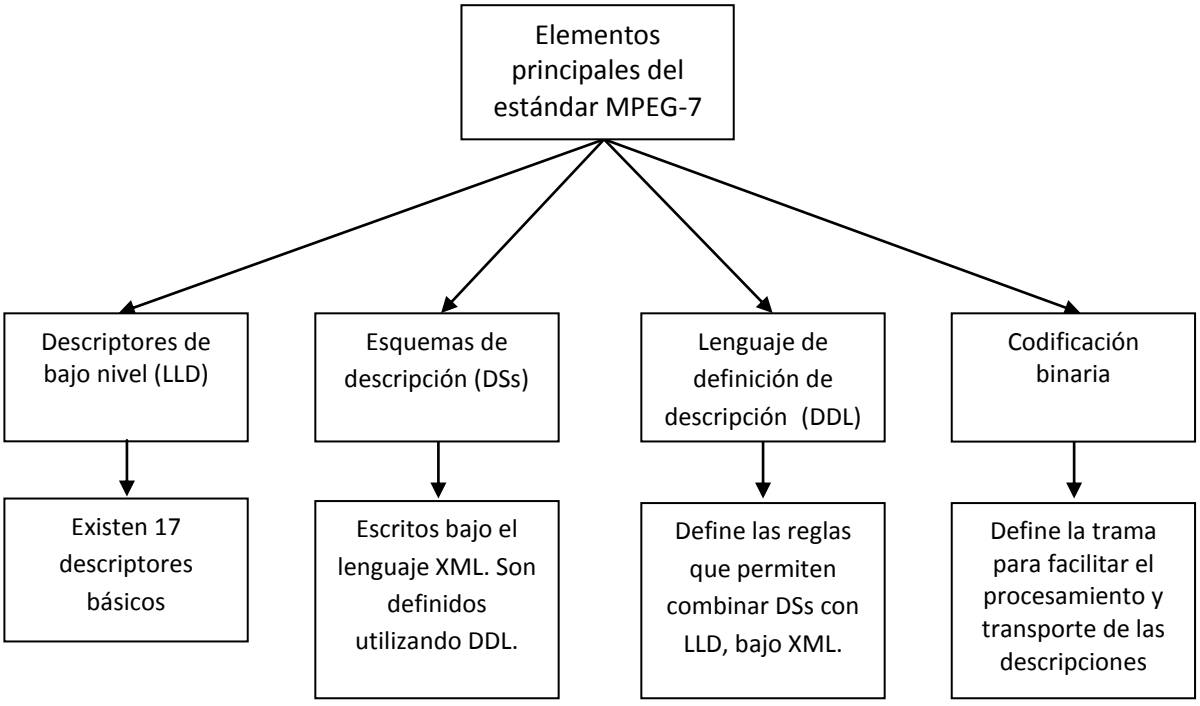


Figura 2.3. Representación gráfica de los elementos principales del estándar MPEG-7

2.4.1. Descriptores de bajo nivel

Existen 17 descriptores (espectrales y temporales), los cuales representan las variaciones en las propiedades del audio en tiempo y frecuencia. Estos son:

- Descriptores básicos. Los cuales muestran la forma de onda y la potencia del audio en el dominio del tiempo.
- Descriptores espectrales básicos. Representan el análisis en tiempo y frecuencia que describen el espectro del audio en términos de su envolvente, centroide, propagación y llanura.
- Descriptores de parámetros de la señal. Los cuales describen la frecuencia fundamental de una señal de audio, así como las frecuencias armónicas de una señal, por lo que sólo se utilizan en señales periódicas o cuasi-periódicas.
- Descriptores temporales de timbre. Se aplican en segmentos de sonidos donde se requiere conocer un timbre o tono característico.
- Descriptores espectrales de timbre. Utilizado en la percepción de timbres musicales, ya que representan características espectrales en un espacio lineal de frecuencia.
- Descriptores de base espectral. Los cuales representan proyecciones en dimensiones más simples de un espacio espectral con dimensiones más complejas. Utilizados en la clasificación de sonidos e indexado descriptivos.

2.4.2. Esquemas de Descripción MPEG-7

Los esquemas descriptores están divididos en 5 grupos de herramientas de descripción de audio que propiamente corresponden a las áreas de aplicación:

- Descripción de firmas de audio
- Descripción de timbres de instrumentos musicales
- Descripción de melodías
- Reconocimiento de sonidos y descripción de indexado
- Descripción de contenido de voz

2.4.3. Lenguaje de definición

El estándar MPEG-7 se basa en el lenguaje Extensible Markup Languaje (XML) o Lenguaje de Marcado Extensible, el cual fue definido por W3C (World Wide Web Consortium) quién hizo su primera publicación el 2 de mayo del 2001 y la segunda publicación el 28 de octubre del 2004[[6, "http://www.w3.org/XML/Schema"] y [2, "*Distributed Multimedia Database Technologies, supported MPEG-7 and by MPEG-21",* pág.31]].

En el esquema donde se codifican los elementos descriptivos para el procesamiento multimedia, el filtrado, y la interacción entre esquemas, donde un documento MPEG-7 que ha recopilado la información descriptiva tiene la siguiente estructura:

```
<Mpeg7>
…
<Time>
<TimePoint>
    2003-03-20T15:30+01:00
</TimePoint>
<Duration>
    P10D
<Duration>
</Time>
…
```

Figura 2.4. Representación del formato XML de un documento MPEG-7

2.4.4. Codificación Binaria

MPEG-7 define una estructura de codificación y decodificación y procesamiento, esta estructura es conocida como BiM, la cual permite transmitir un documento XML.

Entre unas de sus propiedades es que cuenta con una alta tasa de compresión al eliminar la redundancia estructural del documento.

Cada documento es dividido y transmitido por partes, cada pieza del documento es llamado unidad de acceso, donde cada unidad de acceso llevará la información completa de un descriptor.

Entre sus ventajas están que de esta manera se permite transmitir la información de un descriptor necesitado, ahorrando ancho de banda y en caso de que se modifique la información de un descriptor, sólo se tendrá que transmitir la información de este mismo y no de todo el documento.

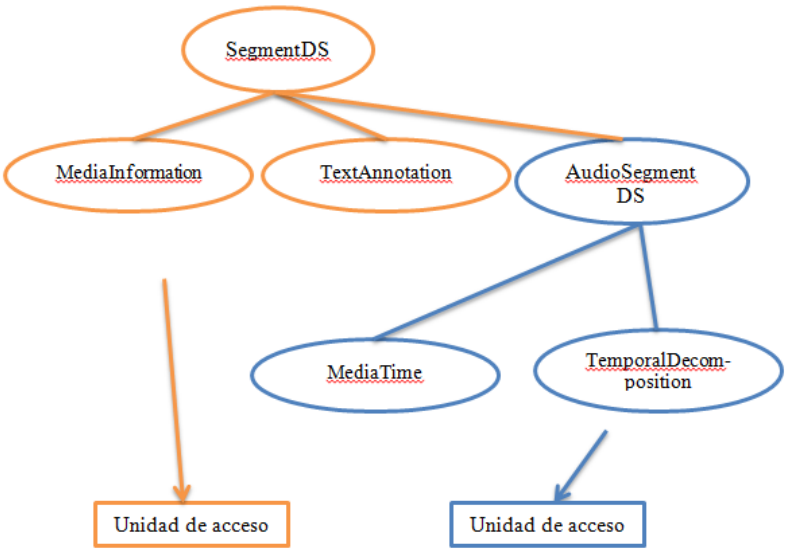


Figura 2.5. Estructura y transmisión de un documento XML

Bibliografía

1. **MOREAU, Nicolas, SIKORA, Thomas,** *MPEG-7 Audio and beyond,* Wiley, Alemania, 2005, p 281.
2. **KOSCH, Harald,** *Distributed Multimedia Database Technologies, supported MPEG-7 and by MPEG-21,* CRC Press, EstadosUnidos, 2004, p 260.
3. **ANGELIDES, Marios C, AGIUS Harry**, *The Handbook of MPEG Applications*, Wiley, Gran Bretaña, 2011, p 531.
4. **WATKINSON, John***, The MPEG Handbook*, Focal Press, 2da edición, Gran Bretaña, 2004, p 435.
5. **CHANG Wo, NIST**, Organismo de consulta de información, normatividad y librerías del estándar MPEG-7, http://m7itb.nist.gov/M7Validation.html, consultada marzo de 2011.
6. **W3C,** Organismo de consulta de información y normatividad del esquema XML, http://www.w3.org/XML/Schema, consultada marzo de 2011.
7. **JOANNEUM RESEARCH – DIGITAL,** Instituto de Tecnologías de Información y Comunicación ,http://iiss039.joanneum.at/cms/index.php?id=230, consultada marzo de 2011.

Capítulo 3. Esquemas de Descripción y descriptores de audio

3.1. Introducción

De acuerdo a la estructura vista en el capítulo anterior (figura 2.3), la parte 4 del estándar MPEG-7 se divide en 2 estructuras básicas, en la primera de ellas se encuentran los esquemas de descripción (DSs) o herramientas de alto nivel (HLDs), que se utilizan en aplicaciones definidas y en la segunda estructura están los descriptores de bajo nivel (LLDs), los cuales permiten obtener una descripción genérica del audio, permitiendo así tener una mayor flexibilidad en el diseño de nuevas aplicaciones y herramientas descriptivas.

Los descriptores que permiten obtener una descripción genérica de audio constan de 17 descriptores temporales y espectrales (LLDs) que son agrupados en 5 grupos, además de un esquema de series escalables y un descriptor de silencio, como se muestra en la figura siguiente:

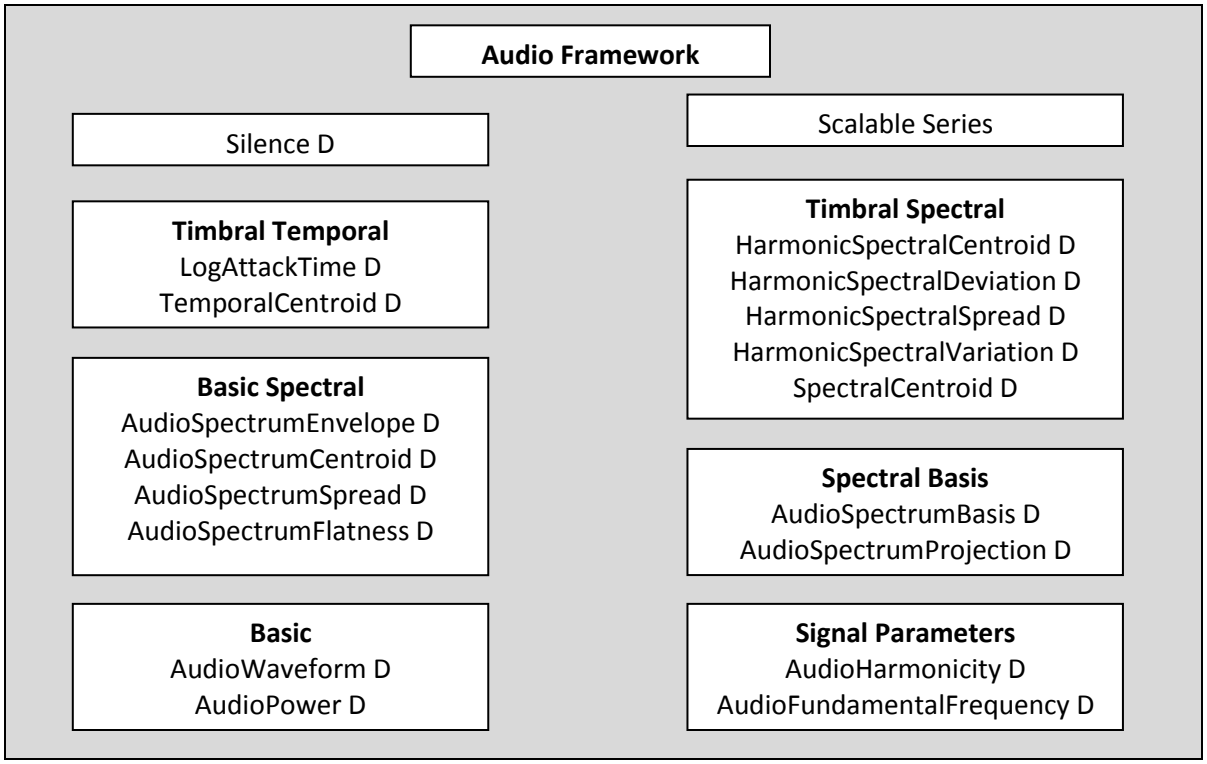


Figura 3.2. Agrupación de las herramientas descriptivas de bajo nivel [3, "*Coding of moving pictures and audio"*].

Los descriptores de audio de alto nivel (HLDs) se comprenden por las siguientes herramientas:

- Sound Recognition

- Instrumental Timbre Description
- Spoken Content Descriptions
- MelodyDescription Tools

3.2. Parámetros básicos

Existen dos maneras de realizar descripciones de una señal de audio bajo el estándar MPEG-7:

La primera de ellas es extrayendo características de segmentos de sonido de longitud variable para comparar y marcar regiones con distintas propiedades acústicas. Para este caso se utiliza el descriptor AudioSegment. Donde un segmento de audio representa un intervalo temporal arbitrario.

La segunda manera es extrayendo características de intervalos regulares, donde se utiliza la herramienta ScalableSeries.

Cada una de estas herramientas almacenan sus datos obtenidos en su descripción correspondiente.

3.2.1. Parámetros en el Dominio del Tiempo

Partiendo de una señal de audio digital, tenemos que en el dominio del tiempo, se utilizarán los siguientes parámetros para la señal de entrada como se muestra en la figura 3.3:

- *n* es el índice del número de muestras en el tiempo
- *s(n)* es la señal de audio
- F_s es la tasa de muestreo de *s(n)*

Los siguientes parámetros son utilizados en cada trama

- *l* es el índice o número de trama
- *hopSize* es el intervalo de tiempo entre dos tramas sucesivas
- N_{hop} corresponde al número de la muestra de la señal en el intervalo de tiempo *hopSize*
- L_w es la longitud de una trama ($L_w \geq hopSize$)
- N_w es el número de muestras de la señal correspondiente a L_w
- L es el número total de tramas en *s(n)*

3.2.2. Parámetros en el Dominio de Frecuencia

Para la extracción de características de audio en el dominio de la frecuencia, se utilizan técnicas, en las que para hacer una estimación o un cálculo, se utilizan técnicas basadas en el traslape de tramas, los siguientes parámetros son utilizados en el dominio de la frecuencia:

- *k* es el índice de frecuencia

- $S_l(k)$ es el espectro extraído de la trama *l* de *s(n)*
- $P_l(k)$ es el espectro de potencia extraído de la trama *s(n)*

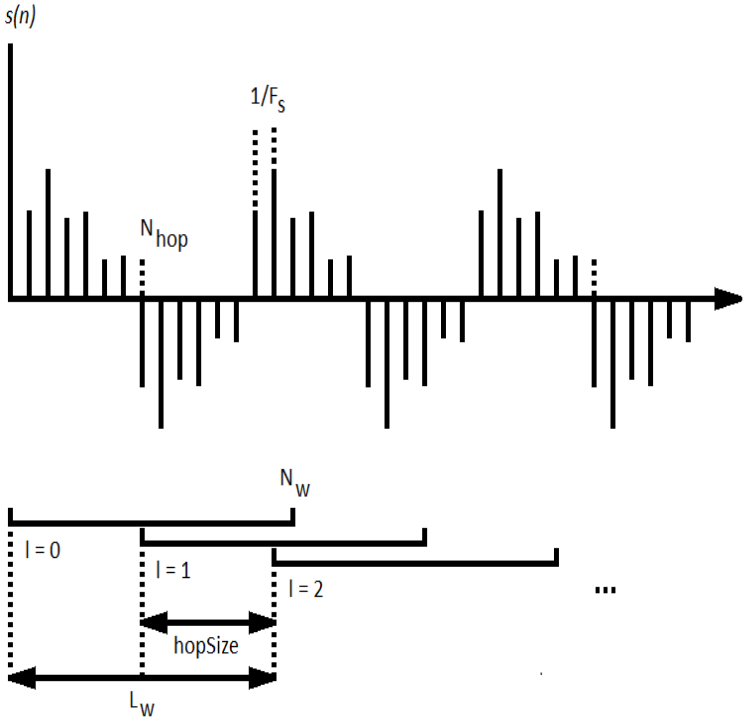


Figura 3.3. Parámetros utilizados en el dominio del tiempo

Para obtener una estimación de una señal en el dominio de la frecuencia, utilizando la Transformada Discreta de Fourier, primero se selecciona una trama de longitud L_w de donde se recomienda que L_w tenga una duración de 30 ms y que el valor de $hopSize$ sea igual a 10 ms. [2, "*MPEG-7 Audio and beyond* ", pág 15].

En el dominio del tiempo, la trama se multiplica por una ventana (por ejemplo una ventana Hamming) punto a punto, la cual permite analizar una porción de la señal de una longitud o tiempo L_w de una señal de tiempo infinito *s(n)*, en el dominio de la frecuencia dicha multiplicación equivale a una convolución de la señal *s(k)* junto con la ventana *w(k):*

$$S(k)= s(k)*w(k)$$

Dicha convolución cambia el espectro de frecuencia, permitiendo reducir en mayor o menor grado las frecuencias armónicas no deseadas debido a las discontinuidades de la señal y así obtener una mejor respuesta en frecuencia.

Después de realizar la convolución, se aplica la DFT, quedando la siguiente ecuación:

$$S_l(k) = \sum_{n=0}^{N_{FT}-1} s\left(n + lN_{hop}\right)w(n)exp^{\left(-j\frac{2\pi nk}{N_{FT}}\right)}(0 \le l \le L-1; o \le k \le N_{FT}-1) \qquad (3.1)$$

De donde podemos observar que $s(n + lN_{hop})$, corresponde a la muestra de la señal dentro de la trama *l*, *w(n)* es la ventana aplicada (Hamming), N_{FT} es el número de coeficientes de la DFT, cuyo tamaño es igual a la potencia siguiente de 2, cuyo valor sea mayor o igual a N_w ($N_{FT} \geq N_w$). Donde los coeficientes agregados debido a esta potencia en el intervalo $N_w + 1$ hasta N_{FT}, tienen un valor igual a 0.

De acuerdo al Teorema de Parseval, la potencia promedio de la trama *l* de análisis es:

$$\overline{P_l} = \frac{1}{E_w}\sum_{n=0}^{N_w-1}\left|s(n + lN_{hop})\right|^2 = \frac{1}{N_{FT}E_w}\sum_{k=0}^{N_{FT}-1}|S_l(k)|^2 \quad (3.2.a)$$

De donde podemos observar que el espectro de potencia P_l, se define como la sumatoria del cuadrado de cada uno de los coeficientes N_{FT}del espectro $S_l(k)$ de la trama *l,* dividido entre la energía de la ventana *w(n)* y el número total de coeficientes N_{FT}. El espectro de potencia $P_l(k)$, de cada coeficiente de la Transformada de Fourier para cada trama es igual a:

$$P_l(k) = \frac{1}{N_{FT}E_w}|S_l(k)|^2, \quad para\ k = 0\ y\ k = \frac{N_{FT}}{2}$$

$$P_l(k) = 2\ \frac{1}{N_{FT}E_w}|S_l(k)|^2, \quad para\ 0\ < k\ \frac{N_{FT}}{2} \quad (3.2.b)$$

La energía de la ventana aplicada está dada por:

$$E_w = \sum_{n=0}^{N_w-1}|w(n)|^2 \quad (3.3)$$

Existe una relación entre cada coeficiente de la Transformada de Fourier y el espectro de frecuencias. Cada coeficiente representa un cierto rango de frecuencias dado por:

$$\Delta F = \frac{F_s}{N_{FT}} \quad (3.4)$$

Para encontrar el coeficiente respectivo al que pertenece una frecuencia *f*, se utiliza la siguiente ecuación:

$$k = round\left(\frac{f}{\Delta F}\right)\left(0 \leq f \leq \frac{F_s}{2}\right) \quad (3.5.a)$$

Por lo que:

$$f(k) = k\Delta F \quad \left(0 \leq k \leq \frac{N_{FT}}{2}\right) \quad (3.5.b)$$

Del siguiente ejemplo, se tomó una fracción de 6 segundos de la canción Erase&Rewind, la cual tiene una tasa de muestreo *(F_s)* de 44.1 kHz. El espectro de potencia será extraído mediante la FFT, donde el valor de *hopsize* es igual a 10 ms y la trama L_w igual a 30 ms, el número de muestras N_w es igual a 1323, N_{FT} es igual a 2^{11}= 2048 ya que 2^{10}= 1024 es menor a N_w. El rango de

frecuencias del espectro de la señal será de 0 a 22.05 kHz donde la frecuencia de Nyquist es de 44.1 kHz.

Por lo que el espectro de frecuencias de la figura 3.5.a y 3.5.b, estará definido por las ecuaciones (3.7) y (3.8):

Donde $\Delta F = \frac{F_s}{N_{FT}} = \frac{44.1\ kHz}{2048}$ =21.53 kHz, redondeándolo a 22 kHz

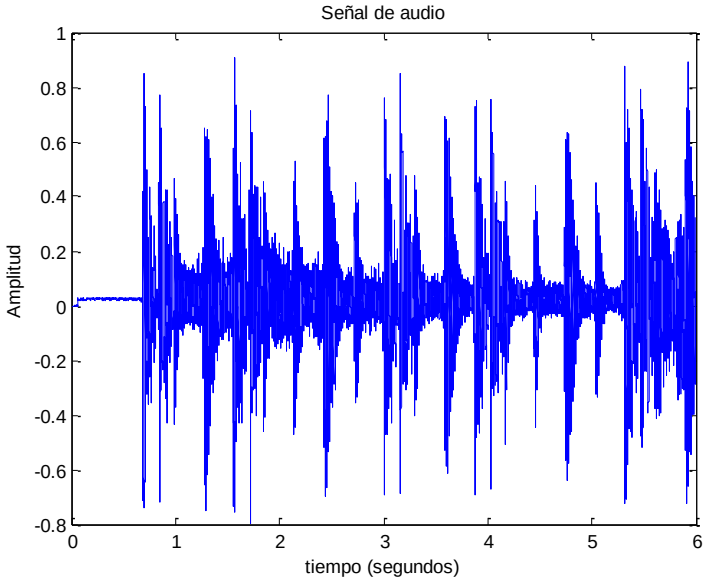


Figura 3.4. Señal de entrada en función del tiempo (Erase and Rewind, 44.1 kHz)

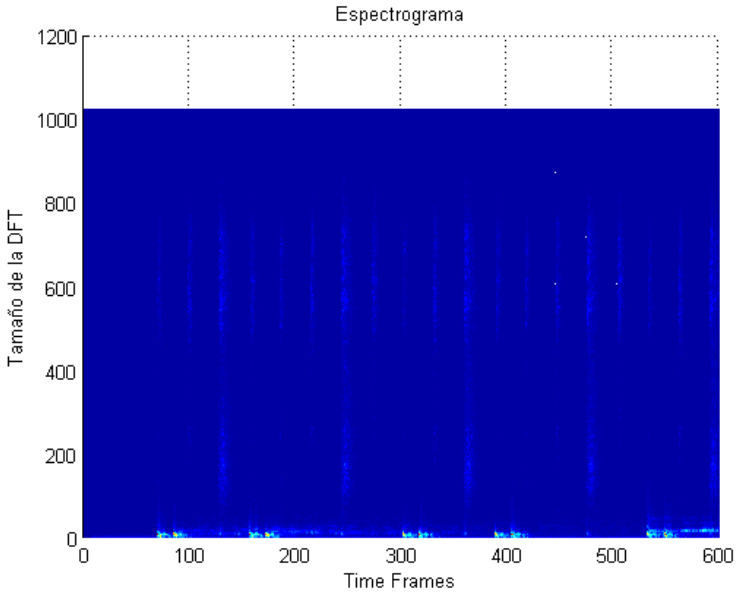


Figura 3.5.a) Espectrograma de la señal de la figura 3.4, donde se observa el tamaño de la FFT ($\frac{N_{FT}}{2} = 1024$), donde para un extracto de 6 segundos, se tiene un total de 598 tramas con una duración de 30 ms cada una y un *hopSize* igual a 441 muestras por trama.

