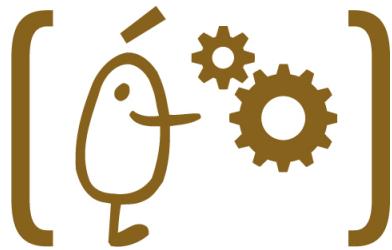




VNIVERSITAT
DE VALÈNCIA

Procesamiento de señales de audio en música

Autores: Nicolás Camañes Antolín, Javier Hinarejos Albero, Diego Lacomba Fañanas, Jose Marquez Algaba, Luis Rioja Gallo



Máster en Ciencia de Datos

Resumen

En el presente trabajo se introducirá el procedimiento clásico de análisis de señales de audio en música, con el objetivo de aplicar y entender las técnicas de extracción de características en este tipo de señales así como los resultados obtenidos. Dentro del marco teórico, encontramos características tanto del dominio temporal como del frecuencial, siendo necesaria la introducción de conceptos como la transformada de Fourier clásica o la de tiempo reducido entre otros. De forma complementaria, una vez obtenidas algunas de las principales características de un conjunto de canciones, se realizará un clustering con tal de agrupar las canciones por géneros musicales.

Índice

1. Introducción	4
2. Fundamentos teóricos	4
2.1. Extracción de características	4
2.1.1. Características de Dominio Temporal	5
2.1.2. Características de Dominio Frecuencial	16
2.1.3. Características de Dominio Temporal-Frecuencial	17
2.1.4. Características de Dominio Cepstral	34
3. Implementación y resultados obtenidos	37
3.1. K-means Clustering (no Supervisado)	37
3.2. Supervisado (Random Forest Classifier)	39
4. Conclusión	41
Referencias	42

Índice de figuras

1.	Waveform de diferentes canciones	6
2.	Longitud de marco	8
3.	Aplicación de la ventana de Hann a una señal.	8
4.	Eliminando la fuga espectral mediante funciones ventana y superposición de frames.	9
5.	Amplitud Envelope de diferentes canciones	11
6.	RMS	13
7.	ZCR	15
8.	Espectros frecuenciales	18
9.	Comparación entre los espectrogramas de las diferentes canciones	20
10.	Comparación entre espectrograma, armónico y percusión	21
11.	Comparación entre el número de bandas de mel	22
12.	Banco de filtros de Mel	23
13.	Comparación del espectrograma de mel en función de la canción	23
14.	Band Energy Ratio de las canciones. $F = 2048$	25
15.	Comparación del centroide espectral en función de la canción	26
16.	Comparación del rolloff espectral en función de la canción	28
17.	Comparación del ancho de banda espectral en función de la canción	29
18.	Comparación de la planitud espectral en función de la canción	31
19.	Comparación de la frecuencia cromática en función de la canción.	32
20.	Comparación del tempograma en función de la canción.	33
21.	Proceso para obtener el cepstrum a partir de una señal de audio	35
22.	Fases para la obtención de los coeficientes MFCCS	35
23.	Representación gráfica de los MFCCs de una canción	36
24.	Método del codo	37
25.	Coeficiente de Silhouette e índice de Davis-Bouldin	38
26.	Clustering $K = 5$	38
27.	Gráficos de dispersión de diferentes variables	39

Índice de tablas

1.	Precisión	40
2.	Precisión	40

1. Introducción

Mediante la realización de este proyecto se pretende realizar una investigación acerca de las características que se pueden extraer de una señales de audio musical y comparar las posibles diferencias entre distintos géneros musicales. La extracción de las diferentes características se realizará en Python haciendo uso de la librería Librosa. Esta librería está diseñada para facilitar la extracción de características de señales de audio, así como para su manipulación y análisis.

El trabajo está dividido en dos fases diferentes. En primer lugar, una explicación de los fundamentos teóricos necesarios para el entendimiento de las diferentes características que se pueden extraer de una señal de audio junto a la posible influencia del género en el valor de éstas. En segundo lugar, se realizará la extracción de algunas de las características explicadas para un conjunto de canciones con distintos géneros musicales, generando un dataframe para finalizar el proyecto aplicando técnicas de machine learning, como son el análisis de componentes principales o algoritmos de clasificación no-supervisada (clustering).

El conjunto de canciones del que se dispone consta de cinco géneros musicales diferentes: Jazz, Ópera, Rap, Reggae y Rock. Cada uno de los géneros consta con 20 canciones diferentes. Por tanto, se dispone de un total de 100 canciones. Para la realización de este trabajo, hemos utilizado como inspiración el siguiente artículo web [3].

El dataset creado, el Notebook gráficos generados para la explicación de los fundamentos teóricos y el Notebook con las implementaciones en Python correspondientes con este proyecto se encuentran disponibles en el siguiente repositorio de GitHub [2].

2. Fundamentos teóricos

2.1. Extracción de características

El proceso mediante el cual una señal de audio es procesada con el fin de extraer diferentes características de dicha señal que permitan describirla y analizarla es denominado **Feature Extraction**. Una forma de clasificar las diferentes características (*features*) es en función de su dominio, distinguiendo tres categorías diferentes: Dominio Temporal, Dominio Frecuencial y Dominio Cepstral.

Es importante tener en cuenta que de cara a emplear las variables para la implementación de algoritmos de *machine learning*, será interesante que las variables empleadas en el modelo, no presenten mucha correlación entre ellas. Puesto que esto aportaría redundancia para el modelo.

En este apartado, lo que se pretende es mostrar los fundamentos teóricos y metodologías para la extracción de diferentes características para cada una de las categorías nombradas previamente. Además, se realizará la comparación de dichas características entre cinco canciones. Una canción

representativa de cada género musical del *dataset* creado para este proyecto. Dichas canciones serán las siguientes:

1. **Jazz:** *How High The Moon* de *Ella Fitzgerald*
2. **Ópera (Aria):** *Nessun Dorma* de *Giancomo Puccini* (*versión Pavarotti*)
3. **Rap:** *Big Poppa* de *The Notorious B.I.G*
4. **Reggae:** *Get Up Stand Up* de *Bob Marley*
5. **Rock:** *The Trooper* de *Iron Maiden*

2.1.1. Características de Dominio Temporal

La producción de sonido hace que las moléculas de aire circundantes vibren, manifestándose en regiones alternas de compresión (alta presión) y rarefacción (baja presión). Estas compresiones y rarefacciones viajan a través del medio y llegan a nuestros oídos, permitiéndonos percibir el sonido tal como es. Por lo tanto, la propagación del sonido implica la transmisión de estas variaciones de presión a lo largo del tiempo.

La representación en el dominio del tiempo del sonido implica capturar y analizar estas variaciones de presión en diferentes intervalos de tiempo mediante el muestreo de la onda de sonido en puntos discretos en el tiempo. Cada muestra representa el nivel de presión del sonido en un momento específico. Al trazar estas muestras, obtenemos una forma de onda que muestra cómo cambia el nivel de presión del sonido con el tiempo.

Podemos representar esta onda (o señal analógica) con dos ejes. El eje horizontal representa el tiempo, mientras que el eje vertical representa la amplitud o intensidad del sonido, generalmente escalada para ajustarse entre -1 y 1, donde los valores positivos indican compresión y los valores negativos indican rarefacción. A continuación, en los gráficos de la Figura 1 podemos ver las 5 canciones representadas en este formato. En este momento ya podemos apreciar una gran diferencia entre dos géneros musicales como son el rock y la ópera. Por un lado el rock mantiene una intensidad alta constantemente mientras que el aria presenta una intensidad mucho menor hasta la parte del climax final.

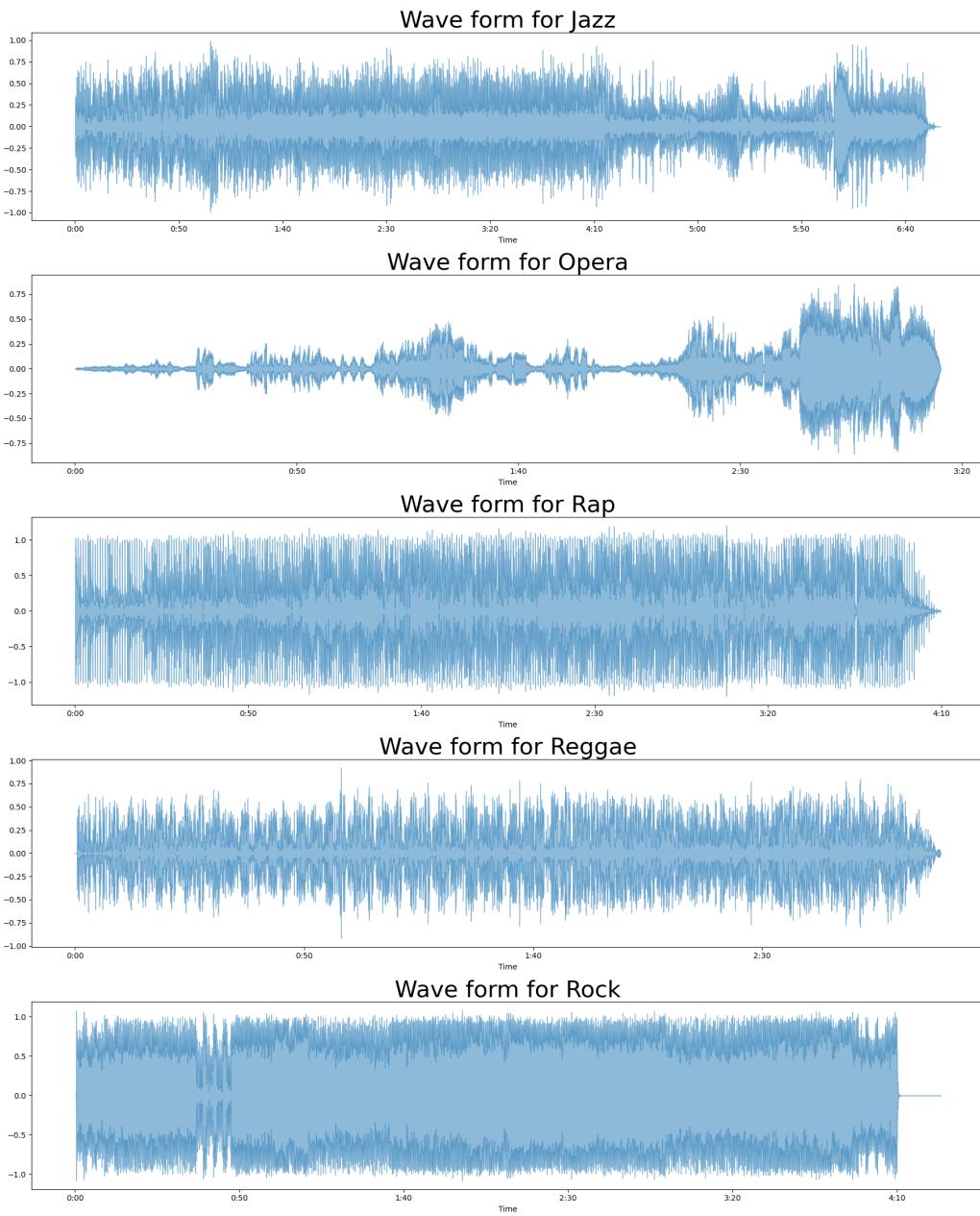


Figura 1: Waveform de diferentes canciones

Para procesar y almacenar estas señales analógicas de manera digital, se deben convertir en una representación discreta. Aquí es donde entra el muestreo, para obtener capturas de pantalla de la onda de sonido en intervalos de tiempo discretos y uniformemente espaciados. La tasa de muestreo determina con qué frecuencia se toman muestras de la señal analógica y se mide en muestras por segundo o hertzios (Hz).

Una tasa de muestreo más alta significa una representación más precisa pero requiere más recursos de memoria, mientras que una tasa de muestreo más baja significa una representación menos precisa pero requiere menos recursos de memoria.

Al elegir una tasa de muestreo adecuada para la conversión analógica a digital, es importante conocer el teorema de muestreo de Nyquist-Shannon. Este teorema establece que, para capturar y reconstruir con precisión una señal analógica y evitar el *aliasing*, la tasa de muestreo (F_m) debe ser al menos el doble de la frecuencia más alta presente en la señal de audio (f_{max}), conocida como la Tasa/Frecuencia de Nyquist (N_f).

$$F_m \geq 2 \cdot f_{max}$$

Es por esto que la tasa típica de muestreo en los CD rom es de 44,1 KHz. Aplicando el teorema de Nyquist, la frecuencia máxima recuperable es de 22,05 KHz, un valor ligeramente superior a la frecuencia máxima reconocible por los humanos (20KHz aproximadamente). De esta forma se garantiza que no se pierde ninguna información relevante para el ser humano.

