UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

TRABAJO FIN DE GRADO

DETECCIÓN DE MÚSICA EN CONTENIDOS MULTIMEDIA MEDIANTE RITMO Y ARMONÍA

Diego de Benito Gorrón
Tutor: Joaquín González Rodríguez

Junio 2017

DETECCIÓN DE MÚSICA EN CONTENIDOS MULTIMEDIA MEDIANTE RITMO Y ARMONÍA

AUTOR: Diego de Benito Gorrón
TUTOR: Joaquín González Rodríguez

AUDIAS – Audio, Data Intelligence and Speech
Departamento de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio 2017

Resumen (castellano)

Este Trabajo Fin de Grado se encuadra en el ámbito de la segmentación de audio, más concretamente en la detección de contenidos musicales en señales de audio. Al ser la musicalidad una propiedad de las señales de audio más subjetiva que, por ejemplo, la presencia de voz hablada, es necesario establecer qué propiedades objetivas de la señal de audio influirán en la decisión sobre presencia o ausencia de música. Para el desarrollo de este trabajo, se toman como referencias de la musicalidad de un audio la presencia de un pulso rítmico en su evolución temporal y la aparición de armonía o cromaticidad en sus componentes frecuenciales.

El sistema desarrollado en este TFG toma las decisiones de detección de música a partir de los dos componentes citados: ritmo y armonía. Para ello, cuenta con dos detectores dedicados a cada uno de los componentes, que pueden funcionar conjuntamente para detectar la presencia de contenidos musicales, pero también por separado para segmentar la señal en función del ritmo o de la armonía.

El detector de ritmo se basa, principalmente, en la periodicidad de la energía localizada de la señal de audio. Esta periodicidad puede cuantificarse mediante la construcción y el análisis de matrices de autocorrelación, que contienen la evolución de la función autocorrelación a lo largo de la duración de la señal de audio.

Por otra parte, el detector de armonía o cromaticidad parte del cálculo del cromagrama, una representación espectral basada en la Transformada de Fourier de Tiempo Corto, o *Short-Time Fourier Transform* (STFT). La peculiaridad de un cromagrama frente a un espectrograma es que acumula las componentes frecuenciales pertenecientes a una misma nota musical, permitiendo observar si existe una distribución del espectro que favorezca ciertas notas, denotando la presencia de armonía.

En el desarrollo de este trabajo también se incluyen pruebas de rendimiento sobre la base de datos ATVS-Radio, que contiene 25 horas de audio etiquetado según la presencia de música y de voz.

Palabras clave

Audio, música, voz, ruido, segmentación, detector, multimedia, ritmo, armonía, autocorrelación, cromagrama, espectrograma, recuperación de información musical.

Abstract (English)

This Bachelor Thesis is framed within the area of audio segmentation, as it is focused in the detection of musical contents in audio signals. Musicality is a more subjective property of audio signals than, for example, speech activity, so it is necessary to define which objective properties of the signal will be relevant when assessing whether there is music present in it or not. The features evaluated as traces of musicality are the presence of a rhythmic beat and the harmony (chromaticity) found in the frequency spectrum of the signal.

Our system bases its decisions in both components: rhythm and harmony. For that purpose, it uses two specific detectors, each one focused in one of the components. These detectors can work together to detect different kinds of musical contents, but they can also run separately to segment the audio based only on rhythm or harmony.

The rhythm detector mainly evaluates the periodicity found in the local energy of the audio signal. This periodicity can be quantified building and analyzing the autocorrelation matrix of the audio. These matrices show the evolution of the autocorrelation function along the audio signal.

On its side, the harmony detector starts from the chromagram matrix of the audio, a spectral representation derived from the Short-Time Fourier Transform (STFT). Unlike the spectrogram representation, a chromagram clusters the spectral components belonging to the same musical note, showing if the spectral distribution stimulates certain notes more than others, suggesting the existence of harmony.

The realization of this Bachelor Thesis also includes some performance tests using the ATVS-Radio database, which contains 25 hours of audio with music and speech activity tags.

Keywords

Audio, music, voice, segmentation, detector, multimedia, rhythm, harmony, autocorrelation, chromagram, spectrogram, music information retrieval.

Agradecimientos

A mi familia y amigos, por su confianza y apoyo incondicional. No hay palabras para describir todo lo que me aportáis.

A mi tutor, Joaquín González, por ofrecerme la oportunidad de embarcarme en este proyecto y por la ayuda que me ha ofrecido durante su desarrollo. También a los profesores Daniel Ramos, Doroteo Torre y Javier Franco, por incentivar mi interés en el tratamiento de señales de audio y hacer posible la consecución de este TFG.

A mis compañeros de laboratorio en AUDIAS, en especial a Álvaro por sus consejos para el desarrollo de este trabajo.

Además, también doy las gracias a todas las personas que me han inculcado el interés por la música, a mis compañeros de grupo y mis profesores, Javier y Mario.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1 Introducción

1.1 Motivación

La detección de música en contenidos multimedia es de interés por varios motivos. Si bien la presencia de contenidos musicales en una señal de audio puede ser, por sí sola, una información relevante sobre la señal para determinadas aplicaciones, la principal motivación para el desarrollo de este TFG es obtener un detector que pueda complementar a otros sistemas de segmentación de audio, como detección de voz o identificación de locutores.

La presencia de contenidos musicales influye drásticamente, por ejemplo, en las aplicaciones de detección y reconocimiento de voz. Será de utilidad, entonces, que una aplicación de procesamiento de audio pueda disponer de información sobre la presencia o ausencia de música a lo largo de la evolución temporal de la señal.

Además, el de la música es un caso especial, ya que las señales de audio consideradas "musicales" pueden tener características muy diversas, y en muchas ocasiones depende de la subjetividad del oyente o de la intención del creador determinar si la naturaleza de una determinada señal de audio es o no musical. Sabiendo esto, el sistema de detección que se plantea tiene como objetivo principal aquellos contenidos musicales en los que se puedan encontrar componentes de ritmo y/o de armonía, que además son los que más consecuencias pueden acarrear en otras aplicaciones de segmentación de audio.

1.2 Objetivos

La principal meta de este TFG es el diseño y desarrollo de un sistema de detección de música que permita llevar a cabo, para un fichero de audio dado, una segmentación temporal del mismo en función de la presencia o no de contenido musical.

Se busca también que el sistema entregue un valor que represente la fiabilidad de la presencia de música a lo largo de la duración del fichero, de manera adicional a la decisión binaria (presencia o ausencia de música).

Además, este trabajo también tiene como objetivo la comprobación del sistema de detección descrito sobre una gran variedad de casos, para lo cual se utilizará la base de datos ATVS-Radio, que contiene 25 horas de audio etiquetado.

Como objetivo complementario al desarrollo del detector, se pretende identificar casos particulares que permitan estudiar el comportamiento del sistema al completo y de los detectores de ritmo y de armonía por separado.

1.3 Organización de la memoria

Para la organización de esta memoria se ha escogido la siguiente división en capítulos:

- **Capítulo 1: Introducción.**

  En este primer capítulo se describe brevemente el interés que existe detrás de la detección de música en contenidos multimedia, se explica la problemática que entrañan los contenidos musicales y se enumeran los objetivos marcados para el desarrollo de este TFG. Además, se detalla la estructuración de la memoria en capítulos, esbozando los contenidos que se pueden encontrar en cada uno de ellos.

- **Capítulo 2: Estado del arte.**

  En este capítulo tienen cabida los aspectos relativos a las tecnologías existentes en detección de música y a la recuperación de información musical (*Music Information Retrieval*, MIR), mencionando en cada caso las diferencias y similitudes con el sistema propuesto en este TFG. También se formulan algunas de las herramientas matemáticas de mayor relevancia en el desarrollo de este TFG, aportando muestras gráficas de su aplicación a señales de audio.

- **Capítulo 3: Diseño y desarrollo.**

  Este capítulo incluye una descripción del entorno experimental utilizado y documenta el proceso de diseño de cada uno de los detectores individuales (de ritmo y de armonía). Para cada uno de ellos, se explica el planteamiento inicial pensado y qué problemas se han encontrado a la hora de llevar dichos planteamientos a cabo, además de las soluciones que se han ideado. También se describe la combinación de ambos detectores, incluyendo resultados gráficos.

- **Capítulo 4: Pruebas y resultados.**

  Se describen en este capítulo las distintas pruebas que se han realizado para evaluar el funcionamiento del detector, y se recopilan y analizan los resultados obtenidos.

- **Capítulo 5: Conclusiones y trabajo futuro.**

  Para finalizar la memoria, se incluyen algunas reflexiones sobre el trabajo llevado a cabo y se plantean otras aplicaciones en las que podría ser de utilidad.

2 Estado del arte

En comparación con la detección y segmentación de voz, la detección de contenidos musicales en audio es un campo con menor y más reciente desarrollo.

Aunque en los últimos años ha tenido lugar un gran ascenso en la popularidad de servicios de "detección de música" como, por ejemplo, Shazam o SoundHound, merece la pena mencionar sus diferencias con el sistema propuesto en este TFG. Entre las tecnologías que utilizan estas aplicaciones están las denominadas *audio fingerprinting* y *query-by-example*: cuentan con una base de datos que guarda características de un gran número de canciones, y realiza comparaciones con las características del audio de entrada, mostrando al usuario si existe alguna coincidencia [1].

Se trata, entonces, de sistemas muy específicos, de alto nivel semántico, que no nos pueden dar una respuesta a las cuestiones –menos específicas– que definen el detector que buscamos: ¿Existen contenidos musicales en esta señal de audio? ¿En qué intervalos de tiempo?

2.1 Obtención de información en señales de música

La recuperación de información musical, o *Music Information Retrieval* (MIR), es el campo de investigación que se dedica al estudio de las señales musicales con el fin de extraer características significativas de las mismas, ya sea partiendo de meta-datos o del contenido de la señal [2].

Estas características pueden buscarse a muchos niveles distintos de especificidad: tendríamos el nivel más alto en el *audio fingerprinting*, mencionado anteriormente, que en MIR se utiliza con el objetivo de identificar una interpretación de una canción concreta, en muchos casos a partir de un fragmento grabado de esa misma interpretación con cierto nivel de ruido añadido. Otros usos de las técnicas de MIR con alta especificidad tienen que ver con la detección de plagios y las infracciones de derechos de autor [3].

Bajando en la escala de especificidad, encontramos aplicaciones como la detección automática del género musical o incluso del intérprete: aunque son tareas para las que típicamente se han utilizado los meta-datos incluidos en los ficheros de audio, conseguir aproximaciones basadas en el contenido de la propia señal es interesante para fines relacionados, por ejemplo, con sistemas de recomendación musical.

La detección de presencia de música, o segmentación música-voz, también es objeto de investigación de las técnicas de *Music Information Retrieval*, pese a tratarse de una tarea de muy baja especificidad: no tenemos interés en conocer datos como el género musical, la tonalidad o el intérprete, simplemente distinguir en qué fragmentos de la señal aparecen contenidos musicales.

Para tareas de especificidad baja, las estrategias MIR suelen requerir la extracción de características de alto nivel de la señal: líneas melódicas, acordes o seguimiento del pulso

rítmico. En el caso particular de la detección de presencia de música, sólo hace falta determinar si esas características están presentes y con qué nivel de fiabilidad.

2.2 Características del audio en señales musicales

Las señales musicales muestran características de interés tanto en el dominio temporal como en el frecuencial. Una buena aproximación inicial es que el dominio temporal nos aporta información relacionada con el pulso rítmico, mientras que el dominio espectral nos permite acceder a la información melódica y armónica [4].

El pulso rítmico está estrechamente ligado al *tempo*, que en música se mide habitualmente en pulsaciones por minuto, o *beats per minute* (bpm). El rango de valores de bpm que podemos encontrar en música es muy amplio, pero podríamos acotarlo entre aproximadamente 60 bpm (indicados como *larghetto* o *adagio*) y valores de 200 bpm, o incluso más allá (*presto* o *prestissimo*). Viéndolo de otra manera, en un segundo de señal esperaríamos encontrar desde una hasta cinco pulsaciones o acentos principales, aunque siempre desde un punto de vista teórico: en la realidad, la música no se limita a acentuar aquellos instantes que coinciden con pulsaciones del metrónomo, sino que la sensación rítmica muchas veces nace de la combinación de diversos patrones de acentuación.

El espectro de frecuencias audibles, considerado típicamente desde los 20Hz hasta los 20kHz, contiene información de la señal musical relativa a la armonía y la melodía. Sin embargo, debemos tener en cuenta que la señal musical no es estacionaria, y por tanto la información espectral sólo será de utilidad si estudiamos su evolución en el tiempo. Es más, las características espectrales en función del tiempo pueden aportar también información relativa a los patrones rítmicos de la señal.

A continuación, se detallarán algunas de las técnicas que sirven como base para extraer las características relevantes de la señal de audio, tanto en el dominio temporal como en el espectral.

2.2.1 Energía localizada y autocorrelación

Sea la energía de una señal discreta real de duración finita (*N* muestras):

$$E_x = \sum_{n=0}^{N-1} x^2[n]$$

La energía localizada de una señal discreta *x[n]* es la señal, también discreta, resultante de calcular la energía de *x[n]* en distintas ventanas de tiempo [5]. El enventanado empleado para el cálculo de la energía localizada suele ser de forma rectangular:

$$w[n] = \begin{cases} 1, & n \in [0, L-1] \\ 0, & en\ el\ resto\ de\ casos \end{cases}$$

Donde *L* es la duración en muestras de la ventana.

Podemos expresar la señal energía localizada como:

$$E_x[m] = \sum_{n=0}^{N-1} [x[n] \cdot w[n - P \cdot m]]^2$$

Donde P es la separación en muestras entre ventanas consecutivas, y m es el nuevo eje temporal, con menos muestras que el eje original. Adicionalmente, conociendo la expresión matemática de la ventana rectangular, se puede desarrollar la ecuación:

$$E_x[m] = \sum_{n=P \cdot m}^{P \cdot m + L - 1} x^2[n]$$

El cálculo de la energía localizada de una señal de audio nos permite obtener una evolución temporal de su amplitud:

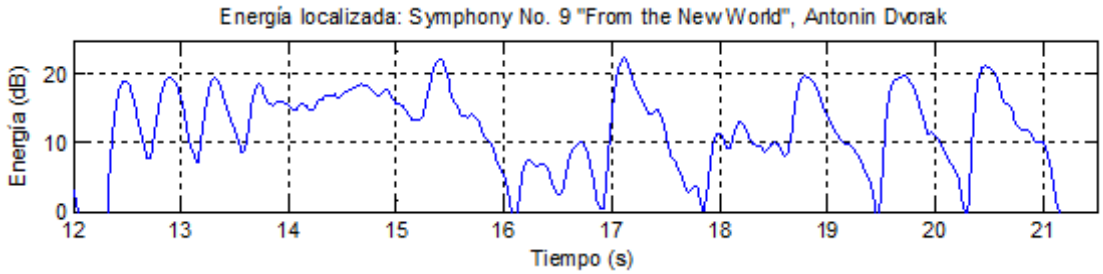


Figura 2-1: Energía localizada en un extracto de la Sinfonía del Nuevo Mundo, de Antonin Dvorak

La figura representa la energía localizada en un fragmento de música sinfónica. Pese a que no se observa un pulso constante (por ejemplo, marcado por un instrumento de percusión), la señal de energía muestra cierta periodicidad en el tiempo, que puede ser utilizada para obtener información acerca del pulso rítmico.

Uno de los aparatos matemáticos más útiles a la hora de estimar la periodicidad de una señal es la función de autocorrelación localizada. Esta función compara una señal temporal con versiones desplazadas de la misma, de manera que muestra máximos locales en aquellos desplazamientos que coincidan con la duración de los periodos de la señal. La función de autocorrelación localizada de la señal $x[n]$, $R_x[k]$, puede expresarse como:

$$R_x[k] = \sum_{n=0}^{N-1-|k|} x[n] \cdot x[n + |k|]$$

Donde k representa el desplazamiento en muestras, que puede tomar valores desde 0 hasta N-1 (también se pueden considerar desplazamientos con $k < 0$, dando lugar a una simetría par). El máximo absoluto de la función de autocorrelación se encuentra en $k = 0$, donde la expresión coincide con la de la energía de la señal. Se suele normalizar la señal de autocorrelación para ubicar este máximo absoluto en $R_x[0] = 1$:

$$R_{x,norm}[k] = \frac{R_x[k]}{E_x}$$

Para ilustrar el aspecto de la función de autocorrelación, a continuación se muestra su aplicación a una ventana de tres segundos de la energía localizada mostrada anteriormente. Para realzar la periodicidad, como paso previo al cálculo de la autocorrelación se ha calculado la primera diferencia de la señal de energía localizada.