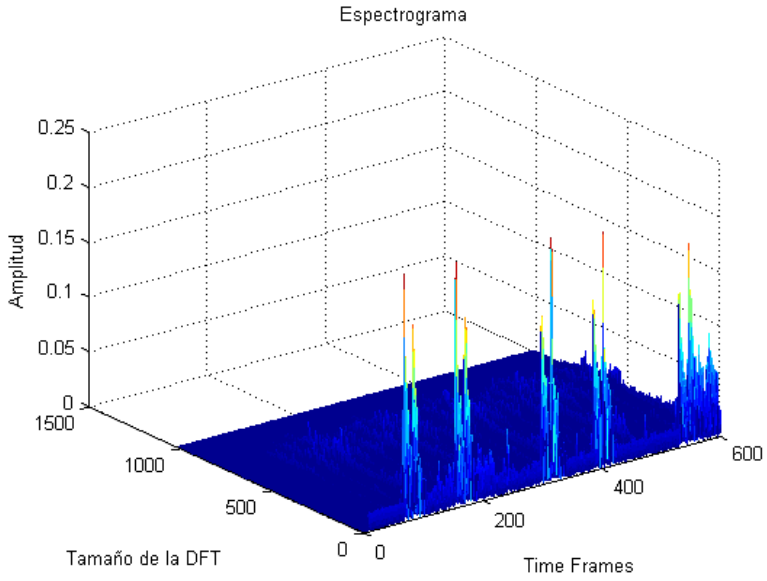


Figura 3.5.b) Espectrograma de la señal de la figura 3.4., donde se observa la amplitud de cada una de las frecuencias de la FFT de la figura 3.5.

Aplicando el Teorema de Parseval de la ecuación (3.2), se obtiene la siguiente gráfica:

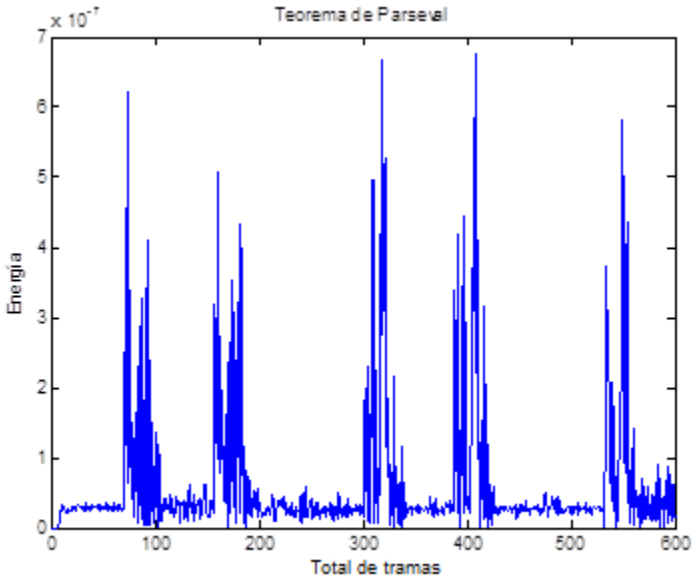


Figura 3.6. Gráfica de la ecuación 3.2. (Teorema de Parseval), donde se aprecia la energía de cada trama.

3.3. Series escalables

Una descripción de series escalables (ScalableSeries), es una manera estandarizada de representar características de un descriptor LLD de manera resumida, en forma de series de vectores, extraídos de las tramas de sonido en intervalos de tiempo regulares (L_w) cuyos atributos son:

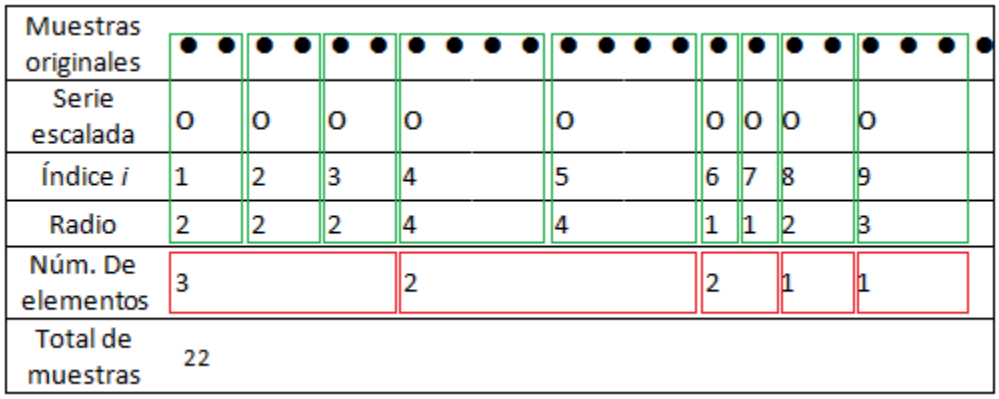


Figura 3.7. Representación gráfica de la estructura de una serie escalable

Como se puede observar en la figura, tenemos un total de 22 muestras totales, de estas muestras tenemos una serie escalada que consta en tomar una muestra como referencia o índice, con un radio que es igual al número de muestras a considerar por cada punto escalado (recuadros verdes), de ahí cada punto escalado con misma longitud o radio se agrupa en otra variable llamada número de elementos (recuadros rojos).

Escalado (Scaling): es una bandera que especifica cómo será escalada la señal original (muestras originales), en caso de que no se escale una señal original, esta bandera no se declara.

Número de muestras (totalNumOfSamples): indica el número total de muestras de la señal original, antes de realizarse alguna operación de escalamiento. En la imagen se seleccionaron 22 muestras, en los ejemplos de las figuras 3.5 y 3.6 se seleccionaron un total de 264600 muestras, las cuales representan 6 segundos de un extracto de canción muestreada a 44.1 kHz.

Radio (Ratio): es un número que indica cuantas muestras contiene cada trama, como se puede observar en la figura, cada trama (asociada con el número de índice) comprende un Radio que es igual al número de muestras originales contenidas dentro de esta misma. Para los ejemplos de las figuras 3.5 y 3.6, el Radio (*scaling Ratio*) tiene un valor igual a 441 muestras comprendidas en un tiempo de 10 ms a una frecuencia de muestreo (Fs) igual a 44.1 kHz.

Número de elementos (*numOfElements*): índica el número total de las tramas de la serie escalable, teniendo en cuenta que índica el número de tramas consecutivas cuyo Radio es del mismo valor, como se observa en la figura anterior, en las figuras 3.5 y 3.6 se puede observar que dicho valor es igual a 600, de donde:

$$numofElements = \frac{totalNumofSamples}{Ratio}$$

Por lo que una serie escalada representa el número de tramas (número de elementos), su respectivo índice, y cuantas muestras contiene cada trama (Radio).

3.3.1. Series de escalares

El estándar MPEG-7, comprende un descriptor de Series Escalares (*SeriesOfScalar*), el cual representa los valores mencionados anteriormente, y los cuales son utilizados en los descriptores temporales LLDs, cuyos atributos son:

Raw. Contiene la serie original en caso de que no se haya escalado dicha señal. Como se mencionó anteriormente la bandera *Scaling* indica si existe una operación de escalado.

Weight. Esta serie es opcional, en caso de existir, cada *weight* corresponde a la muestra de la serie original. Este parámetro se utiliza para controlar el escalado.

Min, Max y *Mean.* Caracterizan los valores de una trama de la serie escalada. Para Min se toma el valor mínimo del total de muestras de una trama, para Max se toma el valor máximo de las mismas muestras de la trama, y para Mean se toma el valor promedio de estas mismas muestras. Estos atributos no se toman en cuenta en caso de no existir escalado.

Variance. Cada elemento de este vector corresponde al de una trama, y considera los pesos (*weight*) si se encuentran presentes. Este valor no se considera cuando no existe escalamiento.

Random. Es un vector que selecciona un valor aleatorio del total de muestras contenidas en una trama, por lo que debe de existir escalamiento.

*First.*Este vector contiene el valor de la primera muestra de cada trama, por lo que no se considera si no existe escalamiento.

Last. Este vector contiene el valor de la última muestra de cada trama, por lo que no se considera si no existe escalamiento.

A continuación se exponen las fórmulas, para cada uno de los vectores mencionados anteriormente.

Considerando que en una trama L_w la cual contiene un número de muestras igual al Radio o N_w, asumiremos que si *Raw* contiene el número total de muestras de una señal la cual no ha sido escalada y *l* es el índice o número de trama, donde *Lo(i)* es la primer muestra de cada trama y *Hi(i)* la última muestra de cada trama, tenemos la siguiente ecuación:

$$lHi(i) = lLo(i) + ratio - 1 \quad (3.6)$$

Definiendo que *Raw'(l)*, solo contendrá el número de muestras contenidas en una trama. Las respectivas ecuaciones son:

$$Min(i) = min_{l=lLo(i)}^{lHi(i)} Raw'(l) \tag{3.7}$$

$$Max(i) = max_{l=lLo(i)}^{lHi(i)} Raw'(l) \tag{3.8}$$

Cuando no se especifica el peso (*weight*), el promedio y la varianza se define de la manera siguiente:

$$Mean(i) = \frac{1}{ratio} \sum_{l=lLo(i)}^{lHi(i)} Raw'(l) \tag{3.9}$$

$$\text{Variance(i)} = \frac{1}{\text{ratio}} \sum_{\text{l=lLo(i)}}^{\text{lHi(i)}} [\text{Raw'(l)} - \text{Mean(i)}]^2 \tag{3.10}$$

Cuando se especifica el peso (*weight*), el promedio y la varianza se define de la manera siguiente:

$$Mean(i) = \frac{\sum_{l=lLo(i)}^{lHi(i)} W(l) Raw'(l)}{\sum_{l=lLo(i)}^{lHi(i)} W(l)} \tag{3.11}$$

$$Variance(i) = \frac{\sum_{l=lLo(i)}^{lHi(i)} [Raw'(l) - Mean(l)]^2}{\sum_{l=lLo(i)}^{lHi(i)} W(l)} \tag{3.12}$$

El peso seleccionado es el resultado de la sumatoria

$$W(i) = \frac{1}{ratio} \sum_{l=lLo(i)}^{lHi(i)} W(l) \tag{3.13}$$

3.4. Descriptores básicos

De acuerdo a la figura 3.2., existen dos descriptores básicos, los cuales nos brindan una descripción temporal de la señal de audio, estos descriptores son: Forma de onda de audio (AudioWaveform) y Energía de Audio (AudioPower).

3.4.1. Forma de onda de audio (Audio Waveform)

El descriptor "Forma de onda de audio" describe la envolvente de la señal utilizando los valores mínimos y máximos de cada trama de la señal, donde cada trama tiene un tamaño $L_w = hopSize$. Estos valores son declarados con las siguientes variables:

- minRange= es el valor mínimo de la señal para cada trama
- maxRange= es el valor máximo de la señal para cada trama

La función utilizada para obtener estos valores y poder representar la figura 3.7. es:

```
[Raw,maxValues,minValues,notfirstvalues,varianceScalewiseValues]=AudioWav
eFormD(music,totalSampleNum,scalingRatio,elementNum,Weight_flag,Weight,Wr
ite_flag,rootFirst,XMLFile);
```

Al utilizarla en Matlab, se declara de la siguiente manera (ver anexo):

```
[Raw,maxValues,minValues,notfirstvalues,varianceScalewiseValues]=AudioWav
eFormD(music,totalSampleNum,scalingRatio,elementNum,0,0,1,1,carp);
```

Dónde:

- *music* es un extracto de música de 6 segundos, cuya frecuencia de muestreo (F_s) es de 44.1 kHz.
- Durante los 6 segundos existen 264600 muestras, siendo este el valor de *totalSampleNum*.
- *scalingRatio* es igual a 441 muestras, este valor se obtiene multiplicando el tiempo de *hopsize (10 ms)* por la frecuencia de muestreo (F_s).
- *elementNum* es igual a 598, y representa el número de tramas de la señal de audio (señal escalada), donde este valor se obtiene dividiendo el número total de muestras (*totalSampleNum)* entre*ScalingRatio.*
- *Weight_flag* y W*eight* con valores igual a 0, permite, obtener los valores mínimos y máximos, necesitados para este ejemplo.
- *Write_flag* con valor igual a 1, permite guardar los resultados en un archivo XML, con valor 0, no guarda dichos resultados.
- *rootFirst* con valor ajustado a 1, sólo es utilizado cuando scalingRatio es un número par, y en este caso permite obtener la serie *rootFirstValues*.
- *XMLFile* es el directorio donde se creará y guardara el descriptor en formato XML

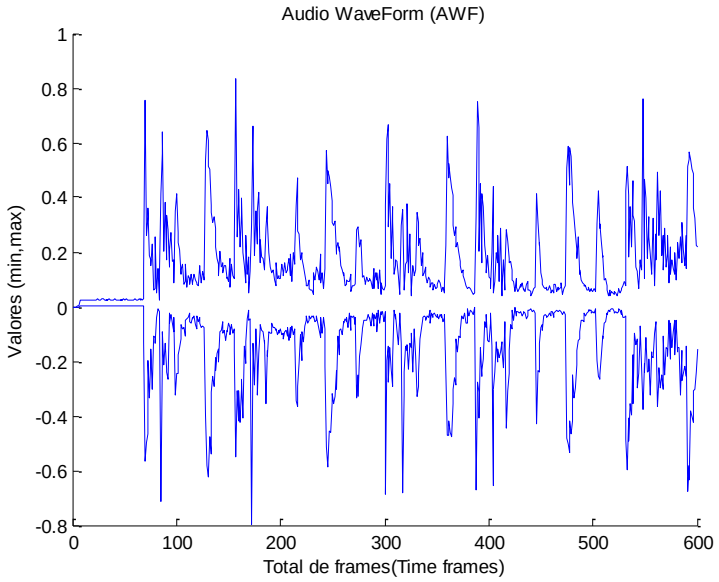


Figura 3.7. Gráfica de la forma de onda de la señal de audio de la figura 3.4. Señal de entrada (Erase and Rewind), muestreada a 44.1 kHz.

3.4.2. Energía de Audio (Audio Power)

Este descriptor describe de manera temporal, la potencia instantánea de cada trama, cada uno de los coeficientes (AP) obtenidos de cada trama ($L_w = hopSize$), se calculan mediante la siguiente ecuación:

$$AP(l) = \frac{1}{N_{hop}} \sum_{n=0}^{N_{hop}-1} \left| s(n + lN_{hop}) \right|^2 \ (0 \leq l \leq L-1) \qquad (3.14)$$

[2, "*MPEG-7 Audio and beyond* ", pág. 34].

De acuerdo a la nomenclatura establecida en la sección 3.2.1., *L* representa el número total de tramas (de los ejercicios anteriores, es igual a 598), N_{hop} es el número de muestra dentro de una trama, *l* es el número de trama, por lo que esta ecuación nos representa el valor promedio del cuadrado de la sumatoria de cada una de las muestras dentro del intervalo $(0 \leq n \leq N_{hop} - 1)$. Siguiendo con la señal de audio tomada anteriormente como ejemplo, tenemos la representación gráfica de dicho descriptor.

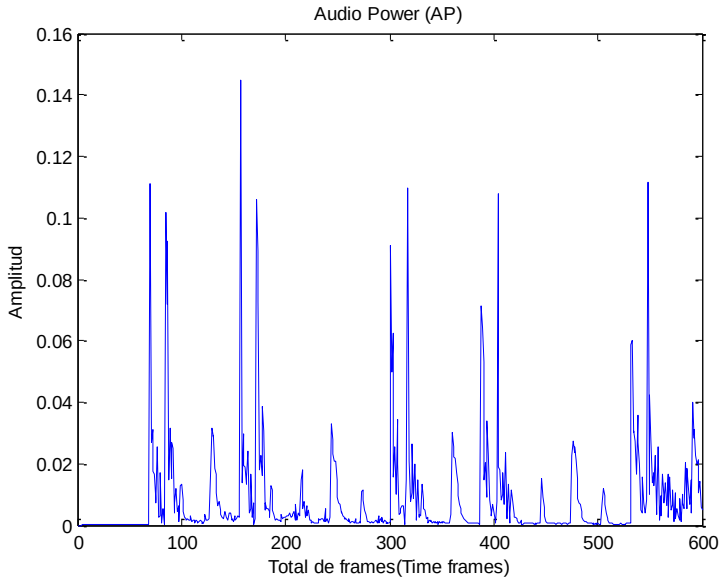


Figura 3.8. Representación gráfica del descriptor Audio Power

Donde, la función correspondiente es:

```
AudioPower_SeriesOfScalar =
AudioPowerType(auData,totalSampleNum,samplingRate,scalingRatio,elementNum
,weight,XMLFile)

[AudioPower]=AudioPowerD(music,totalSampleNum,Fs,scalingRatio,elementNum,
1,carp);
```

De dónde:

- *Fs* es la frecuencia de muestreo
- Los demás valores son los mismos que se declaran en la función AudioWaveForm

3.5. Descriptores básicos de espectro

Dentro de este grupo se encuentran 4 descriptores LLD, los cuales proporcionan una descripción de la frecuencia en forma logarítmica del espectro de potencia, estos descriptores parten de un análisis simple en tiempo y frecuencia de una señal de audio. Para poder realizar dicho análisis se utiliza un traslape de tramas, como se muestra en la figura 3.3.

3.5.1. Envolvente del Espectro de Audio (Audio Spectrum Envelope)

Envolvente del Espectro de Audio (Audio Spectrum Envelope) es el descriptor principal dentro del grupo, ya que de él se toman algunos resultados para los demás descriptores. Es un vector qué describe el espectro de potencia de cada una de las bandas establecidas de acuerdo a una

resolución espectral *r* en una distribución logarítmica (base 2) de una señal de audio, por lo que puede ser empleado para generar un espectrograma de dicha señal.

Dicha resolución espectral (*r*) está definida como el número de bandas de frecuencia por octava, dentro del intervalo (*loEdge* a *hiEdge*), existen 8 posibles resoluciones que van desde 1/16 de octava, siguiendo con 1/8 de octava, ¼, ½, 1, 2, 4, hasta 8 octavas.

$$\mathrm{r}= 2^{j} \text{ octavas} \quad -4\leq \mathrm{j} \leq 3 \tag{3.15}$$

Los límites de frecuencia de cada banda ésta dado por:

$$\mathrm{Edge} = 2^{\mathrm{rm}} * 62.5 \text{ Hz} \quad (0 \leq \mathrm{m} \leq B_{in}) \tag{3.16}$$

Donde *r* es la resolución espectral y *m* es el número de banda dentro del intervalo (*loEdgeahiEdge*).

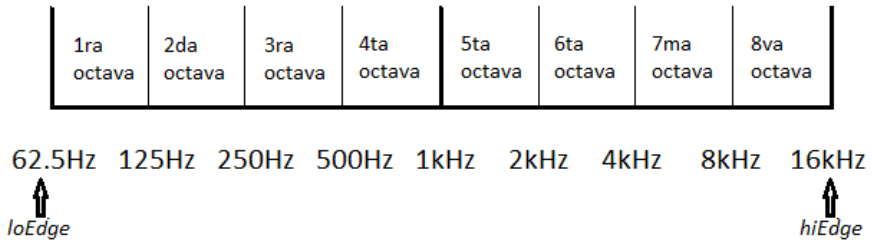


Figura 3.9. Rango de frecuencias (*loEdge-hiEdge*) y las bandas de dichas frecuencias que coinciden con el número de octavas para una resolución *r* igual a 1.

Los valores por default para *hiEdge* y *loEdge* son 16 kHz y 62.5 Hz, respectivamente, el número de bandas que existen en este intervalo ésta dado por: B_{in}=8/r. Los límites de frecuencia inferior (loF_b) y superior (hiF_b) para cada banda son:

$$\begin{matrix} loF_b = loEdge \; x \; 2^{(b-1)r} \\ hiF_b = loEdge \; x \; 2^{br} \end{matrix} \quad (1 \leq b \leq B_{in}) \tag{3.17}$$

La suma de los coeficientes de energía $P(k)$ en cada banda de frecuencias *b* de cada trama, limitados por loK_by hiK_b, se conoce como un coeficiente Audio Spectrum Envelope (ASE), y se calcula de la siguiente manera:

$$ASE(b) = \sum_{k=loK_b}^{hiK_b} P(k), \; (1 \leq b \leq B_{in}) \tag{3.18}$$

Adicionalmente a estos coeficientes ASE, se agregan dos coeficientes más, uno que contendrá la sumatoria de los coeficientes de energía $P(k)$ desde los 0 Hz hasta los 62.5 Hz (ver figura 3.11, (1)) y el segundo coeficiente ASE contendrá la energía de los coeficientes $P(k)$ comprendidos desde los 16 kHz hasta $F_s/2$ (ver figura3.11, (3)). Por lo que el número total de coeficientes será $B = B_{in} + 2$.

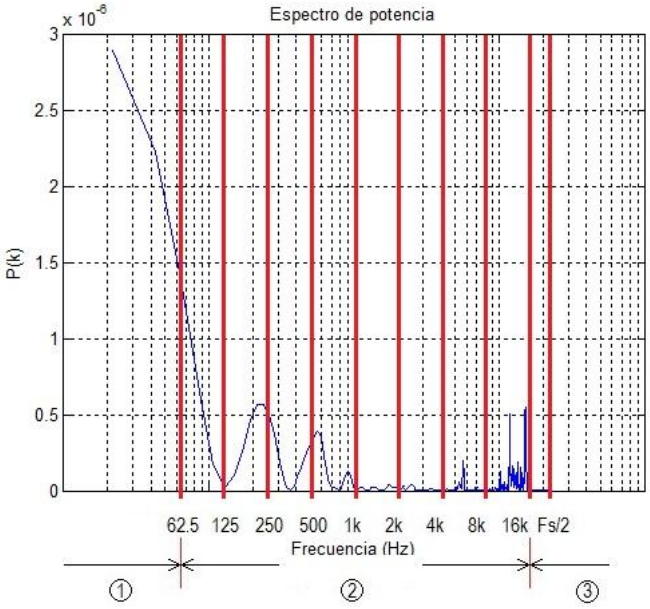


Figura 3.11. Extracción de los coeficientes de energía $P_l(k)$ en el intervalo $\left(0: k: \frac{N_{FT}}{2}\right)$ para la trama 450, separada por las bandas de frecuencia para una resolución espectral (*r*) igual a 1. La sumatoria de la energía $P_l(k)$ en cada banda es igual a un coeficiente *ASE(b).*

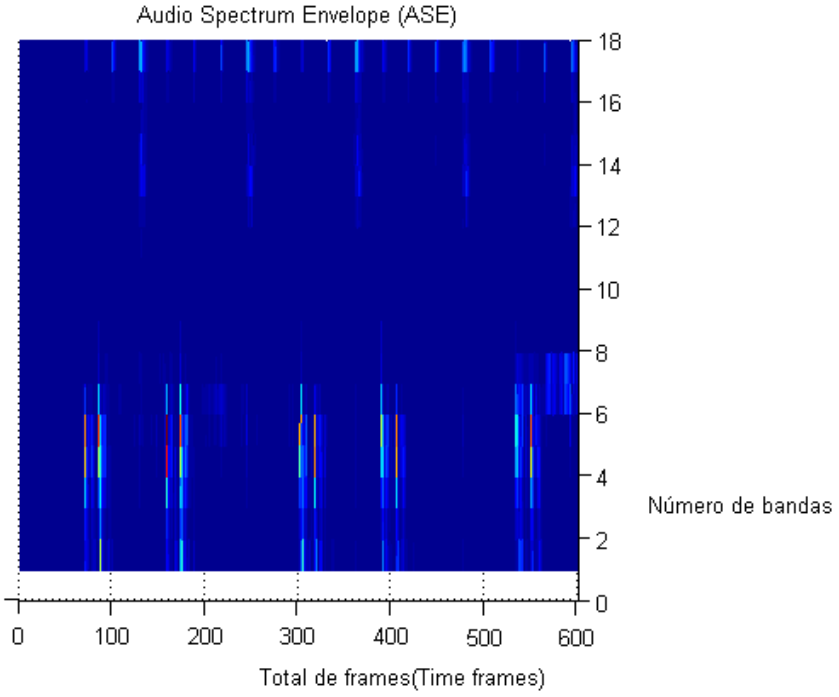


Figura 3.12. Representación gráfica de los coeficientes ASE del extracto de música, para una resolución espectral de ½.