Hablemos de dos parámetros cruciales para extraer características: el tamaño del marco (*frame length*) y la longitud de salto (*hop length*). Después de procesar una señal, se divide en marcos con cierto tamaño y superposición. Esto es esencial para capturar la variación temporal en las características de la señal. Los métodos tradicionales de extracción de características dan un solo número como resumen, pero esto borra la información temporal. La solución es dividir la señal en marcos, calcular características como la media para cada marco y obtener un resumen de características dependiente del tiempo.

1. **Tamaño del Marco (*Frame*):** Describe el tamaño de cada marco. Si por ejemplo trabajamos con una frecuencia de muestreo de 44.1 KHz (frecuencia típica en los CD rom), cada sample equivale a 0.0227ms, un número mucho menor que la resolución temporal propia de los humanos (alrededor de 10ms). Con los *frames* se pretende obtener la suficiente duración temporal para apreciar los eventos acústicos. Por ejemplo, si el tamaño del marco es 1024, lo equivalente a un intervalo temporal de 23.245ms para una frecuencia de muestreo de 44.1 Hz, incluyes 1024 muestras en cada marco y calculas las características necesarias para cada conjunto de estas 1024 muestras. En general, se recomienda que el tamaño del marco sea una potencia de 2. La razón detrás de esto es porque la Transformada Rápida de Fourier requiere que los marcos tengan un tamaño que sea una potencia de 2.
2. **Longitud de Salto:** Se refiere al número de muestras por el cual se avanza un marco en cada paso a lo largo de la secuencia de datos, es decir, el número de muestras que desplazamos hacia la derecha antes de generar un nuevo marco. La longitud de salto, por lo tanto, determina la superposición entre marcos de audio consecutivos. En la Figura 2 podemos ver un ejemplo gráfico de este proceso.

Para mitigar el impacto de un fenómeno llamado “fuga espectral”, que ocurre al convertir una señal de su dominio de tiempo a su dominio de frecuencia y que produce componentes de alta frecuencia no presentes en la señal original, se aplica una función de ventana a cada marco, multiplicando la señal por la función ventana correspondiente sobre el marco, lo que resulta en la pérdida de datos alrededor de los bordes de cada marco, como podemos ver en la Figura 3. Por lo tanto, a menudo se eligen longitudes de salto intermedias para preservar las muestras en los bordes, lo que resulta en diferentes grados de superposición entre marcos. Un ejemplo de este procedimiento se puede observar en la Figura 4. Una de las funciones ventana más conocida es la función de Hann que se define como sigue:

$$w(k) = 0,5 \cdot (1 - \cos(\frac{2\pi k}{K-1})), k = 1 \dots K$$

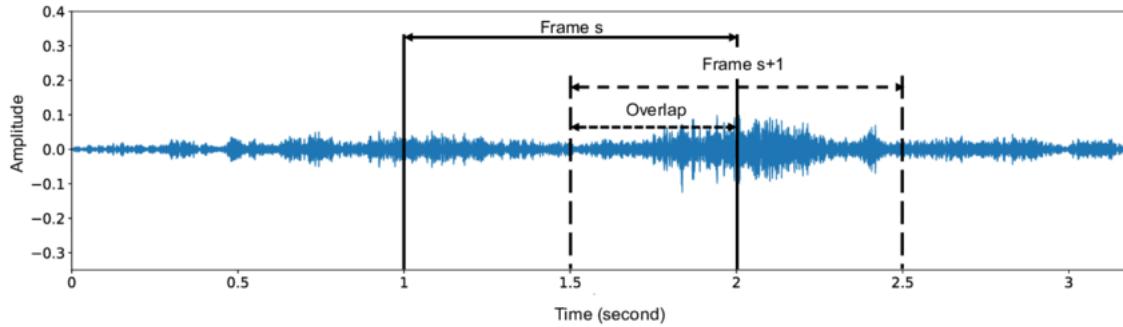


Figura 2: Longitud de marco.

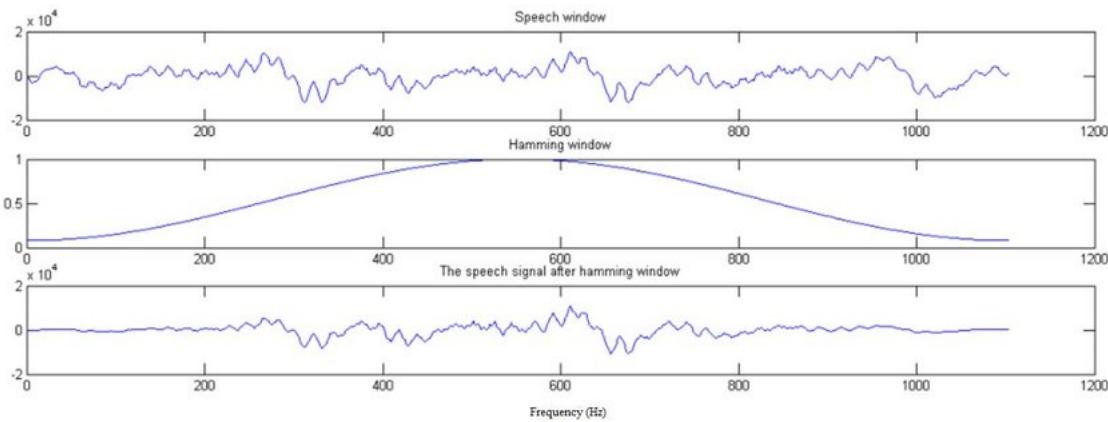


Figura 3: Aplicación de la ventana de Hann a una señal.

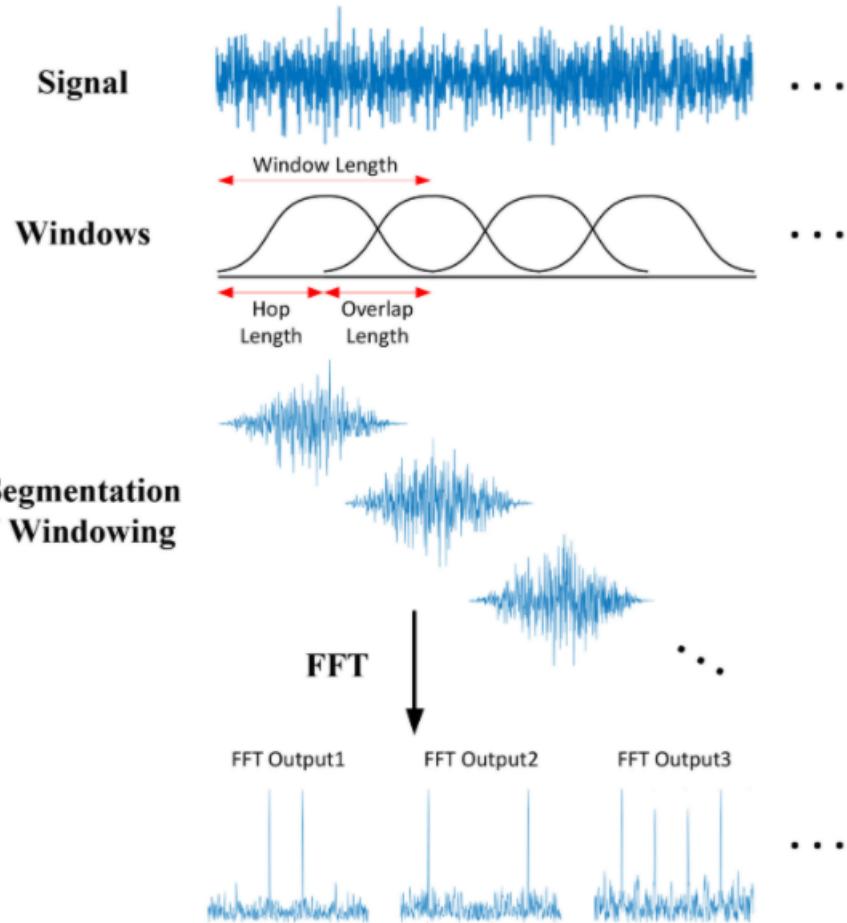


Figura 4: Eliminando la fuga espectral mediante funciones ventana y superposición de frames.

Previamente a la definición matemática de algunas características, aclaramos ciertas notaciones:

- x_i : amplitud en la muestra i .
- $x(t)$: amplitud en el instante t .
- K : tamaño de un frame en muestras.
- H : tamaño de un salto entre frames.

Como fuente de inspiración para la realización de este apartado se ha empleado el siguiente paper. [1]

2.1.1.1. Amplitude Envelope

En español, la envolvente de una señal o AE (del inglés, *Amplitude Envelope*), es simplemente el valor máximo de la amplitud en cierto intervalo de tiempo (frame). Esta característica nos proporciona una idea aproximada sobre el volumen de la señal. Un problema del AE es que es sensible a los outliers. Matemáticamente, podemos definir el envolvente de la amplitud del frame k-ésimo como:

$$AE_k = \max_{i=k, K}^{(k+1) \cdot K - 1} x_i$$

Algunas de las aplicaciones de esta característica son:

- Onset detection: detectar el inicio de una nota musical, una palabra, o un evento sonoro en general.
- Clasificación por género musical.

A continuación, en la Figura 5, podemos observar el envoltorio de la señal correspondiente a canciones de diferentes géneros para un tamaño de frame de 1024 y un tamaño de salto de 512.

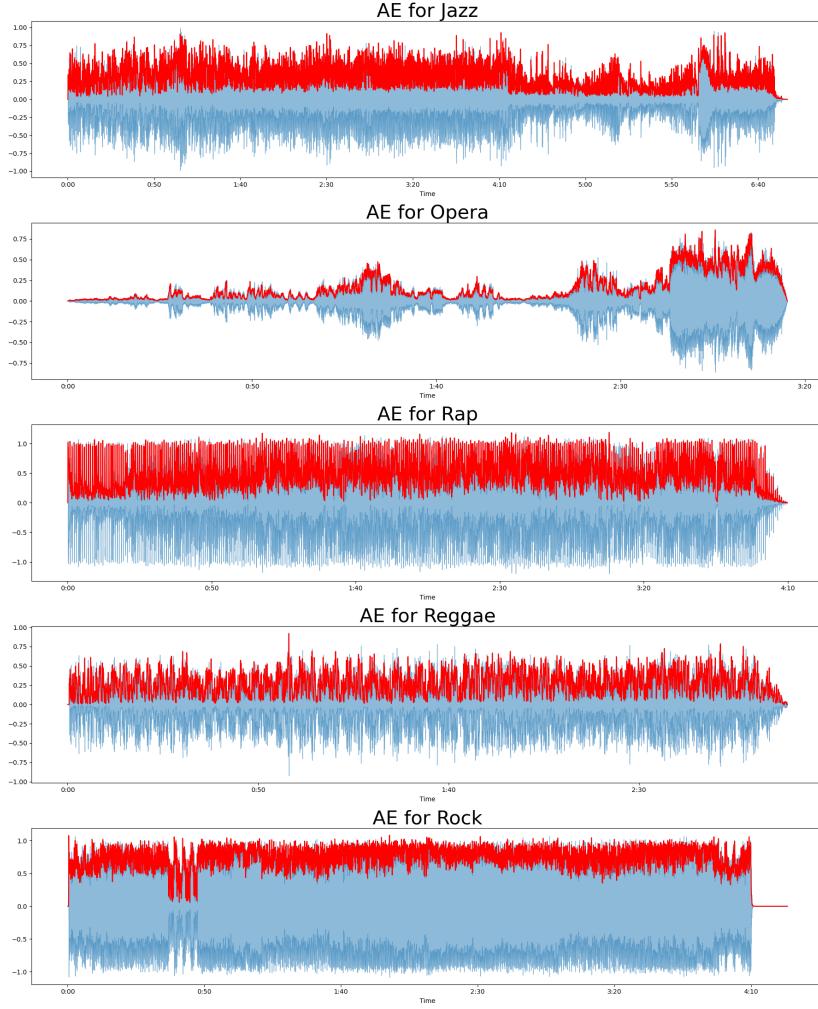


Figura 5: Amplitude Envelope de diferentes canciones

2.1.1.2. Root-Mean Square Energy (RMSE)

La energía E de una señal continua se define como el área bajo la magnitud al cuadrado de la señal. Para señales de audio equivale a como de alto suena la señal. Para una señal continua en un intervalo de tiempo T se define como

$$E_c = \int_0^T |x(t)|^2$$

Y para una señal discreta con K muestras

$$E_d = \sum_{i=1}^K |x_i|^2$$

El root-mean-square energy (RMS), o valor eficaz, es otra medida importante en el análisis de señales. Se calcula tomando la raíz cuadrada del valor promedio de la energía de la señal. Para una señal continua $x(t)$ en un intervalo de tiempo T , el RMSE se calcula de la siguiente manera:

$$RMS = \sqrt{\frac{1}{T} E_c}$$

En el caso de una señal discreta con K muestras, el RMS se calcula como:

$$RMS = \sqrt{\frac{1}{K} E_d}$$

El RMS es útil porque proporciona una medida de la “amplitud efectiva” de la señal. Mientras que la amplitud pico da la magnitud máxima de la señal, el RMS tiene en cuenta la magnitud de todos los valores y proporciona una medida más representativa de la energía o potencia de la señal. En el caso de señales sinusoidales, el valor RMS es directamente proporcional a la amplitud de la señal.