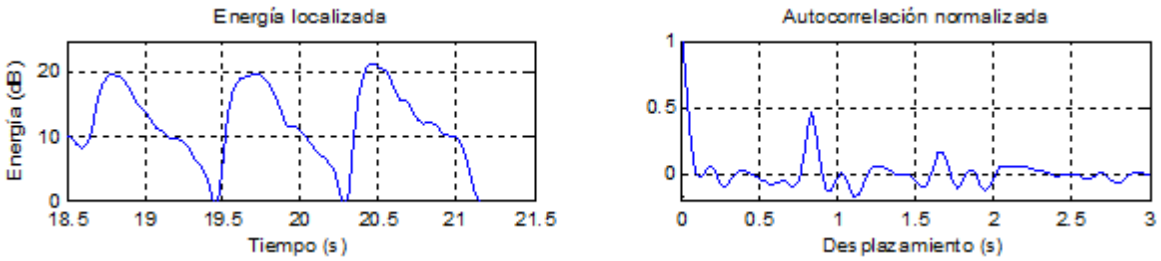


Figura 2-2: Energía localizada (izquierda) y autocorrelación normalizada (derecha) para tres segundos de señal

En la gráfica que representa la autocorrelación (derecha), el desplazamiento de la señal se muestra en segundos. El máximo más llamativo, exceptuando el de desplazamiento nulo, se ubica en algo menos de un segundo. La gráfica de la izquierda nos permite observar que ese valor de tiempo coincide con el periodo que marca la evolución temporal de la energía.

Mediante el uso de la energía localizada y de la función de autocorrelación, tal y como se ha visto, es posible la extracción de características rítmicas de una señal de audio.

2.2.2 Short-Time Fourier Transform: Espectrogramas y cromagramas

La *Short-Time Fourier Transform (*STFT) es una variación de la Transformada de Fourier que cobra especial relevancia cuando queremos estudiar no solamente las componentes frecuenciales de una señal, sino cómo evolucionan éstas en el tiempo [6].

Su implementación consiste en un enventanado de la señal y una aplicación de la Transformada de Fourier (por ejemplo una FFT, o *Fast Fourier Transform*) a cada una de las tramas de señal resultantes del enventanado.

El aspecto final de la STFT, incluyendo su resolución en tiempo y en frecuencia, depende en gran medida de las características de la ventana escogida. La ventana rectangular tiene un espectro con grandes lóbulos secundarios, lo que en muchos casos la hace inadecuada para esta transformada por el efecto de lobulado espectral [5].

Una ventana usada habitualmente es la de Hamming, de contorno más suave. Los valores de longitud y desplazamiento de la ventana tendrán efecto en la resolución temporal y frecuencial, siempre con un compromiso entre ambas: aumentar la resolución frecuencial (tomando más muestras en cada ventana) nos hará perder resolución temporal, y viceversa.

La representación del módulo de la STFT de una señal suele denominarse espectrograma. Esta clase de representaciones son utilizadas habitualmente para mostrar el espectro de una señal de audio.

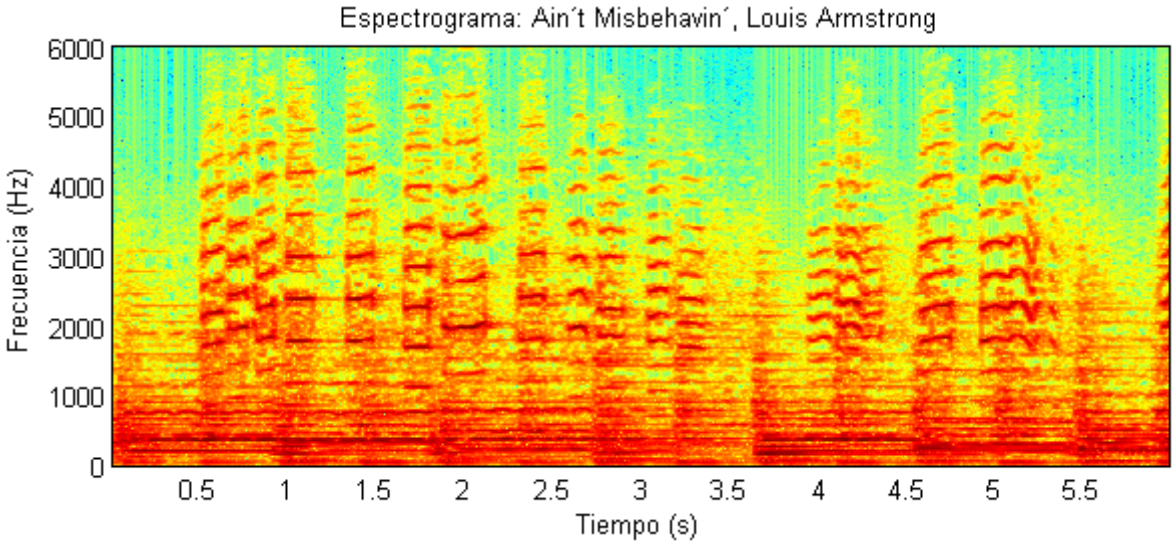


Figura 2-3: Espectrograma de un extracto de la canción *Ain't Misbehavin'*, de Louis Armstrong

De la representación del espectrograma derivan los llamados cromagramas. Su propósito es reducir la información del espectrograma a unas pocas filas de valores, habitualmente doce. El fundamento de los cromagramas es la organización de las frecuencias de las notas musicales en la escala cromática. Cada una de las filas del cromagrama recoge la información espectral de las frecuencias que corresponden a las doce notas musicales (una por fila) de la escala cromática en las diferentes octavas [7].

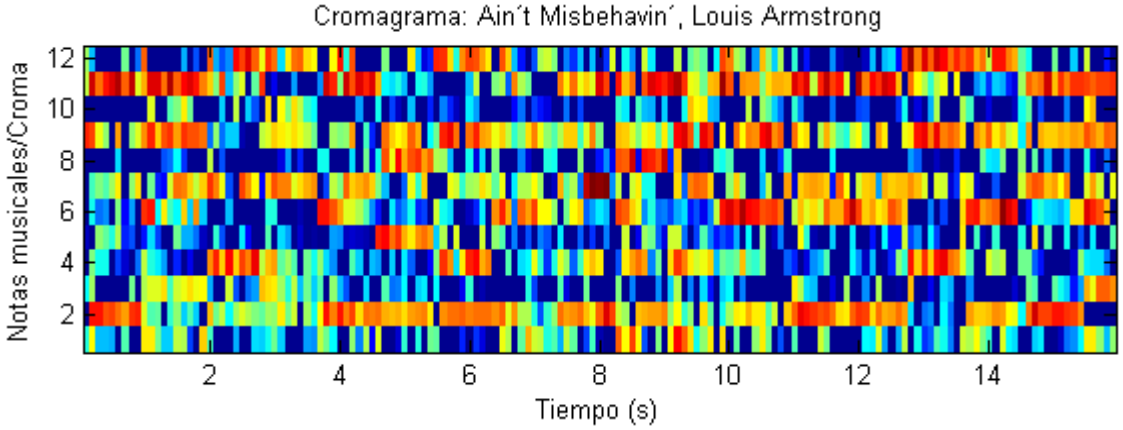


Figura 2-4: Cromagrama de un extracto de la canción *Ain't Misbehavin'*, de Louis Armstrong

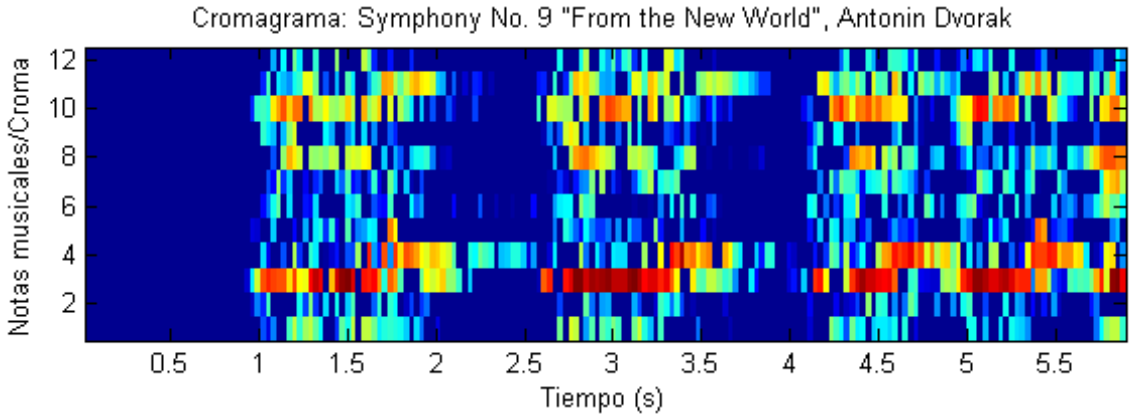


Figura 2-5: Cromagrama de un extracto de la *Sinfonía del Nuevo Mundo*, de Antonin Dvorak

La información del espectrograma, y más particularmente la del cromagrama, puede ser utilizada para inferir características de alto nivel de la señal de música, como la tonalidad en la que se encuentra la pieza musical (utilizando, por ejemplo, modelos estadísticos de la frecuencia de aparición de cada nota) o la sucesión de acordes (conjuntos de notas simultáneas).

Los cromagramas calculados por el detector desarrollado en este TFG están basados en la implementación propuesta por LabROSA bajo la licencia GNU-GPL [8].

3 Diseño y desarrollo

El diseño del detector está pensado para aprovechar las características que pueden ser extraídas de las señales de audio, tanto en el dominio temporal como en el espectral, muchas de las cuales ya han sido formuladas y explicadas en el Capítulo 2.

Como se ha mencionado anteriormente, la detección de música está limitada por la definición que demos de "musicalidad", que es en sí un concepto subjetivo. Por esto, se ha decidido, como primera aproximación, definir como música aquellos contenidos de audio que cumplan alguna de las condiciones siguientes:

i. Mantienen un pulso rítmico aproximadamente constante y dentro de un cierto margen de velocidad o *tempo*.

ii. Muestran contenidos frecuenciales en su espectro en los que ciertas notas musicales se ven favorecidas frente a otras.

La primera condición (i) es evaluada por el detector de ritmo (apartado 3.2), mientras que la segunda (ii) se evalúa en el detector de armonía o croma (apartado 3.3). De manera posterior al trabajo de estos detectores, es necesaria la combinación de los resultados de ambos para que el sistema tome decisiones en cuanto a la presencia o ausencia de música.

A lo largo del desarrollo del sistema se han detectado algunos matices necesarios para las condiciones planteadas, que serán descritos en los apartados correspondientes.

3.1 Entorno experimental: Base de datos ATVS-Radio

Las pruebas del detector de música (Capítulo 4) se han realizado, principalmente, sobre ficheros pertenecientes a la base de datos de audio ATVS-Radio. Esta base de datos cuenta con un total de 50 horas de programas de radio de varias cadenas españolas, dispuestos en ficheros *wav* de un canal (mono) y una hora de duración, muestreados a 16 kilohercios y cuantificados mediante PCM con 16 bits por muestra.

La primera media hora de cada fichero está etiquetada manualmente con el programa Wavesurfer, en función de la presencia o ausencia de voz, música o publicidad, entre otros datos. Los programas contenidos en la base de datos son [9]:

Programa	Cadena	Grabación	Etiquetado
El Pirata y su banda	Rock FM	6:00 – 7:00h	6:00 – 6:30h
Hoy por hoy	Cadena SER	9:00 – 10:00h	9:00 – 9:30h
Julia en la Onda	Onda Cero	18:00 – 19:00h	18:00 – 18:30h
La mañana	COPE	10:00 – 11:00h	10:00 – 10:30h
Más de uno	Onda Cero	9:00 – 10:00h	9:00 – 9:30h

Tabla 3-1: Contenido de la base de datos

La base de datos ATVS-Radio recorre prácticamente todo el espectro de posibilidades que buscamos para probar el detector: audio totalmente musical (canciones con ritmo y armonía), intervalos que contienen exclusivamente voz de uno o varios locutores y también fragmentos en los que voz y música de distintos tipos se solapan, dando lugar a escenarios más complejos para la detección.

Si bien se pretende dar flexibilidad en cuanto a los parámetros del audio de entrada, el diseño del sistema ha sido orientado especialmente hacia el formato de los ficheros de la base de datos.

Durante el desarrollo y las pruebas del sistema han sido de utilidad las etiquetas de música y de voz. Se han creado herramientas para facilitar la lectura de estas etiquetas y su uso posterior en MATLAB. Además, las pruebas realizadas con el detector de música han servido para detectar algunas etiquetas erróneas y corregirlas.

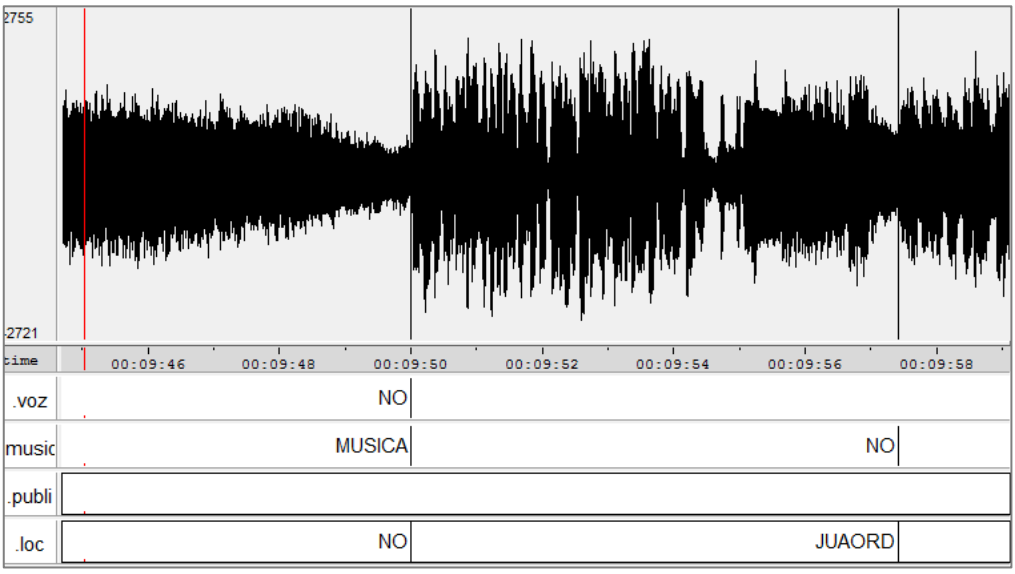


Figura 3-1: Etiquetado de la base de datos en Wavesurfer

3.2 Detector de ritmo

El funcionamiento del detector de ritmo está basado, esencialmente, en la periodicidad de la energía localizada de la señal de audio (apartado 2.2.1). Como mostraba la Figura 2-2, a partir de la energía localizada de una señal podemos obtener información de su periodicidad si calculamos la señal de autocorrelación de su primera diferencia.

El procedimiento básico del detector de ritmo consiste en el cálculo, por ventanas de tiempo, de esta señal de autocorrelación, lo que permite construir lo que llamaremos matriz de autocorrelación de la señal de audio, que contiene la evolución de la autocorrelación respecto al tiempo. En esta matriz esperamos encontrar la información que determine si una señal tiene o no un pulso rítmico a lo largo del tiempo.

En la Figura 3-2 se muestran las gráficas correspondientes a la forma de onda, la energía localizada y la matriz de autocorrelación obtenida para un fragmento de la canción *Paranoid*, de Black Sabbath.