Como se puede observar en la figura 3.12., existen 18 bandas de acuerdo a la ecuación $B = B_{in} + 2$, y sus límites de frecuencia de cada banda se limitan por la ecuación (3.17).

Banda	Rango de frecuencias (Hz)	Banda	Rango de frecuencias (Hz)
1	**0 a 62.5**	10	**1000 a 1414.21**
2	**62.5 a 88.38**	11	**1414.21 a 2000**
3	**88.38 a 125**	12	**2000 a 2828.42**
4	**125 a 176.77**	13	**2828.42 a 4000**
5	**176.77 a 250**	14	**4000 a 5656.85**
6	**250 a 353.55**	15	**5656.85 a 8000**
7	**353.55 a 500**	16	**8000 a 11313.70**
8	**500 a 707.10**	17	**11313.70 a 16000**
9	**707.10 a 1000**	18	**16000 a 22050**

Tabla 3.1., donde se muestra el rango de frecuencias perteneciente a cada banda para una resolución espectral de media octava.

3.5.2. Centroide del Espectro de Audio (Audio Spectrum Centroid)

Este descriptor proporciona en donde se concentra la mayor parte de la energía (centroide) del espectro de potencia referenciado a 1kHz, cuando el valor de la figura es igual a "0". Esto quiere decir que de la gráfica de la figura 3.13 nos indica en que rango de frecuencias se concentra la mayor parte de la energía por cada trama. Donde los coeficientes del espectro debajo de los 62.5 Hz son sumados y se representan por un solo coeficiente (ecuación 3.20, para $k = 0$) para evitar una componente DC, estos coeficientes menores e iguales al índice K_{low}, son los que serán sumados:

$$K_{low} = floor\left(\frac{62.5}{\Delta F}\right) \tag{3.19}$$

Esto da como resultado un nuevo espectro de potencia $P'(k')$:

$$P'(k') = \begin{cases} \sum_{k=0}^{K_{low}} P(k) & para\ k = 0 \\ P(k + K_{low}) & para\ 1 \leq k \leq \frac{N_{FT}}{2} - K_{low} \end{cases} \tag{3.20}$$

Y las frecuencias correspondientes de dicho espectro, ésta dado por:

$$f'(k') = \begin{cases} 31.25\ Hz & para\ k = 0 \\ f(k + K_{low}) & para\ 1 \leq k \leq \frac{N_{FT}}{2} - K_{low} \end{cases} \tag{3.21}$$

Como se puede observar en la ecuación anterior el valor nominal para la representación de los coeficientes por debajo de los 62.5 Hz es igual a 31.25 Hz ($f'(0) = 31.25\ Hz$).

Finalmente, el descriptor ASC está definido por la siguiente ecuación:

$$ASC = \frac{\sum_{k=0}^{\left(\frac{N_{FT}}{2}\right)-K_{low}} log_2\left(\frac{f'(k')}{1000}\right) P'(k')}{\sum_{k=0}^{\left(\frac{N_{FT}}{2}\right)-K_{low}} P'(k')} \tag{3.22}$$

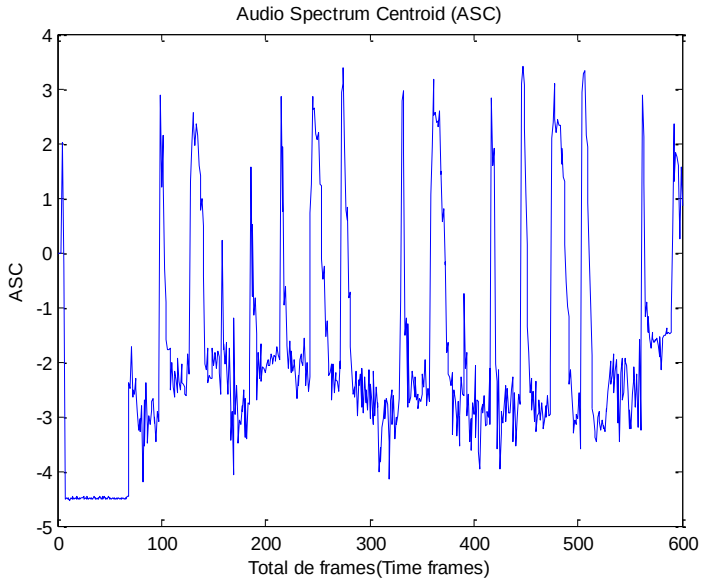


Figura 3.13. Representación gráfica del descriptor ASC para el extracto de música de los ejemplos anteriores.

En la figura anterior, se puede observar que el espectro de frecuencias del extracto de música prevalece en un valor de -2 a -3, que es el rango de frecuencias audibles de los 250 Hz a los 120 Hz aproximadamente lo que indica que este extracto de música ésta dominado por instrumentos musicales que emiten bajas frecuencias

3.5.3. Propagación del Espectro de Audio (Audio Spectrum Spread)

Este descriptor nos indica cómo se encuentra distribuida la energía espectral alrededor de su centroide. Un valor bajo significa que la mayor parte de la energía del espectro se encuentra concentrado en su centroide, mientras que un valor alto indica que la distribución de energía se encuentra en un amplio rango de frecuencias, esto permite reconocer si existe un tono puro o ruido, los valores ASS están dados por:

$$ASS = \sqrt{\frac{\sum_{k=0}^{\left(\frac{N_{FT}}{2}\right)-K_{low}} \left[log_2\left(\frac{f'(k')}{1000}\right) - ASC\right]^2 P'(k')}{\sum_{k=0}^{\left(\frac{N_{FT}}{2}\right)-K_{low}} P'(k')}} \quad 3.23)$$

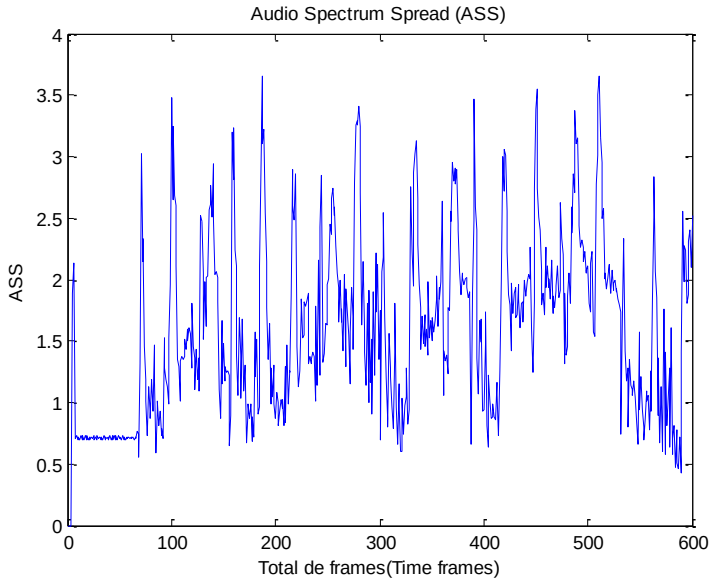


Figura 3.14. En esta representación se puede observar, que hay una dispersión de la distribución de la energía espectral, en la mayor parte de las tramas.

3.5.4. Planitud del Espectro de Audio (Audio Spectrum Flatness)

Este descriptor consiste en una serie de valores que expresan la relación que existe entre el espectro de potencia de una señal respecto a una señal de ruido blanco. Un espectro "plano" puede corresponder a un impulso de la señal o ruido, como se observa en la figura de abajo para las primeras tramas. Donde los valores de los coeficientes ASF cuyo valor es alto, significa o refleja ruido en la señal o que no existe un tono en particular presente en dicha banda, un valor bajo en dichos coeficientes indican una estructura armónica de espectro o un tono dentro de esa banda.

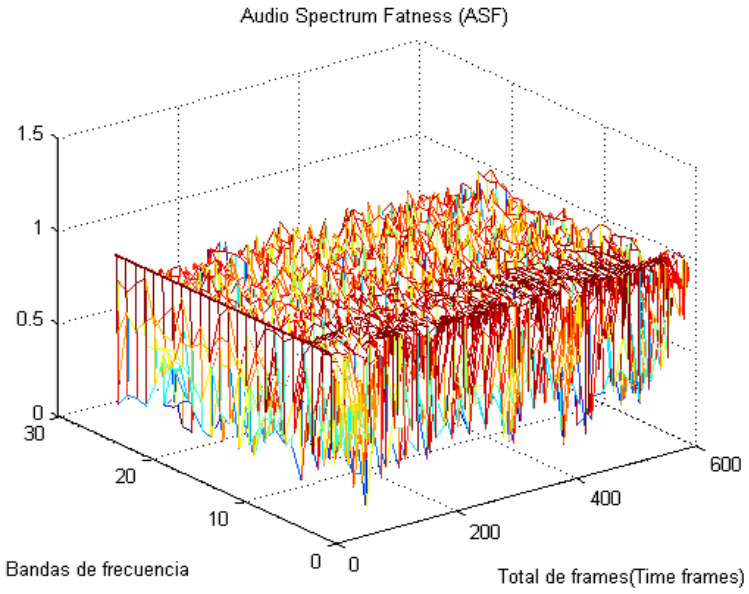


Figura 3.15 Representación de los valores obtenidos del descriptor, donde se muestra un espectro plano en las primeras tramas correspondiente a un impulso, y los diferentes valores ASF que corresponden a un tono en el caso de que dicho valor sea bajo, y a ruido en el caso de que no exista un tono especifico.

Las bandas de frecuencias tienen una resolución de ¼ por octava, y los valores de *loEdge* y *hiEdge,* para las bandas ésta dado por las siguientes ecuaciones:

$$loEdge = 2^{\frac{1}{4}n} x\ 1kHz$$

$$hiEdge = 2^{\frac{1}{4}B}\ x\ loEdge$$

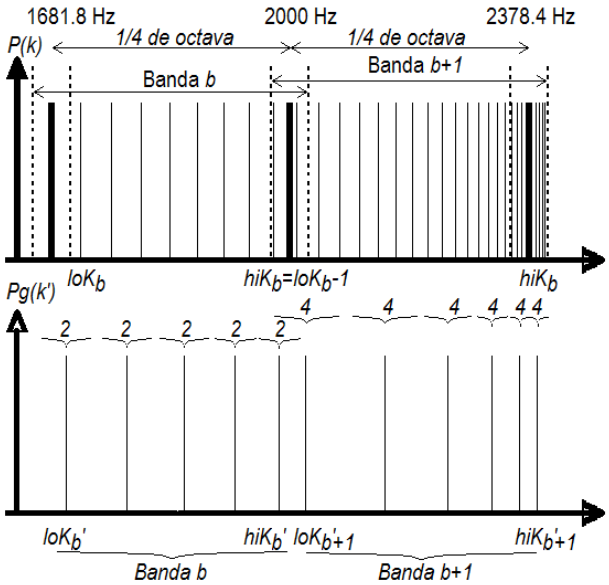
De donde *n* tiene un valor igual a "-8", ya que el valor mínimo recomendado para *loEdge* es igual a 250 Hz, el valor de *B* es determinado para que el valor *hiEdge* no sobrepase la frecuencia de Nyquist, y así se pueda realizar un cálculo correcto.

El valor de los límites de las bandas de frecuencia son modificados de tal forma que entre una banda y otra, exista un traslape de dichas bandas en un 10 %.

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

$$hiF_b = 0.95\ x\ loEdge\ x\ 2^{\frac{1}{4}b} \qquad (1 \leq b \leq B)$$

Para disminuir el procesamiento y realizar el cálculo de los coeficientes ASF, se hace un promedio de cada $2n+1$ coeficientes de energía $P(k)$, donde *n* indica el número de banda a partir de 1kHz, obteniendo un sólo coeficiente de energía $P_g(k')$.



Finalmente, para cada banda *b*, el coeficiente ASF es calculado por la siguiente ecuación:

$$ASF(b) = \frac{\sqrt[hiK'_b - loK'_b + 1]{\prod_{k'=loK'_b}^{hiK'_b} P_g(k')}}{\frac{1}{hiK'_b - loK'_b + 1}\sum_{k'=loK'_b}^{hiK'_b} P_g(k')} \tag{3.24}$$

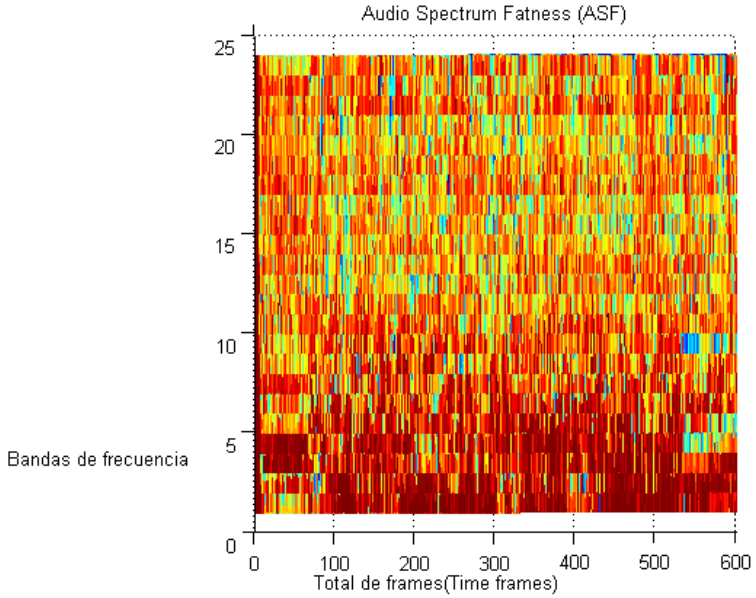


Figura 3.16. Extracción del vector ASF, donde se muestra un total de 24 bandas de frecuencia en un rango de 6 octavas donde *loEdge*=250 Hz y *hiEdge*=16 kHz. Un tono de color más claro tiende a ser un tono o un armónico, un tono de color más oscuro tiende a ser ruido, para cada banda de frecuencia.

Bibliografía

1. **KOSCH, Harald,** *Distributed Multimedia Database Technologies, supported MPEG-7 and by MPEG-21,* CRC Press, EstadosUnidos, 2004, p 260.
2. **MOREAU, Nicolas, SIKORA, Thomas,** *MPEG-7 Audio and beyond,* Wiley, Alemania, 2005, p 281.
3. **DE MALLORCA, Palma,** *Coding of moving pictures and audio,* octubre de 2004, http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm#3._Detailed_technical_description_of_the_MPEG-7_Technologies, consultada marzo de 2011.
4. **JOANNEUM RESEARCH – DIGITAL,** Instituto de Tecnologías de Información y Comunicación, http://iiss039.joanneum.at/cms/index.php?id=230, consultada marzo de 2011.

Capítulo 4. Métodos de clasificación

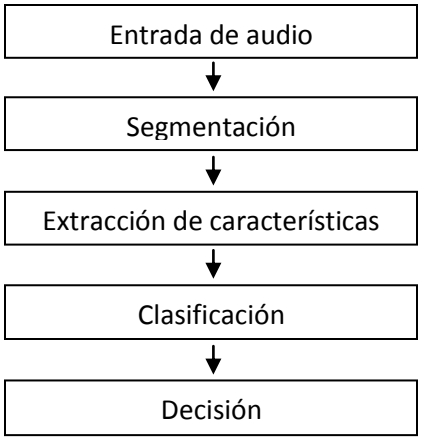
4.1. Introducción

Una vez que se han extraído los descriptores de bajo nivel bajo el estándar MPEG-7, es posible identificar ciertas características en el dominio del tiempo y de frecuencia, que mediante un método de clasificación es posible realizar un indexado de audio.

Con las características espectrales de los diferentes archivos de audio, es posible una identificación basada en contenido ya que estas características son propias de cada sonido y tienen una variación específica durante el tiempo que puede ser visto como una huella digital.

Utilizando un método de clasificación es posible calcular el nivel de similitud de un sonido o si pertenece a una cierta clase de sonido.

Generalmente un sistema de clasificación consiste de 4 etapas como se muestra en la figura:



Donde en la etapa de segmentación, la señal de audio es dividida en tramas, para después extraer sus características (descriptores vistos en el capítulo anterior) bajo el estándar MPEG-7.

El método de clasificación del estándar MPEG-7 se basa en las características espectrales de Audio Spectrum Projection (ASP), de donde se obtiene Audio Spectrum Basis (ASB) que consiste en una matriz de dimensión reducida pero que contiene las características principales de ASP, esto se realiza para fines prácticos y para reducir la cantidad de procesamiento que se tomaría con una matriz como ASP.

MPEG-7 emplea cuatro métodos de reducción para obtener el descriptor ASB: Singular Value Decomposition (SVD), Independient Component Analysis (ICA), Principal Component Analysis (PCA) y/o Discrete Cosine Transform (DCT). Comparando el algoritmo ICA con PCA, ICA ofrece una

mejor exploración de características y la precisión es mayor. [3*, "A Tonal Features Exploration Algorithm with Independent Component Analysis"*]

Hoy en día, el rápido incremento en la cantidad de bases de datos multimedia (audio), exige que los métodos de clasificación permitan ser eficientes, automáticos y que permitan la recuperación del sonido con una alta probabilidad de coincidencia.

La clasificación basada en contenido y la recuperación de audio pueden presentar dos problemas dependiendo de la selección de las características espectrales y/o temporales de los descriptores que se analizarán y el método de clasificación utilizado para estas características seleccionadas, [4, "*Classification of Audio Data using a Centroid Neural Network"*].

Existen cuatro principales métodos de clasificación, los cuales se basan en modelos estadísticos, estos métodos son:

- Gaussian Mixture Model (GMM)
- Hidden Markov Model (HMM)
- Neural Networks (NN)
- Support Vector Machines (SVM)

La selección de los vectores característicos y la selección del clasificador son críticos en el diseño del sistema de clasificación del sonido, ya que esto se verá reflejado en costos y tiempo proporcionalmente a la cantidad de información multimedia (audio) que se encuentre en la base de datos cuando dicho clasificador este funcionando.

Por lo que dos aspectos fundamentales de un sistema de clasificación son: encontrar un esquema de características extraídas adecuado y la selección de un algoritmo de clasificación eficiente, [5, "*A Study of Audio Classification on Using Different Feature Schemes with Three Classifiers"*].

En este trabajo de investigación se utilizarán 4 descriptores (ASC, ASE, ASS y los coeficientes MFCC no incluidos en el estándar) y los métodos de clasificación SVM y ANN.

4.2. Maquina de Vectores de Soporte (Support Vector Machine)

Generalmente este clasificador traslada y separa los diferentes puntos contenidos en un vector de entrenamiento a un hiperplano de forma que estos puntos se van distribuyendo linealmente de acuerdo a una clase correspondiente, por lo que genera un espacio de gran dimensión, este tipo de clasificador utiliza N*(N-1)/2 clasificadores por cada **N** géneros de audio. Este clasificador utiliza grupos de datos pequeños para su entrenamiento, ya que al incrementarse la cantidad de datos, su rendimiento y exactitud va decayendo; es un clasificador de tipo binario ya que distribuye los datos en dos clases y su clasificación se basa en un algoritmo o Kernel el cual mediante decisiones va construyendo dicho hiperplano constituido de las dos clases separadas mediante un margen

máximo, utilizando el principio de inducción de minimización de riesgo estructural (SRM), esto es minimizar el error cuadrático medio, encontrando un hiperplano óptimo.

El hiperplano óptimo debe cumplir la siguiente condición $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, para poder separar los datos, donde $\boldsymbol{w} \in R^N$, siendo $\boldsymbol{w} = \propto_i * \gamma_i * x_i$, y donde $b \in R$. La función de clasificación de un punto desconocido es definido como:

$$f(x) = sign\,(\boldsymbol{w} \cdot \boldsymbol{x} + b) = sign\left(\sum_{j=1}^{l} \alpha_j\ y_j\, \boldsymbol{x_j} \cdot \boldsymbol{x}\right) \tag{4.1}$$

Para realizar el proceso de clasificación, primero el contenido de audio es segmentado en tramas como se observa en la figura siguiente, una vez que el audio se ha segmentado en tramas, por lo general con una duración de 30 ms, se realiza el proceso de extracción de características a través de los descriptores del estándar MPEG-7, después los valores obtenidos de estos descriptores serán vistos como un punto definido de un vector característico. SVM analiza cada punto del vector característico a través de un Kernel, buscando separar estos datos en un hiperplano definido por las dos clases existentes, donde para cada trama o punto vectorial (x_j) existirá una clase (y_j), la función $f(x)$, será etiquetado con una clase -1 o +1

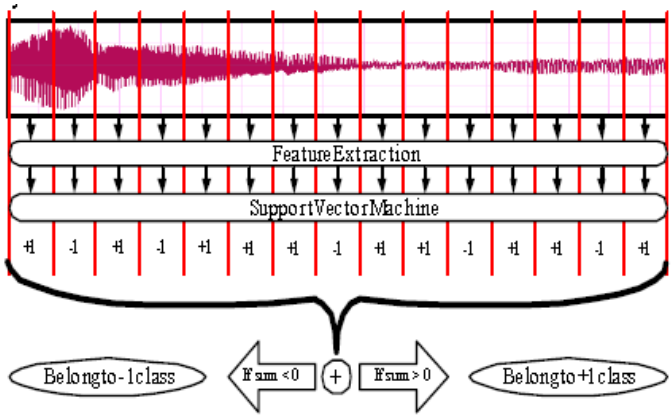


Figura 4.1. Clasificación de audio basada en tramas utilizando SVM (tomada del artículo *Environmental Sound Classification Using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor*)

Finalmente, de la ecuación siguiente, si la suma de estas etiquetas $f(x)$, es mayor que cero o a partir de un umbral definido, el archivo de sonido es clasificado como clase +1, en caso contrario será de clase -1.

$$f(x) = sign\left(\sum_{j=1}^{l} \alpha_j\ y_j\, k(x_i, x_j) + b\right) \tag{4.2}$$

De donde:

- *l= es el número de vectores de soporte*

- α_j *=es un coeficiente determinado (coeficiente de Lagrange) para un vector de soporte* (x_j, y_j)*, siendo*(x_i, y_i) *un vector o muestra aleatoria.*
- $k(x_i, x_j)$*=es la función Kernel*
- y_j *=describe la clase en que se encuentra* $\boldsymbol{x_j}$*, donde* $y_j \in \{-1, +1\}$
- b determina la distancia media que existe entre dos vectores de soporte de las clases existentes:

$$b = -\frac{1}{2}\boldsymbol{w}\,[x_i + x_i'] \tag{4.3}$$

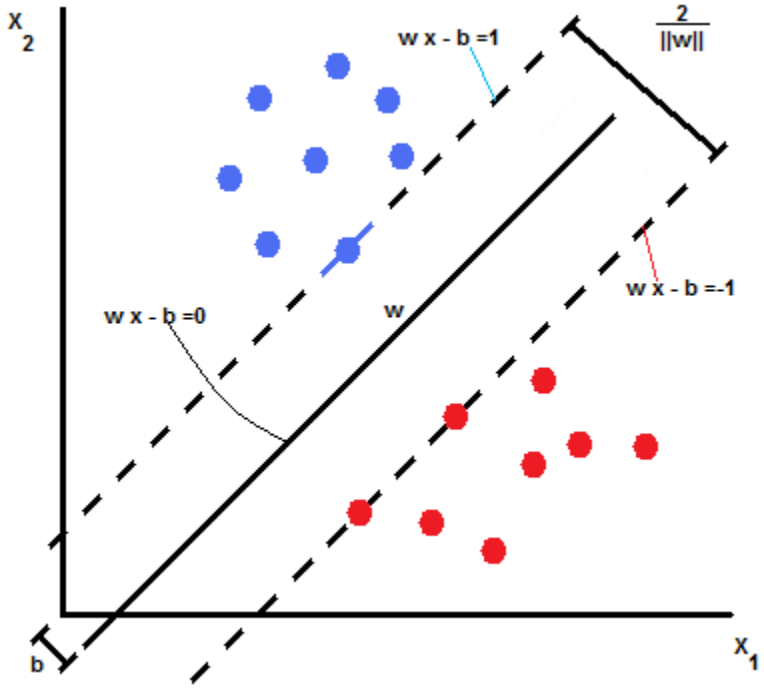


Figura 4.2. Traslación de los valores de entrenamiento en un hiperplano donde se muestran dos clases distintas.