Algunas de las aplicaciones de esta característica son:

- Segmentación de audio.
- Clasificación por género musical

A continuación obtenemos el RMS de las diferentes canciones para un tamaño de frame de 1024 y con un salto de 512. En rojo podemos ver el RMSE sobre la forma de onda de la canción. Se puede observar como presenta mucha menos variabilidad que la onda original. Como hemos mencionado previamente, esto es debido a que el RMS considera todos los valores de la señal y no únicamente el máximo.

2.1.1.3. Tasa de cruces con cero (ZCR)

La tasa de cruces con cero es la tasa de cambios de signo a lo largo de una señal, es decir, la tasa a la que la señal cambia de positivo a cero a negativo o de negativo a cero a positivo. Esta función se ha utilizado mucho tanto en el reconocimiento de voz como en la recuperación de información musical, siendo una función clave para clasificar los sonidos de percusión.

La tasa de cruces con cero (ZCR) se define formalmente como:

$$ZCR = \frac{1}{K-1} \sum_{i=1}^{K-1} 1_{\mathbb{R}<0}(x_i \cdot x_{i-1})$$

Donde:

- $1_{\mathbb{R}<0}(x_i \cdot x_{i-1})$ es una función indicadora que toma el valor de 1 si $x_i \cdot x_{i-1}$ es negativo (indicando un cambio de signo), y 0 en caso contrario.

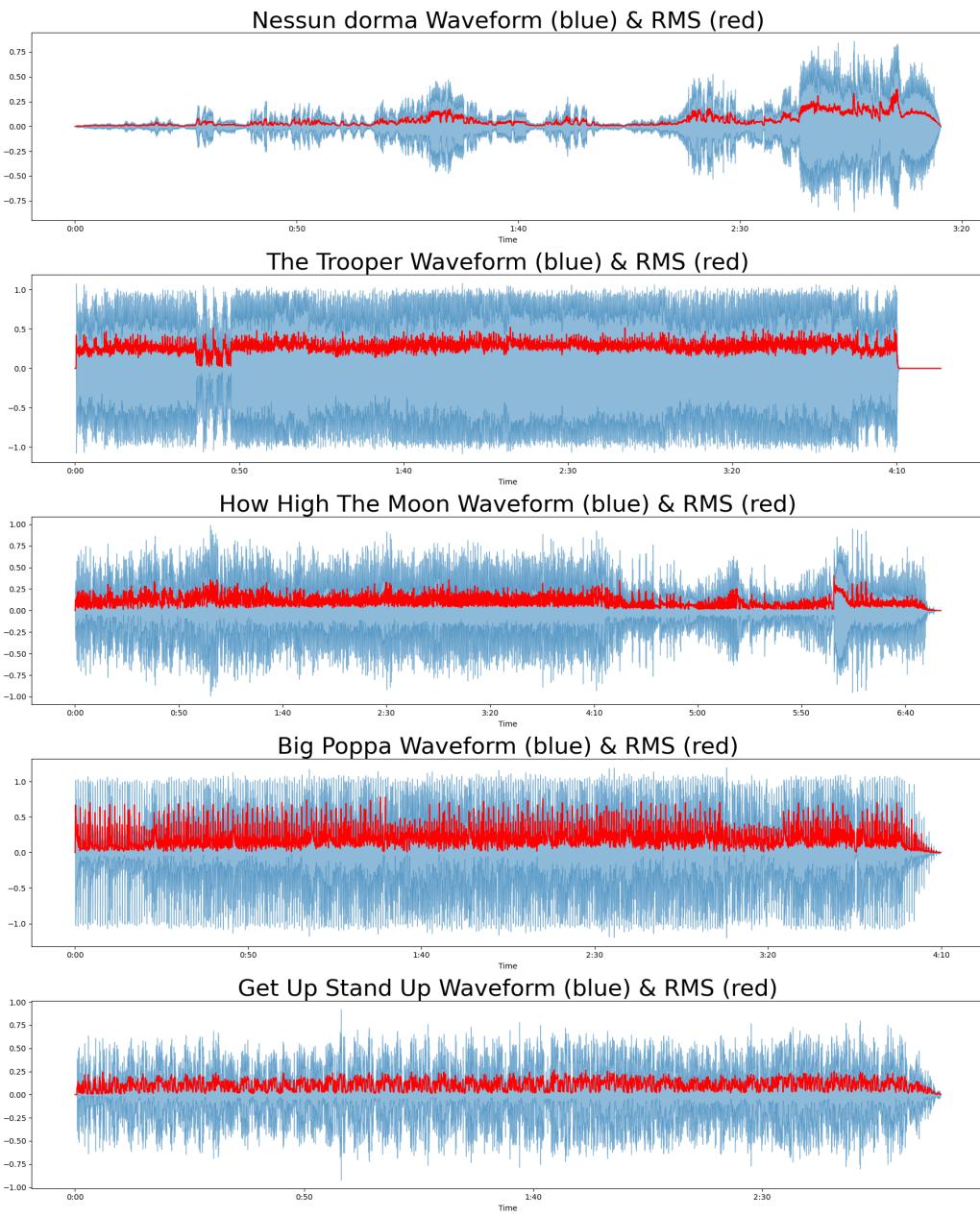


Figura 6: RMS

- La suma se realiza sobre todos los tiempos i desde 1 hasta $K-1$.
- La división por $K - 1$ normaliza la suma para obtener la tasa promedio de cruces con cero por unidad de tiempo.

Algunas de las aplicaciones de esta característica son:

- Reconocimiento de sonidos agudos (bajo ZCR) vs percusión (alto ZCR).
- Estimación de tono monofónico (nota).
- Detectar si una señal contiene voz o no.
- Clasificación por género musical.

Normalmente la música rock tiene un zcr alto respecto a otros géneros como la música clásica debido a los instrumentos de percusión. Para la realización de esta sección se ha usado como inspiración la siguiente tesis [4]

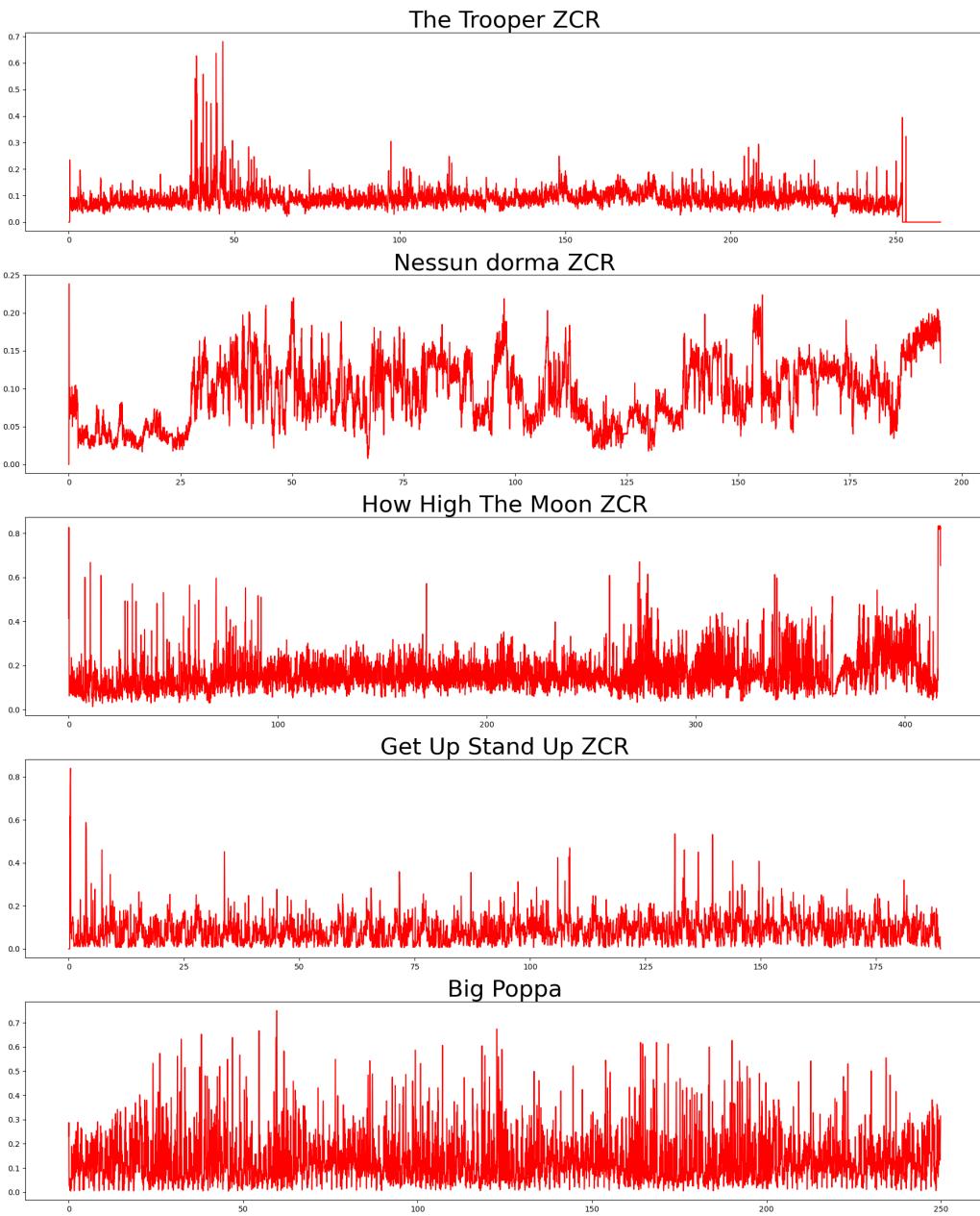


Figura 7: ZCR

2.1.1.4. Crest Factor

El *crest factor*, factor de cresta en castellano, es una característica que se emplea para evaluar cuánto varía la amplitud de una señal de audio a lo largo del tiempo. Para calcularla, se divide el valor máximo de la amplitud de la señal entre el RMS. Mediante este cálculo, se obtiene informa-

ción representativa sobre la fluctuación de la señal, hecho que tiene una influencia directa sobre la calidad del audio. Matemáticamente, el factor de cresta para el frame k-ésimo se define como:

$$CF_k = \frac{\max_{i=k \cdot K}^{(k+1) \cdot K - 1} |x_i|}{\sqrt{\frac{1}{K} \sum_{i=k \cdot K}^{(k+1) \cdot K - 1} x_i^2}} \quad (2.1)$$

Un valor de factor de cresta mayor indica una mayor diferencia entre el máximo y la media de la amplitud, suceso normalmente asociado a señales más dinámicas con mayores variaciones en la amplitud. Un valor de cresta menor, sugiere un sonido más uniforme o lo que es lo mismo, una señal más comprimida. En géneros como el jazz o la ópera, apreciaremos un mayor crest factor respecto a géneros más actuales como el rap, el cual presenta menos dinamismo. Esta característica ha sido explicada puesto que aporta información pero no ha sido incluida en implementación realizada en este proyecto.

2.1.2. Características de Dominio Frecuencial

Otra forma de representar una señal es mediante su dominio de frecuencias. Para ello, la señal se descompone en las frecuencias que la constituyen mediante una transformación. En lugar de observar la amplitud de la señal en varios puntos en el tiempo, examinamos las amplitudes de los diferentes componentes de frecuencia que constituyen la señal. Cada componente de frecuencia representa una onda sinusoidal de una frecuencia particular y al combinar estos componentes, podemos reconstruir la señal original en el dominio del tiempo.