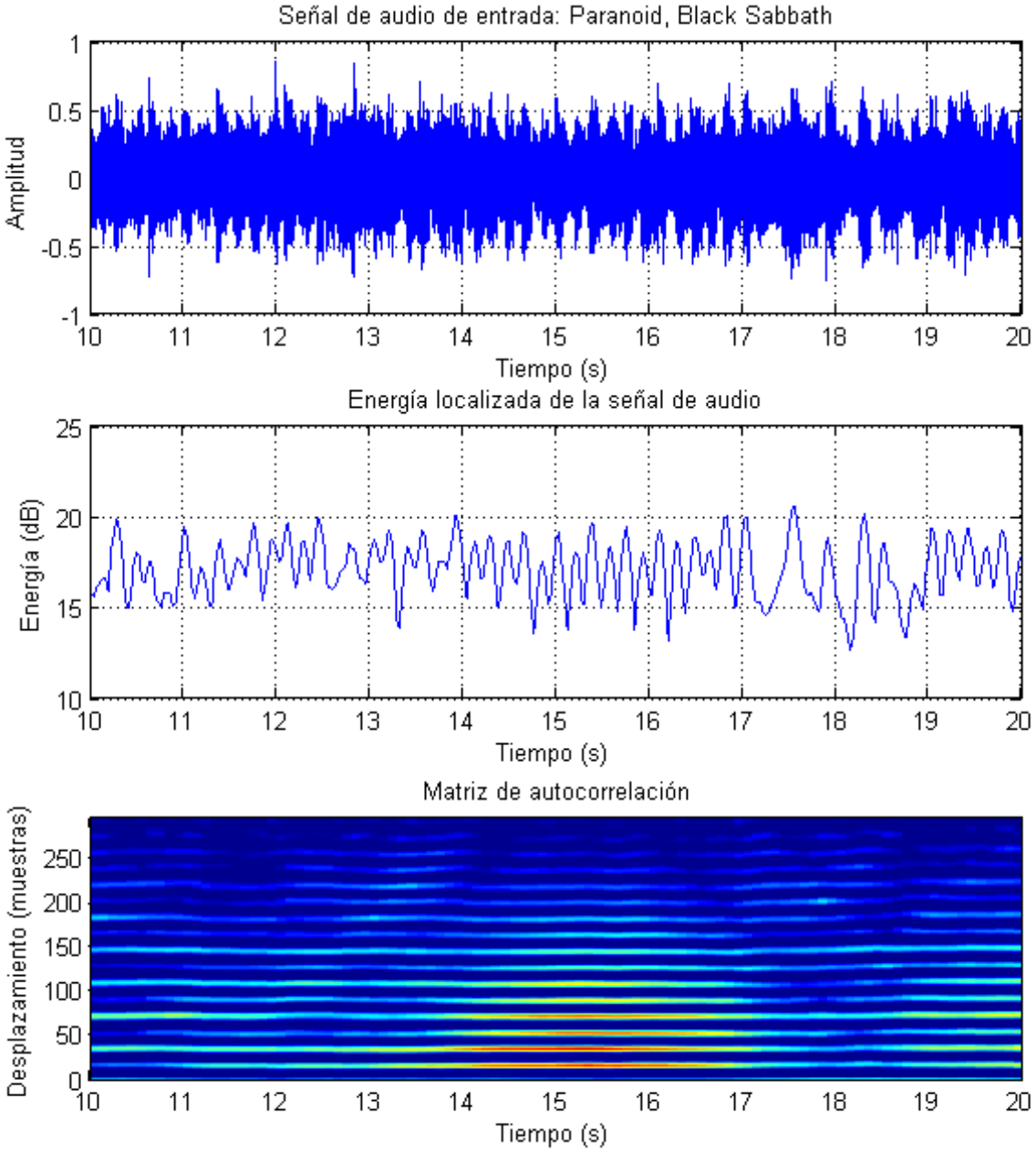


Figura 3-2: Obtención de la matriz de autocorrelación para un extracto de la canción *Paranoid*, de Black Sabbath

El aspecto de la señal de energía localizada indica la presencia de un pulso rítmico bastante marcado. La consecuencia es observable en la matriz de autocorrelación. Para entender la información que nos da esta matriz, debemos tener en cuenta que está construida uniendo como columnas sucesivas funciones de autocorrelación normalizada, descritas en el apartado 2.2.1.

Estas autocorrelaciones han sido calculadas sobre ventanas de la señal de energía localizada (sin olvidar el paso previo de la primera diferencia). Las franjas horizontales que se pueden observar en la matriz son los máximos locales de estas funciones de autocorrelación, que a grandes rasgos mantienen su posición a lo largo del tiempo. Esto nos indica que el pulso rítmico es claro y de tempo constante.

El otro caso a analizar es el de una señal de audio que sólo contenga voz hablada. En principio, la voz hablada no tiene un pulso rítmico marcado, por lo cual esperamos un resultado muy distinto al de la figura anterior (3-2). En la Figura 3-3 podemos ver los resultados para este segundo caso.

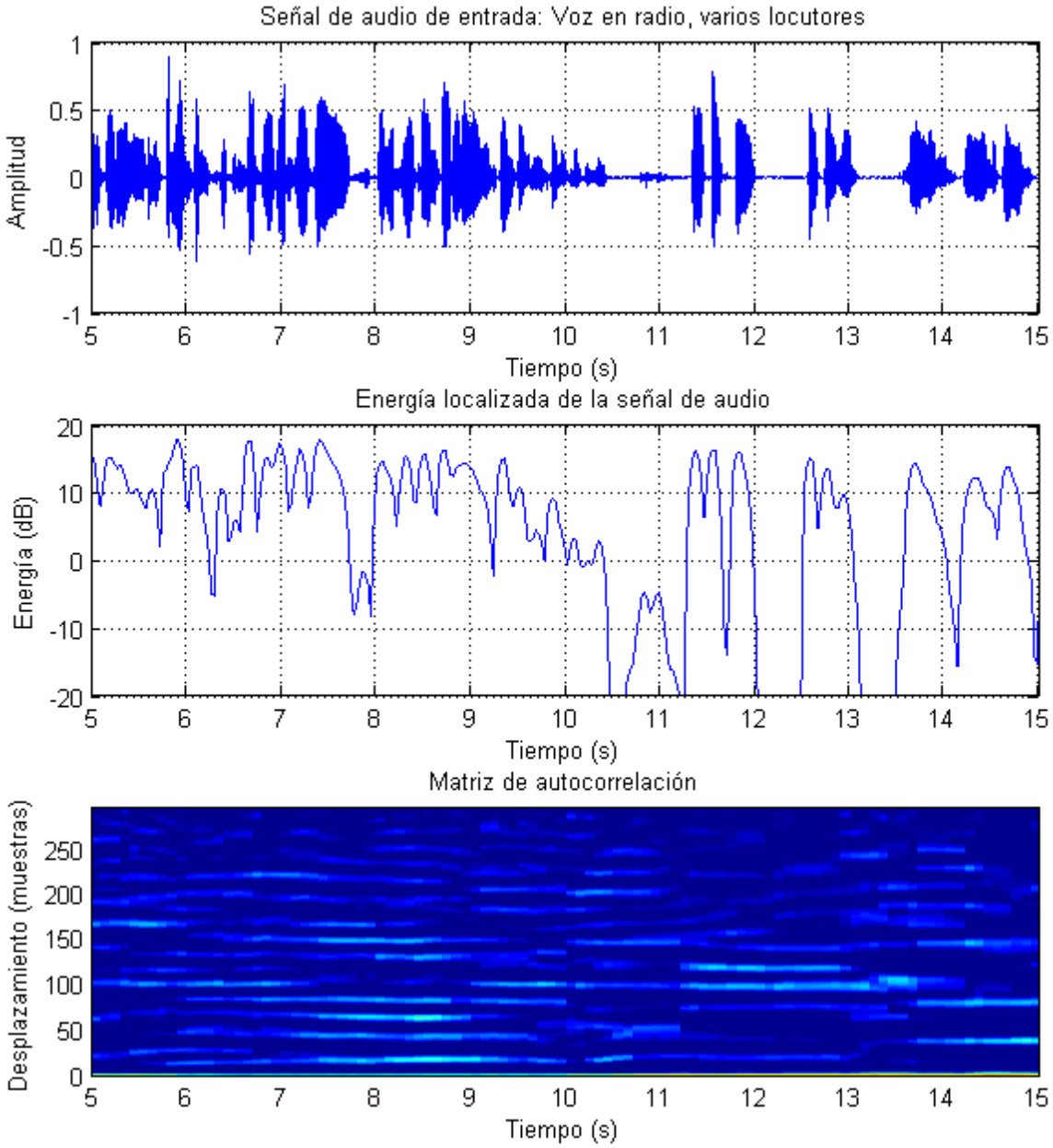


Figura 3-3: Matriz de autocorrelación para un fragmento de audio con voz hablada de varios locutores, tomado de la base de datos de radio

Lo que se observa esta vez es que, aunque la voz hablada no tenga el mismo ritmo que podríamos encontrar en una canción con percusión, sí existen pequeños intervalos de tiempo en los que la energía localizada parece seguir patrones periódicos, creando en algunas zonas de la matriz de autocorrelación algunas franjas horizontales similares a las encontradas en música en la Figura 3-1, si bien estas aparecen con menor regularidad e intensidad.

Uno de los problemas más graves de la obtención de esta matriz era la respuesta obtenida ante pequeñas variaciones de amplitud en la energía, como las que podríamos encontrar por ejemplo en una señal de naturaleza ruidosa. La Figura 3-4 muestra el aspecto de la matriz de autocorrelación obtenida para una señal de ruido blanco Gaussiano.

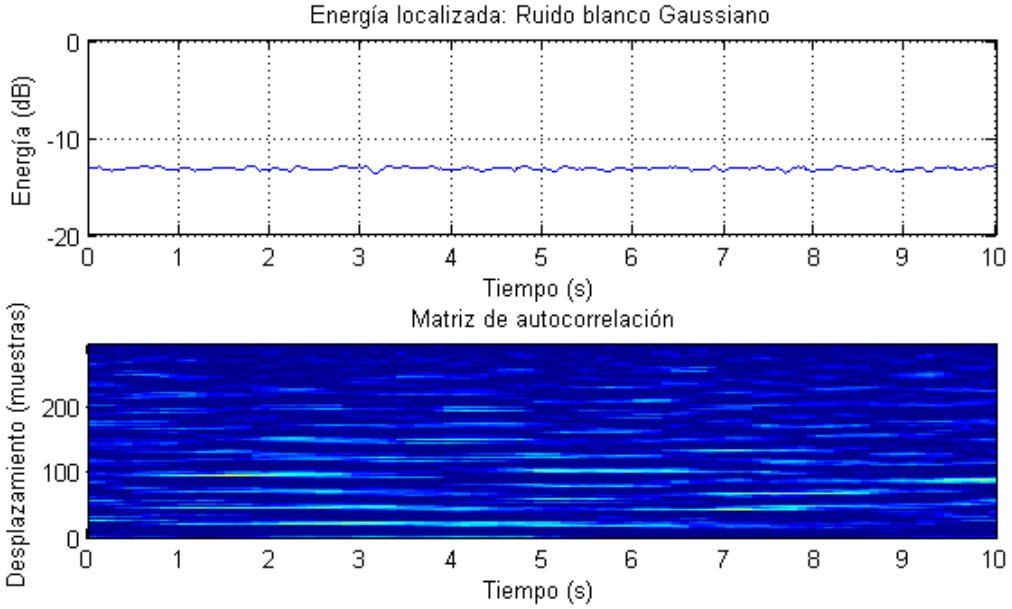


Figura 3-4: Matriz de autocorrelación para una señal de ruido blanco Gaussiano

Hasta ahora, la matriz de autocorrelación muestra un aspecto muy característico ante la presencia de un pulso rítmico marcado, pero no es tan robusta como cabría esperar frente a señales de voz hablada o incluso de ruido, para las que también se obtienen, en ciertos momentos, franjas horizontales.

Un mecanismo sencillo para evitar que las pequeñas variaciones de energía sean interpretadas como ritmo (como ocurre en la Figura 3-4) consiste en determinar un umbral para estas variaciones. Como se ha explicado anteriormente, un paso previo al cálculo de la autocorrelación es aplicar a la señal de energía la operación de primera diferencia. Es tras este paso cuando se puede implementar con mayor facilidad el umbral para las variaciones de energía, que deje intactas variaciones de cierta amplitud, como las visibles en la Figura 3-2, pero anule las variaciones pequeñas que no se corresponden en realidad con la existencia de ritmo.

Para determinar el valor del umbral se ha observado el aspecto que toma la primera diferencia de la energía localizada en señales de música, voz y ruido. Como ejemplo, en la Figura 3-5 se muestran las gráficas correspondientes a algunos segundos de las señales ilustradas en figuras anteriores.

Observando distintos casos, se ha encontrado que un buen valor para el umbral es 0,35 dB. La Figura 3-6 ilustra cómo afecta este umbral a la señal diferencial de energía en los casos de música y de ruido blanco. En las gráficas centrales se han marcado en negro aquellas muestras de la señal cuyo valor absoluto no supera el umbral. Las gráficas inferiores muestran que, efectivamente, las variaciones de energía del ruido blanco se anulan por completo, mientras que en el caso de la señal de música se mantiene la periodicidad.

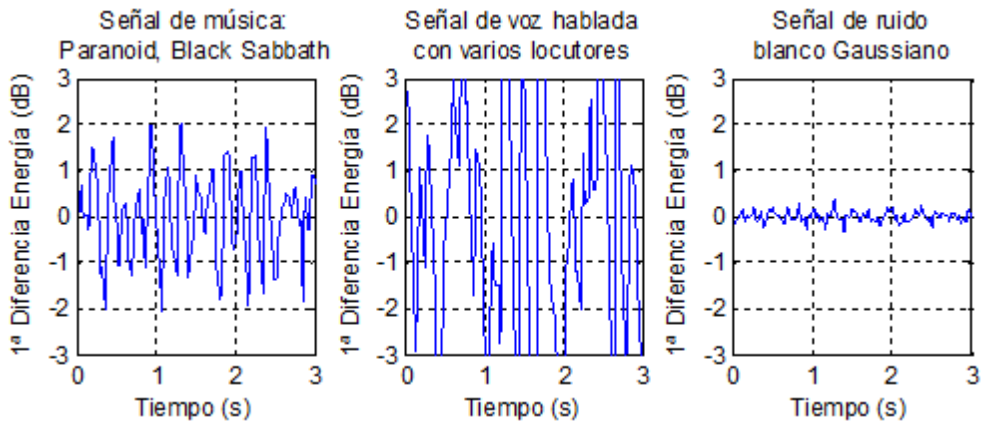


Figura 3-5: Primera diferencia de la energía en señales de música, voz hablada y ruido

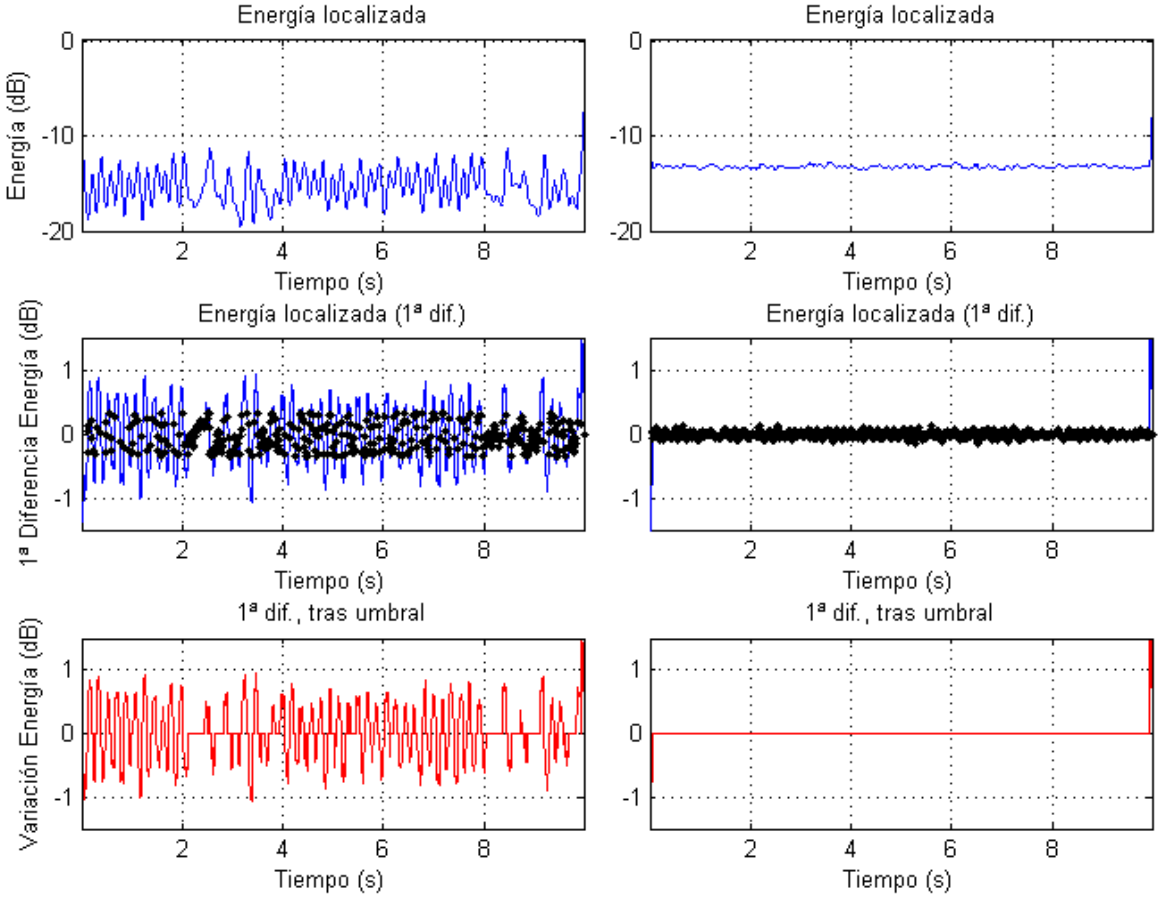


Figura 3-6: Proceso de umbralización de la variación de energía en música (izquierda) y ruido blanco (derecha)

En la Figura 3-5 se aprecia que la estrategia de umbralización, tal y como está planteada, no es de utilidad frente a las variaciones de energía de las señales de voz hablada, que alcanzan mayor amplitud que en el caso de la señal de ruido o de muchas señales de música.