La condición de optimización recae en encontrar el plano que maximice la distancia entre el hiperplano y el vector de soporte más cercano, esto implica que se tienen que encontrar los valores de w y b tal que se cumplan las siguientes condiciones:

$$w \cdot \boldsymbol{x} - b = 1$$
$$w \cdot \boldsymbol{x} - b = -1$$

$$\text{Ó}\ \ d(w,b) = \min(x_i | y_i = 1)\frac{wx_i+b}{|w|} - max\,(x_i | y_i = -1)\frac{wx_i+b}{|w|} = \frac{2}{|w|} \tag{4.4}$$

De acuerdo a la figura anterior, la distancia máxima será $2/|w|$, y por consiguiente, para maximizar la distancia es necesario minimizar $|w|$. Teniendo en cuenta todos los vectores de

soporte x_i que coinciden con la primera clase $w \cdot x_i - b \geq 1$ y los vectores de soporte x_i que coinciden con la segunda clase $w \cdot x_i - b \leq -1$, se cumple la condición:

$$y_i(w \cdot x_i - b) \geq 1 \text{ , para todo } 1 \leq i \leq n \tag{4.5}$$

Denotando un vector $\boldsymbol{z} = \varphi(\boldsymbol{x})$ y teniendo su espacio correspondiente, se realiza un mapeo utilizando la función φ a un espacio característico Z, donde $\varphi \in R^N$. Esta función φ utiliza los puntos del espacio vectorial de entrada, para calcular un producto punto y trasladar dichos puntos a un espacio característico Z. Dicha función es la función Kernel.

$$\boldsymbol{z_i} \cdot \boldsymbol{z_j} = \varphi(\boldsymbol{x_i}) \cdot \varphi(\boldsymbol{x_j}) = k(x_i, x_j) \tag{4.6}$$

Para que una función pueda ser aplicada como una función Kernel, esta debe de cumplir el Teorema de Mercer, algunas funciones típicas que son utilizadas como funciones Kernel son:

Lineal $k(x_i, x_j) = \boldsymbol{x_i} \cdot \boldsymbol{x_j}$
Polinomial $k(x_i, x_j) = (\gamma \cdot \boldsymbol{x_i} \cdot \boldsymbol{x_j} + c)^d$
Base radial $k(x_i, x_j) = -\gamma \cdot |\boldsymbol{x_i} - \boldsymbol{x_j}|^2$

Donde γ es la desviación estándar (δ^2), Quedando la función kernel de base radial exponencial como:

$$k(x, x_j) = \exp\left(-\frac{|x - x_j|^2}{\delta^2}\right) \tag{4.7}$$

Referencias

1. **Robert J. Elliotl, Lakhdar Aggoun, John B. Moore,** *Estimation and Control, Hidden Markov Models,* Springer, New York, Estados Unidos.
2. **Nianyi chen,** *Support Vector machine in Chemistry*, World Scientific, Singapur, 2004.
3. **Hsin-Lung Hsieh and/Din-Yuen Chan,** *A Tonal Features Exploration Algorithm with Independent Component Analysis*, Department of Information Engineering, I-Shou University, Chayi, Taiwan, artículo publicado en IEEE, artículo publicado, 2004.
4. **Dong-Chul Park**, *Classification of Audio Data using a Centroid Neural Network*, Dept. of Electronics Engineering, Myong Ji University, Yong In, KOREA artículo publicado, 2010.
5. **Van Feng, Huijing Dou, Yanzhou Qian***, A Study of Audio Classification on Using Different Feature Schemes with Three Classifiers***,** School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing, China, artículo publicado, 2010.
6. **Jia-Ching Wang, Jhing-Fa Wang, Kuok Wai He, and Cheng-Shu Hsu**, *Environmental Sound Classification Using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor***,** Department of Electrical Engineering, National Cheng Kung University, University Road, Tainan, Taiwan, 2006.
7. **CHERM,** *Applications of Support Vector Machines in Chemistry,* SVM Org, 2007, http://www.support-vector-machines.org/SVM_soft.html, consultada septiembre de 2011.
8. **NIRAJAN, Mahesan,** *Information: signals, Images, Systems***,** Organismo, http://www.isis.ecs.soton.ac.uk/resources/svminfo/, consultada agosto de 2011.
9. **MANGASARIAN, Olvi. L, MUSICANT, David R,** *Lagrarian Support Vector Machine,* Universidad Wisconsin Madyson, 2001, http://www.cs.wisc.edu/dmi/lsvm/, consultada agosto de 2011.
10. **CANU,Stephanie,** *SVM and Kernel Methods Matlab Toolbox, 2007,* http://asi.insa-rouen.fr/enseignants/~arakotom/toolbox/index.html, consultada marzo de 2011.
11. **SHAWE-TAYLOR, John, CRISTIANINI, Nello,** *Support Vector Machines and other kernel-based learning methods, Press, Universidad de Cambrigde, 2000,* http://www.support-vector.net/software.html, consultada abril de 2011.

Capítulo 5. Estrategia de reconocimiento

El proceso de reconocimiento de patrones se basa en tres enfoques posibles: estadístico, sintáctico y neural, un enfoque estadístico se basa en la medición y decisión a través de un modelo probabilístico, el sintáctico se basa en el análisis de descripciones estructurales de ciertos patrones, y el neural, es basado en el entrenamiento del sistema con un grupo de datos que han sido previamente almacenados. En el caso de reconocimiento de imágenes, se ha demostrado que tanto los enfoques sintácticos y neuronales son de mayor utilidad comparado con el modelo estadístico.

5.1. Experimentos y resultados

5.1.1. Características generales

En este proyecto de investigación se utilizó un total de 604 archivos de audio de los cuales 264 son del género hard-rock, 240 son archivos de música clásica y 100 son de música electrónica, de estos tracks se utilizaron 213 archivos de hard-rock, 100 de electrónica y 195 de clásica para hacer un modelo patrón para el período de entrenamiento y poder realizar la clasificación de géneros correspondientes.

Los archivos de sonido soportados por Matlab son de extensión *.wav con las siguientes características: frecuencia de muestreo (Fs) de 44100 Hz, codificación a 16 bits, con calidad estéreo, por lo que todos los archivos que no tienen estas características, como archivos *.mp3, son convertidos previamente para su procesamiento, para que después puedan ser utilizados en la extracción de los descriptores establecidos dentro del estándar MPEG-7 que serán utilizados en la etapas de entrenamiento y clasificación junto con la obtención de los coeficientes MFCC (Mel Frecuency Cepstral Coeficients). Para la extracción de estas características, se utilizó una muestra de 10 segundos del archivo total de audio, esta muestra contiene un total de 998 tramas, cada trama tiene una duración de 30 ms y se tiene un traslape entre cada trama de 10 ms, estos intervalos de tiempo son típicos y se obtienen mejores resultados en el procesamiento mencionado en estudios previos [6, "*Aplicación de las máquinas de soporte vectorial al reconocimiento de hablantes"*].

En el proceso de extracción de características se utilizaron los descriptores de bajo nivel de la parte 4 del estándar MPEG-7, las características y/o descriptores que se utilizaron para este proyecto son: Audio Spectrum Centroid, Audio Spectrum Spread , Audio SpectrumEnveloped y los coeficientes MFCC, que cuentan con la característica de ser coeficientes no correlacionados entre si debido a la aplicación de la Transformada Coseno Discreta, esto es de gran ayuda ya que cada

valor no depende de su vecino y se pueden manipular de tal manera que al modificar un solo coeficiente, este no afecta en el total de la reconstrucción de la señal.

5.2. Procedimiento de clasificación

De acuerdo a [8, *"A Study of Audio Classification on Using Different Feature Schemes with Three Classifiers"*], establece un modelo de clasificación basado en 5 grupos, de los cuales, dos grupos son más importantes, el primero de ellos utiliza los coeficientes MFCC y el segundo grupo utiliza los valores de los descriptores ASS y ASC, teniendo una precisión de cada grupo del 92 % y 58 % respectivamente, y al utilizar de manera combinada los valores MFCC, ASS y ASC llega hasta una precisión total del 94 %, utilizando el método de clasificación SVM, el cual presenta una mayor exactitud a comparación de los métodos de clasificación Fisher Kernel y Potencial Function, que en este mismo artículo presentan una precisión utilizando los valores de MFCC, ASS y ASC de forma combinada del 90.67 % y 74 % respectivamente.

De igual forma al observar que SVM presenta una mayor precisión en la clasificación de audio, se optó por utilizar este método de clasificación en este trabajo de investigación.

El procedimiento que se llevo a cabo consiste en la obtención de un patrón modelo para cada género de música, basado en el promedio, desviaciones estándar y varianza de los diferentes descriptores ASS y ASC, así como de los coeficientes MFCC; como medida de distancias se utilizaron las distancias Euclidianas y de Hanning; como método de clasificación de audio se utilizó Support Vector Machine y hasta cierto punto en una decisión más fina se utilizó un modelo Neuronal Network.

5.3. Coeficientes MFCC

Para la obtención de los coeficientes MFCC se llevaron a cabo los siguientes pasos:

1. Como se ha mencionado en capítulos anteriores, se realiza el entramado de la señal de audio, la cual es convolucionada por una ventana Hamming del mismo tamaño, esto se realiza debido a que al realizar la Transformada Discreta de Fourier (DFT) de un trama se evitan que aparezcan frecuencias no deseadas y exista una distorsión en la obtención del espectro, el tamaño o duración de cada trama tiene un intervalo de 30 ms y el traslape entre estas ventanas es de 10 ms.
2. Se aplicó la (DFT) a cada trama, de 2^N coeficientes, siendo 2^N mayor o igual que el número de muestras de cada trama. Por ejemplo para nuestro caso la Frecuencia de muestreo es de 44.1 kHz por lo que cada 10 ms es equivalente a 441 muestras, entonces para un total de 30 ms, el número total de muestras será de 1323, cumpliendo la condición 2^N>=1323, N tiene un valor de 11, 2^11=2048 coeficientes DFT para cada trama. Como la primera mitad del total de coeficientes DFT son los mismos valores que la

segunda mitad de coeficientes DFT, entonces sólo se toma en cuenta la primera mitad de coeficientes que contiene la DFT para el análisis espectral.

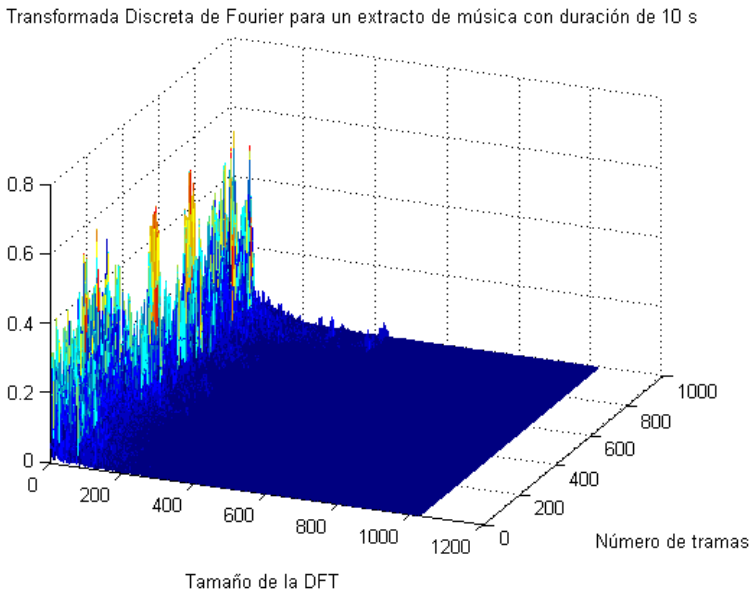


Figura 5.1. Transformada discreta de Fourier aplicada a un extracto de audio con duración de 10 ms

3. Al haber realizado la DFT, se realiza un ventaneado triangular, multiplicando punto a punto los coeficientes de la DFT por la amplitud de la ventana correspondiente de un banco de filtros, el resultado de la multiplicación de cada ventana por un número determinado de coeficientes DFT es promediado con el fin de disminuir el tamaño del vector de la DFT, obteniéndose un coeficiente C_x que formará parte de un Vector Espectral (*V*) donde *x* es el número de ventanas aplicadas, con este procedimiento también se muestra más a detalle el rango de frecuencias que contienen mayor información del espectro. Este banco de filtros es definido por la forma de los filtros, la escala de frecuencias a la que son ubicados, y el rango de frecuencias, (frecuencia inicial, frecuencia media y frecuencia final), estos filtros están ubicados de acuerdo a la escala de frecuencias Mel, obtenida a partir de la siguiente ecuación:

$$f_{MEL} = 1000 * \frac{\log\left(1+\frac{f_{lineal}}{1000}\right)}{\log(2)} \qquad (5.1)$$

Para nuestro caso se utilizó un banco de filtros de 40 ventanas triangulares, de acuerdo a la fórmula anterior se obtuvo la siguiente tabla de frecuencias MEL:

Número de ventana	Tamaño de la ventana	flineal	fMEL
	No aplica	0	0
	No aplica	631.132682	28.7235496
1	49	1068.71275	48.6383047
2	37	1403.99923	63.8975649
3	30	1675.871	76.2707511

4	25	1904.54284	86.6778608
5	21	2101.87961	95.6588767
6	19	2275.44319	103.557948
7	17	2430.35055	110.607954
8	15	2570.22596	116.973839
9	14	2697.73043	122.776709
10	14	2814.8755	128.108112
11	13	2923.21877	133.038934
12	11	3023.99095	137.625188
13	10	3118.18143	141.911901
14	10	3206.59749	145.935814
15	10	3289.90642	149.727297
16	10	3368.66615	153.31174
17	9	3443.3479	156.710589
18	8	3514.35323	159.94212
19	9	3582.02721	163.022038
20	8	3646.66846	165.963934
21	7	3708.53719	168.779648
22	8	3767.86147	171.479562
23	8	3824.84235	174.072825
24	7	3879.65796	176.567544
25	6	3932.46684	178.970935
26	7	3983.41076	181.28945
27	7	4032.61695	183.528878
28	6	4080.20007	185.694439
29	6	4126.26377	187.790849
30	6	4170.90211	189.822389
31	6	4214.20064	191.792954
32	6	4256.23746	193.706096
33	6	4297.08402	195.565068
34	6	4336.80588	197.372854
35	6	4375.46332	199.132197
36	5	4413.11189	200.845625
37	5	4449.80292	202.515475
38	6	4485.58391	204.143908
39	5	4520.49891	205.732928
40	5	4554.58885	207.284399

Tabla 5.1. dónde se muestra el número total de bancos y sus correspondientes rangos de frecuencia lineal y MEL.

De tal forma que el Banco de filtros MEL considera sólo los primeros 208 coeficientes de la DFT para la obtención de los Vectores Espectrales.

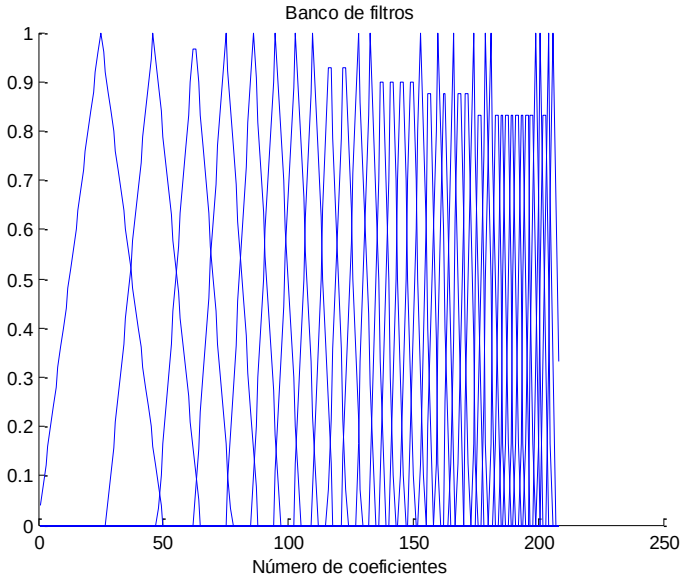


Figura 5.2. Banco de filtros con 40 ventanas triangulares, para el análisis de un espectro de 1024 puntos, este banco de filtros sólo considera los primeros 208 puntos de la DFT para la obtención de los Vectores Espectrales.

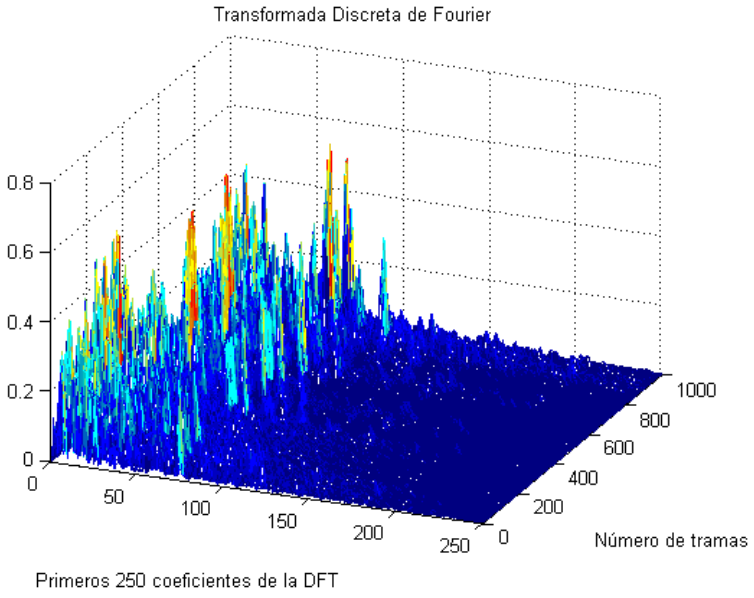


Figura 5.3. Muestra de los primeros 250 coeficientes de la DFT.

4. Una vez que se han obtenido los coeficientes $C_x´s$, se obtiene el logaritmo de cada uno de estos coeficientes y se multiplica por 20 para obtener los valores en decíbeles (dB).

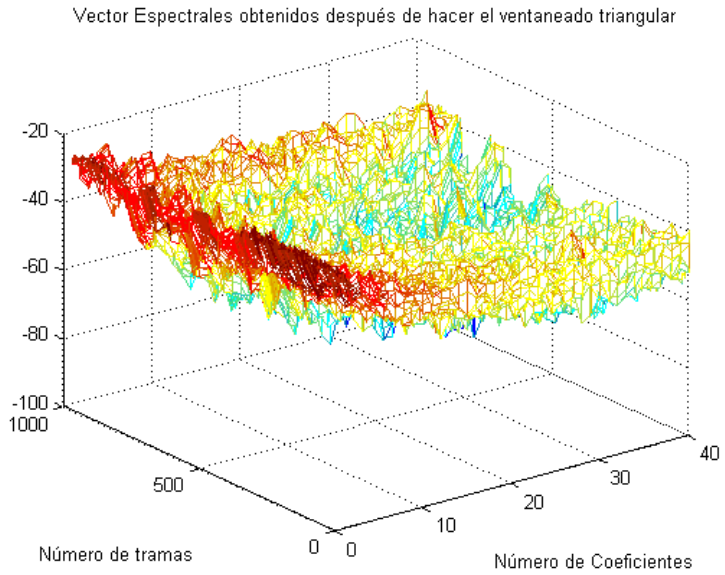


Figura 5.4. Vectores espectrales obtenidos después del ventaneado triangular a lo largo del tiempo (tramas).

5. Como último paso, se realiza la transformada Cepstral, aplicando la Transformada Coseno Discreta (DCT) a cada vector espectral, el tamaño de la Transformada Cepstral es igual al número de coeficientes $C_x´s$.

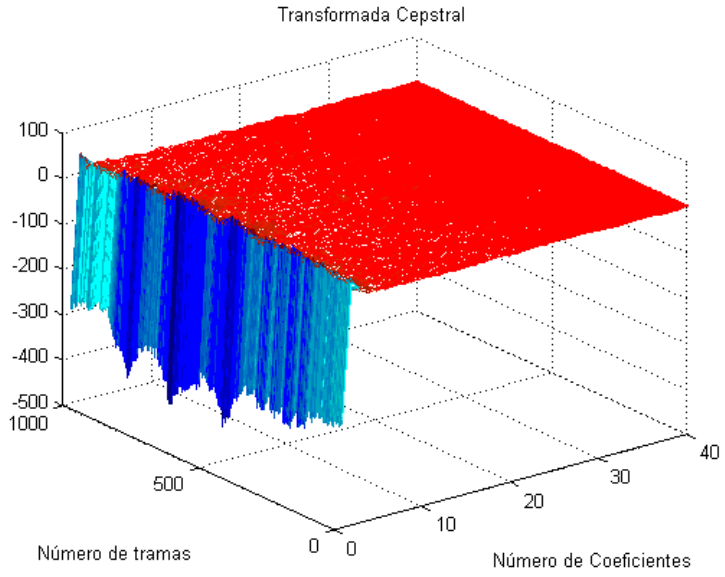


Figura 5.5. Transformada Cepstral obtenida después de aplicar la DCT a los Vectores Espectrales del análisis anterior.

5.3.1. Tarea # 1. Número de coeficientes MFCC necesarios para la reconstrucción de adecuada de los Vectores Espectrales

Al obtenerse los coeficientes MFCC los cuales son no correlacionados entre sí (Transformada Cepstral), en [8, *"A Study of Audio Classification on Using Different Feature Schemes with Three Classifiers"*] se menciona que para la fase de entrenamiento se utilizan los coeficientes 1´s, 5´s y 12´s MFCC en el SVM, por lo que los primeros 12 coeficientes son tomados en cuenta. A continuación hacemos un análisis de la importancia de cada uno de estos coeficientes para el proceso de clasificación.