2.1.2.1. Spectrum y Transformada de Fourier

La herramienta matemática más empleada para obtener el dominio frecuencial de una señal es la transformada de Fourier. La transformación de Fourier toma la señal como entrada y la descompone en una suma de ondas seno y coseno de diferentes frecuencias, cada una con su propia amplitud y fase. Cuanto mayor sea el peso de la frecuencia, más similaridad tendrá su función seno correspondiente con la señal original. La representación resultante es lo que constituye el espectro de frecuencia. Matemáticamente, la transformada de Fourier de una señal continua se define de la siguiente manera:

$$h(f) = \int_{-\infty}^{\infty} x(t)e^{-2\pi jft} dt \quad (2.2)$$

Dado que el sonido procesado digitalmente es discreto, dada una secuencia discreta x_i , $i = 0, \dots, N - 1$ podemos definir la transformada discreta de Fourier (DFT) análoga:

$$h(f) = \sum_{i=0}^{N-1} x_i e^{-j(2\pi/N)if} \quad (2.3)$$

El resultado obtenido para las distintas frecuencias evaluadas se puede representar en un gráfico que se conoce como espectro frecuencial de una señal, donde el eje de las abscisas corresponde al rango de frecuencias y el eje de las ordenadas a su magnitud. En la Figura 8 podemos ver los espectros obtenidos para las 5 canciones. Se puede apreciar como la ópera y el jazz trabajan con un rango más amplio de frecuencias debido a la mayor diversidad de instrumentos empleados.

2.1.3. Características de Dominio Temporal-Frecuencial

Cuando aplicamos la transformada discreta de Fourier obtenemos las frecuencias que componen la señal original. El problema que surge es que conocemos “el qué” pero no “el cuándo”, es decir, no conocemos qué frecuencias están presentes en un intervalo de tiempo concreto.

2.1.3.1. Espectrograma y Short-Time Fourier Transform

La idea general para resolver el problema planteado anteriormente es considerar pequeños segmentos de la señal (*frames*) y aplicar FFT (del inglés, *Fast Fourier Transform*) a cada uno de ellos. Como se ha mencionado previamente, con el objetivo de evitar el *Spectral Leakage*, se aplica una función ventana al trozo de señal contenido en el *frame* previamente a la FFT con cierto número de *samples* de la señal superpuestas en cada frame con tal de evitar la pérdida de información. Esto es lo que se conoce como STFT (del inglés, *Short-Time Fourier Transform*). La formulación matemática de la STFT es:

$$S(m, k) = \sum_{n=0}^{N-1} x(n + mH) \cdot w(n) \cdot e^{-i2\pi m \frac{k}{N}}$$

donde:

- n es el sample
- m es el frame
- k es la frecuencia
- N es el tamaño del *frame*
- H es el tamaño de salto
- $w(n)$ es la función ventana

Como podemos ver se obtiene un resultado en función del tiempo (m) y la frecuencia (k). En otras palabras, el resultado de aplicar la DFT a la señal modificada por la función ventana para la frecuencia k en el frame m . La función ventana empleada típicamente es la ventana de Hann. Mientras que en la DFT el número de valores que puede tomar k (frecuencias evaluadas) es

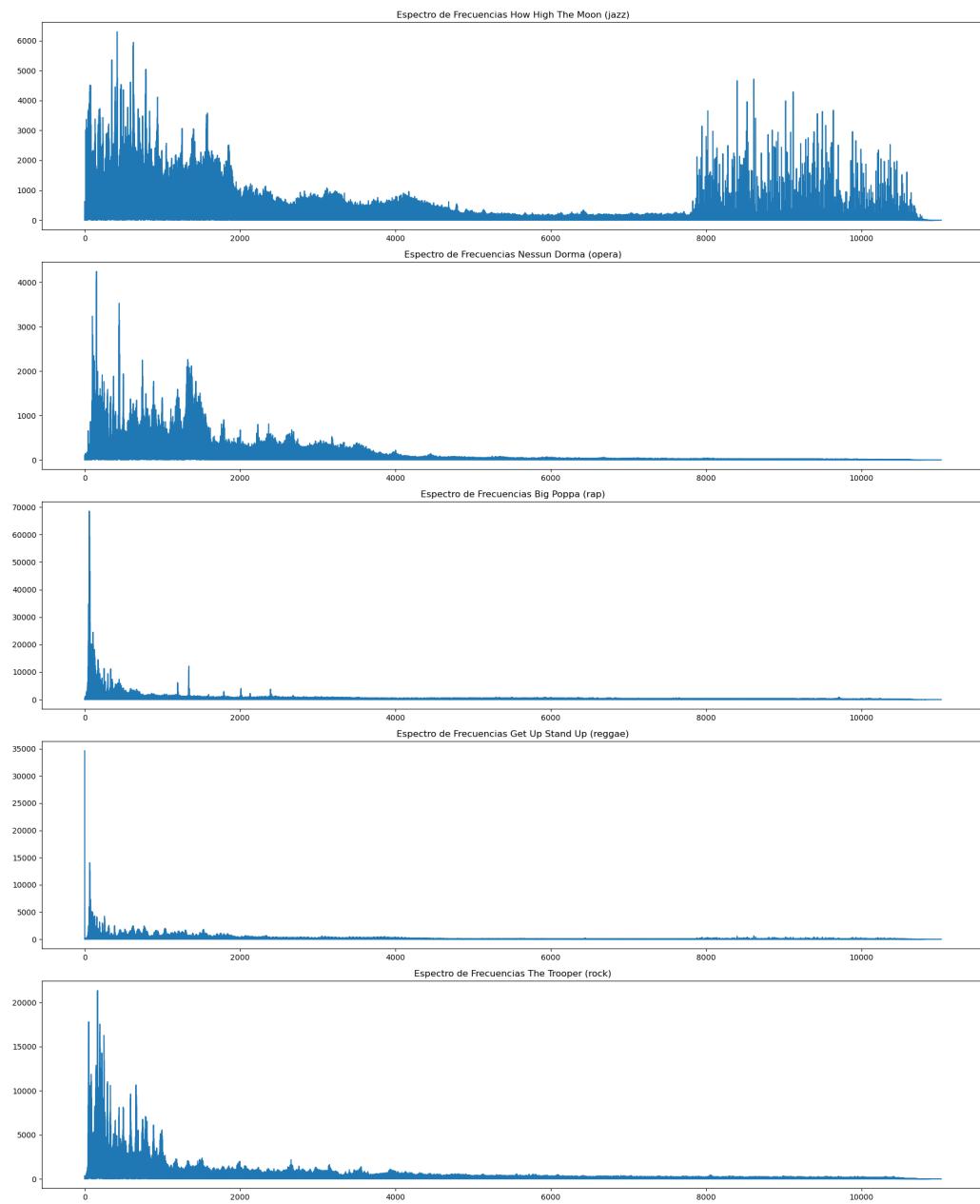


Figura 8: Espectros frecuenciales

igual al número de *samples* de la señal, en la STFT no se consideran la mitad de las frecuencias posibles por el efecto “espejo”(simetría entorno a la frecuencia de Nyquist) obtenido en el resultado de la DFT. Por lo tanto, k tomará valores hasta $\frac{\text{Tamaño frame}}{2} + 1$.

Como el resultado de la DFT contiene coeficientes complejos (con componente imaginaria), a la hora de visualizar el resultado se aplica la transformación $Y(m, k) = |S(m, k)|^2$. Posteriormente se transforma de una representación lineal de la amplitud a decibelios (dB) mediante $10\log_{10}(Y)$ para poder interpretar el resultado mediante un *heatmap*, ya que la forma que tenemos los humanos de percibir las frecuencias no es lineal, sino logarítmica.

Un aspecto importante a considerar es que un tamaño grande de *frame* aumenta la resolución frecuencial pero disminuye la resolución temporal. De forma adversa, un tamaño de frame menor, disminuye la resolución frecuencial y aumenta la temporal. Es decir, para localizar fenómenos breves se emplean tamaños de frame pequeños y para localizar frecuencias de forma más nítida se emplean tamaños de frame grandes. Esto es debido a que la resolución temporal y la frecuencial son inversamente proporcionales, es decir, al aumentar el tamaño de frame evaluamos más frecuencias posibles pero en un intervalo de tiempo mayor. Esto es lo que se conoce como principio de incertidumbre.

En la gráfica de la Figura 9 se ha generado el spectrograma correspondiente a cada una de cinco las canciones. Como se observa en la figura, la ópera presera frecuencias mucho más bajas en comparación con el resto de géneros musicales analizados, siendo el rock el género que posee las frecuencias más altas seguido del Reggae.

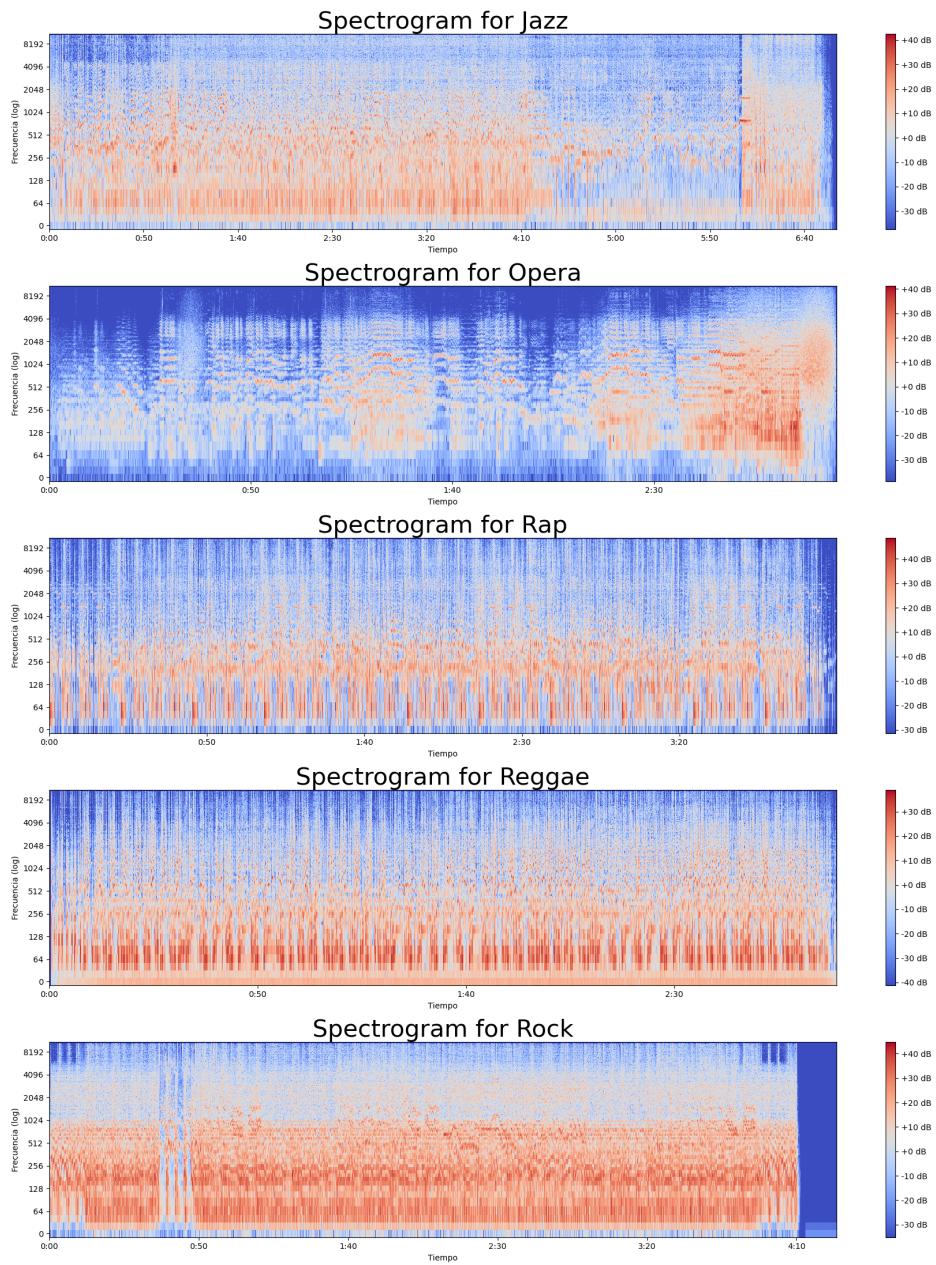


Figura 9: Comparación entre los espectrogramas de las diferentes canciones

De ahora en adelante vamos a definir algunas convenciones matemáticas previamente a la definición de algunas de las características que explicaremos.

- $m_t(n)$: magnitud de la señal en la frecuencia n y frame t
- N : numero de frecuencias en el espectro

2.1.3.2. Armónicos y Percusión

Un sonido armónico es lo que percibimos como un sonido tonal, lo que nos permite escuchar melodías y acordes como las notas de un piano. El prototipo de un sonido armónico se traduce en una línea horizontal en una representación del espectrograma.