Durante el desarrollo del detector de ritmo, se ha comprobado que una manera de reforzar la robustez de las matrices de autocorrelación consiste en duplicar el proceso: construir una

segunda matriz a partir de la autocorrelación de cada una de las columnas. En esta nueva matriz, las franjas constituidas por picos en las funciones de autocorrelación son todavía más claras y permiten un seguimiento más fiable, como se verá a continuación.

Sobre esta nueva matriz de autocorrelación, de cara a detectar la presencia de pulso rítmico podemos establecer algunos criterios:

a) Existen máximos locales marcados en cada columna.
b) Estos máximos locales mantienen su posición a lo largo del tiempo.
c) Los máximos de cada columna se encuentran en posiciones equidistantes.

Los dos primeros criterios, como se ha ido viendo a lo largo de este capítulo, responden a la existencia de periodicidad en la energía localizada de la señal. El tercer criterio, sin embargo, se introduce tras observar una excepción en la continuidad de los máximos.

Se ha observado que la matriz de autocorrelación de ciertas canciones muestra la aparición y desaparición de máximos en algunas zonas. También se ha comprobado, mediante la escucha de los ficheros de audio en cuestión, que la aparición y desaparición de estos máximos coincide con cambios de sección en la canción (por ejemplo, de estrofa a estribillo) en los que la separación temporal entre acentos rítmicos (típicamente, golpes de plato en la batería) se duplica o se reduce a la mitad.

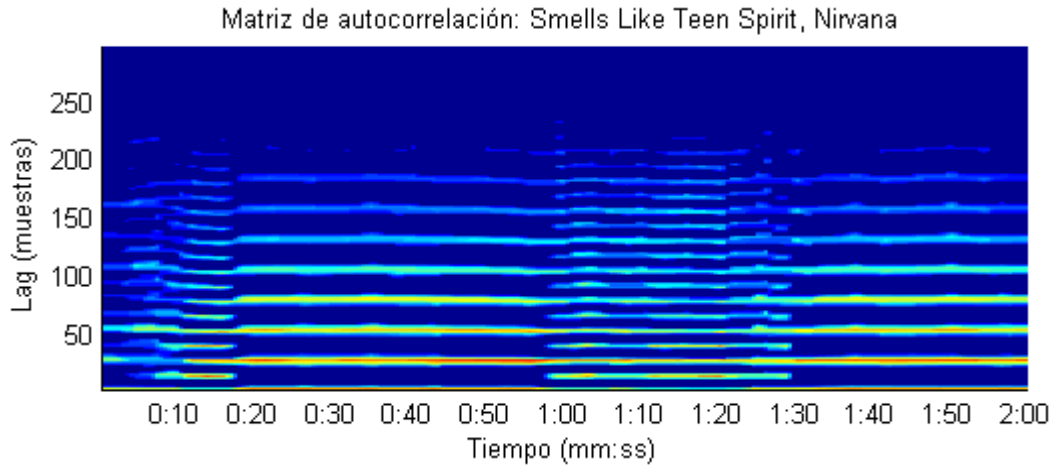


Figura 3-7: Matriz de autocorrelación para un fragmento de la canción *Smells Like Teen Spirit,* de Nirvana

En la Figura 3-7 se aprecia un ejemplo de este fenómeno. Las zonas que presentan la mitad de franjas (0:18 – 0:58 y 1:30 – 2:00) coinciden con las estrofas de la canción, mientras que el resto del tiempo se corresponde con los estribillos, en los que los acentos rítmicos duplican su velocidad.

Este recurso, bastante habitual en canciones de distintos estilos, invalidaba el planteamiento inicial de seguimiento de máximos. La solución que se ha encontrado es seguir a lo largo de la matriz los tres primeros máximos y también la relación (cociente) entre sus posiciones en la columna. En la siguiente figura (3-8) se muestra el resultado del seguimiento de los máximos para la misma matriz de autocorrelación.

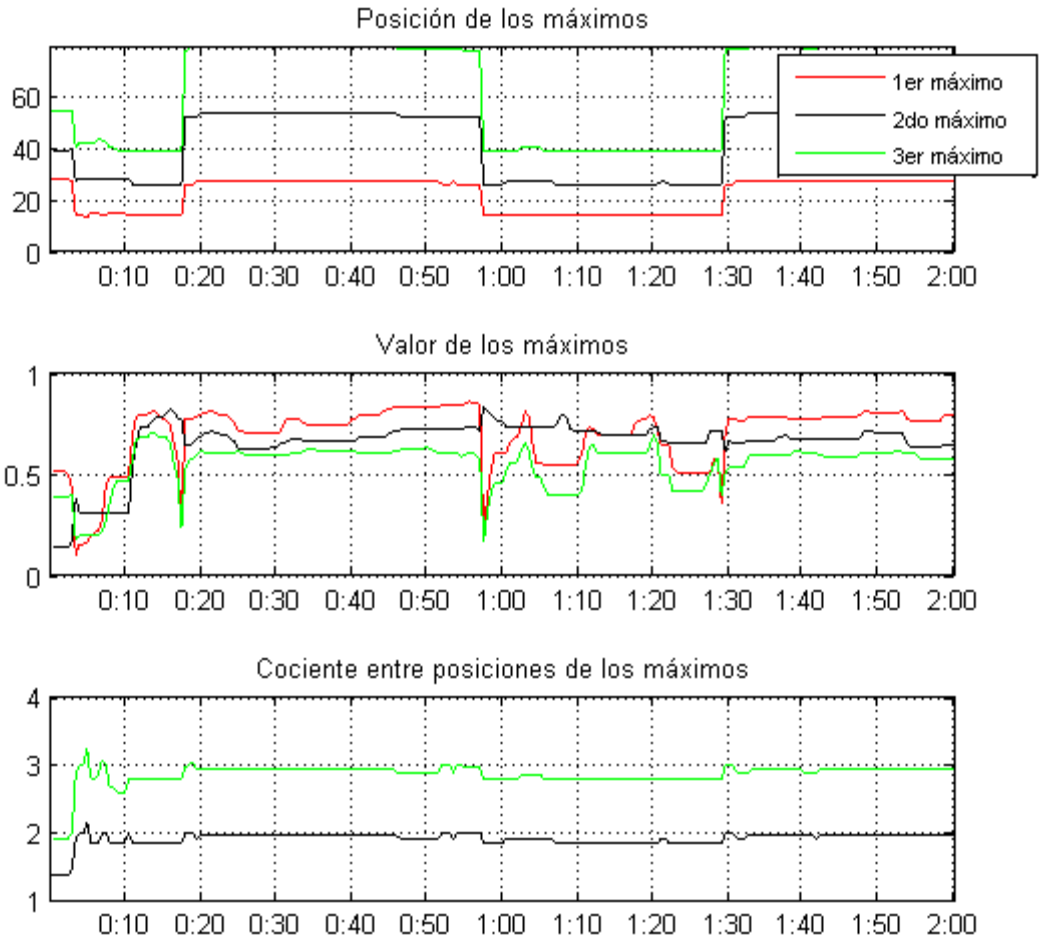


Figura 3-8: Seguimiento de máximos a lo largo de una matriz de autocorrelación

La gráfica inferior de la Figura 3-8 muestra en negro el cociente entre las posiciones del segundo y el primer máximo y en verde el cociente entre el tercero y el primero. Asumiendo que los máximos aparecerán en posiciones equidistantes, los valores esperados para estos cocientes son, respectivamente, 2 y 3. La figura también nos confirma que el cálculo de cocientes es una buena solución frente al problema que se nos planteaba, ya que no se ven afectados por la aparición y desaparición de máximos.

Para combinar toda esta información en una señal de puntuación o *score* que evalúe la presencia de ritmo en la señal, se han utilizado valores basados en las siguientes características:

- Continuidad de las posiciones de los tres primeros máximos en un intervalo corto de tiempo (3 segundos). Se calcula mediante la desviación típica de los cocientes entre posiciones en dicho intervalo, para evitar el problema visto en la Figura 3-7.
- Valores de amplitud de los tres primeros máximos y desviación típica de todos los valores de cada columna. La aparición de máximos muy marcados da lugar a una desviación típica mayor.
- Acercamiento de los cocientes entre posiciones de los máximos a sus valores esperados (2 y 3), calculado mediante distribuciones normales centradas en dichos valores.

$$Punt_{ritmo} = log_{10}(P_{máximos} + P_{cocientes} + P_{desv.típ.})$$

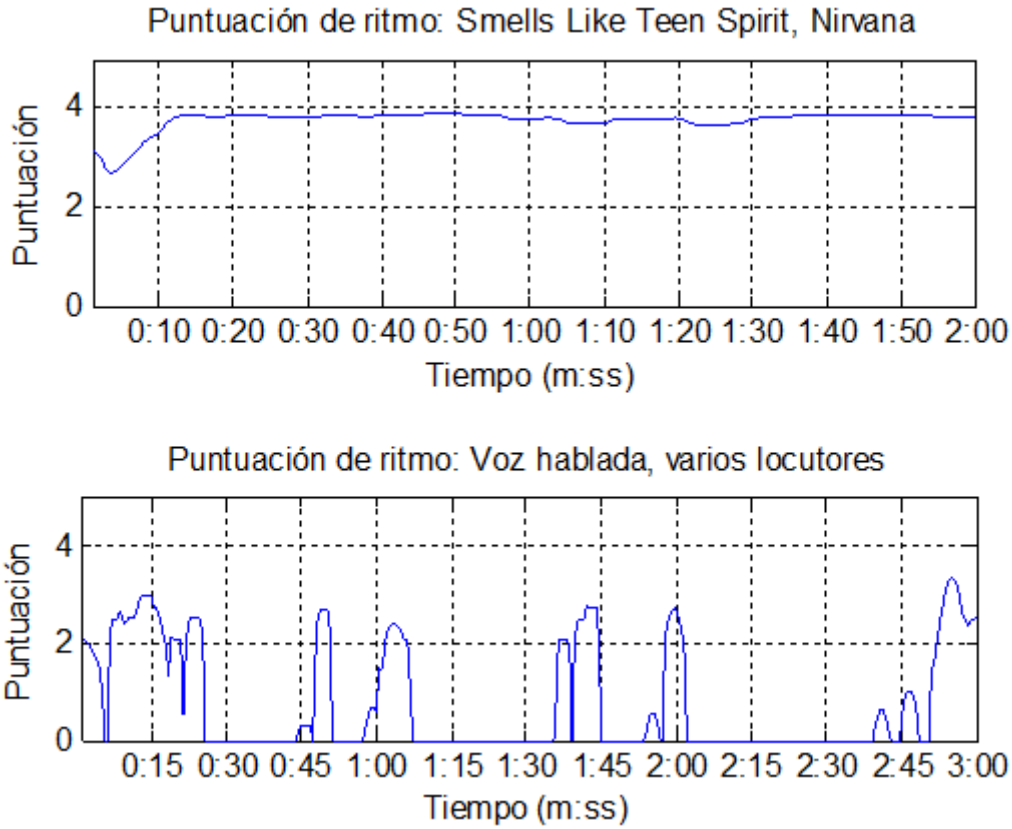


Figura 3-9: Puntuación de ritmo obtenida para música (arriba) y voz hablada (abajo)

El detector arroja algunos picos de puntuación durante señales de voz hablada. Analizando varias ocasiones en las que sucede esto, se ha visto que a veces estos picos están originados simplemente por golpes de voz alineados en intervalos regulares de tiempo, es decir, un fraseo más rítmico de lo habitual. En otros casos, se ha visto que la risa de un locutor también puede crear estas puntuaciones rítmicas altas.

Como, en efecto, se trata de presencia de ritmo, podemos considerar que no es un fallo de diseño del detector. Sin embargo, esto dificulta en gran medida determinar la presencia de un ritmo musical que aparezca de fondo en una señal de voz hablada, al menos basando nuestro detector en la energía localizada de la señal. Para hacer frente a este problema, el detector de ritmo incluye la posibilidad de filtrar la señal con un filtro banda eliminada diseñado con la herramienta FDATool. Este filtro trata de anular las frecuencias en las que más presencia tiene la voz respecto a los componentes rítmicos de la música, la banda de 100 a 7000 Hz.

Este filtrado mejora en muchos casos la detección de ritmo cuando existe solapamiento con voz. El diseño está orientado a señales muestreadas a 16kHz, en las que la máxima frecuencia que encontramos es 8kHz. Sin embargo, la voz hablada tiene componentes con frecuencias entre los 7 y 8kHz [5], que quedan intactas tras este filtro. En señales con mayor frecuencia de muestreo, como 22,05kHz, sería posible anular de forma más completa la voz.

Como el filtrado puede deteriorar la detección de ritmo en música limpia, el detector de ritmo cuenta con la posibilidad de calcular una puntuación para el audio filtrado y otra para el audio original, siendo la puntuación final el máximo de ambas en cada instante de tiempo.