Como punto número uno, se realizó la Transformada Inversa Coseno Discreta (IDCT), utilizando una rutina programada en Matlab, en la cual sólo utilizamos un cierto número de coeficientes MFCC para la reconstrucción de los vectores espectrales y realizamos una correlación para ver la similitud entre los vectores espectrales reconstruidos y los vectores espectrales considerando los 40 coeficientes. De donde se obtuvo la gráfica siguiente:

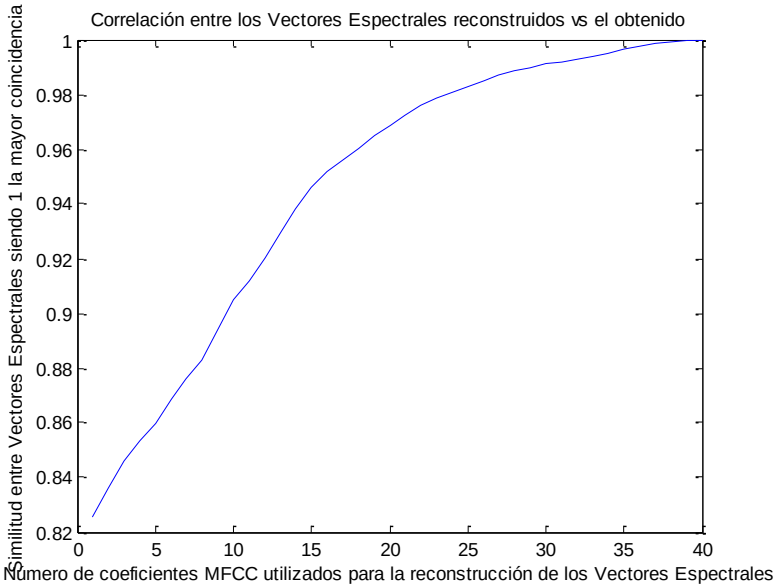
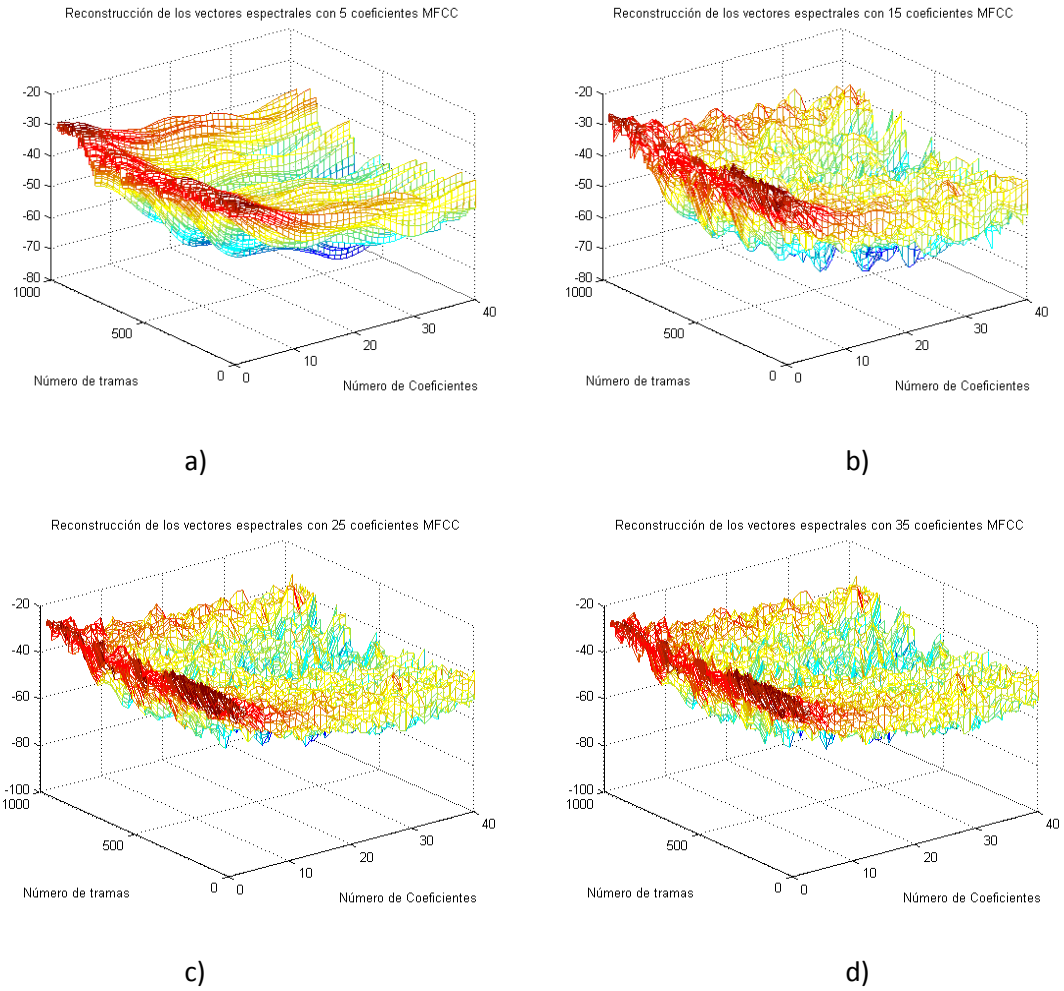


Figura 5.6. Correlación de los Vectores Espectrales reconstruidos con una cierto número de coeficientes MFCC con el vector completo de coeficientes MFCC.

Como se observa en esta gráfica, tan sólo utilizando el primer coeficiente MFCC, se recupera en un 82.55 % la matriz de los vectores espectrales, conforme se van aumentando el número de coeficientes MFCC para realizar la IDCT y obtener los Vectores Espectrales, la correlación o similitud entre los Vectores Espectrales recuperados y el obtenido aumenta casi de manera constante hasta utilizar un total de 15 coeficientes, al seguir aumentando coeficientes MFCC para la reconstrucción de los Vectores Espectrales, podemos observar que la similitud ya no aumenta de manera constante, lo que indica que los primeros 15 coeficientes contienen la mayoría de la información y permiten una recuperación óptima para la reconstrucción de la señal.

A continuación se muestra de manera gráfica la similitud entre Vectores Espectrales recuperados y el obtenido, utilizando diferentes cantidades de coeficientes MFCC.



Figuras 5.7. a), b), c) y d) Reconstrucción de los Vectores Espectrales con 5, 15 25 y 35 coeficientes MFCC

5.3.2. Tarea # 2. Obtención de los coeficientes MFCC que contienen la mayor información espectral

Como segundo paso, una vez que se ha determinado cuantos coeficientes son suficientes para una reconstrucción óptima de una señal, se identificará que coeficientes de los primeros 15 coeficientes MFCC, contienen una mayor información y cuanto afectan en la reconstrucción de la señal, una vez que son determinados estos coeficientes, estos coeficientes representativos de la señal son los que se utilizarán para nuestro clasificador SVM.

.

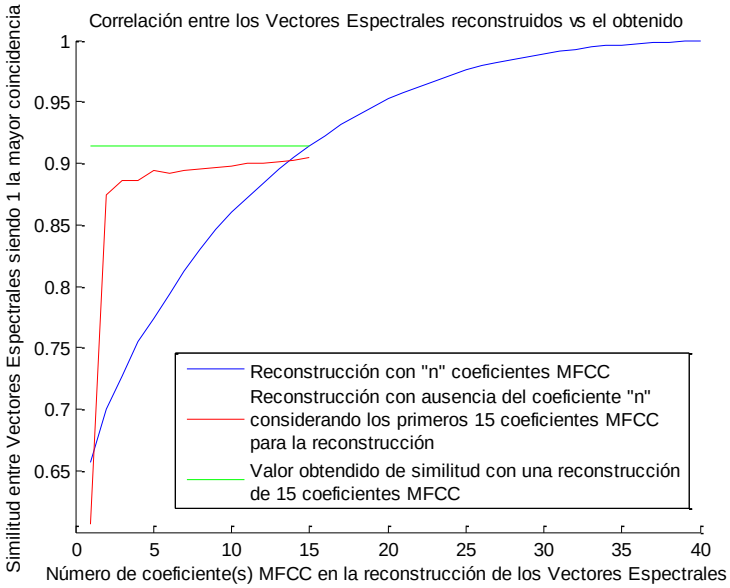


Figura 5.8. Similitud existente utilizando diferentes coeficientes MFCC para un género de música de tipo hardrock.

Como podemos observar en esta gráfica, de otra canción de tipo hardrock, la línea verde indica el valor de correlación que existe entre la reconstrucción de los Vectores Espectrales tomando en cuenta los primeros 15 coeficientes MFCC, con un valor del 91.35 % de similitud, la línea roja nos indica el valor de la correlación utilizando estos primeros 15 coeficientes pero con la ausencia de un coeficiente “n”, como podemos observar con la ausencia del primer coeficiente, el valor de la correlación decae hasta un valor de 60.68 %, también se puede observar que en la ausencia del coeficiente 4 y 6, el valor de correlación decae, por lo que a conclusión, los coeficientes 1, 4 y 6 representan una parte importante de la información para la recuperación de los Vectores Espectrales para este ejemplo, también comparando contra la línea azul, podemos observar que sólo utilizando el coeficiente 1 para la reconstrucción de estos vectores, este coeficiente contiene la mayor información que utilizando los 14 coeficientes del 2 al 15 para la reconstrucción de dichos vectores.

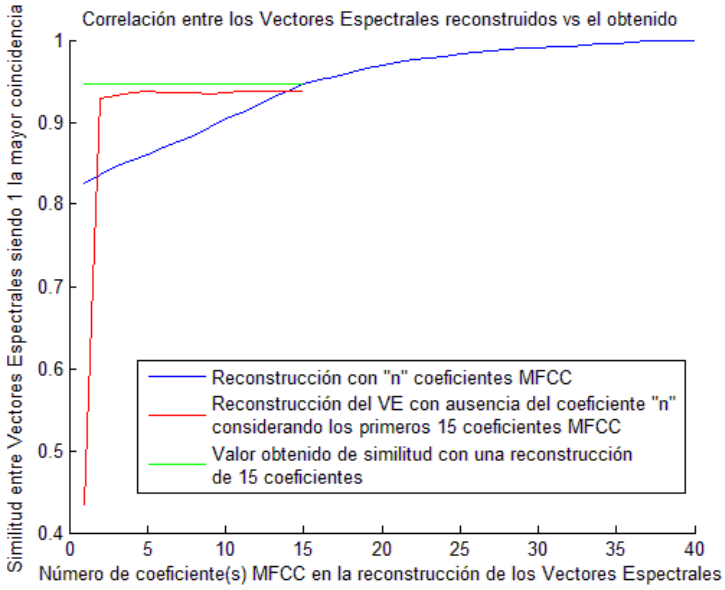


Figura 5.9. Similitud existente utilizando diferentes coeficientes MFCC para un género de música de tipo clásico.

Como podemos observar en esta gráfica, de una música tipo clásica, podemos observar cuanto es lo que afecta la ausencia de un cierto coeficiente MFCC en la reconstrucción de los VE, por ejemplo, sin tener en cuenta el primer coeficiente MFCC, podemos observar que utilizando sólo los 14 coeficientes restantes, se pueden recuperar los VE en un 43%, mientras que utilizando el coeficiente 1 y los coeficientes del 3 al 15, la recuperación de los VE sube hasta un 92.7 %, también se puede notar que en ausencia del coeficiente 9 y 10 el porcentaje de correlación vuelve a decaer

5.3.3. Tarea # 3. Coeficientes MFCC importantes de la BD

Ahora que se han obtenido los valores anteriores para el total de las canciones de la base de datos, se hace un promedio de todos estos valores, obteniéndose la siguiente gráfica:

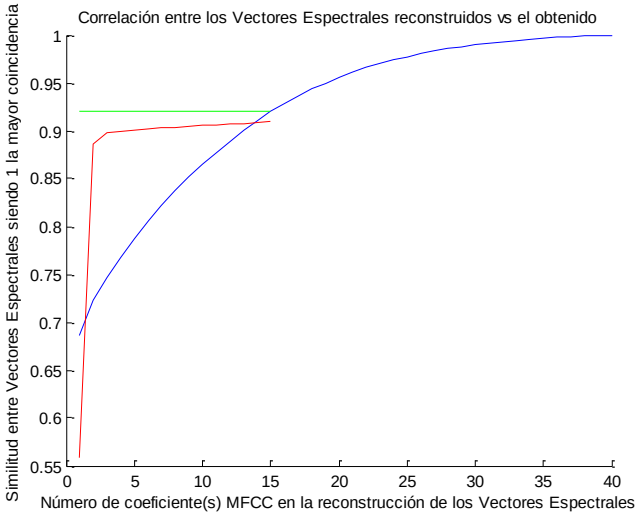


Figura 5.10. Promedio de la similitud existente utilizando diferentes coeficientes MFCC aplicada a toda la base de datos.

Maximizando el área de interés, podemos observar que, como se ha mencionado anteriormente el coeficiente 1 es primordial en la recuperación de la señal, también podemos observar que al eliminar el coeficiente 4 o 5, el valor de la correlación se mantiene en un 90% y al eliminar el coeficiente 6, podemos observar que dicho valor de correlación sube a un 90.2% por lo que se puede concluir que el coeficiente 5 tiene información relativamente representativa en la reconstrucción de los VE, de igual forma que con el coeficiente 11, como se observa en la siguiente figura:

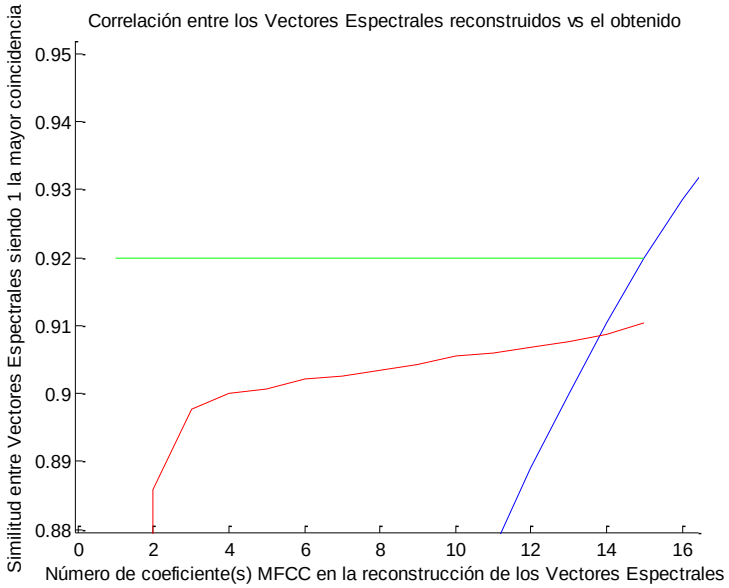


Figura 5.11. Expansión del área de interés de la figura anterior.

Como tercer paso, una vez que se han identificado los coeficientes que representan la mayor información de los Vectores Espectrales que en este caso son los coeficientes 1´s, 5´s y 11´s, se procede a la comparación de ellos de acuerdo al género de música. Para esto se seleccionan sólo los coeficientes MFCC 1´s y 5´s de cada una de las tramas de una canción.

Como podemos observar en la siguientes tres figuras, existe una cierta similitud en los valores de los coeficientes MFCC 1´s, de acuerdo al género de música, como primer figura se muestran los coeficientes para el género de música Hardrock, la segunda figura muestra el género Clásico y la tercera figura muestra el género Electrónica.

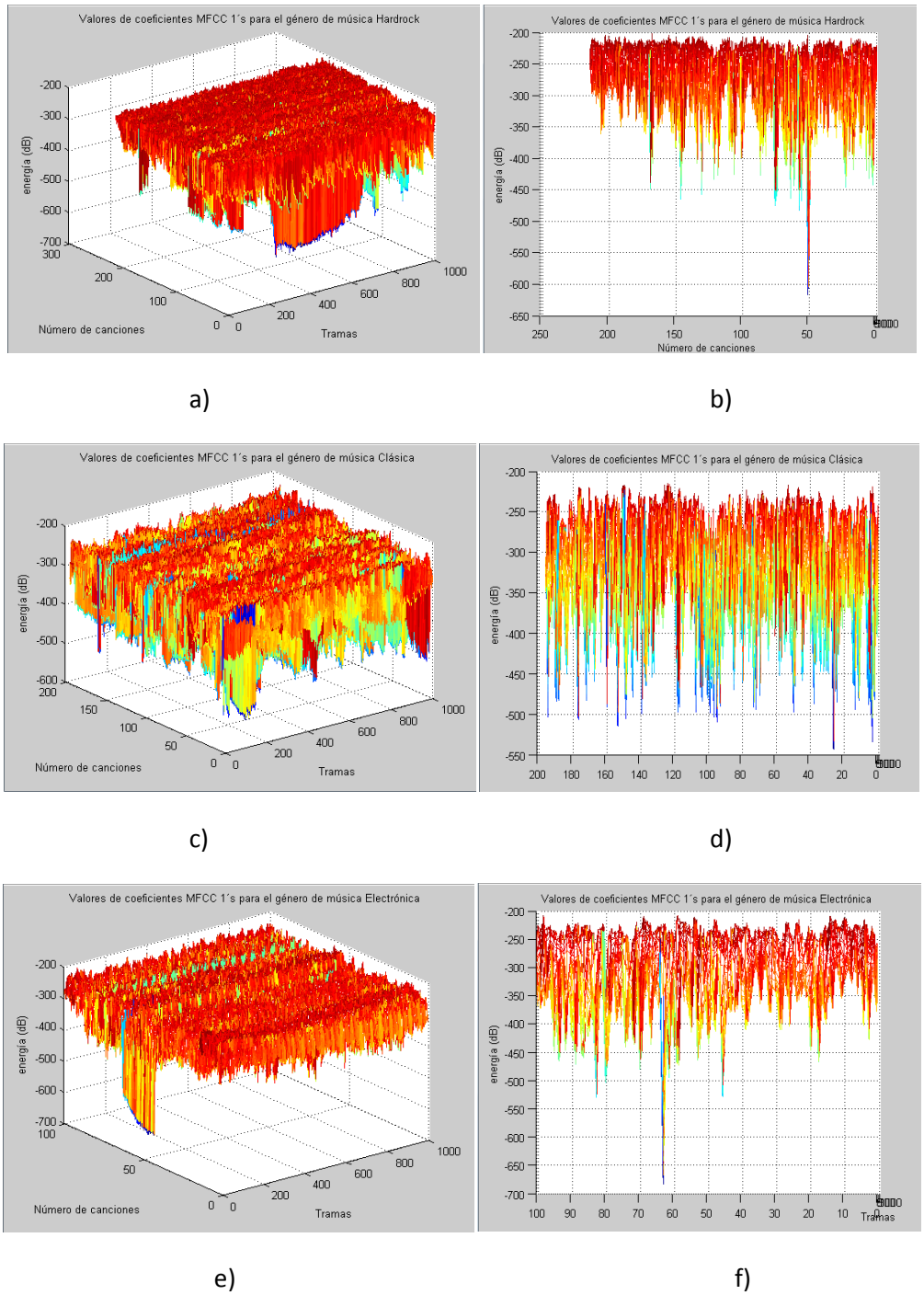


Figura 5.12. Matrices generadas con los coeficientes MFCC 1´s, a) Hardrock, c) Clásica y e) Electrónica. Energía contenida de los coeficientes MFCC 1´s b) Hardrock, d) Clásica y f) Electrónica.

Referencias

1. **Lonnie C. Ludeman Hoboken**, *Random Processes, filtering, estimation and detection,* Wiley-Interscience, New Jersey, Estados Unidos , 2003.
2. **Nianyi chen,** *Support Vector machine in Chemistry,* World Scientific. Singapur, 2004.
3. **Thomas P. Barnwell III, Monson Hayes,** *Digital Filtering,* Wiley, Estados Unidos, 1993.
4. **Muhammad Sarfraz,** *Intelligent Recognition*, John Wiley & Sons, England, 2005.
5. **Willi-Hans Steeb**¸ *Mathematical Tools in Signal processing with C++ & Javasimulations,* World Scientific, University of Johannesburg, South Africa, 2005.
6. **Juan Gabriel Pedroza Bernal**, *Aplicación de las máquinas de soporte vectorial al reconocimiento de hablantes*, tesis en maestría en ingeniería, Universidad Autónoma Metropolitana, México D.F. 2007.
7. **Enrique Prieto Labrador***, Estudio comparativo de parámetros espectrales para clasificación de audio,* tesis en ingeniería, Universidad Carlos III de Madrid, Madrid, 2008.
8. **Van Feng, Huijing Dou, Yanzhou Qian,** *A Study of Audio Classification on Using Different Feature Schemes with Three Classifiers*, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing, China, artículo publicado, 2010.

Capítulo 6. Entrenamiento SVM

6.1. Obtención de los vectores de entrenamiento

Una vez que se han detectado cuales son las diferencias de energía de dichos coeficientes, se procede a obtener una media de cada una de las tramas por cada género de música, esta media es la que se utilizara para los vectores de entrenamiento del SVM. Una Máquina Vectorial utiliza dos vectores de entrenamiento. En la siguiente gráfica se muestran los promedios de los coeficientes 1´s y 5´S para los tres géneros de música.

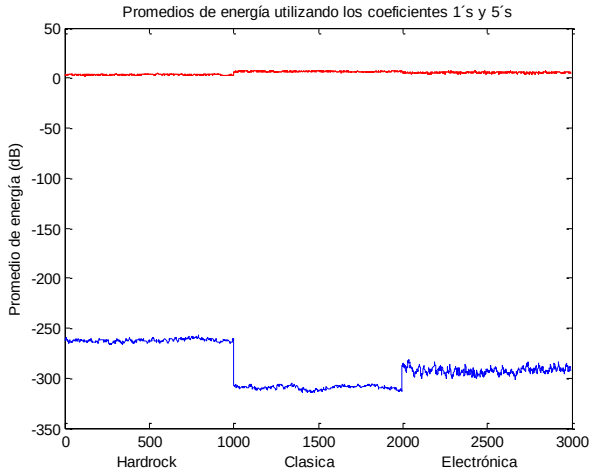
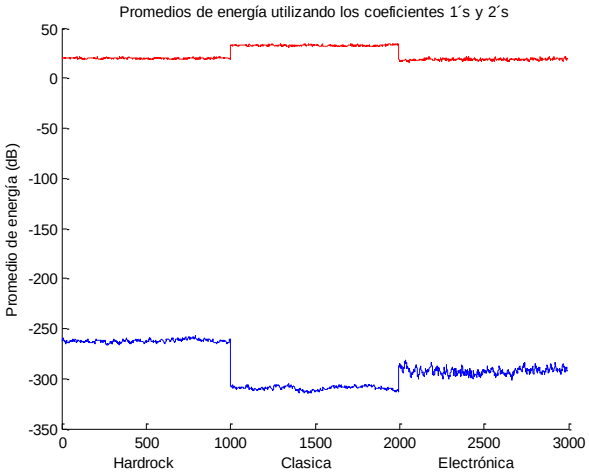


Figura 6.1. Promedio de los coeficientes MFCC para cada género de música, los coeficientes MFCC 1´s están representados de color azul, los coeficientes 5´s están representados de rojo.

Como podemos observar, la distancia existente entre los promedios de los coeficientes 1´s es notoria mientras que no existe una gran diferencia para los promedios de los coeficientes 5´s, el objetivo es encontrar entre las distintas clases una distancia que permita diferenciar o clasificar lo mejor posible a un archivo de audio, por lo que se busca encontrar vectores de entrenamiento que contengan valores promedio que puedan tener la mayor distancia posible entre los diferentes géneros de música, que permitirá disminuir la probabilidad de error en la etapa de clasificación.

Figura 6.2. En esta imagen se representan los promedios de los coeficientes MFCC 1´s (línea azul) y 2´s (línea roja) de los distintos géneros de música, en donde se puede observar una clara diferencia entre los valores promedio de cada género de música.



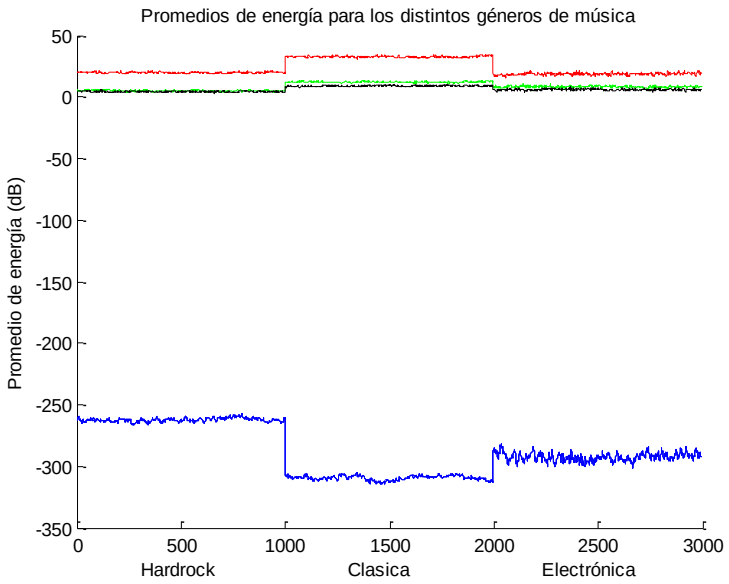


Figura 6.3. Al realizar una subrutina, se busco encontrar aquellos vectores que mostraran en sus valores una mayor diferencia entre los distintos géneros

En la figura anterior se muestran los valores promedio de los diferentes géneros de música para los primeros 4 coeficientes MFCC. La línea de color azul representa las medias de los coeficientes 1 de los tres géneros de música mientras que la línea de rojo representa la media para los coeficientes 2, de verde para los coeficientes 3 y de negro para los coeficientes 4.

Como conclusión podemos ver que entre mayor sea el coeficiente, menor será la distancia y esto ocasionará que aumente la probabilidad de error a la hora de clasificar un track. Por lo que, finalmente se seleccionaron los coeficientes 1´s y 2´s para los vectores de entrenamiento y clasificación de audio y para la primera etapa de entrenamiento y clasificación del SVM, se seleccionaron estos coeficientes por grupos: Hardrock y Clásica, Hardrock y Electrónica, y Clásica y electrónica.