Por otro lado, un sonido de percusión es lo que percibimos como un aplauso o el sonido de un tambor. El prototipo de un sonido de percusión se traduce en una línea vertical en una representación del espectrograma.

Para obtener estas características se descompone el espectrograma en frecuencias fundamentales (harmonics) y en aquellas zonas con energía más alta (percussive).

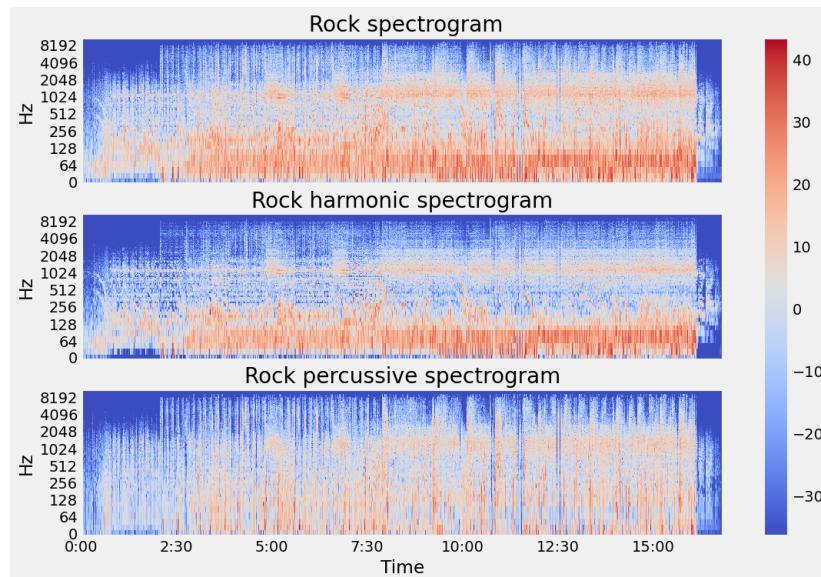


Figura 10: Comparación entre espectrograma, armónico y percusión

2.1.3.3. Espectrograma de Mel

El término *Mel* proviene de la palabra *Melody*, en castellano *Melodía*. Esta característica de una señal de audio consiste en representar la energía espectral de una señal de audio. Pero en lugar de aplicar la escala lineal, como se realiza en el espectrograma, se emplea la escala de mel. El motivo que hace importante esta característica es que el oído humano no aprecia de igual forma los cambios de frecuencia de frecuencias altas que de frecuencias bajas. Nos resulta mucho más fácil diferenciar entre dos sonidos de frecuencias bajas que se llevan x frecuencias entre ellos, que entre dos sonidos de frecuencias altas que se llevan x sonidos entre ellos. Más concretamente, se puede entender como que los humanos percibimos las frecuencias logarítmicamente. Las fórmulas para obtener la frecuencia de mel a partir de la frecuencia de la señal y su inversa se corresponden con las siguientes:

$$m = 2595 \log\left(1 + \frac{f}{500}\right)$$

$$f = 700(10^{m/2595} - 1)$$

El proceso de obtener esta característica de una señal tendrá como objetivo convertir las frecuencias del espectrograma original a frecuencias de mel. Para ello, se deben llevar a cabo los siguientes pasos: elegir el número de bandas de mel, construir los bancos de filtros de mel y aplicar los bancos de filtros de mel al espectrograma.

En primer lugar, para las bandas de mel, se puede comprobar mediante el siguiente gráfico la influencia que tiene en el análisis del espectrograma la elección del número de bandas de mel. En nuestro caso usaremos 128 puesto que ya es un número suficientemente grande como para apenas perder información.

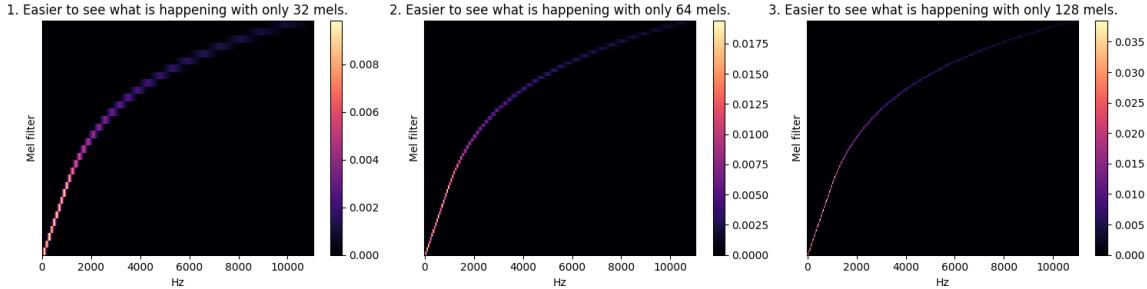


Figura 11: Comparación entre el número de bandas de mel

En segundo lugar, el banco de filtros mel consiste en un conjunto de filtros triangulares uniformemente distribuidos entre el mínimo y máximo de las frecuencias en la escala mel, tal y como se aprecia en la figura 12. Esto conlleva a que, al obtener los valores en la escala lineal de dichos filtros, las frecuencias centrales siguen una distribución logarítmica. Y, mediante dichos se consigue mapear las frecuencias lineales a las frecuencias mel. Cada uno de los filtros de mel cubre un rango específico de frecuencias y se superpone con los demás. Estos filtros capturan la

energía en diferentes regiones de frecuencia y ayudan a modelar cómo percibimos las frecuencias en términos de la escala Mel.

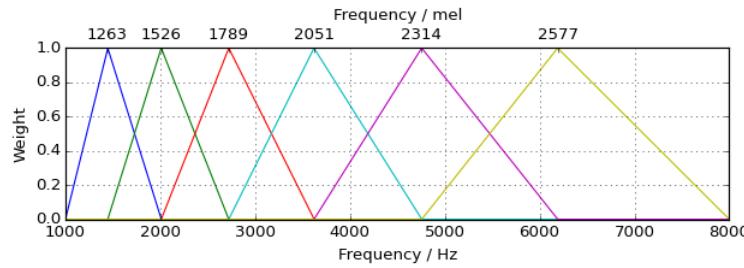


Figura 12: Banco de filtros de Mel

Por último, para finalizar con la obtención de esta característica se deberá aplicar el banco de filtros a la señal que se esté tratando de analizar. En nuestro caso, se ha generado el espectrograma de mel para las 5 canciones. Correspondiente con la Figura 13

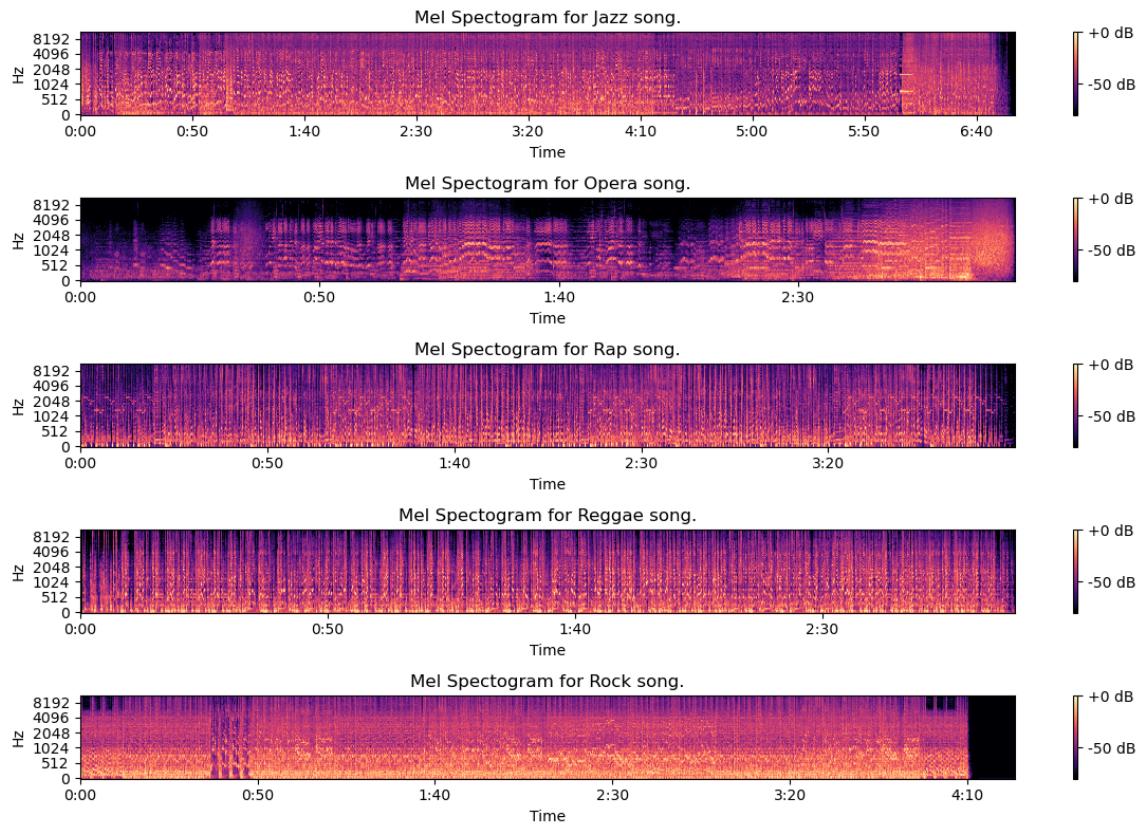


Figura 13: Comparación del espectrograma de mel en función de la canción

Como se puede observar en la figura 13, la canción de rock se observan unas frecuencias de mel mucho más altas que en las de la canción de ópera. Esto es debido a que en una canción de rock son mucho más frecuentes sonidos de guitarras y baterías que producen frecuencias más altas mientras que en las canciones de ópera predominan más sonidos procedentes de voces líricas y diversos instrumentos de cuerda con melodías más bajas y sonidos más suaves.

2.1.3.4. Band Energy Ratio

Esta característica nos aporta información entre la relación entre las bandas de frecuencia altas y bajas. Se puede entender como una medida de la dominancia de las frecuencias bajas. Matemáticamente se define como:

$$BER_t = \frac{\sum_{n=1}^{F-1} m_t(n)^2}{\sum_{n=F}^N m_t(n)^2} \quad (2.4)$$

donde F es la frecuencia divisora (del inglés, *split frequency*), la que se establece como la frontera entre frecuencias altas y bajas. Como podemos ver, la ecuación también está en función del tiempo (frame t). El resultado es la relación entre frecuencias altas y bajas en un determinado frame.

Algunas de las aplicaciones de esta característica son:

- Distinción entre señales de música o voz.
- Clasificación por género musical.