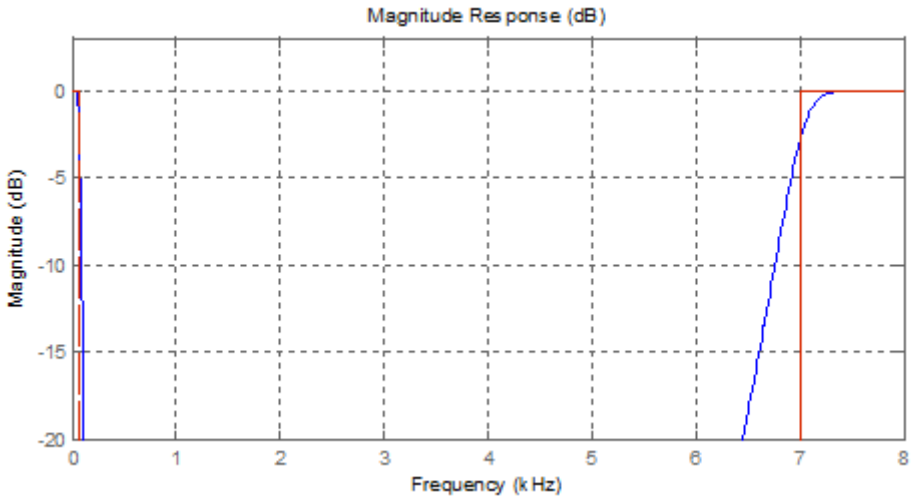


Figura 3-10: Módulo de la respuesta en frecuencia del filtro diseñado

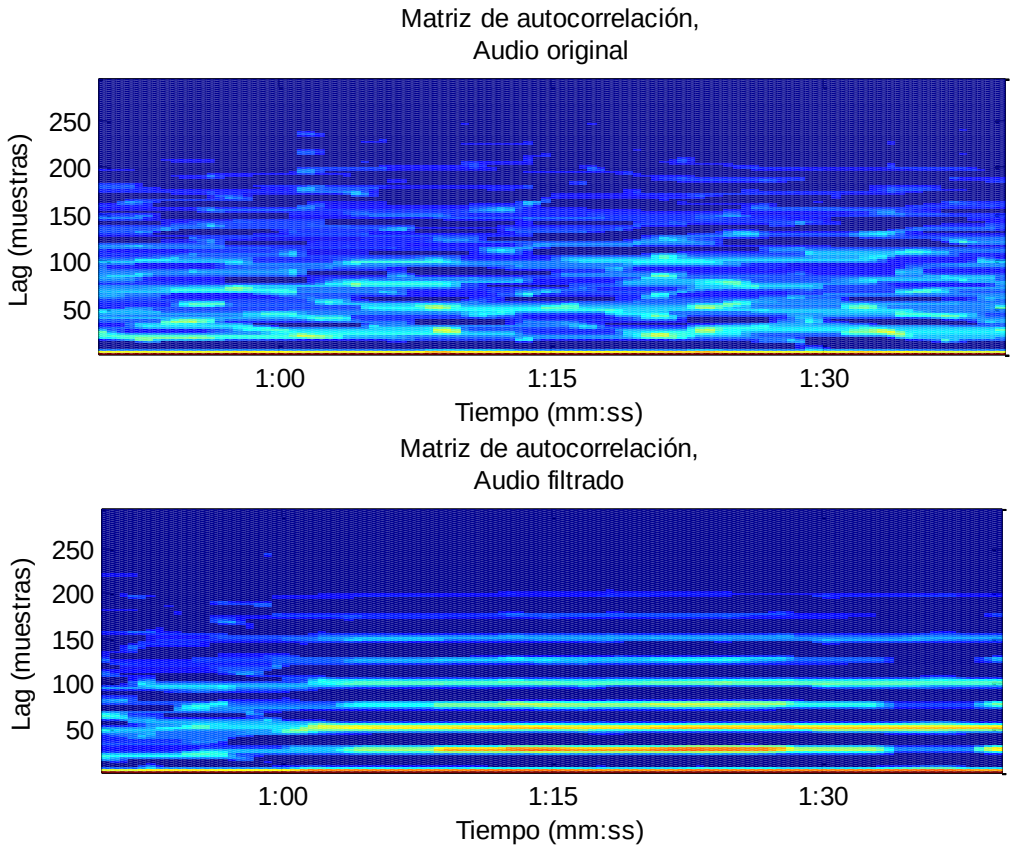


Figura 3-11: Matrices de autocorrelación para un mismo fragmento de audio (solapamiento de voz y ritmo), antes y después de aplicar el filtrado

3.3 Detector de armonía

El propósito del detector de armonía es otorgar a una señal de audio dada una puntuación que evalúe la existencia de una ordenación musical en su espectro de frecuencias. La siguiente figura trata de ilustrar la diferencia en la organización del espectro en una señal musical y otra de voz hablada.

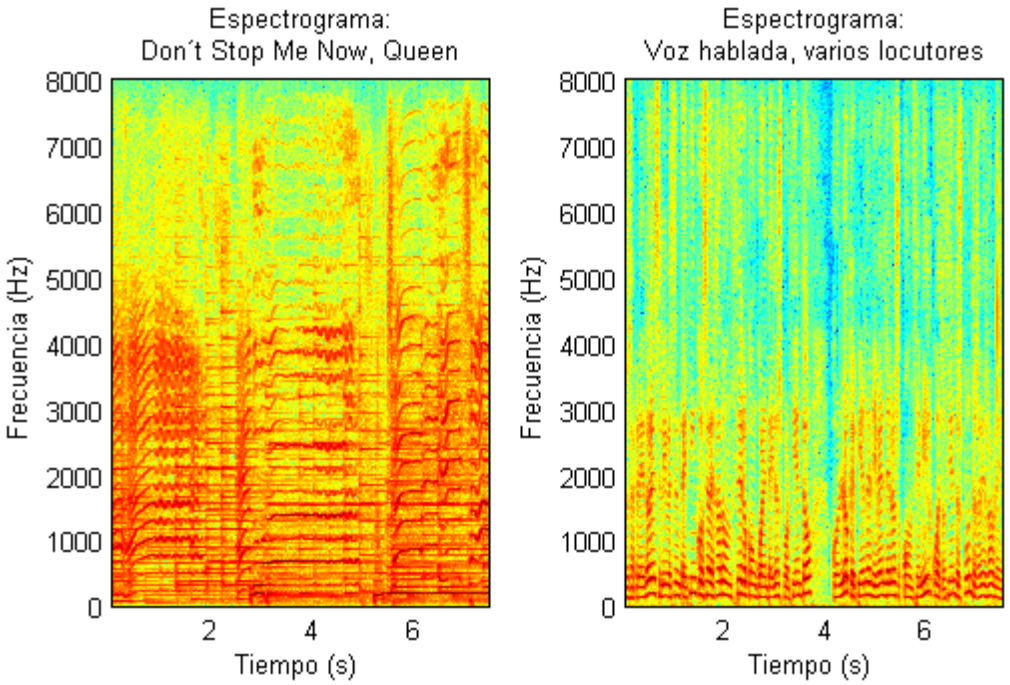


Figura 3-12: Espectrogramas para una pieza musical (izquierda) y voz hablada (derecha)

Recordando el apartado 2.2.2 de esta memoria, la información contenida en los espectrogramas permite obtener los llamados cromagramas, que resumen el espectro entero de la señal en doce valores, cada uno de ellos correspondiente a una nota musical. Esto es posible gracias a que existe una correspondencia directa[1] entre cada nota musical y la frecuencia fundamental que representa.

La Figura 3-13 muestra las representaciones en forma de cromagrama correspondientes a los espectrogramas de la Figura 3-12. La presencia de armonía (o su ausencia) se vuelve muy notable en los cromagramas. Mientras el cromagrama de la izquierda, correspondiente a un fragmento de música, deja clara una distribución del espectro en torno a ciertas notas musicales, el cromagrama de la derecha, extraído de una señal de voz hablada, no presenta esta característica.

La frecuencia fundamental (o *pitch*) de la voz hablada varía de forma continua según la entonación de cada frase y, aunque puede coincidir con la frecuencia correspondiente a una nota musical, generalmente lo hace de manera circunstancial y sin continuidad en el tiempo.

[1] La norma ISO 16:1975 establece como estándar de afinación musical la correspondencia entre 440Hz y la nota La_4.

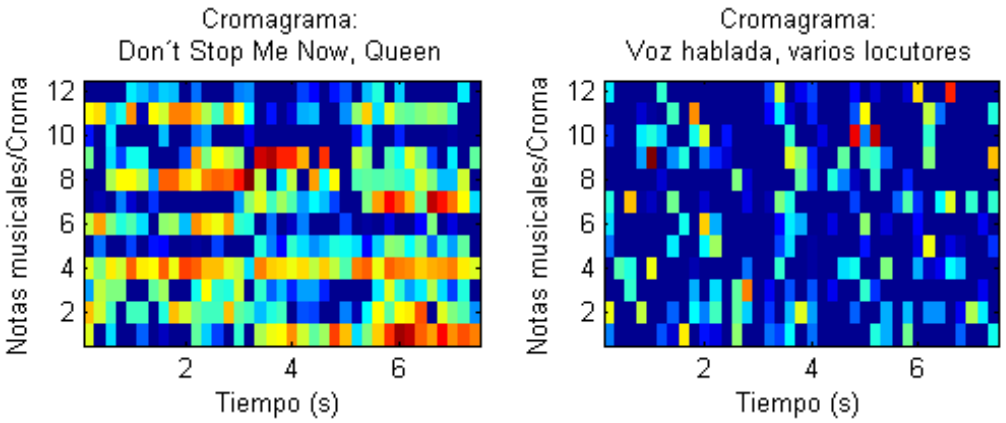


Figura 3-13: Cromagramas para una pieza musical (izquierda) y voz hablada (derecha)

El proceso ideado para extraer la información de presencia de armonía separa, a lo largo del cromagrama, las doce notas en "activas" e "inactivas". Para ello, busca el valor mínimo y el máximo de cada columna y establece un umbral entre ambos. A continuación, calcula la media de los valores que superan el umbral y la de aquellos que quedan por debajo: la distancia entre los dos valores medios obtenidos es un indicador de la presencia de armonía.

Además, de manera previa al análisis de las columnas, se aplica un procesado al cromagrama con el fin de atenuar los picos que no muestran continuidad temporal (visibles en la Figura 3-13, derecha) y reforzar aquellos que se mantienen durante cierto tiempo. Este acondicionamiento se lleva a cabo mediante la operación apertura, que combina los operadores erosión y dilatación para eliminar máximos locales. En la figura 3-14 se ilustra el proceso sobre cromagramas de música y voz, utilizando un código de colores común para resaltar el efecto en las amplitudes (mostradas en escala logarítmica).

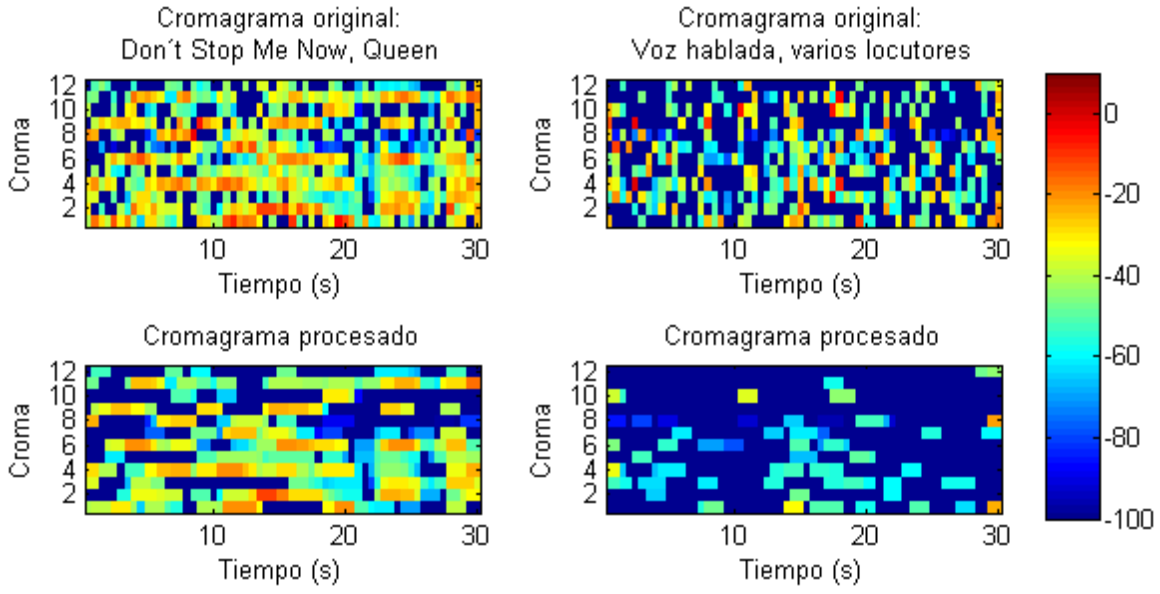


Figura 3-14: Procesado de cromagrama mediante la operación apertura

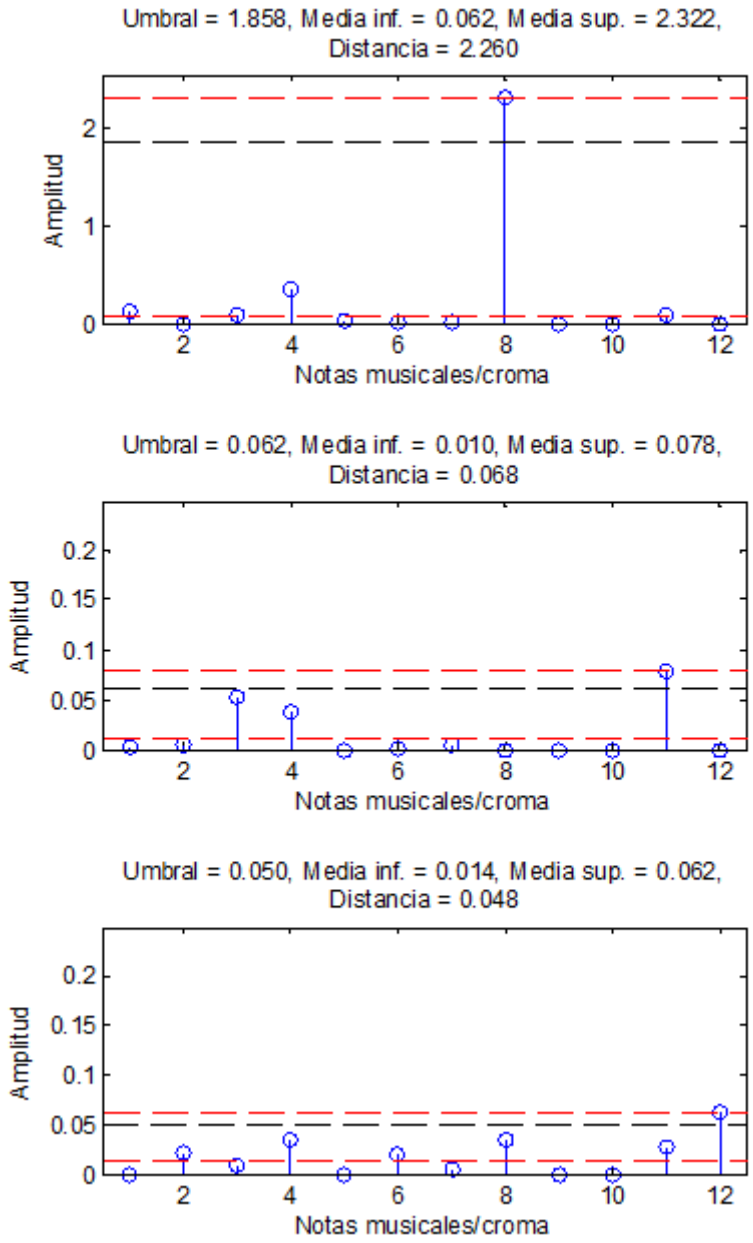


Figura 3-15: Detalle del análisis de columnas de un cromagrama en señales de música (gráfica superior), voz (central) y ruido blanco (inferior)

La Figura 3-15 muestra más en detalle el análisis por columnas de los cromagramas. Se muestran los doce valores de la columna en unidades naturales y se marcan mediante líneas discontinuas el umbral (en negro) y las medias de los valores superiores e inferiores (en rojo).

El umbral se ha calculado como:

$$Umbral = min(col) + 0{,}8 \cdot [\max(col) - \min(col)]$$

Donde *col* es el vector de doce elementos que contiene los valores hallados en la columna del cromagrama.

Aplicando este análisis a señales de música, de voz hablada y de ruido, se han sacado ciertas conclusiones:

a. La distancia entre medias es un buen indicador de la presencia de melodía y armonía, dando valores mayores en secciones instrumentales o sin percusión.

b. Las señales de música no suelen mostrar más de tres o cuatro notas activas (que superan el umbral) simultáneamente. En señales de voz y, sobre todo, de ruido, es más común encontrar un mayor número de notas simultáneas. En la figura 3-16 se muestran algunos resultados gráficos obtenidos al calcular el número de notas activas a lo largo de los cromagramas de distintos ficheros.

c. El movimiento del cromagrama, esto es, el número de notas que se activan o desactivan de una columna a la siguiente, muestra un comportamiento similar al del número de notas activas. Es un resultado razonable, ya que las notas activas en una señal musical con armonía tienden a mantenerse durante cierto periodo de tiempo. Además, en ocasiones se observa, en señales de música, que el movimiento tiene picos en intervalos regulares de tiempo.