6.2. Fase de entrenamiento

Como conclusión en la fase de entrenamiento se crearon y seleccionaron 2 estructuras SVM, que serán utilizadas para la clasificación de audio, previamente se hicieron una serie de pruebas y análisis para seleccionar la estructura o estructuras SVMs que nos permitieran clasificar lo mejor posible estos tres tipos de géneros. Donde finalmente el diagrama de flujo para la etapa de entrenamiento es el siguiente:

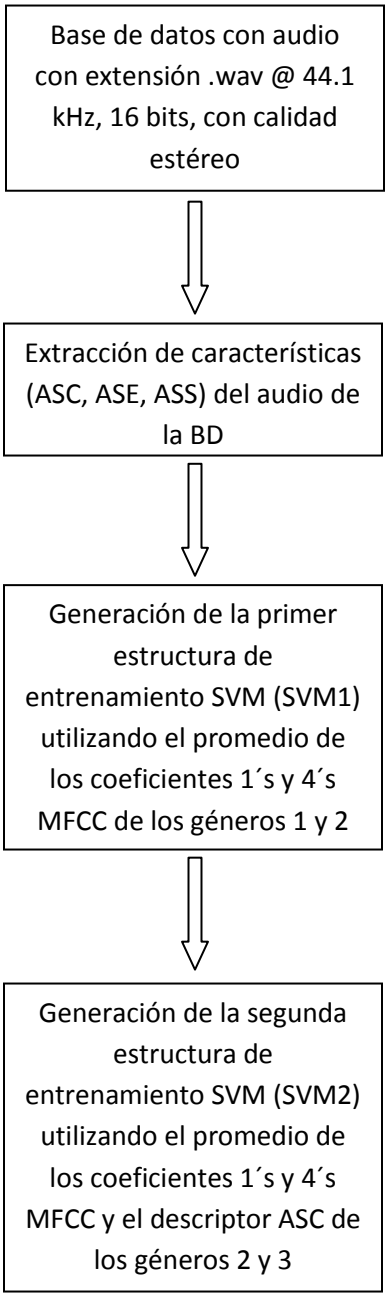


Diagrama a bloques del proceso de entrenamiento y obtención de las estructuras SVMs para la clasificación de audio.

6.2.1. Prueba 1. Entrenamiento con promedio de los coeficientes MFCC 1´s y 2´s de cada género de música

Como primera etapa de prueba de entrenamiento, y como se mencionó anteriormente, se utilizará el SVM con vectores de entrenamiento de sólo dos géneros de música (por ejemplo, Hardrock y Clásica), y primeramente se utilizarán estos dos géneros, ya que dichos coeficientes tienen una distancia mayor entre sí. Al realizar nuestra fase de entrenamiento con estos dos géneros, se genera el siguiente mapeo de valores en el hiperplano del SVM:

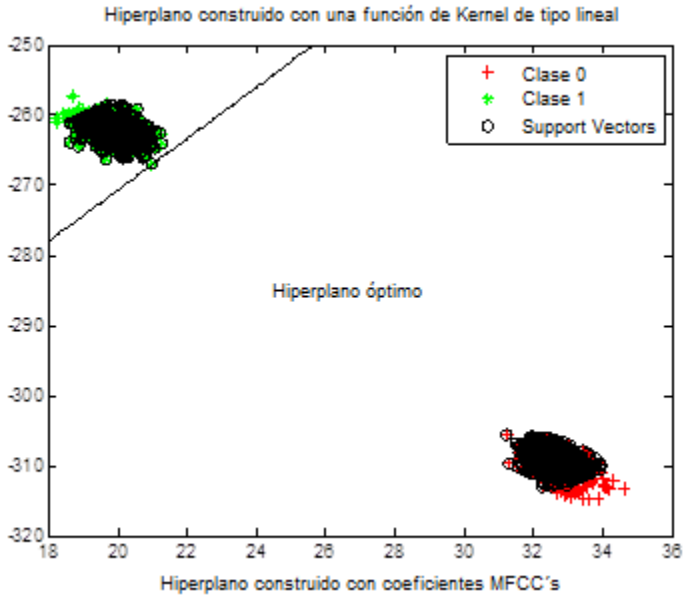


Figura 6.4. Estructura generada con valores de los coeficientes MFCC 1´s y 2´s de los géneros de música hardrock y clásica

Como podemos observar en el Espacio Vectorial anterior, se distinguen los dos géneros de música, donde los puntos rojos representan los valores del género de música Clásico y los puntos verdes al género de música Hardrock.

En la siguiente figura, se realizó una estructura SVM utilizando las medias de los coeficientes MFCC 1´s y 2´s de los géneros de música hardrock y electrónica.

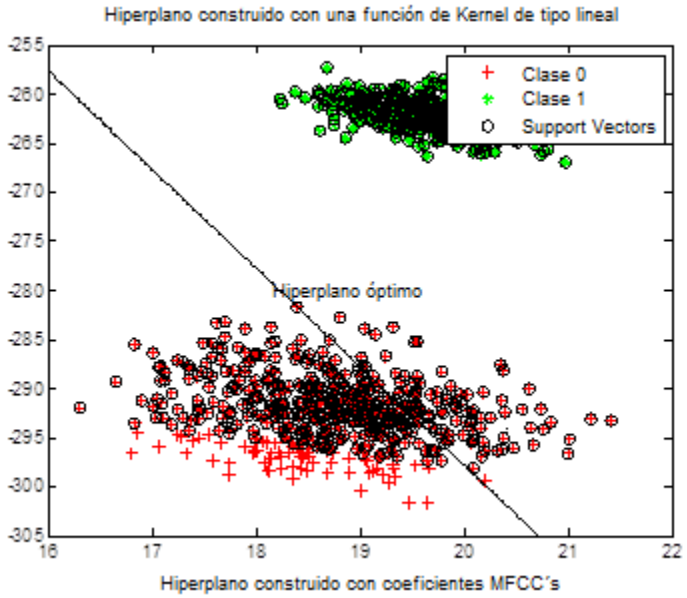


Figura 6.5. Estructura generada con valores de los coeficientes MFCC 1´s y 2´s de los géneros de música hardrock (puntos verdes) y electrónica (puntos rojos)

Y como prueba final se utilizó para la fase de entrenamiento los coeficientes de los géneros de música clásico y electrónico, generándose el siguiente espacio vectorial.

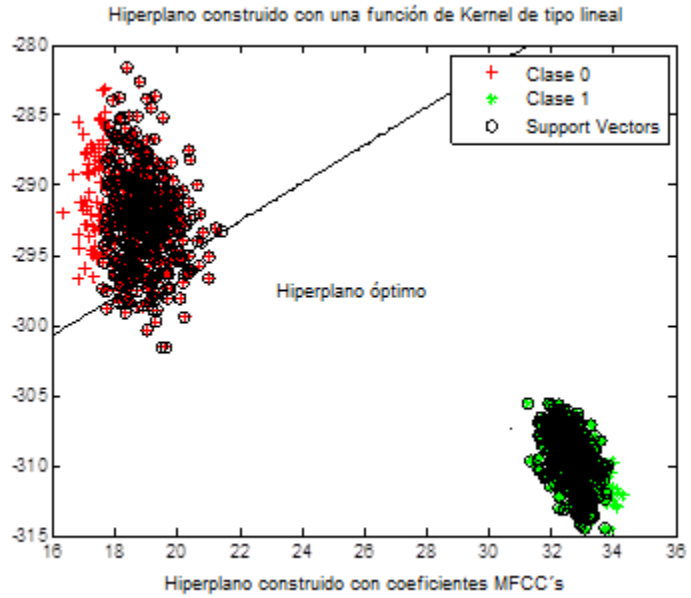


Figura 6.6. Estructura generada con valores de los coeficientes MFCC 1´s y 2´s de los géneros de música clásica y electrónica

Como conclusión se puede observar que agrupando los géneros de música en pares para la fase de entrenamiento, es posible generar los vectores de soporte, que identificarán si un valor en la etapa de clasificación pertenece a uno u otro género de música.

6.2.2. Prueba 2. Entrenamiento con coeficientes MFCC "puros"

Como prueba 2, se llevará a cabo el mismo procedimiento que en la prueba 1, lo único que cambiará es la forma de estructurar los valores de los vectores que serán ingresados al entrenamiento del SVM.

En este caso los valores contenidos en ambos vectores del SVM, serán los mismos coeficientes 1´s y 2´s escogidos de manera aleatoria de cada una de las tramas de cada una de las canciones de las dos clases de géneros que serán ingresadas en la fase de entrenamiento, con lo que se formará dos macrovectores para dicha fase. Los dos macrovectores obtenidos para esta fase de entrenamiento del SVM son:

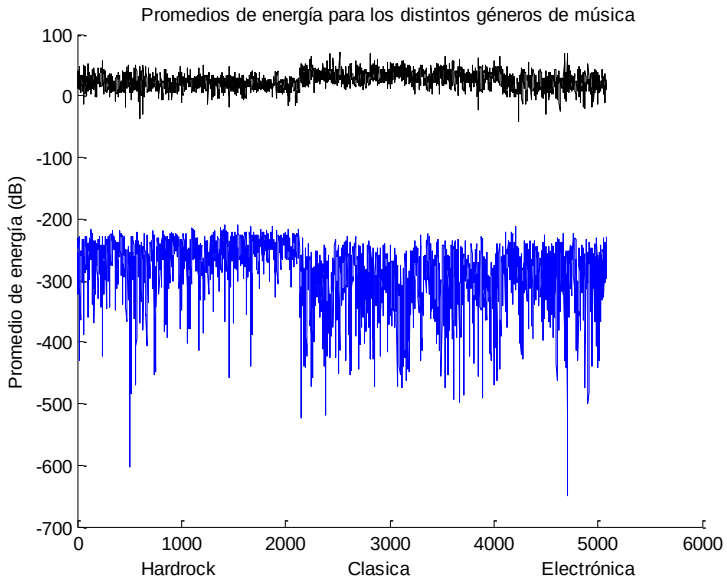


Figura 6.7. Vectores de entrenamiento

En esta figura podemos observar que prácticamente no existe una distancia que pueda diferenciar a los géneros de música clásico y electrónico, y que la distancia de estos dos géneros comparando con el género hardrock no es lo suficientemente diferenciable, por lo que el mapeo de los valores de los vectores en el espacio vectorial estarán traslapados y no se podrán diferenciar las dos clases, por consiguiente no se pueden obtener los vectores de soporte en la siguiente estructura SVM:

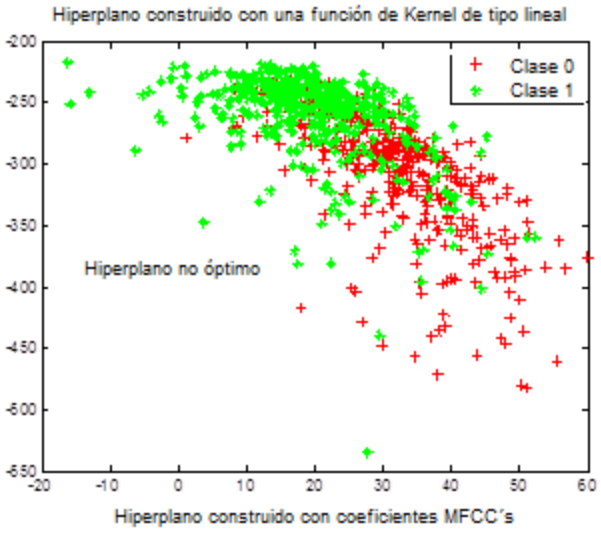


Figura 6.8. Estructura generada con valores de los coeficientes MFCC 1´s y 2´s seleccionados aleatoriamente de los géneros de música hardrock y clásica.

Como se puede observar en la figura anterior, los valores de los coeficientes contenidos en los dos macrovectores de entrenamiento hacen que al ser trasladados a un espacio vectorial, estos valores se traslapen impidiendo establecer una distancia entre estos dos grupos, con lo cual no se pueden generar los vectores de soporte y por consiguiente no se puede realizar una clasificación.

6.2.3. Prueba 3. Entrenamiento con coeficientes MFCC en combinación con el descriptor ASC

En esta prueba, se busca maximizar la distancia que existe entre los tres géneros de música, de tal forma que se procedió a utilizar una combinación de los valores del descriptor ASC con los coeficientes MFCC. Creándose los siguientes vectores de entrenamiento:

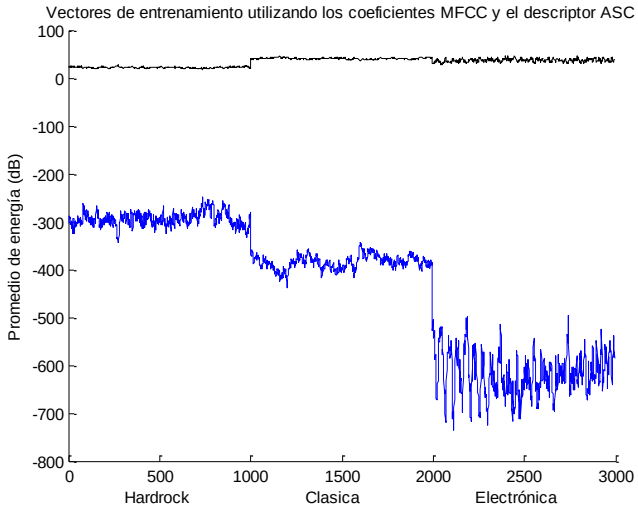


Figura 6.9. Vectores de entrenamiento utilizando una combinación entre los valores de los coeficientes MFCC y los valores de los descriptores ASC.

Como se puede observar en la figura anterior, al aplicar una combinación de los coeficientes MFCC y el descriptor ASC, se puede observar que las distancias entre los distintos géneros de música varían de tal forma que estas se pueden diferenciar. Y al aplicar una fase de entrenamiento SVM se puede mejorar la clasificación.

Como primera fase de esta prueba, se construyó los vectores de entrenamiento utilizando los valores de los géneros Hardrock y Clásica, obteniéndose el siguiente Espacio Vectorial:

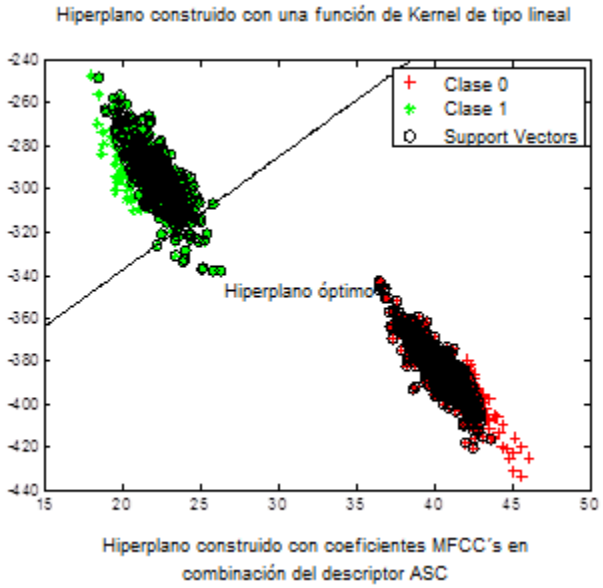


Figura 6.10. Estructura generada con valores de los coeficientes MFCC 1´s y 2´s combinados con los valores del descriptor ASC, de los géneros de música hardrock y clásica.

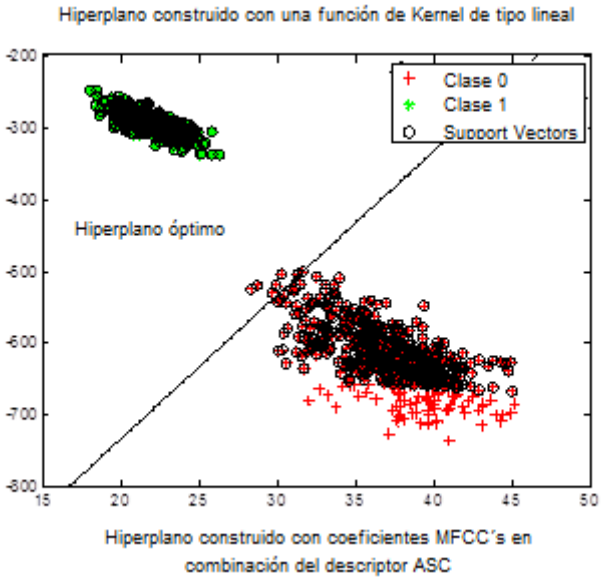


Figura 6.11. Estructura generada con valores de los coeficientes MFCC 1´s y 2´s combinados con los valores del descriptor ASC, de los géneros de música hardrock y electrónica.

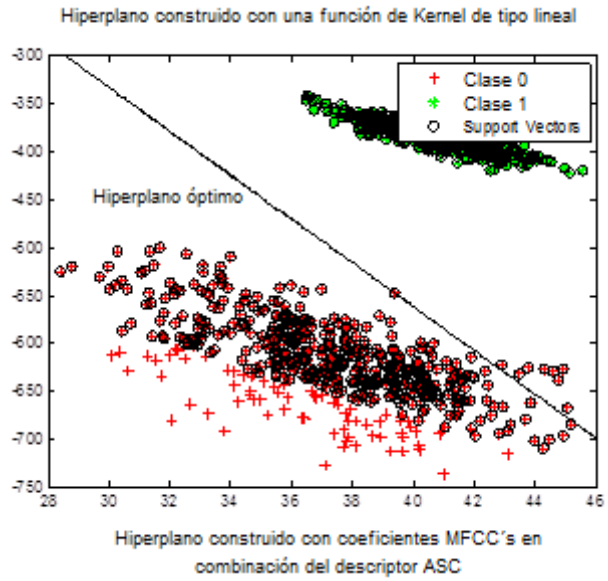


Figura 6.12. Estructura generada con valores de los coeficientes MFCC 1´s y 2´s combinados con los valores del descriptor ASC, de los géneros de música clásica y electrónica.

6.2.4. Prueba 4. Entrenamiento con coeficientes MFCC en combinación con el descriptor ASC y ASS

En esta prueba, al igual que en la prueba anterior, se busca maximizar la distancia que existe entre los tres géneros de música, de tal forma que se procedió a utilizar una combinación de los valores de los descriptores ASC y ASS con los coeficientes MFCC. Creándose los siguientes vectores de entrenamiento:

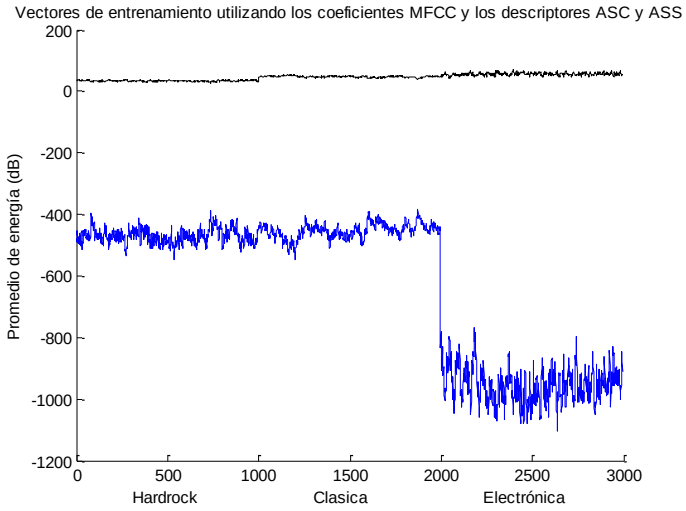


Figura 6.13. Vectores de entrenamiento utilizando una combinación entre los valores de los coeficientes MFCC y los valores de los descriptores ASC y ASS.

Como se puede observar en la figura anterior, al aplicar una combinación de los coeficientes MFCC y los descriptores ASC y ASS, se puede observar que las distancias entre los géneros de música Hardrock y Clasica, son bastante similares, en cambio la distancia del género electrónico varía de tal forma que este género se puede diferenciar de los dos restantes.

Como primera fase de esta prueba, se construyó los vectores de entrenamiento utilizando los valores de los géneros Hardrock y Clásica, obteniéndose el siguiente Espacio Vectorial:

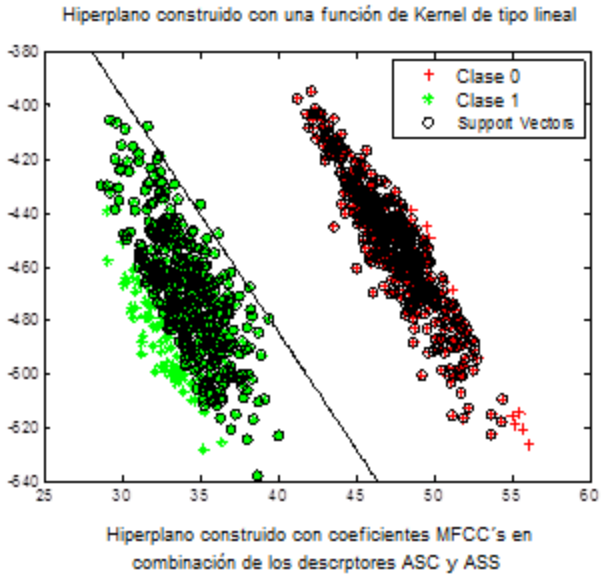
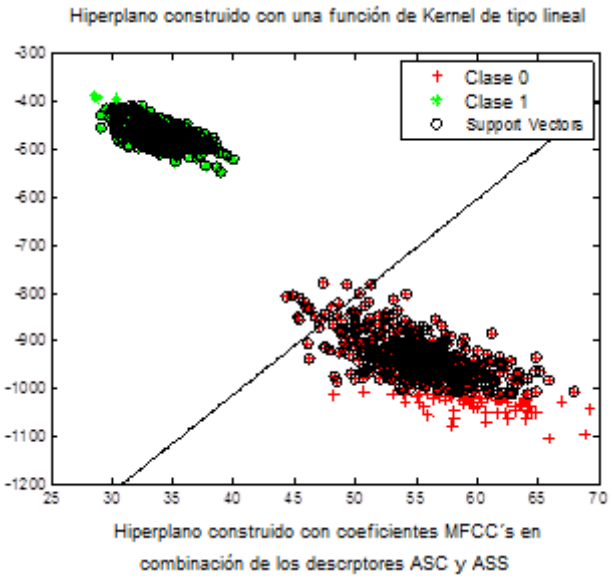


Figura 6.14. Estructura generada con valores de los coeficientes MFCC 1´s y 2´s combinados con los valores de los descriptores ASC y ASS, de los géneros de música hardrock y clásica.

Figura 6.15. Estructura generada con valores de los coeficientes MFCC 1´s y 2´s combinados con los valores de los descriptores ASC y ASS, de los géneros de música hardrock y electrónica.



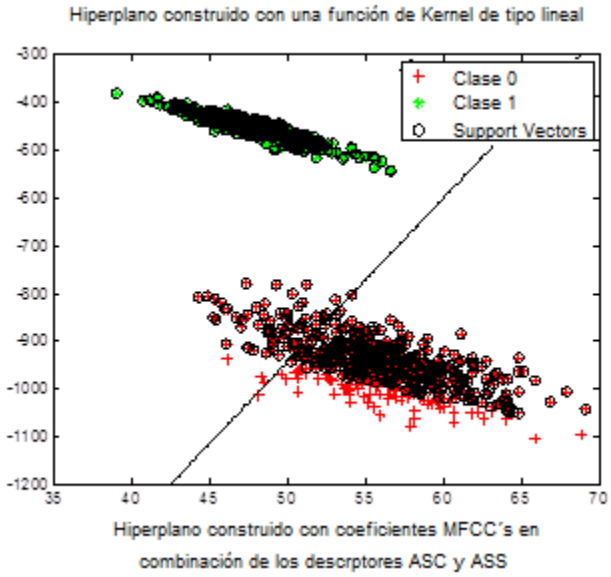


Figura 6.16. Estructura generada con valores de los coeficientes MFCC 1´s y 2´s combinados con los valores de los descriptores ASC y ASS, de los géneros de música clásica y electrónica.