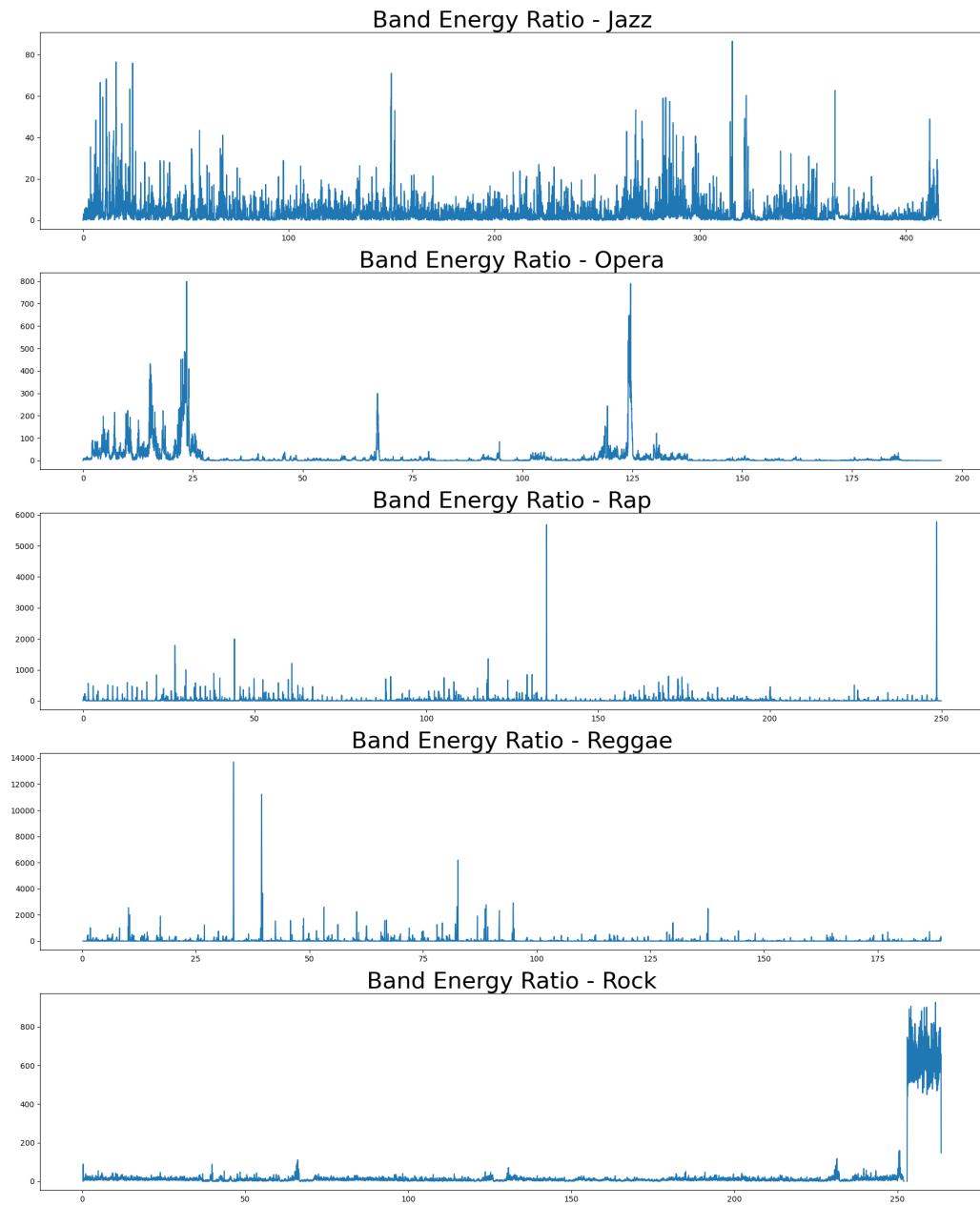


Figura 14: Band Energy Ratio de las canciones. $F = 2048$

2.1.3.5. Spectral Centroid

El centroide espectral se puede definir brevemente como el centro de gravedad de la magnitud del espectro, es decir, el rango de frecuencias donde la mayor parte de la energía de la señal está concentrada. Se puede interpretar como la medida del “brillo” de un sonido. Una señal

brillante tendrá la mayor parte de la energía concentrada en frecuencias altas. Matemáticamente se define como:

$$SC_t = \frac{\sum_{n=1}^N m_t(n) \cdot n}{\sum_{n=1}^N m_t(n)} \quad (2.5)$$

Vemos que consiste en la media ponderada de las frecuencias en un determinado frame.

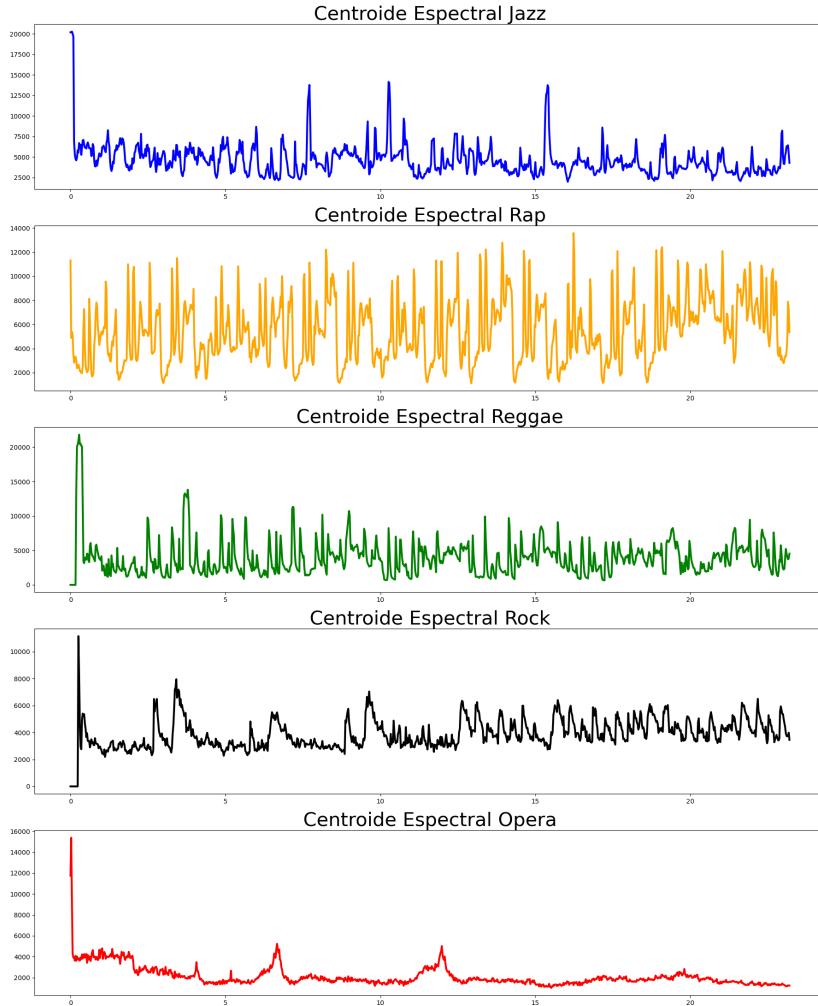


Figura 15: Comparación del centroide espectral en función de la canción

La gráfica más significativa es la de la ópera, ya que prácticamente se encuentra un pico al inicio de la pista. Esto se debe a que en este instante, únicamente suenan los violines y las voces más agudas. Esto indica una concentración pronunciada de energía en ciertas frecuencias. Por otra parte, la gráfica que muestra más variación es la de rap. Esto se podría explicar de la siguiente manera, dicha canción posee una base que muestra sonidos diferentes en la mezcla.

Algunas de las aplicaciones de esta característica son:

- Clasificación de audio (pájaro, motor, instrumento, ...).
- Clasificación por género musical.

En el contexto de procesamiento de señales de audio, por ejemplo, el centroide espectral se utiliza para caracterizar la distribución de frecuencias en una señal.

2.1.3.6. Spectral Rolloff

Caída espectral: Un extractor de características que extrae el punto de caída espectral. Esta es una medida de la cantidad de sesgo a la derecha del espectro de potencia.

El punto de caída espectral es la fracción de contenedores en el espectro de potencia en la que el 85 % de la potencia se encuentra en frecuencias más bajas.

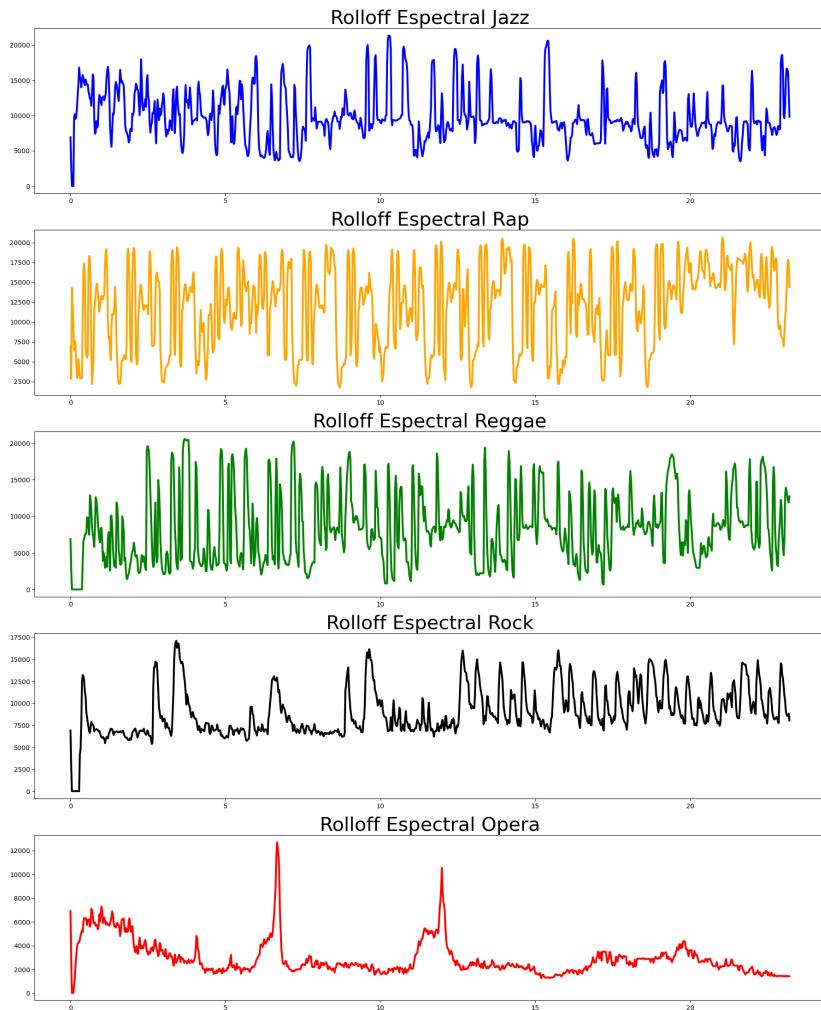


Figura 16: Comparación del rolloff espectral en función de la canción

Respecto a esta gráfica, quería destacar la gran variabilidad que se observa principalmente en los géneros de rap y reggae. Por ejemplo, en la canción de reggae Get Up Stand Up ocurre este suceso porque Bob Marley utiliza distintos instrumentos como la guitarra, el bajo, la batería y el teclado junto con su voz para crear un sonido distintivo. Esa gran variabilidad de instrumentos se muestra en el rolloff espectral.

2.1.3.7. Spectral Bandwidth

Esta característica es una derivación de la anterior. Consiste en el rango espectral alrededor del centroide o de otra forma, la varianza del centroide. Tiene una correlación directa con el timbre percibido. Matemáticamente se define como:

$$BW_t = \frac{\sum_{n=1}^N |n - SC_t| \cdot m_t(n) \cdot n}{\sum_{n=1}^N m_t(n)} \quad (2.6)$$

Consiste en la media ponderada de las distancias de las frecuencias al centroide espectral. Para entender como cambia el BW en función de como la energía está distribuida a lo largo de la señal, si la energía está repartida a lo largo del rango frecuencial, el BW es mayor, por otro lado, si la energía de la señal está acumulada en un rango pequeño de frecuencias, el BW es menor. Una de las aplicaciones de esta característica es la clasificación por género musical.

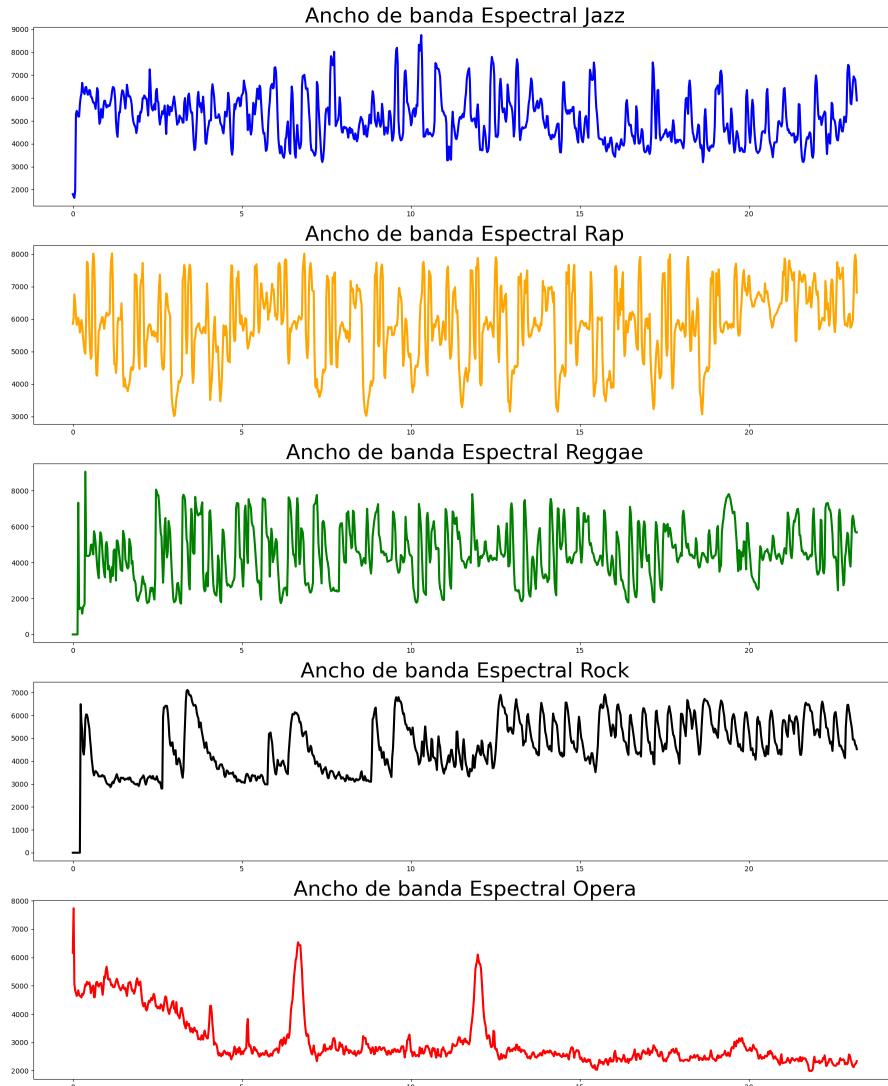


Figura 17: Comparación del ancho de banda espectral en función de la canción

En este caso, nos centraremos en la canción de Rock, "The Trooper". Esta famosa canción de

Iron Maiden, también catalogada como heavy metal, muestra un ancho de banda perceptiblemente alto debido a la combinación de guitarras distorsionadas, una batería potente y voces intensas con el principal objetivo de capturar la energía de dicha canción.