Una aproximación para cuantificar el contenido armónico del cromagrama puede basarse, entonces, en la distancia entre medias, y adicionalmente valorar positivamente que el número de notas activas y el movimiento del cromagrama no sea elevado.

En pruebas posteriores, se detectó un problema adicional: ocurre al analizar señales que, pese a no ser musicales en absoluto, tienen un *pitch* constante y se mantienen en el tiempo. Es el caso de los tonos de señalización de las llamadas telefónicas o de ruidos de banda estrecha. Para tratar estos casos, la solución que se ha hallado es hacer uso de la información de movimiento del cromagrama.

Aunque en las conclusiones antes enumeradas se indicaba que un bajo movimiento del cromagrama es típico en las señales de música, un intervalo de tiempo demasiado largo con un movimiento de cromagrama especialmente bajo puede indicar que estamos ante una señal del tipo que se describe en el párrafo anterior. Sin embargo, también es posible que se trate de música, ya que existen secciones en las que la armonía se mantiene constante (por ejemplo, una nota final que se prolongue en el tiempo).

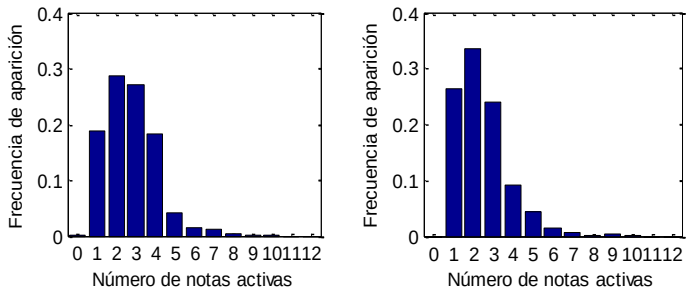
De esta manera, al encontrar un caso así puede ser de utilidad la información de contexto: si alrededor de la zona en la que encontramos un cromagrama "estático" no hemos detectado armonía (también podríamos recurrir a la información de ritmo), lo más probable es que no se trate de música. Incluso si no estuviéramos ante un ruido o ante un tono de llamada, sino que la señal fuera, por ejemplo, una nota aislada de piano en un entorno no musical, tendría sentido que el detector diese una respuesta negativa.

Incorporando todos estas observaciones a la expresión de puntuación final, se obtiene:

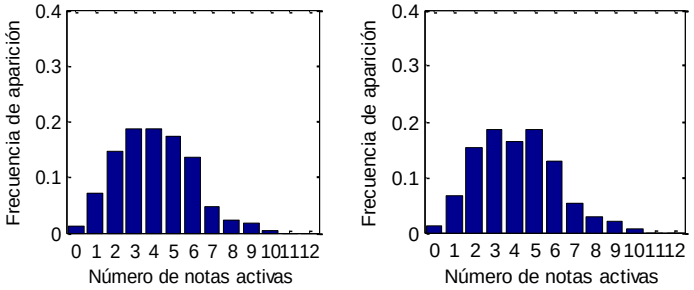
$$Punt_{armonía} = log_{10}(Variabilidad \cdot P_{dist.medias} + P_{movimiento})$$

Donde *Variabilidad* representa un valor binario (0 ó 1) que se desactiva si el movimiento del cromagrama se mantiene por debajo de 2 durante más de 5 segundos.

i)



ii)



iii)

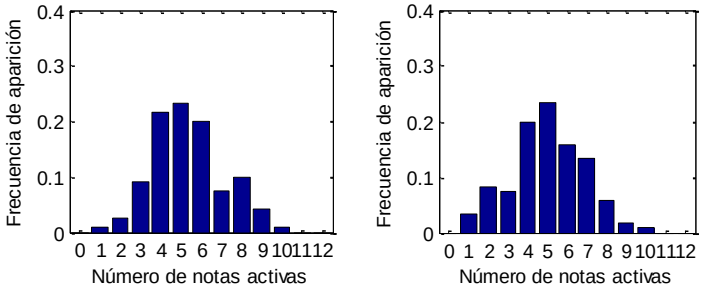


Figura 3-16: Histogramas de número de notas activas para distintos cromagramas de música (i), voz hablada (ii) y ruido (iii)

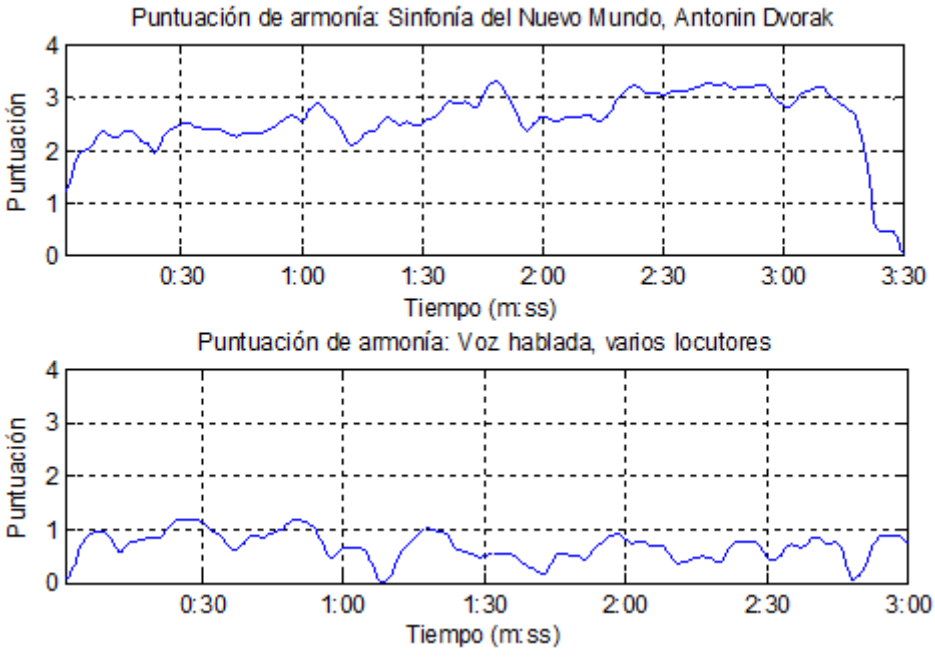


Figura 3-17: Puntuación de armonía obtenida para música (arriba) y voz hablada (abajo)

3.4 Combinación de los detectores

El trabajo individual del detector de ritmo y el de armonía nos puede dar una respuesta parcial respecto a la presencia de música en una señal de audio. Sin embargo, al unir los resultados de cada uno de los detectores cubrimos un número mayor de casos: podemos encontrar señales musicales con poca intensidad rítmica o, al contrario, música en la que la armonía no juega un papel relevante.

De cara a combinar ambos detectores, una herramienta de utilidad son las gráficas de dispersión (*scatter plots*). Como cada uno de los detectores nos devuelve una señal temporal con su evaluación de ritmo o armonía (según el caso), en una gráfica de dispersión bidimensional podemos representar el valor de ambas puntuaciones a lo largo del tiempo.

Las puntuaciones individuales de los detectores de ritmo y croma están sincronizadas: se da un valor cada medio segundo de señal. Para mostrar el funcionamiento combinado de los detectores, se muestran gráficas de dispersión en señales de voz, de ruido y de música de distintos estilos: pop, acústica, rock, electrónica, hip-hop y sinfónica.

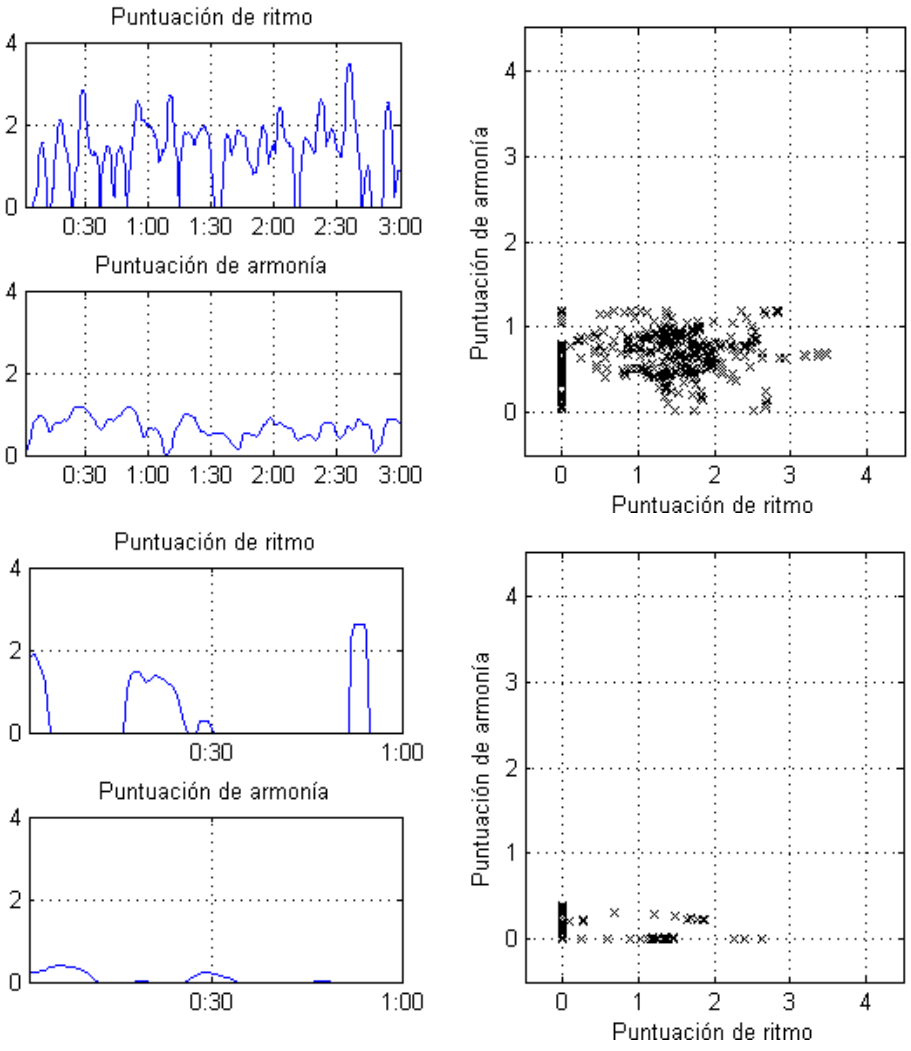


Figura 3-18: Gráficas de dispersión para una señal de voz (superior) y una señal de ruido rojo (inferior)

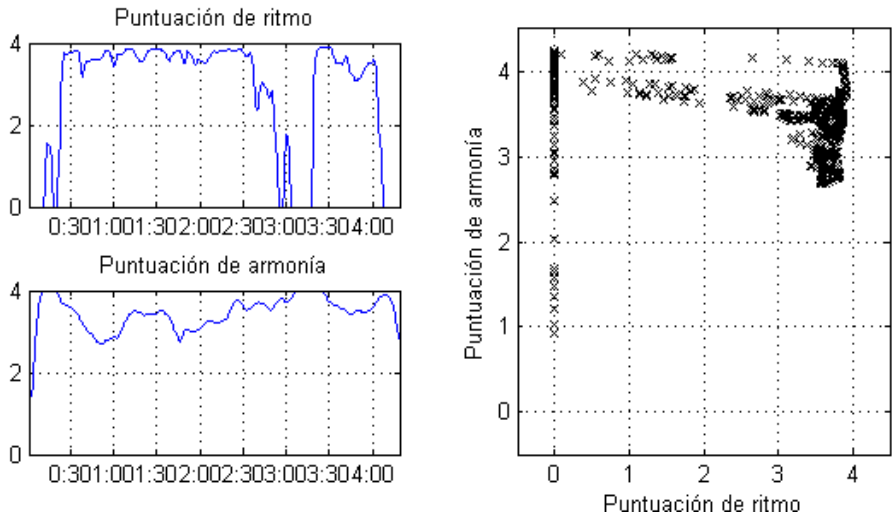


Figura 3-19: Gráfica de dispersión para la canción *Clocks*, de Coldplay (pop)

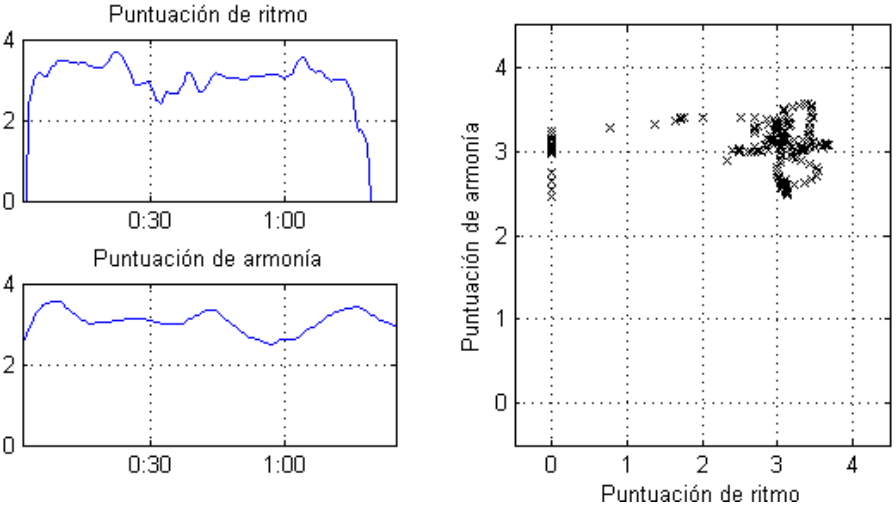


Figura 3-20: Gráfica de dispersión para la canción *Pigs on the Wing*, de Pink Floyd (folk /acústica)

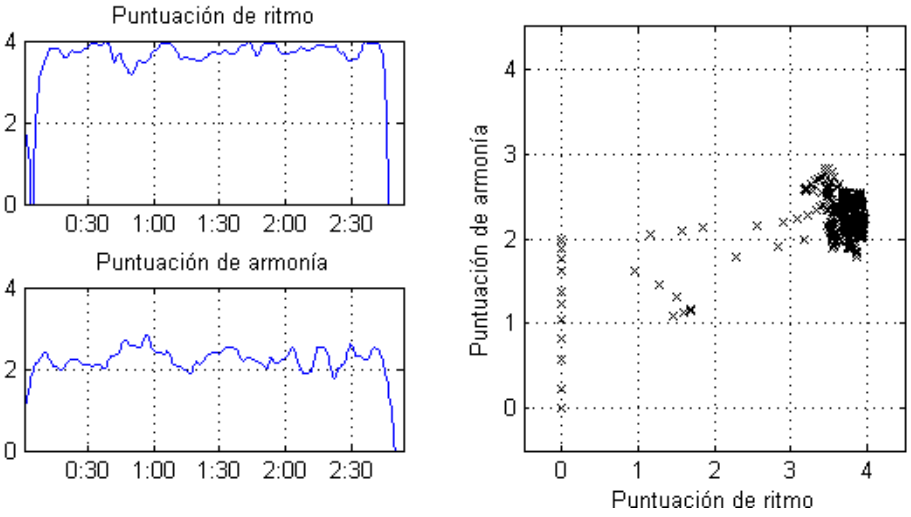


Figura 3-21: Gráfica de dispersión para la canción *Paranoid*, de Black Sabbath (rock)

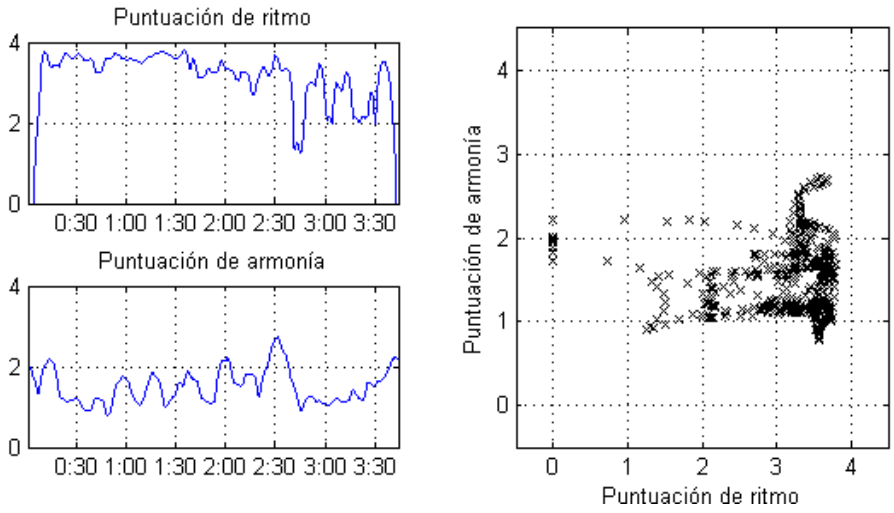


Figura 3-22: Gráfica de dispersión para la canción *Harder, Better, Faster, Stronger*, de Daft Punk (electrónica)