6.2.5. Prueba 5. Entrenamiento con los coeficientes MFCC en combinación con el descriptor ASS

En esta prueba, al igual que en la prueba anterior, se busca maximizar la distancia que existe entre los tres géneros de música, de tal forma que se procedió a utilizar una combinación de los valores del descriptor ASS con los coeficientes MFCC. Creándose los siguientes vectores de entrenamiento:

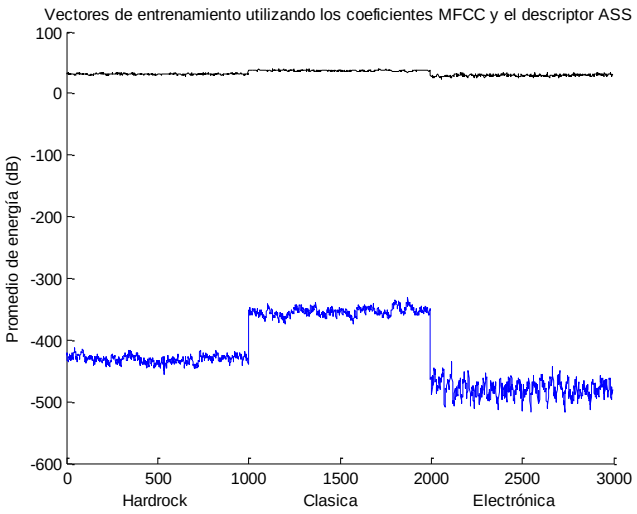


Figura 6.17. Vectores de entrenamiento utilizando una combinación entre los valores de los coeficientes MFCC y los valores del descriptor ASS.

Como se puede observar en la figura anterior, al aplicar una combinación de los coeficientes MFCC y el descriptor ASS, se puede observar que las distancias entre los géneros de música Hardrock y Electrónica, son bastante similares, en cambio la distancia del género clásico varía de tal forma que este género se puede diferenciar de los dos restantes.

Como primera fase de esta prueba, se construyó los vectores de entrenamiento utilizando los valores de los géneros Hardrock y Clásica de la misma forma que en las pruebas anteriores, en donde se obtuvo el siguiente hiperplano:

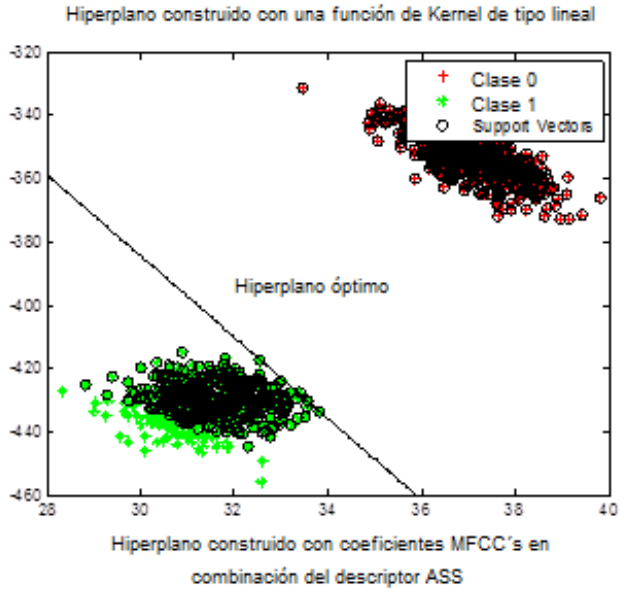
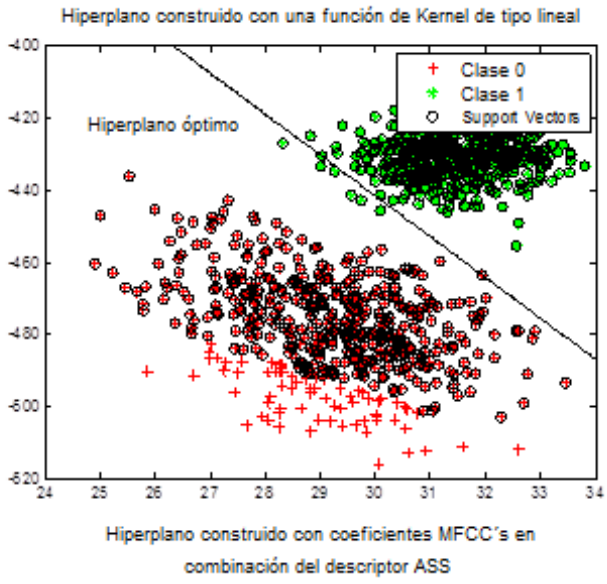


Figura 6.18. Estructura generada con valores de los coeficientes MFCC 1´s y 2´s combinados con los valores del descriptor ASS, de los géneros de música hardrock y clásica.

Figura 6.19. Estructura generada con valores de los coeficientes MFCC 1´s y 2´s combinados con los valores del descriptor ASS, de los géneros de música hardrock y electrónica.



Como podemos observar en esta figura los valores se encuentran dispersos y estos dos grupos tienen una distancia menor, ya que como se observa en el vector de entrenamiento, sus promedios casi tienen valores similares.

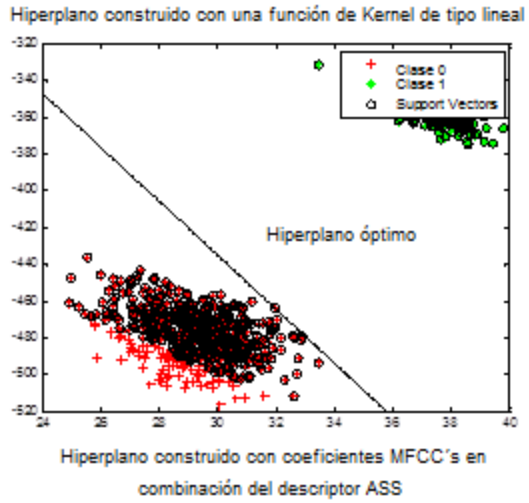


Figura 6.20. Estructura generada con valores de los coeficientes MFCC 1´s y 2´s combinados con los valores del descriptor ASS, de los géneros de música clásica y electrónica.

Como se ha observado en las pruebas anteriores, en las que se han utilizado los coeficientes MFCC 1´s y 2´s, con alguna combinación con los descriptores ASC y/o ASS, el vector de entrenamiento para el caso de los coeficientes MFCC 2´s disminuye de tal manera que no se puede apreciar una diferencia favorable entre cada uno de los géneros para poder clasificar el audio y obtener una mayor eficiencia a la hora de clasificar un archivo de audio, por lo que en las siguientes pruebas lo que se busca es sólo utilizar los coeficientes MFCC 1´s para generar ambos vectores de entrenamiento con una combinación con el descriptor ASC y/o el descriptor ASS de tal manera que sean distintos entre ellos, esto es sólo utilizar los vectores de color azul de las diferentes pruebas anteriormente realizadas para ser ingresados en los dos vectores que son requeridos en la fase de entrenamiento.

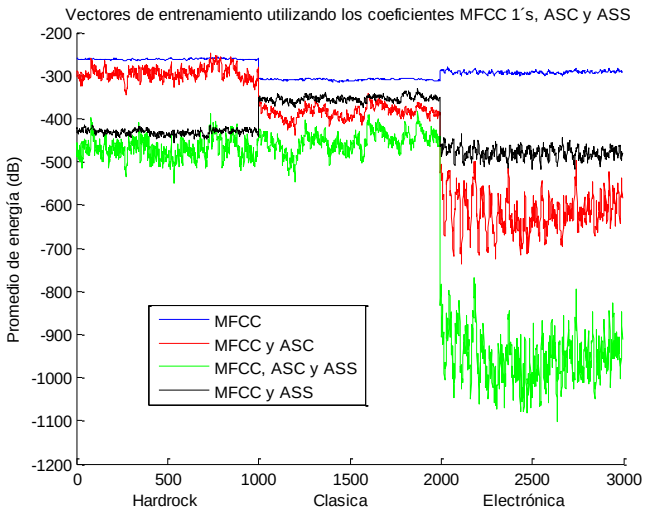


Figura 6.21. Vectores de entrenamiento

6.2.6. *Prueba 6. Entrenamiento utilizando distintas distancias Euclidianas utilizando diferentes descriptores*

Analizando los vectores de entrenamiento de la figura anterior, se realizarán pruebas con los vectores que muestren una distancia significativa de tal manera que se puedan diferenciar cada uno de los géneros, esto es, se realizarán pruebas con los vectores de entrenamiento que tienen una combinación con algún descriptor.

Donde en la prueba 6 se utilizaron los vectores de entrenamiento (MFCC y ASC) y (MFCC y ASS)

6.2.7. Prueba 7. Obtención de las características técnicas necesarias que se deben tener en el dispositivo móvil, limitación de formatos de audio.

En esta prueba lo que se busca es encontrar cuales son las limitaciones técnicas que se presentan en los dispositivos móviles, por lo cual esta prueba consiste en grabar y /o convertir la base de datos multimedia de extensión *.wav @ 44.1kHz a extensión *.amr @ 8 kHz, que es el formato que normalmente tienen los equipos celulares para la grabación de sonidos y a extensión *.mp3 @ 44.1kHz y 8 kHz que es un formato muy utilizado y es soportado por la mayoría de los dispositivos; una vez que se grabaron y/o convirtieron los archivos multimedia a extensión *.amr y *.mp3, se procede a volverlos a convertir a *.wav para poder ser procesados en Matlab, estas conversiones se hacen ya que se pierde cierta calidad en la señal de audio, adicionándose ruido, que de alguna manera estará simulando la grabación de un audio por parte del usuario en su dispositivo móvil.

6.2.8. Prueba 8. Simulación en caso de que sólo se logren grabar 5 segundos.

Simulando que el usuario sólo logre grabar un extracto de audio en un tiempo de 5 s y utilizando la música convertida a los diferentes formatos mencionados en la prueba anterior y de acuerdo a los resultados obtenidos, se procederá a completar el vector de clasificación de 10 s, por lo que se repetirá el audio dos veces.

6.3. Fase de clasificación

Primer paso para la clasificación de un archivo de audio

Al haber generado las estructuras de entrenamiento, se procede a sustituir los valores de los vectores de entrenamiento por los respectivos valores de coeficientes MFCC, o la combinación de estos con los descriptores ASC y ASS del archivo de audio que será clasificado, en la etapa de clasificación, SVM genera un vector llamado classes, que es de tamaño igual al número de muestras ingresadas en la fase de entrenamiento y clasificación, este vector nos indica a que clase pertenece cada muestra ingresada de un archivo de audio. Por lo tanto en la fase de entrenamiento se establece un umbral, del cual se decide si con “n” número de muestras etiquetadas con cierta clase pertenecen a un grupo o al otro.

El primer género será etiquetado como clase 1 y el segundo género como clase 0, entonces si un track es definido que es de clase 1, el proceso de clasificación será concluido como se observa en el diagrama de flujo, en el caso de que el track haya sido definido como clase 0, se procederá a utilizar la segunda estructura SVM, que permitirá clasificar este archivo de audio, en cualquiera de los dos géneros restantes. Por lo que el diagrama de flujo es el siguiente:

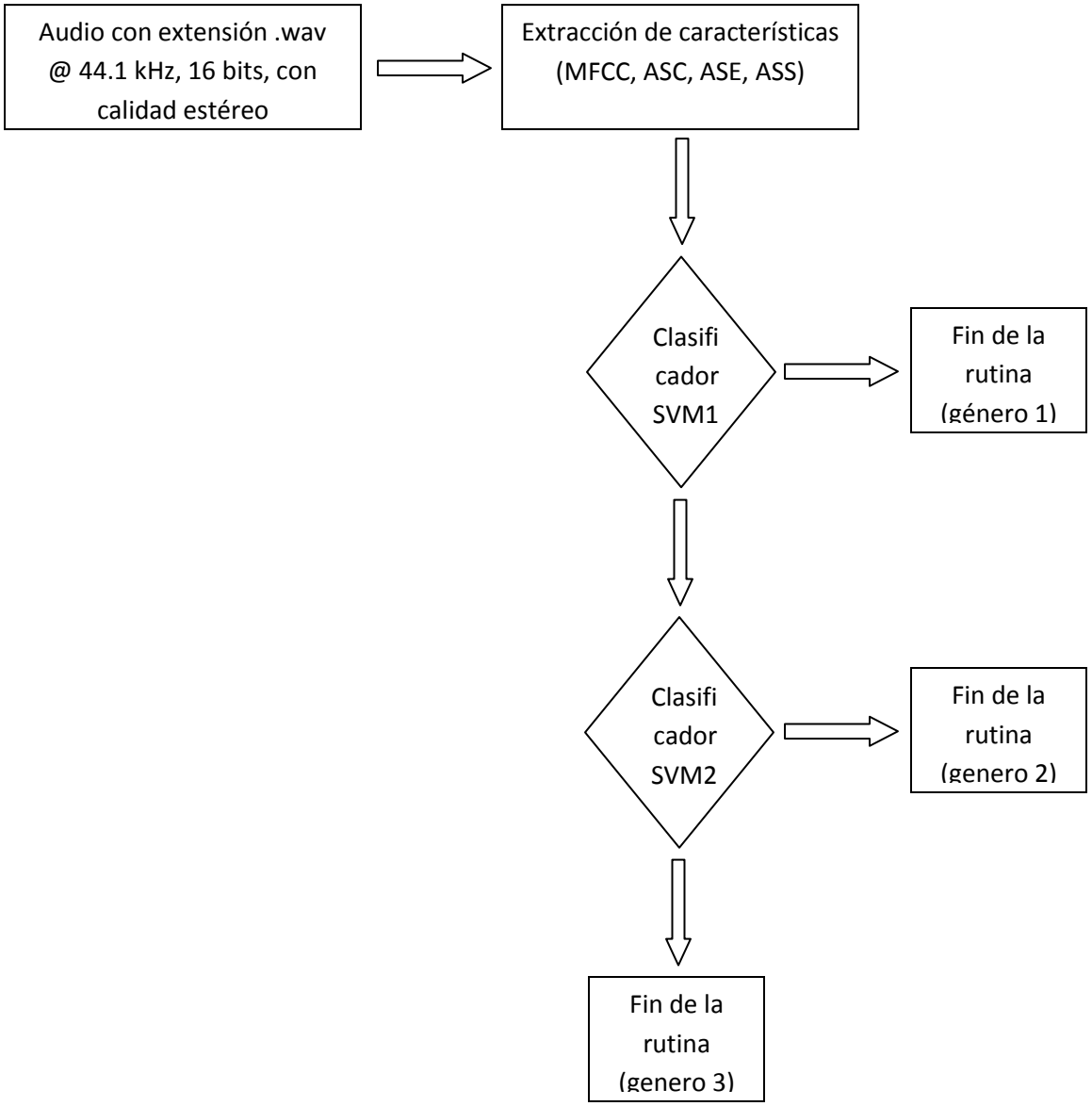


Diagrama a bloques del proceso de clasificación para una petición de búsqueda de música basada en su contenido.

Referencias

1. **Van Feng, Huijing Dou, Yanzhou Qian,** *A Study of Audio Classification on Using Different Feature Schemes with Three Classifiers*, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing, China, artículo publicado, 2010.
2. **GARCIA DIAZ, Elkin Eduardo,** *Adaboost aplicado a clasificación de fonemas, publicación del* Laboratorio de Sistemas Paralelos y arquitectura Computacional, 2007, http://www.capsl.udel.edu/~egarcia/Papers/CWCAS06.pdf, consultada abril de 2011.
3. **MATEOS GARCIA, Ismael,** *Máquinas de Vectores de Soporte (SVM) para reconocimiento de locutor e idioma, Universidad Autónoma de Madrid, febrero de 2008, tesis de ingeniería,* http://www.coit.es/pub/ficheros/presumen_banesto_3ca00e7d.pdf, consultada septiembre de 2011.
4. **PEDROZA, Gabriel,** *Aplicación de las máquinas de soporte vectorial al reconocimiento de hablantes,* Universidad Autónoma Metropolitana, junio de 2007, tesis de maestría en Ciencias y TI, http://mcyti.izt.uam.mx/Tesis_egresados/TesisGabrielPedroza.pdf, consultada agosto de 2011.

Capítulo 7. Resultados

Como se mencionó en el capítulo segundo, este trabajo de investigación se baso bajo la siguiente arquitectura:

Figura 7.1. Arquitectura general del aplicativo y servicio de identificación de música

De esta arquitectura se puede observar dos partes principales que conforman nuestro sistema, la primera de ellas es del lado del usuario en donde se puede apreciar que estos mismos podrán hacer uso del aplicativo desde distintos dispositivos móviles, de donde dependiendo de las características del dispositivo y de las condiciones en que se encuentre el usuario, podrá hacer uso de la aplicación Web de reconocimiento de música utilizando diferentes formatos de audio, por ejemplo, en el caso de que el usuario este trabajando en una PC, un iPad o tableta electrónica, podrá utilizar los formatos comúnmente utilizados en estos dispositivos *.wav y *.mp3 a 44.1 kHz, y como estos dispositivos se conectan a la red LAN o WiFi, no se tiene problema por cuanta información se enviará a través de la red, en el caso de que el usuario se encuentre en una avenida, este podrá grabar 5 segundos del audio desde su dispositivo celular en los formatos

soportados por estos dispositivos *.amr y *.mp3 a 8 kHz, y después podrá hacer uso del servicio de la telefonía celular, enviando el archivo multimedia a través de un mensaje MMS o vía GPRS.

En el caso de que el usuario utilice el servicio, este podrá utilizar una aplicación Web que automáticamente buscará el medio por el cual se enviará la información, ya sea por la red LAN, WiFi, GPRS o MMS, además que este aplicativo será capaz de reconocer el formato del archivo multimedia enviado, para después ser procesado y arrojar la información que necesita el usuario.

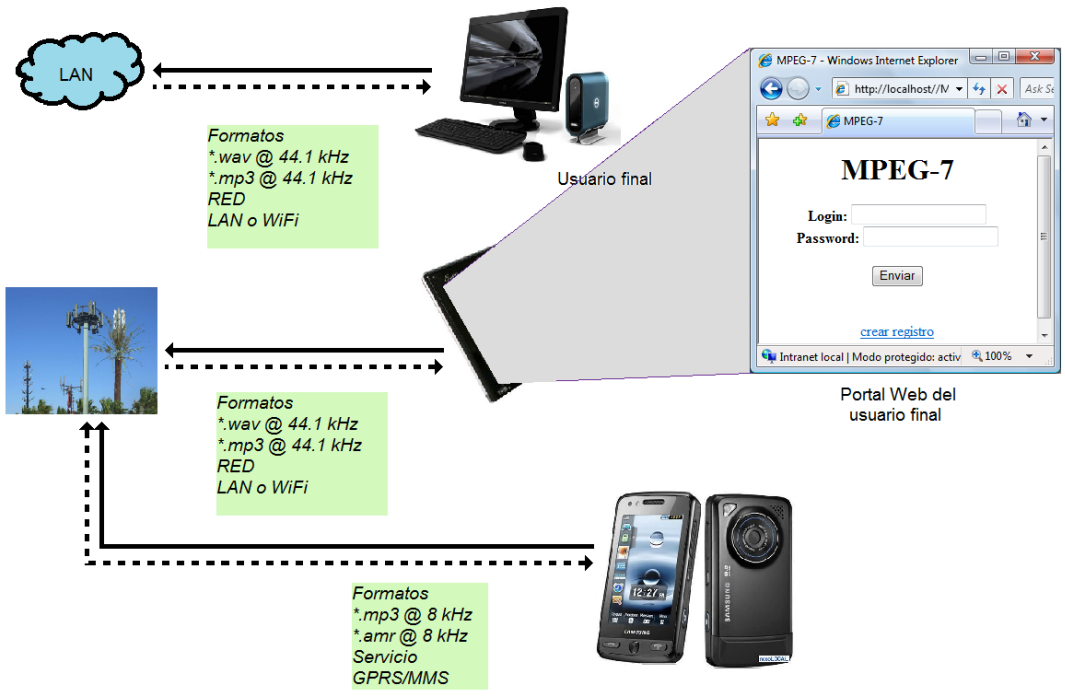


Figura 7.2. Características técnicas de acuerdo al dispositovo móvil.

El sistema se conformara principalmente de 3 bloques o secciones, la primera consta de los servidores Web, los cuales darán alojamiento al aplicativo Web codificado en PHP, este bloque proporcionará la interfaz y la interacción entre el usuario final y las herramientas del estándar MPEG-7 que nos hacen posible identificar la música.

El segundo bloque dará alojamiento a la base de datos multimedia y está conformado por un conjunto de servidores; esta base de datos está administrada de tal manera que existe una carpeta por cada archivo de música, en la cual se alojan cada uno de los descriptores que han sido previamente procesados por la herramienta MATLAB.

Finalmente el bloque de almacenamiento de datos y codificación, contendrá la información de cada usuario y sus peticiones, al igual que en el bloque anterior, cada usuario cuenta con una carpeta en la cual se aloja el archivo multimedia que envió a través del dispositivo móvil y así mismo en esta carpeta se guardarán los descriptores correspondientes y sus resultados.

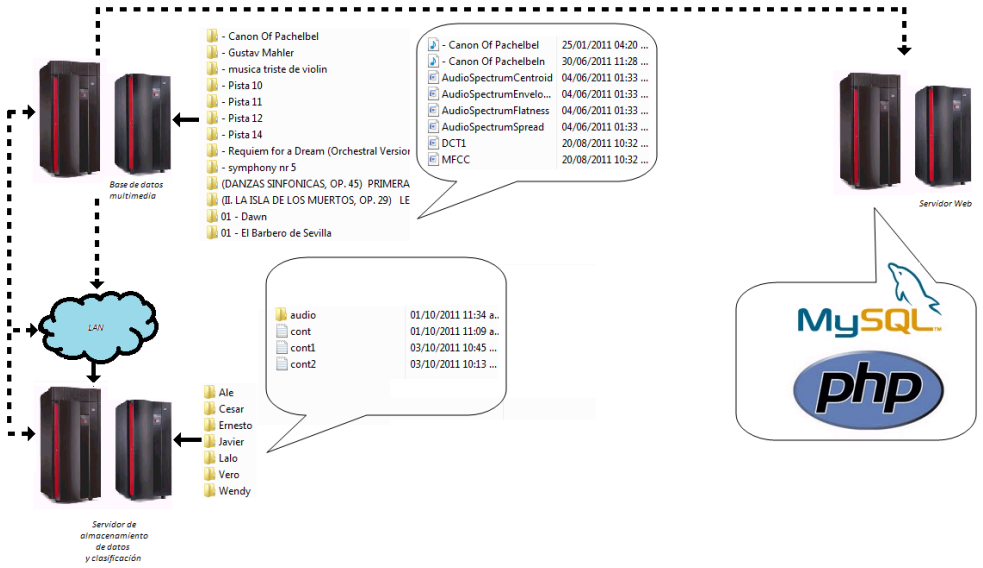


Figura 7.3. Almacenamiento y tareas realizadas por cada uno de los bloques que conforman el servicio de identificación de música.

En la figura siguiente se muestra un ejemplo donde se observan los puntos de un vector de clasificación con los valores descriptivos de un audio de género hardrock distribuidos en el hiperplano previamente entrenado con los coeficientes MFCC como se observa en la figura 6.4, donde en esta estructura podemos observar que la mayoría de las muestras pertenecen o se encuentran dentro del hiperplano de la clase 1 (hardrock), de aquí nosotros definimos un umbral para determinar con cuantos puntos pertenecerá a una clase o a otra. Por ejemplo si en la misma figura observamos que las muestras de otro track perteneciente a este mismo género hardrock se han clasificado como clase 0 (clásica) y sobrepasan el umbral, este archivo será clasificado de manera errónea como este segundo género clásico.