2.1.3.8. Spectral Flatness

La planitud espectral es una medida utilizada en el procesamiento de señales digitales para caracterizar un espectro de audio. La planitud espectral generalmente se mide en decibelios y proporciona una manera de cuantificar en qué medida un sonido se parece a un tono puro, en lugar de ser parecido a un ruido.

El significado de tonal en este contexto tiene que ver con la cantidad de picos o estructura resonante en un espectro de potencia, a diferencia del espectro plano de un ruido blanco. Una planitud espectral alta indica que el espectro tiene una cantidad similar de potencia en todas las bandas espectrales; esto sonaría similar al ruido blanco y el gráfico del espectro parecería relativamente plano y suave. Una planitud espectral baja indica que la potencia espectral se concentra en un número relativamente pequeño de bandas; esto normalmente sonaría como una mezcla de ondas sinusoidales y el espectro parecería “puntiagudo”.

La planitud espectral se calcula dividiendo la media geométrica del espectro de potencia por la media aritmética del espectro de potencia.

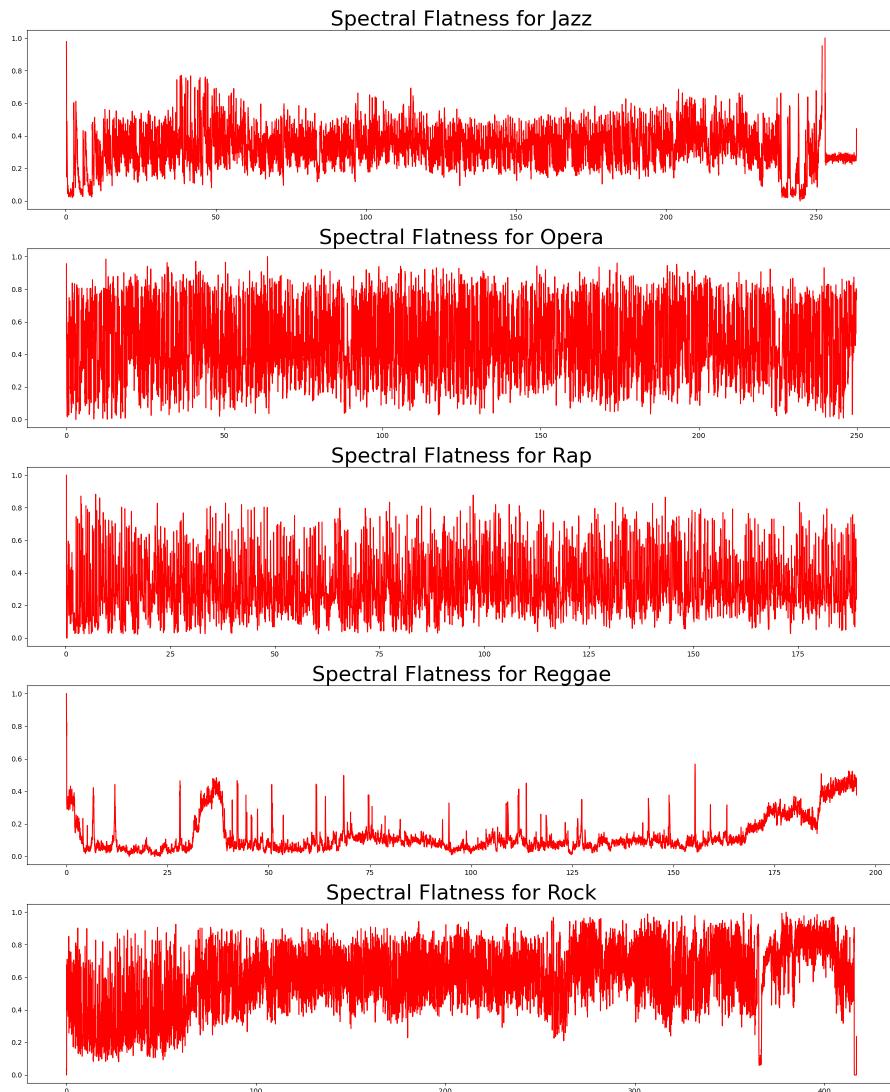


Figura 18: Comparación de la planitud espectral en función de la canción

Tras la representación de estas gráficas, destaca principalmente la planitud espectral del reggae. Esto tiene sentido, ya que esta canción sugiere una mezcla que respeta la importancia de las frecuencias bajas, medias y altas, buscando un equilibrio tonal que se ajuste al estilo relajado de Marley. Esto contribuye a la naturaleza relajada pero rítmica del reggae.

2.1.3.9. Frecuencia cromática

La frecuencia cromática se relaciona estrechamente con las doce clases de tonos diferentes. Las características basadas en croma, que también se conocen como “perfils de clase de tono”, son una herramienta poderosa para analizar música cuyos tonos se pueden categorizar de manera

significativa (a menudo en doce categorías) y cuya afinación se aproxima a la escala de temperamento igual. Una propiedad principal de las características cromáticas es que capturan las características armónicas y melódicas de la música, al tiempo que son resistentes a los cambios en el timbre y la instrumentación.

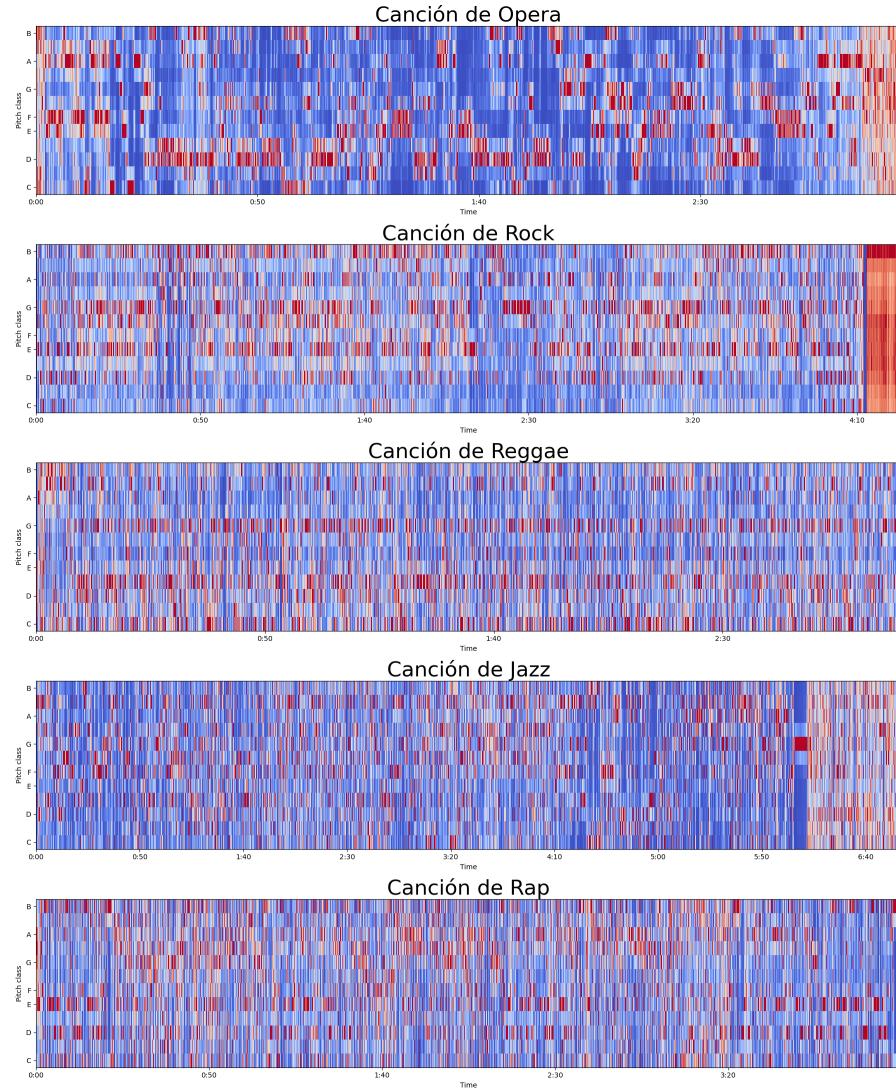


Figura 19: Comparación de la frecuencia cromática en función de la canción.

En el jazz y especialmente en la canción de How High The Moon, la variación en la frecuencia cromática se traduce en una riqueza armónica y melódica, así como en oportunidades para la improvisación creativa y la expresión individual por parte de los músicos. Además, el jazz tiene el distintivo de ser un género en el cual la improvisación es una de sus características principales. También a menudo utiliza alteraciones en los acordes, como séptimas mayores

o menores, novenas, y otras extensiones. Esto contribuye a una variación en la frecuencia cromática y crea tensiones armónicas que se resuelven de manera expresiva. Por tanto, como el jazz es un género que fomenta la creatividad y la expresión individual, los músicos pueden elegir explorar cromatismos de maneras únicas en cada interpretación, lo que contribuye a la singularidad de cada versión de la canción.

2.1.3.10. Tempograma

El tempograma se refiere a la velocidad de una pieza musical. Más precisamente, el tempo se refiere a la frecuencia del ritmo musical y viene dado por el recíproco del período de tiempo. El tempo a menudo se define en unidades de latidos por minuto (BPM).

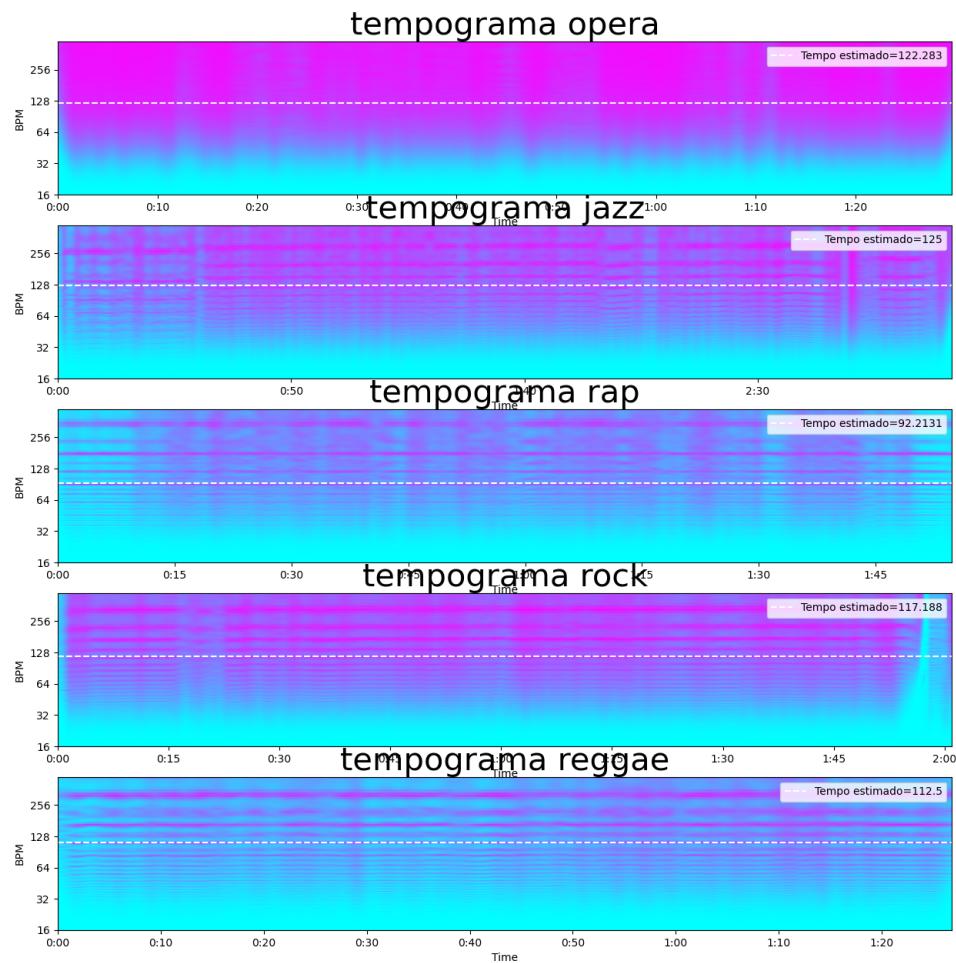


Figura 20: Comparación del tempograma en función de la canción.