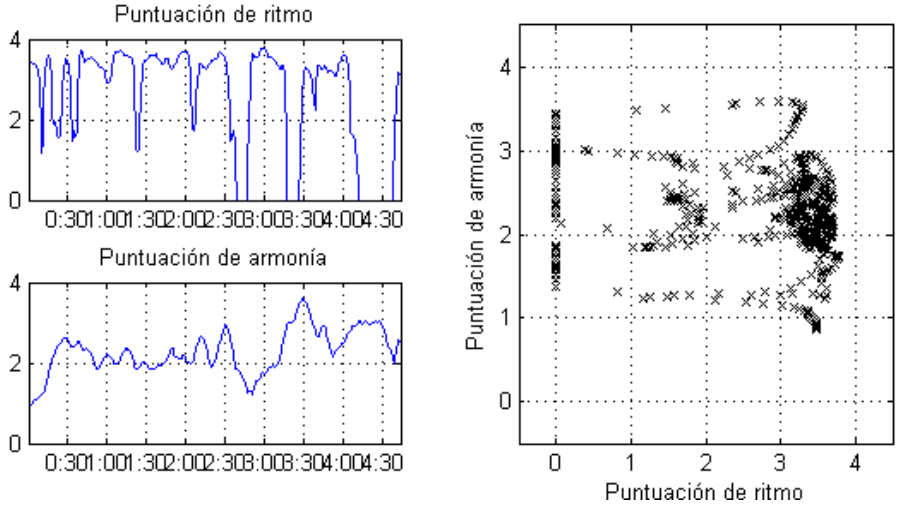


Figura 3-23: Gráfica de dispersión para la canción *Power*, de Kanye West (hip-hop)

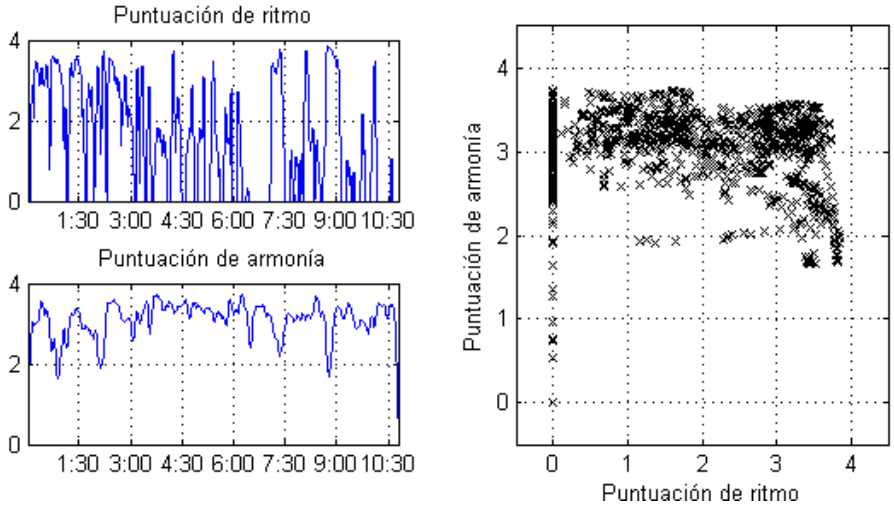


Figura 3-24: Gráfica de dispersión para la *Primavera*, de Antonio Vivaldi (sinfónica/clásica)

Las gráficas de dispersión dejan ver una buena versatilidad del detector combinado en cuanto a estilos musicales. Es razonable que no en todos los casos se obtengan puntuaciones similares: por ejemplo, en música clásica (Figura 3-24) la componente armónica tiene mucho más peso que el ritmo. Sucede al contrario en otras señales analizadas, pertenecientes a estilos en los que la armonía no está tan marcada (Figuras 3-22, 3-23).

Además, durante el transcurso de las canciones analizadas se han encontrado algunas secciones sin ritmo, o de ritmo menos marcado. En estas secciones el detector de armonía no se ve afectado (siempre que la armonía siga presente), incluso llega a dar puntuaciones más altas.

En general, el comportamiento observado al aplicar el detector a señales de música es que, pese a los cambios en las puntuaciones de armonía y ritmo que se suceden a lo largo de la señal, solemos encontrar un valor alto para alguna de sus componentes, o incluso para ambas.

Para combinar de manera efectiva ambos detectores, y en base a las observaciones sobre los distintos ficheros, se calcula una puntuación unidimensional basada en la distancia euclídea de cada punto al origen de la gráfica de dispersión.

$$distancia[n] = \sqrt{\alpha_c \cdot Punt_c{}^2[n] + \alpha_r \cdot Punt_r{}^2[n]}$$

Donde $Punt_c[n]$ y $Punt_r[n]$ son las puntuaciones de armonía (cromaticidad) y ritmo obtenidas por los detectores, y α_c y α_r son dos factores de ajuste. El propósito de los factores de ajuste es ponderar con distinto valor a los ejes de armonía y ritmo, a la vista de que la voz hablada puede alcanzar por sí sola un rango de puntuaciones mayores en ritmo que en armonía.

La decisión tomada por el detector será el resultado de umbralizar la puntuación de distancia en un valor *U*, que se decidirá en el capítulo de Pruebas y Resultados (Figura 4-2). Este valor de umbral nos permite formular ya las expresiones para la decisión binaria (decisión "*hard*": valores 0 ó 1) y para un vector de fiabilidad, resultante de normalizar la puntuación de distancia entre el umbral (decisión "*soft*").

$$Mus[n] = \begin{cases} 1, & distancia[n] \geq U \\ 0, & distancia[n] < U \end{cases}$$

$$Fiab[n] = \frac{distancia[n]}{U}$$

A la vista de las dos fórmulas anteriores, una condición equivalente para la decisión binaria es que la fiabilidad sea mayor o igual que 1.

Se puede llevar a cabo un proceso adicional sobre la decisión del detector, utilizando el valor de fiabilidad. Se trata de una doble umbralización que toma como umbrales de fiabilidad los valores 1 y 0,7 (umbrales fuerte y débil, respectivamente).

El objetivo es considerar que sigue habiendo música en los intervalos temporales cuyos valores de fiabilidad queden situados entre los dos umbrales, siempre y cuando se encuentren intercalados entre zonas con fiabilidad mayor o igual que 1. El uso de doble umbral mejora la detección en situaciones en las que la voz oculta casi por completo la música, que son frecuentes en programas de radio.

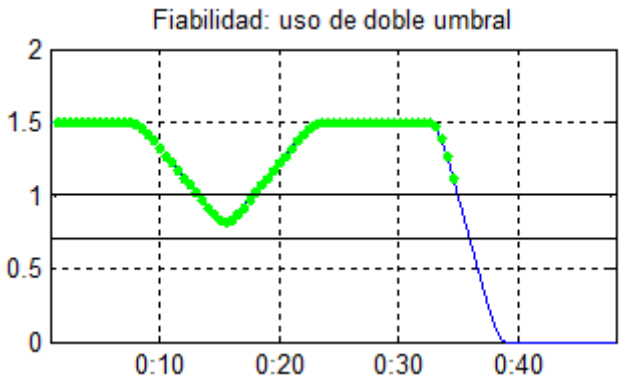


Figura 3-25: Uso de doble umbralización sobre el valor de fiabilidad

La figura ilustra el funcionamiento de la doble umbralización. Los dos umbrales están marcados con rectas horizontales negras, en los valores 1 y 0,7. A lo largo del eje temporal se representa el valor de fiabilidad, marcando en verde los valores que se corresponden con una decisión positiva del detector.

Entre los 10 y 20 segundos aparece un tramo en el que la fiabilidad entra en la zona comprendida entre los dos umbrales. Como no se llega a atravesar el umbral débil, la decisión sigue siendo positiva. No ocurre lo mismo hacia el segundo 35, ya que el descenso de la fiabilidad llega hasta valores inferiores a 0,7.

4 Pruebas y resultados

Las pruebas realizadas sobre el detector de música se dividen en tres fases. En la primera de ellas, se comprueba el funcionamiento del detector en el caso más sencillo, un entorno experimental "limpio" en el que las señales de música y de voz no están solapadas.

La segunda fase de pruebas se ha creado un entorno controlado superponiendo intervalos de música (ritmo y/o armonía) a una de las señales de audio de voz hablada, con relaciones de amplitud entre ellas (SNR) conocidas.

Por último, se han realizado pruebas en un entorno real, con fragmentos de la base de datos ATVS-Radio en los que se suceden secciones con ausencia y presencia de música, ya sea limpia o solapada con voz.

4.1 Pruebas en entorno limpio

Con ayuda de la base de datos ATVS-Radio, se ha creado un conjunto de prueba con 72 minutos de música y 72 minutos de voz hablada por varios locutores. A continuación se representan las gráficas de dispersión obtenidas para música y voz:

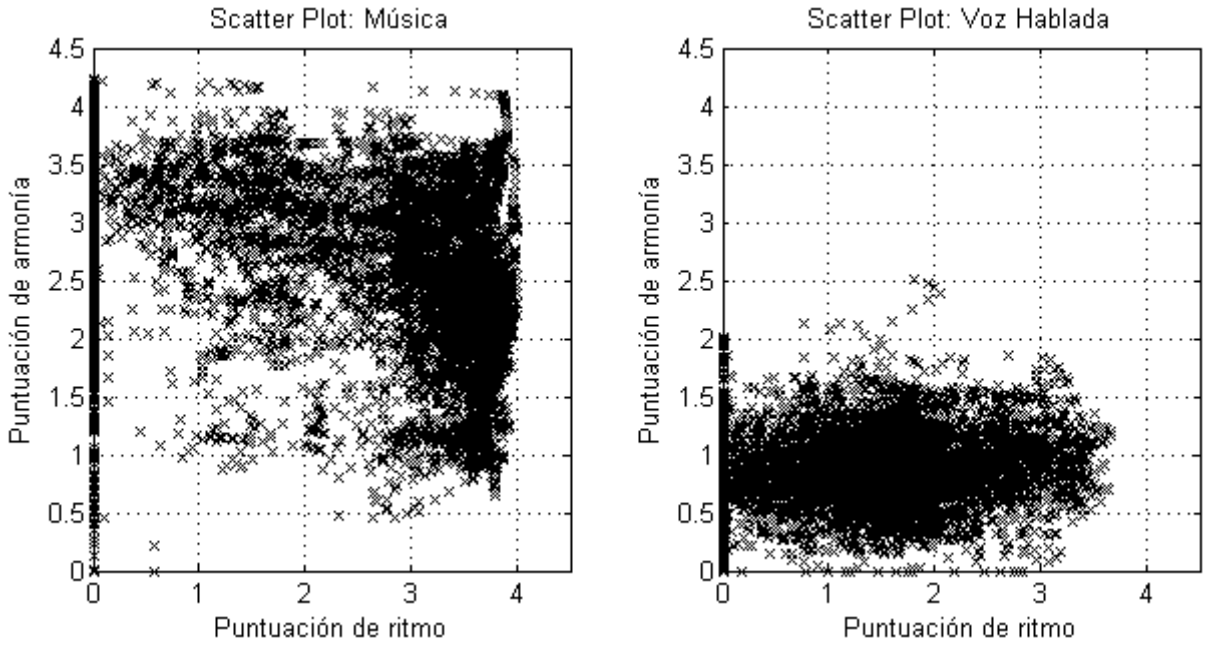


Figura 4-1: Gráficas de dispersión para 72 minutos de música y 72 minutos de voz, condiciones limpias

Parece que, para condiciones en las que música y voz no están solapadas, la decisión del detector puede alcanzar resultados muy buenos. Para tomar esta decisión, primero se ha calculado la puntuación combinada basada en la distancia al origen de la gráfica.

$$distancia[n] = \sqrt{\alpha_c \cdot Punt_c{}^2[n] + \alpha_r \cdot Punt_r{}^2[n]}$$

En la siguiente figura se muestran las funciones de distribución acumulada de los valores de distancia para música y para voz, fijando los parámetros $\alpha_r = 0{,}25$ y $\alpha_c = 1$.

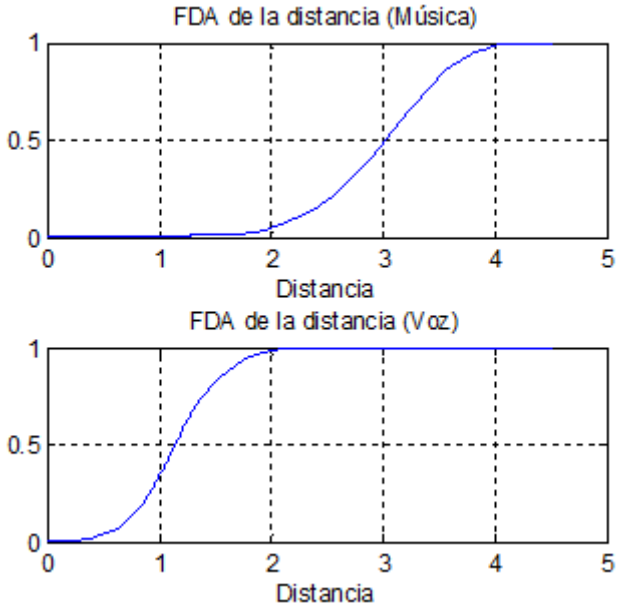


Figura 4-2: Funciones de distribución acumulada de la distancia en música y voz (entorno limpio)

El mejor resultado para el conjunto de prueba se obtiene tomando como criterio para la decisión que la distancia sea mayor o igual a 1,9. Para este conjunto, los resultados obtenidos son:

Pruebas en entorno limpio	
Total aciertos	96,82%
Aciertos positivos	96,19%
Aciertos negativos	97,45%

Tabla 4-1: Resultados de las pruebas en entorno limpio

4.2 Pruebas con solapamiento de voz y música en entorno controlado

La creación del entorno de pruebas controlado se ha llevado a cabo utilizando la DAW (*Digital Audio Workstation*) Reaper. La finalidad de este entorno controlado es observar cómo afecta al detector la relación de volúmenes entre música y voz. Esta relación puede considerarse una SNR (*Signal to Noise Ratio*), al tratarse la voz, en este caso, de un ruido que obstaculiza la detección de música.

Para generar los ficheros de audio de prueba, se han utilizado tres elementos:

- Fragmento de siete minutos de voz hablada por varios locutores, extraído de la base de datos ATVS-Radio.
- Pista de percusión de un minuto de duración, generada en Reaper.
- Pista de armonía de un minuto de duración, generada también en Reaper.

El objetivo de las pistas de percusión y armonía es imitar la base musical que aparece de fondo, en muchas ocasiones, durante los fragmentos hablados en programas de radio.

Los ficheros creados tienen la siguiente estructura:

Inicio	Fin	Contenido
0:00	1:00	Voz
1:00	2:00	Voz + Ritmo
2:00	3:00	Voz
3:00	4:00	Voz + Armonía
4:00	5:00	Voz
5:00	6:00	Voz + Ritmo + Armonía
6:00	7:00	Voz

Tabla 4-2: Estructura de los ficheros de prueba del entorno controlado

Siguiendo el formato descrito, se han generado cuatro ficheros con valores de SNR de 0dB, -6dB, -15dB y -20dB. La estructura de los ficheros busca representar los distintos escenarios de solapamiento entre música y voz (con ritmo, con armonía o con ambas componentes) y la transiciones del detector en los momentos en los que la música aparece o desaparece.