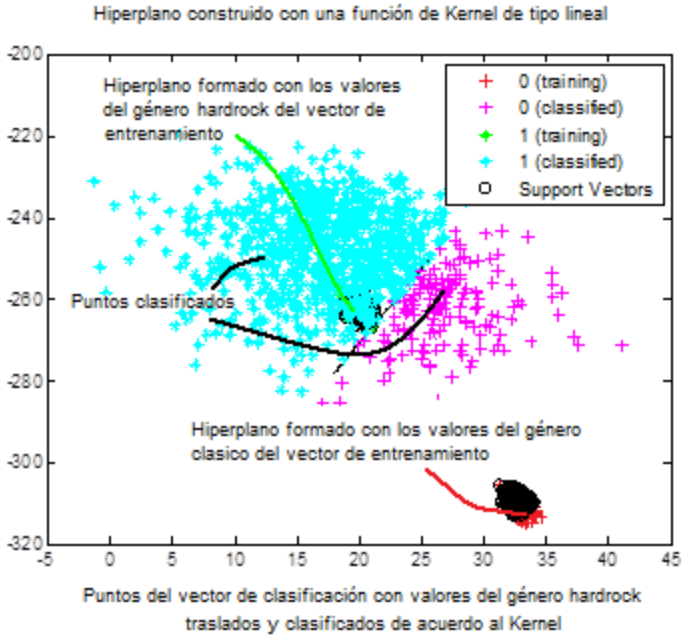


Figura 7.4. Ejemplo del proceso de clasificación donde para un extracto de música de tipo hard rock, indica que la mayoría de los puntos del vector corresponden al género 1 (hard rock), el género 0 corresponde a la música clásica.

Al realizar las subrutinas y clasificando el audio de la misma base de datos de acuerdo a las pruebas realizadas anteriormente, obtenemos los siguientes resultados:

Utilizando SVM con los diferentes vectores de entrenamiento y diferentes estructuras se obtiene:

	SVM1			SVM2		
Género	Hardrock	Clásica	Electrónica	Hardrock	Clásica	Electrónica
Hardrock	77.46 %	22.54 %	N/A	N/A	N/A	N/A
Clásica	17.43 %	82.57 %	N/A	N/A	N/A	N/A
Electrónica	36 %	64 %	N/A	N/A	N/A	N/A
Hardrock	60 %	N/A	40 %	N/A	N/A	N/A
Clásica	94.87 %	N/A	5.12 %	N/A	N/A	N/A
Electrónica	22 %	N/A	78 %	N/A	N/A	N/A
Hardrock	N/A	7.98 %	92.02 %	N/A	N/A	N/A
Clásica	N/A	55.38 %	44.62 %	N/A	N/A	N/A
Electrónica	N/A	19 %	81 %	N/A	N/A	N/A

	SVM3			SVM4		
Género	Hardrock	Clásica	Electrónica	Hardrock	Clásica	Electrónica
Hardrock	57.74 %	42.25 %	N/A	49.76 %	50.23 %	N/A
Clásica	22.05 %	77.94 %	N/A	6.66 %	93.33 %	N/A
Electrónica	15 %	85 %	N/A	82 %	18 %	N/A
Hardrock	61.97 %	N/A	38.02 %	62.91 %	N/A	37.08 %
Clásica	24.10 %	N/A	75.89 %	49.74 %	N/A	50.25 %
Electrónica	20 %	N/A	80 %	17 %	N/A	83 %
Hardrock	N/A	6.57 %	93.42 %	N/A	63.84 %	36.16 %
Clásica	N/A	55.38 %	44.61 %	N/A	47.17 %	52.83 %
Electrónica	N/A	12 %	88 %	N/A	18 %	82 %

	SVM5					
Género	Hardrock	Clásica	Electrónica			
Hardrock	47.88 %	52.11 %	N/A			
Clásica	32.30 %	67.69 %	N/A			
Electrónica	95 %	5 %	N/A			
Hardrock	61.03 %	N/A	38.96 %			
Clásica	93.33 %	N/A	6.66 %			
Electrónica	22 %	N/A	78 %			
Hardrock	N/A	2.81 %	97.18 %			
Clásica	N/A	68.71 %	31.29 %			
Electrónica	N/A	5 %	95 %			

Tablas 7.1., comparativas de resultados obtenidos utilizando las diferentes estructuras SVMs de entrenamiento de las pruebas 1-5, en donde se muestra el porcentaje de éxito obtenido durante la fase de clasificación utilizando el mismo audio de la Base de datos multimedia con que se realizo el entrenamiento.

De la prueba 6 se puede concluir que utilizando las distancias Euclidianas, no es suficiente para llegar a una buena aproximación de éxito, ya que de las diferentes pruebas hechas y de los resultados obtenidos, no se tiene una buena aproximación comparándolo con la estructura 4 SVM, la cual ha tenido una mejor probabilidad de éxito.

De acuerdo al artículo [referencia 1 del capítulo anterior, "A Study of Audio Classification on Using Different Feature Schemes with Three Classifiers"] y comparándolo con este trabajo de investigación, que al igual utiliza los coeficientes MFCC para su proceso de clasificación y reconocimiento de música, podemos observar que se obtuvieron porcentajes de probabilidad de clasificación similares, por lo que dichos coeficientes MFCC son de gran importancia y utilidad en el procesamiento, reconocimiento e identificación de música.

De la prueba 7 al realizar el mismo procedimiento de comparación y utilizando de igual manera la estructura 4 SVM para la fase de clasificación, en el caso del formato convertido de *.amr a *.wav se tiene una probabilidad de éxito alrededor del 38 % para los casos de hardrock, para la música clásica del 91.8 % y para el caso de la música electrónica del 76 % y para el formato convertido de *.mp3 @ 44.1 kHz a *.wav se tiene una probabilidad de éxito alrededor del 50.7 % para los casos de hardrock, para la música clásica del 92.83 % y para el caso de la música electrónica del 82 %, por lo cual, se concluye que el formato *.mp3 @ 44.1 kHz podría ser ideal para el uso en esta aplicación, sólo si se mejora o se encuentra una mejor estructura que logre una mayor discriminación utilizando de diferente manera los descriptores que se están considerando o en su caso utilizando otra combinación de descriptores; en el caso del formato *.amr sería una limitante en dispositivos móviles ya que por lo mismo que agrega ruido a la señal, disminuye su eficiencia en los resultados de probabilidad de encontrar un género.

	SVM4 para *.amr @ 8 kHz	SVM4 para *.mp3 a 8 kHz	SVM4 para *.mp3 @ 44.1	SVM4 para *.wav @ 44.1 kHz
Género	Probabilidad	Probabilidad	Probabilidad	Probabilidad
Hardrock	38.49 %	30.9 %	50.7 %	50.24 %
Clásica	91.8 %	95.38 %	92.83 %	93.34 %
Electrónica	76 %	63 %	82 %	82 %

Tabla 7.2., comparativa de los resultados obtenidos en el proceso de clasificación utilizando diferentes formatos de audio a diferentes tasas de muestreo, donde se observa que la probabilidad de que se encuentre un género disminuye para los formatos con frecuencia de muestreo a 8 kHz.

De la prueba 8 se puede observar que en el proceso de clasificación utilizando sólo 5 segundos de audio y repitiéndolo 2 veces para completar el vector de 10 segundos, se obtienen porcentajes similares comparando que si se utilizarán 10 segundos del audio a clasificar:

	*.mp3 @ 8 kHz 5 segundos	*.mp3 @ 8 kHz 10 segundos	*.wav @ 44.1 kHz 5 segundos	*.wav @ 44.1 kHz 10 segundos
Género	Probabilidad	Probabilidad	Probabilidad	Probabilidad
Hardrock	69.95 %	69.01 %	51.17 %	50.24 %
Clásica	92.30 %	95.38 %	92.30 %	93.34 %
Electrónica	59 %	63 %	82 %	82 %

Tabla 7.3., comparativa de resultados obtenidos, donde se observa que al utilizar sólo 5 segundos de un extracto de audio, es tiempo suficiente para obtener los mismos resultados que utilizado 10 segundos del mismo audio, los cuales son necesarios en el proceso de clasificación.

Por lo que se puede concluir que sólo grabando 5 segundos del audio, la aplicación no tendría inconvenientes de encontrar el género del archivo de audio

Una vez que se ha clasificado un archivo de audio en un género determinado, el siguiente paso es encontrar los archivos de audio que contengan las mayores coincidencias con el audio en búsqueda que se encuentran contenidos en la propia base de datos, para esto, se realiza una subrutina que consta de dos fases; la primera de ellas, es realizar un matching del valor promedio de cada descriptor ASC y ASS de cada archivo de audio de la BD contra el valor promedio ASC y ASS del archivo en búsqueda, la decisión para que pueda ser considerado un matching es +/- el valor de la desviación estándar obtenida del total de los valores ASC y ASS según sea el género.

Una vez que se han descartado los archivos de audio de la BD cuyos parámetros de búsqueda no entran dentro del rango establecido, se procede a realizar una serie de entrenamientos SVMs, donde se generan diferentes estructuras con los valores de los descriptores ASC y ASS de los archivos que cumplen con los parámetros establecidos anteriormente, esto con el fin de encontrar las estructuras que contengan las mayores coincidencias utilizando los valores ASC y ASS del archivo de búsqueda, estas estructuras que contienen las mayores coincidencias, serán reflejadas en un archivo *.txt conteniendo los nombres de los respectivos archivos multimedia, así como su probabilidad de coincidencia.

Por otra parte, se realizó un aplicativo WEB con la herramienta PHP Editor, con este aplicativo Web se busca aprovechar la librería y herramientas del estándar MPEG-7, para que el usuario final pueda hacer uso de las tecnologías y herramientas disponibles que le permitan realizar búsquedas de audio analizando su contenido del mismo. La siguiente imagen muestra la página principal de dicho aplicativo

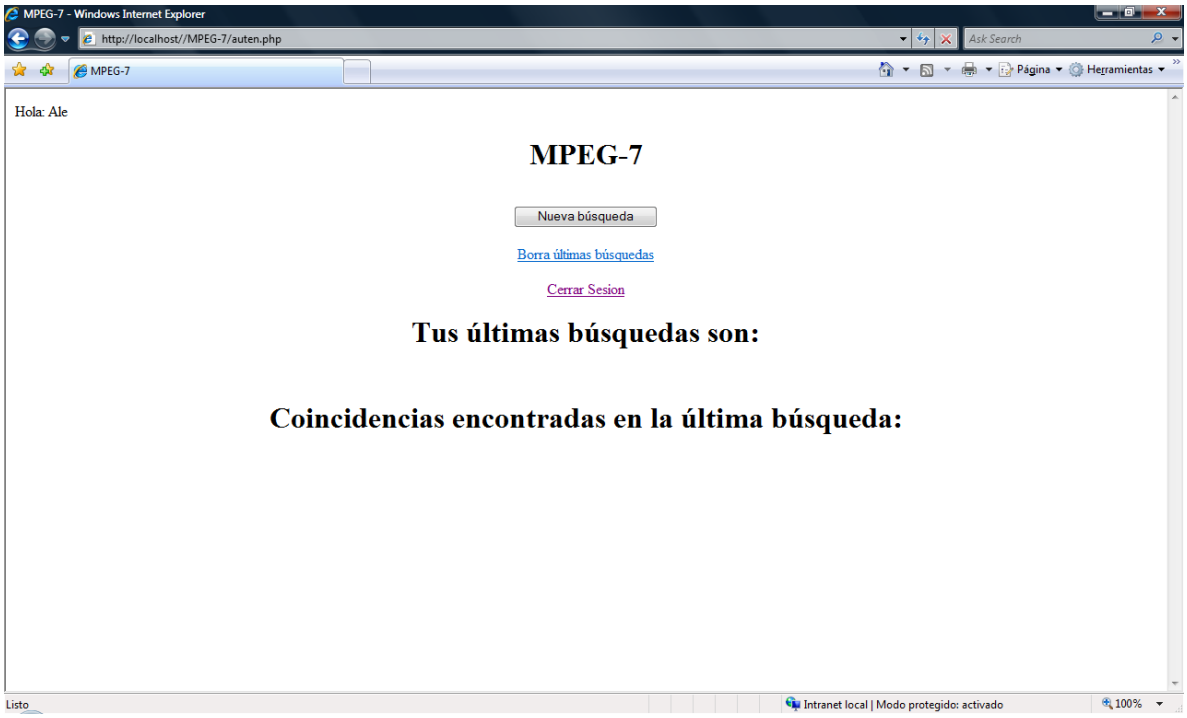


Figura 7.5. Imagen de la ventana principal del aplicativo WEB

En la siguiente figura se muestra que una vez que el usuario a enviado un archivo con extensión *.wav al aplicativo y que este ha sido procesado por las herramientas de la librería MPEG-7 bajo el software de MATLAB, podemos observar que dicho aplicativo Web nos arroja, dos listas, una donde aparecen las últimas búsquedas realizadas por el usuario y otra lista donde aparecen las coincidencias encontradas durante la última búsqueda y su probabilidad de aproximación:

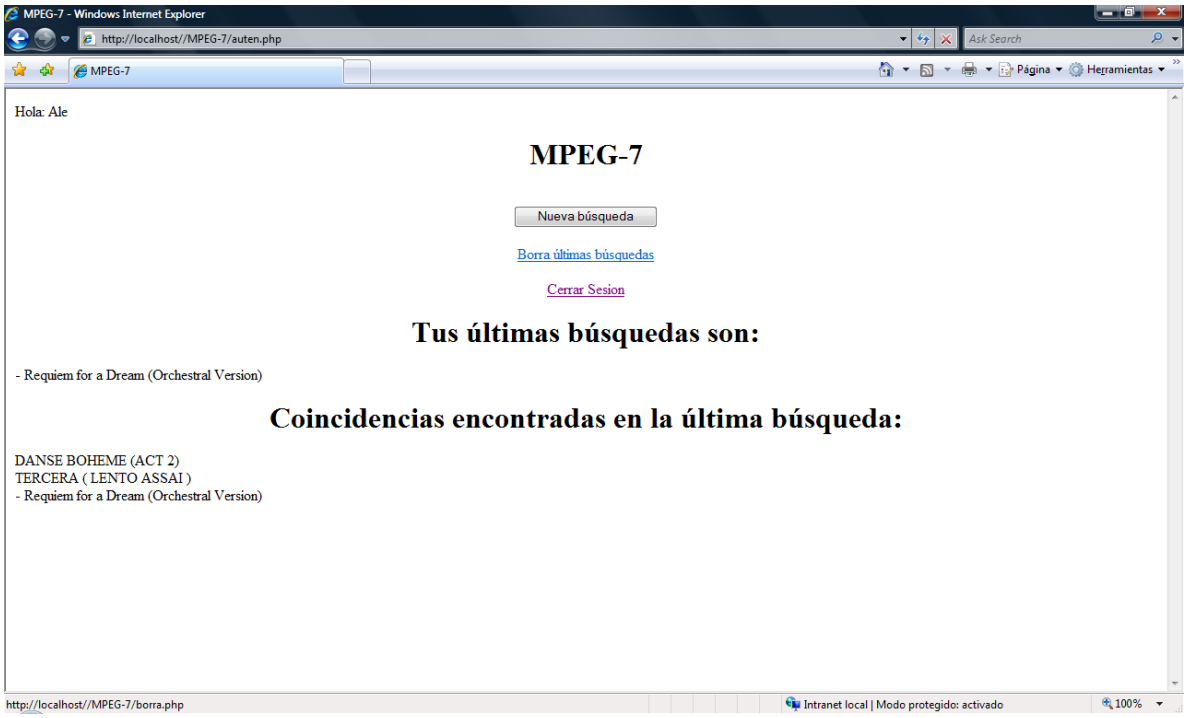


Figura 7.6. Imagen del aplicativo Web donde se muestran los resultados obtenidos al realizar una búsqueda bajo el estándar MPEG-7.

Capítulo 8. Conclusiones

En este trabajo de investigación se logró encontrar una estructura que nos permite clasificar dos géneros de música clásico y electrónico con una probabilidad de éxito del 93 % y 82 % respectivamente. Para el caso de la música hardrock, no se logró encontrar una estructura que nos permitiera discriminar este género de música de las dos restantes, el problema principal que se presenta a la hora de clasificar es el que los valores obtenidos de los descriptores de los diferentes archivos multimedia se encuentran demasiado dispersos y muchas veces tienden a ser similares entre los diferentes géneros de música, por lo que a la hora de realizar la clasificación se obtiene una probabilidad de error mayor, entonces para aumentar la probabilidad de éxito, se tiene que buscar la manera de cómo utilizar los valores de los diferentes descriptores de tal manera que se logre obtener un patrón y que coincida con el género de música al que corresponde, de ser así se lograría incrementar la probabilidad de éxito.

Otro punto de vista que se tiene que tomar en cuenta es hacer un análisis de acuerdo a los resultados de las pruebas 4 y 7, y ver que archivos se enviarán y cuanta cantidad de información se utiliza para el envío de los archivos de acuerdo al servicio, obteniendo lo siguiente:

Haciendo un análisis más detallado en este trabajo de investigación, nuestra estructura de clasificación con la cual se encontraron las mejores probabilidades de clasificación de género (estructura 4 SVM) utiliza los archivos que contienen los coeficientes MFCC, los descriptores ASS y ASC, con un tamaño total de 322 kB, por lo que para el segundo objetivo de este trabajo de investigación, sería ideal que el dispositivo móvil realizará la grabación en un formato *.mp3 @ 44.1 kHz y enviará este archivo a través de la aplicación para su procesamiento y como segunda opción sería que el dispositivo móvil grabara el audio en formato *.wav y realizará el mismo procesamiento para obtener los descriptores, y después enviar esta información para la clasificación de la música, pero también se generaría otra cuestión la cual es la capacidad de procesamiento del dispositivo móvil y cuanto afectaría en su rendimiento durante este tiempo de procesamiento, por lo que lo ideal sería que el dispositivo móvil solo enviará el archivo multimedia al servicio.

Al hacer una comparación entre los tamaños de los diferentes archivos multimedia que existen y que son soportados por los dispositivos móviles (*.amr Y *.mp3), se tiene como referencia extractos de audio de 10 segundos de donde para un formato estándar .wav de 16 bits @ 44.1 kHz se requieren 2 Mb, para un archivo con extensión .amr 7.4 kbps @ 8 kHz se requieren 8.9 Kb, para un archivo con extensión .mp3 128 kbps @ 8000 kHz se requieren 152 Kb , para un archivo con extensión .mp3 128 kbps @ 44100 kHz se requieren 156 Kb y para los archivos con la información de los descriptores con extensión *.xml se requieren 322 Kb, considerando el costo/beneficio que se tiene al utilizar el servicio multimedia (MMS) el cual tiene una capacidad de enviar 764 kB por mensaje a un costo promedio de $ 2.00, sería factible que el dispositivo móvil grabara en el formato *.mp3 @ 44.1 kHz, ya que tiene dos ventajas sobre los demás formatos, la primera de ellas, es que por mantener la misma tasa de muestreo, no se agrega ruido a la hora de

convertir esta grabación a formato *.wav, por lo que los promedios de que se clasifique en el género correcto es aceptable, la segunda ventaja es que debido a su formato, sólo se necesitan 156 kB para una grabación de 10 segundos, y este puede ser enviado en un solo mensaje multimedia.

Como se ha mencionado anteriormente, en el caso de que el dispositivo móvil requiera enviar esta información a través de la red telefónica, se tendría que ver el costo/beneficio de la aplicación, ya que un mensaje multimedia acepta 764 kB, capacidad suficiente para enviar los archivos *.xml o un extracto de música con extensión *.amr o un extracto con extensión *.mp3, pero no para enviar un extracto de música *.wav con duración de 10 segundos, este último formato sería una gran limitante debido al costo. Como se observó en la prueba 7, el formato *.amr necesita 8.9 kB para una grabación de 10 segundos y para el caso del formato *mp3 @ 8 kHz se necesitan entre 152 y 156 kB, y estos se podrían enviar sin algún problema a través de la red telefónica en un solo mensaje multimedia, la desventaja en grabar en estos formatos, es que a la hora de convertirla a formato *.wav, se tienen ciertas pérdidas de información y/o se agrega ruido especialmente cuando existen o se convierten formatos con diferentes tasas de muestreo, lo cual ocasiona que el porcentaje de probabilidad de éxito durante la clasificación disminuya dependiendo del formato a convertir, en base a los resultados obtenidos en especial para el formato mp3 @ 44.1 kHz, esta disminución de probabilidad de éxito, se mantiene dentro del rango comparándola con los resultados obtenidos en una clasificación con archivos de extensión *.wav.

Teniendo en cuenta los puntos anteriores, se puede obtener una tabla comparativa de tamaño/eficiencia, que muestra cuales serían los puntos base de las principales limitaciones técnicas de este aplicativo considerando los distintos formatos multimedia de los dispositivos móviles (*.amr y *.mp3), la información necesaria para el procesamiento, los medios de transmisión y la utilización de la red.

	SVM4 para *.amr @ 8 kHz	SVM4 para *.mp3 a 8 kHz	SVM4 para *.mp3 @ 44.1	SVM4 para *.wav @ 44.1 kHz
Género	Probabilidad	Probabilidad	Probabilidad	Probabilidad
Hardrock	38.49 %	30.9 %	50.7 %	50.24 %
Clásica	91.8 %	95.38 %	92.83 %	93.34 %
Electrónica	76 %	63 %	82 %	82 %

Tabla 8.1., comparativa de los resultados obtenidos en el proceso de clasificación utilizando diferentes formatos de audio a diferentes tasas de muestreo, donde se observa que la probabilidad de que se encuentre un género disminuye para los formatos con frecuencia de muestreo a 8 kHz.

Tipo de archivo	Probabilidad de	Tipo de servicio	
		Red de telefonía celular	Red de Internet

	identificación de género	(dispositivo teléfono celular)		(dispositivo PC, Pad)	
		Ventajas	Desventajas	Ventajas	Desventajas
Archivo multimedia con extensión *.wav @ 44 kHz, 2 MB	82 % - 93 %	Utilizando este tipo de archivo se tiene la mejor probabilidad de identificación en comparación con el formato AMR y MP3	El tamaño de un mensaje multimedia es de 764 kB, no suficiente para mandar un extracto de 10 s de tamaño igual a 2 MB	Estos tipos de dispositivos no tienen limitaciones en el uso de la red	Al tener una gran demanda, se puede saturar la red y hacerla ineficiente
Archivo multimedia con extensión *.amr 8 kHz, 8.9 kB	76 % - 91 %	Se puede enviar este tipo de archivo en un mensaje multimedia.	Utilizando este tipo de archivo se tiene una probabilidad de identificación menor	Estos tipos de dispositivos no tienen limitaciones en el uso de la red	Utilizando este tipo de archivo se tiene una probabilidad menor de identificación a comparación con el formato *.wav
Archivo multimedia con extensión *.mp3 8 kHz,	63 % - 95 %	Se puede enviar este tipo de archivo en un mensaje multimedia.	Utilizando este tipo de archivo se tiene una probabilidad de identificación menor	Estos tipos de dispositivos no tienen limitaciones en el uso de la red	Utilizando este tipo de archivo se tiene una probabilidad menor de identificación a comparación con el formato *.wav
Archivo multimedia con extensión *.mp3 44.1 kHz, 152 kB	82 % - 92.83 %	Se puede enviar este tipo de archivo en un mensaje multimedia, se tiene una probabilidad	Ninguna	Estos tipos de dispositivos no tienen limitaciones en el uso de la red	Utilizando este tipo de archivo se tiene una probabilidad técnicamente similar en comparación

		muy similar a los resultados obtenidos en la clasificación *.wav.			con el formato *.wav
Archivos de descripción con extensión *.xml, 322 kB	80 % - 90 %	Se pueden enviar estos archivos en un mensaje multimedia.	Ver las limitaciones técnicas de cada dispositivo móvil, como procesamiento y memoria.	Hay limitaciones en el uso de los recursos del equipo; en la red no se genera mucho tráfico al haber una gran demanda de usuarios	Ninguna

Tabla 8.2., donde se muestran las ventajas y desventajas presentadas en cada tipo de formato de audio

Finalmente como se ha descrito en este último capítulo, se ha logrado ser capaz de crear una aplicación la cual aprovecha algunas características proporcionadas por el estándar MPEG-7, y como se mencionó anteriormente, cada descriptor nos proporcionará determinados valores que muchas veces tienden a ser similares entre un genero y otro, es por esta razón que nuestra máquina vectorial, al hacer el procesamiento de entrenamiento, no logra encontrar una distancia y los vectores necesarios de soporte que nos permitirán diferenciar entre un género y otro, es por esta razón que este trabajo de investigación queda abierto a futuras investigaciones, en las cuales se haga un análisis más detallado y con mayor profundidad de tales descriptores, para así encontrar características que nos permitan crear un vector de entrenamiento más eficiente, que incluso nos sea capaz de reconocer tonos característicos de cada instrumento.