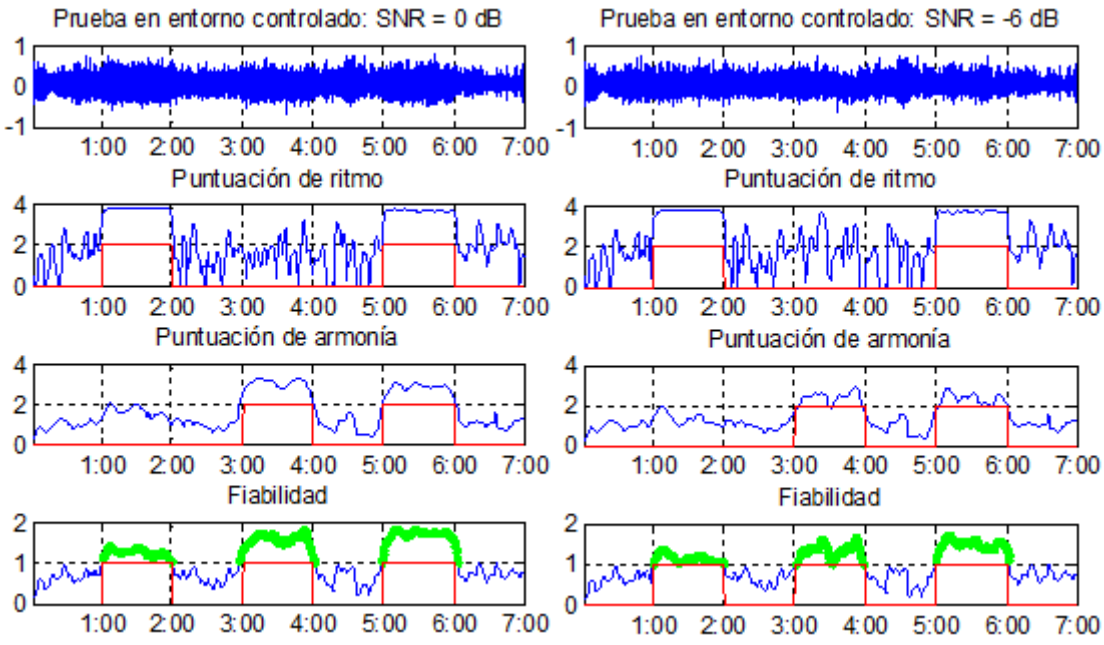


Figura 4-3: Resultados gráficos de las pruebas en entorno controlado (I): SNR de 0 y -6 dB

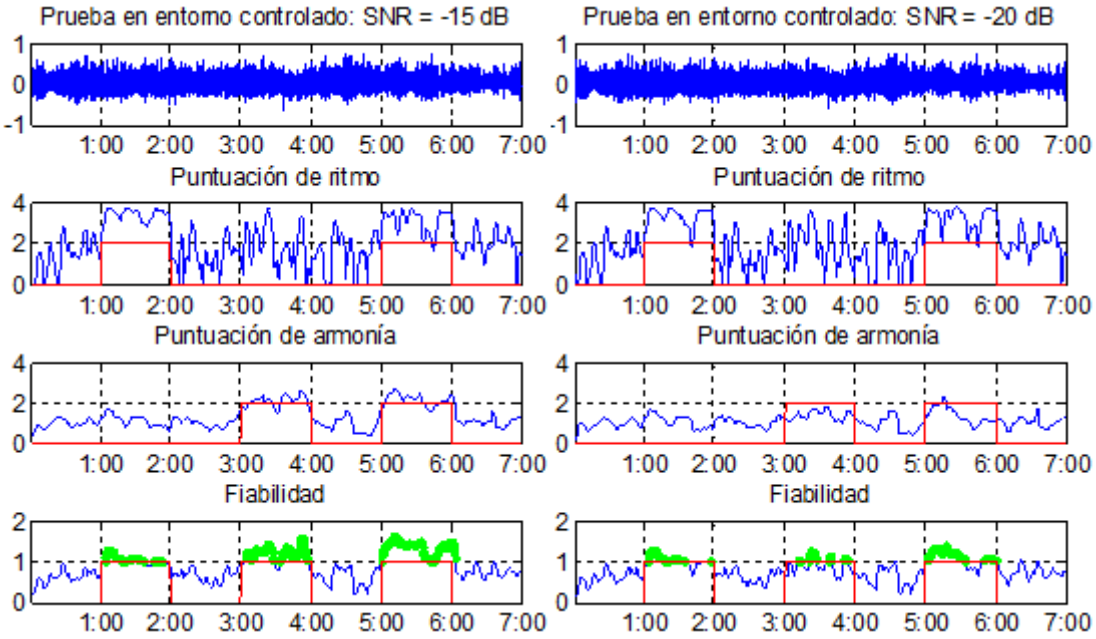


Figura 4-4: Resultados gráficos de las pruebas en entorno controlado (II): SNR de -15 y -20 dB

Los resultados gráficos mostrados en las figuras 4-3 y 4-4 muestran las puntuaciones individuales de los detectores de ritmo y armonía para los distintos ficheros de prueba, además de la salida final del sistema, mostrada como vector de fiabilidad y como decisión binaria: se resaltan en verde los valores en los que la fiabilidad queda por encima de 1 y, por tanto, la decisión binaria es positiva. Los intervalos de tiempo marcados en rojo hacen referencia a las etiquetas manuales (*ground truth*) de presencia de música (o de ritmo y armonía, separadamente, en las gráficas de los detectores individuales).

	SNR = 0dB	SNR = -6 dB	SNR = -15 dB	SNR = -20 dB
Total aciertos	**97,38%**	**97,62%**	**90,71%**	**78,10%**
Aciertos positivos	99,88%	98,45%	91,55%	79,05%
Aciertos negativos	97,50%	99,17%	99,17%	99,05%

Tabla 4-3: Resultados de las pruebas en entorno controlado

Los resultados indican que el detector muestra cierta robustez frente a la presencia de voz, aunque su rendimiento empeora, sobre todo, para valores de SNR inferiores a -15 decibelios.

4.3 Pruebas en medios reales

Tras comprobar los resultados del detector de música en un entorno limpio y en un entorno de solapamiento con SNR conocida, el siguiente paso es ponerlo a prueba ante ficheros de audio pertenecientes a medios reales. Para ello, se ha utilizado un subconjunto de la base de datos ATVS-Radio, formado por diez ficheros de cinco minutos de duración.

El subconjunto de ficheros trata de representar los diversos escenarios en los que voz y música pueden solaparse en los medios reales. Se ha tratado de evitar fragmentos en los que se emite una canción completa, ya que esos casos se asemejan más a las pruebas realizadas en entorno limpio, que se han analizado ya en el apartado 4.1.

Nº	Programa	Total aciertos	Aciertos positivos	Aciertos negativos
1	*El Pirata y su banda* (Rock FM)	**70,83%**	68,41%	100%
2	*Julia en la Onda* (Onda Cero)	**93,00%**	76,57%	94,67%
3	*Julia en la Onda* (Onda Cero)	**83,17%**	95,29%	82,72%
4	*La Mañana* (Cadena COPE)	**81,00%**	84,94%	72,37%
5	*La Mañana* (Cadena COPE)	**74,33%**	76,95%	73,54%
6	*La Mañana* (Cadena COPE)	**79,83%**	79,59%	80,59%
7	*La Mañana* (Cadena COPE)	**69,83%**	73,59%	21,54%
8	*La Mañana* (Cadena COPE)	**75,67%**	67,95%	83,20%
9	*Más de uno* (Onda Cero)	**85,17%**	90,73%	84,09%
10	*Más de uno* (Onda Cero)	**95,33%**	81,54%	95,85%

Tabla 4-4: Resultados de las pruebas en medios reales

Durante la realización de las pruebas se ha observado que los factores que más obstaculizan el trabajo del detector son las SNR bajas (volumen de música muy tenue en comparación con el de la voz) y las transiciones abruptas (un corto periodo sin música entre dos periodos con música, o viceversa).

El problema de las SNR es bastante habitual en los ficheros de audio extraídos de la radio, ya que el mensaje principal lo lleva la voz, mientras que la música sólo tiene una función ambiental. Al escuchar algunos de los fragmentos con atención, la música sólo es ligeramente apreciable durante los silencios que deja la voz. Nosotros, como oyentes, entendemos que la música tiene continuidad aunque no la percibamos constantemente, pero es un proceso que difícilmente puede llevar a cabo un detector basado exclusivamente en ritmo y armonía. No obstante, se ha conseguido mejorar ligeramente el rendimiento en estas situaciones aplicando el método de doble umbralización sobre la puntuación de fiabilidad.

Por otro lado, al utilizar la detección de música, por ejemplo, como mecanismo auxiliar para mejorar sistemas de reconocimiento de voz, los casos en los que la música queda prácticamente oculta no serán problemáticos, ya que en estos intervalos de tiempo la voz se percibe con total claridad.

Las transiciones constituyen el otro principal problema en la detección. Por ejemplo, al encontrar un intervalo de unos pocos segundos en los que sólo hay voz, situado entre dos fragmentos con música, en ocasiones no se llega a detectar correctamente la ausencia de música (esta es la causa de la baja tasa de aciertos negativos en el fichero 7). Aquí encontramos un dilema: se podría pensar en reducir la contextualidad del detector para ganar precisión en transiciones de corta duración, pero la presencia de música es dependiente del entorno temporal de la señal.

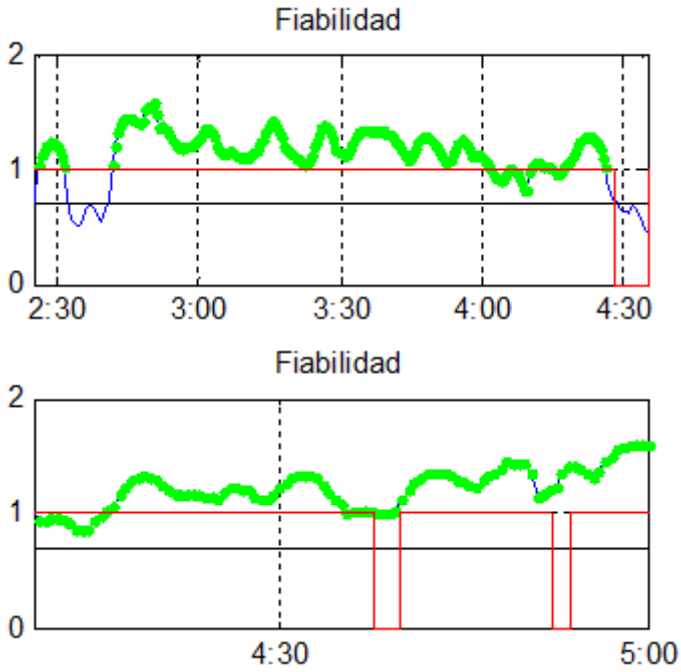


Figura 4-5: Detalle de las situaciones de baja SNR (Fichero 1, superior) y de transiciones abruptas (Fichero 7, inferior)

La gráfica inferior de la Figura 4-5 trata de ilustrar una situación de transiciones abruptas entre sucesivos intervalos con música y sin ella. Los intervalos sin música mostrados duran entre uno y dos segundos, pero el detector no llega a decidir ausencia de música en ningún momento. Estos casos, además, son más complicados al emplear la doble umbralización, ya que la decisión de ausencia de música requiere que el valor de fiabilidad sea, en algún momento, inferior al umbral débil (marcado en negro y de valor 0,7).

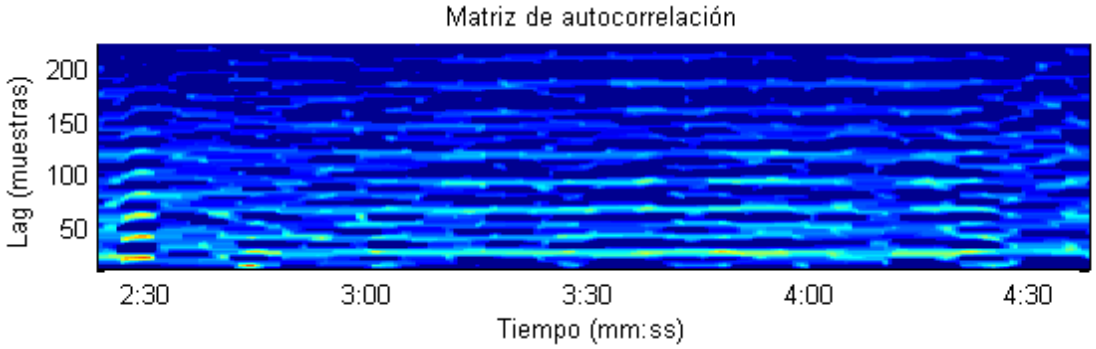


Figura 4-6: Matriz de correlación para el fragmento mostrado en Figura 4-5, superior

En la Figura 4-6 se muestra la matriz de autocorrelación del caso superior de la Figura 4-5. A pesar de poder recuperar el pulso rítmico durante gran parte del fragmento, a partir de 2:30 hay un intervalo en el que se pierde, resultando en una bajada drástica de la fiabilidad en la Figura 4-5 (superior).

5 Conclusiones y trabajo futuro

Durante el desarrollo de este TFG se ha diseñado y programado en MATLAB un sistema completo de detección de música en contenidos multimedia. Este detector de música está formado a partir de la combinación de una puntuación de ritmo y otra de armonía, que funcionan de manera independiente.

A lo largo del proceso de diseño de cada uno de los detectores, se ha podido comprobar la utilidad de herramientas matemáticas como la función de autocorrelación o la *Short-Time Fourier Transform* en el análisis de señales de audio. También se han determinado características de las matrices de autocorrelación y de los cromagramas que permiten evaluar si en la señal existe ritmo o armonía, y que podrían ser utilizadas para obtener información más específica (por ejemplo, el tempo del pulso rítmico o la tonalidad).

Las muestras gráficas de la combinación de los detectores han permitido evidenciar que las componentes de ritmo y de armonía pueden estar presentes en distintas proporciones en música de distintos estilos. Añadiendo algunos datos, se podría tener la base para diseñar un sistema de detección automática de géneros musicales.

También se ha visto que una señal de voz hablada puede obtener puntuaciones de ritmo elevadas en momentos concretos, debido a que se puede encontrar periodicidad en la energía en la voz: por ejemplo, en frases pronunciadas de forma especialmente rítmica o en la risa de los locutores.

Las pruebas realizadas han servido para evaluar el comportamiento del detector en una gran variedad de situaciones e identificar las limitaciones que muestra: el rendimiento baja en las transiciones abruptas y en los tramos en los que la música tiene un volumen muy inferior al de la voz.

Además, de manera transversal al desarrollo del detector y a la realización de las pruebas, se han corregido algunas etiquetas erróneas de la base de datos ATVS-Radio y se han creado herramientas en MATLAB para la lectura de estas etiquetas y la comprobación de errores de detección.

6 Referencias

[1] P. Cano, E. Batlle, E. Gomez, L. de C.T.Gomes and M. Bonnet, "Audio Fingerprinting: Concepts and Applications," Springer-Verlag, 2005, p. 4.

[2] M. Schedl, E. Gómez y J. Urbano, Music Information Retrieval: Recent Developments and Applications, 2014.

[3] D. Ramos, «Recuperación de Información Musical,» de *Tecnologías de Audio*.

[4] M. Müller, D. P.W.Ellis, A. Klapuri y G. Richard, «Signal Processing for Music Analysis,» *IEEE Journal of Selected Topics in Signal Processing*, pp. 2-4, 2011.

[5] J. Ortega-García, «Preprocesado y Parametrización de la Señal de Voz,» de *Tratamiento de Señales de Voz y Audio*, pp. 2-7.

[6] M. Müller, «Fourier Analysis of Signals,» de *Fundamentals of Music Processing*, Springer, 2015.

[7] J. H. Jensen, M. G. Christensen, D. P.W. Ellis y S. H. Jensen, A Tempo-Insensitive Distance Measure for Cover Song Identification Based on Chroma Features, IEEE, 2008.

[8] D. Ellis, «LABRosa - Music Beat Tracking and Cover Song Identification,» [En línea]. Available: https://labrosa.ee.columbia.edu/projects/coversongs/.

[9] B. G. Naranjo, Segmentación de Audio Broadcast, 2016, p. 9.