Cabe destacar que resulta sorprendente el alto BPM (122) que posee la interpretación de Nes-

sum Dorma, ya que el tempo de una canción influye en la percepción emocional de la obra. La pieza está destinada a ser apasionada y dramática, y un tempo excesivamente rápido podría afectar la capacidad de transmitir estas emociones. Además, “Nessun Dorma.” es una pieza desafiante vocalmente, con notas altas y frases melódicas que requieren control vocal y resistencia. Un tempo extremadamente rápido podría aumentar la dificultad técnica para el tenor, ya que tendría que ejecutar pasajes rápidos y difíciles a una velocidad más acelerada.

Por otra parte, también resulta sorprendente el bajo BPM de la canción de rap. Sin embargo, esto se debe a que es una obra que se encuentra dentro de un disco que se llama Slow Flow y por tanto, tiene sentido que el objetivo del artista sea tener un tempo bajo.

2.1.4. Características de Dominio Cepstral

El análisis de las características cepstrales brinda una herramienta de gran utilidad para la extracción de características relevantes de las señales acústicas. Más concretamente, permite obtener información acerca de las características ocultas en el dominio espectral de una señal de audio. Siendo, estas características, de gran utilidad para aplicaciones como son el reconocimiento de voz o la clasificación de señales de audio.

Esta sección del proyecto se enfoca en dos aspectos fundamentales del dominio cepstral: la transformada inversa de Fourier discreta, conocida como cepstrum, y los coeficientes cepstrales de frecuencia Mel.

2.1.4.1. Cepstrum (Transformada Inversa de Fourier Discreta)

La palabra Cepstrum proviene de la palabra Spectrum. Analizando este concepto desde el punto de vista matemático, podría entenderse como un operador que transforma una convolución en el tiempo en una suma en el dominio espectral. Mediante esta operación, se hace posible separar una señal de audio en dos componentes.

$$C(x(t)) = F^{-1}(\log(F(x(t))))$$

Donde $x(t)$ representa la señal de audio y $F(x(t))$ representa el Spectrum. En otras palabras, también puede entenderse como el espectro del logaritmo del espectro de la señal. En la Figura 21 se puede apreciar el proceso de transformación que sufre una señal de audio hasta obtener su Cepstrum correspondiente.

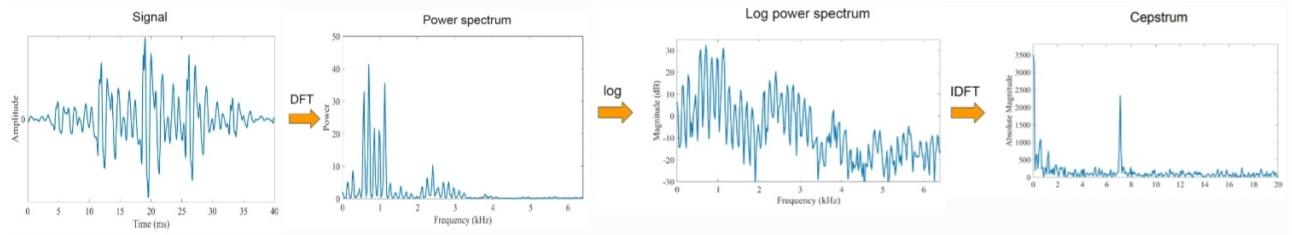


Figura 21: Proceso para obtener el cepstrum a partir de una señal de audio

En la nueva representación de la señal, el eje de abscisas ya no representa la frecuencia, sino que la quefrecuencia (quefrenciy), que es la inversa de la frecuencia. Mientras que, el eje de ordenadas ahora representa la magnitud de las contribuciones de diferentes quefrecencias en la señal analizada. Por último, es importante nombrar que cada uno de los picos del cepstrum reciben el nombre de r-harmónicos, que aportan información valiosa sobre la periodicidad y la estructura espectral de la señal. Para la realización de esta sección se ha usado como inspiración la siguiente tesis [4]

2.1.4.2. Coeficientes Cepstrales de Frecuencia Mel

Los Coeficientes Cepstrales de Frecuencia Mel (MFCCS) son empleados para calcular las características más relevantes del espectro de frecuencias de una señal de audio. El hecho de que estos coeficientes capturen de manera compacta, robusta y apropiada las características más relevantes de una señal de audio, los hace muy útiles para la creación de modelos de procesamiento de señales. Las diferentes fases para la extracción de estos coeficientes aparece representada en la Figura 22.



Figura 22: Fases para la obtención de los coeficientes MFCCS

En primer lugar, se debe aplicar la transformada discreta de Fourier para obtener el Spectrum de la señal de audio. Este proceso ha sido explicado con más detalle previamente en este trabajo. Ya que, se deberá separar la señal en diferentes ventanas, donde se le da un mayor peso a las muestras de los segmentos centrales, de este modo se evitan efectos indeseados de alta frecuencia. Además, gracias al solapamiento (también explicado previamente), no se pierde información como consecuencia del efecto de atenuación de los extremos de cada bloque.

En segundo lugar, se deberá aplicar el logaritmo al Spectrum obtenido en la fase anterior.

En tercer lugar, se aplicará la escala de Mel, explicada en la sección “2.1.3.3 Espectrograma de Mel”. Más concretamente, en esta fase se lo que se hace es aplicar los filtros de mel al log-

espectrograma . De esta forma, al final de este paso habremos dado con el espectrograma de Mel.

En cuarto lugar, se aplicará la Transformada del Coseno Discreta (DCT), esta es una operación equivalente a aplicar la Transformada Inversa de Fourier Discreta. El motivo de usar DCT es que es más eficiente computacionalmente, tiende a decorrelacionar los coeficientes y produce coeficientes muy compactos (menos cantidad de coeficientes necesarios). Matemáticamente se describe de la siguiente forma:

$$X_k = \sum_{n=0}^{N-1} x_n \cos \left(\frac{\pi}{N} k \left(n + \frac{1}{2} \right) \right)$$

Donde N es el número de coeficientes x_n y k puede tomar los valores $k = 1, \dots, N-1$. Y, los coeficientes x_n son los valores de amplitud de los coeficientes espectrales después de pasar por el banco de filtros mel y ser sometidos a una transformada de logaritmo.

La cantidad habitual de coeficientes a calcular es de 12 o 13. Esto es debido a que son los que representan la información más relevante. Además, al seleccionar una cantidad no muy grande de coeficientes se consigue una menor correlación entre ellos facilitando las tareas a algoritmos de clasificación. No solo eso sino que, está probado que 12 o 13 coeficientes son suficientes para representar las características auditivas a las que los humanos son sensibles.

En la figura 23, se ha creado una gráfica para poder representar los 12 MFCCs de una canción de Jazz.

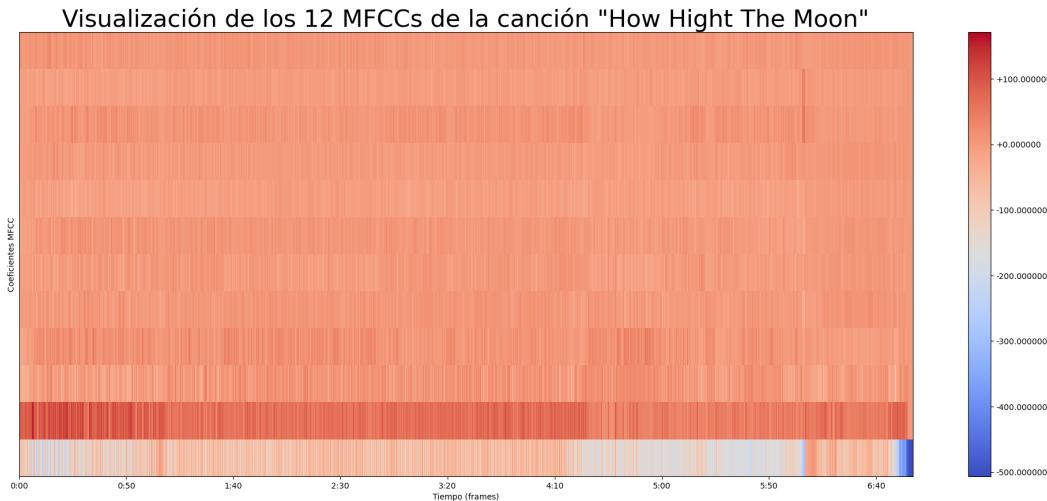


Figura 23: Representación gráfica de los MFCCs de una canción

El eje de ordenadas en la figura anterior representa los coeficientes. Mientras que el eje de abscisas representa el tiempo. Cada una de las franjas verticales que se aprecian en el eje de las abscisas para cada uno de los coeficientes se corresponde con los frames en los que se a

dividido la señal de audio. Para la realización de esta sección se ha empleado como inspiración la siguiente fuente [5]

3. Implementación y resultados obtenidos

En esta sección se pretende explicar el trabajo realizado con el dataset de 100 canciones explicado en la introducción de este documento. Como punto de partida para trabajar con el dataset, lo primero que se ha realizado es una extracción de las características explicada con detalle previamente. Seguidamente, se tratado de hacer uso de dos técnicas diferentes de clasificación. Obteniéndose en la implementación diferentes resultados y conclusiones para cada una de dichas técnicas.

3.1. K-means Clustering (no Supervisado)

Con todas las características (variables) obtenidas, para comprobar si son lo suficientemente descriptivas, se propone realizar un agrupamiento (clustering), de manera que se formen grupos diferenciables del mismo género, para ello se elige el algoritmo de K-means de scikit-learn.

Antes que nada se debe de convertir algunos de los datos que existen de cada variable en un único valor que represente cada una de ellas a lo largo del tiempo, optándose por el valor medio como el más representativo para cada una.

Antes de generar el clustering, se comprueba mediante el método del codo el número de grupos más idóneo en función de la distancias a los centroides,

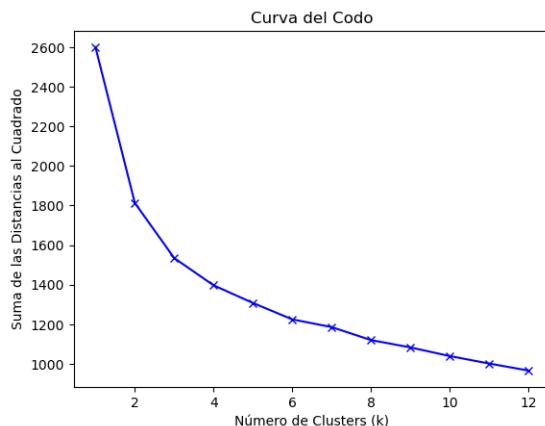


Figura 24: Método del codo

donde inicialmente no se observa un punto de inflexión con claridad como para establecer un

número determinado de grupos, es por ello que se estudian el coeficiente de Silhouette y el índice de Davies-Bouldin,

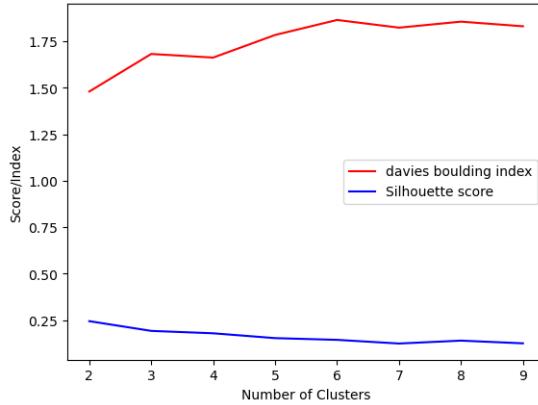


Figura 25: Coeficiente de Silhouette e índice de Davis-Bouldin

en el cual, parece que entre 3 y 6 grupos es el número k de clusters que explicaría mejor un agrupamiento de nuestro conjunto de datos.

No obstante, y debido a que el objeto es comprobar si seríamos capaces de diferenciar cada uno de los géneros usados en este estudio, se entrena el modelo con 5 clusters ($k = 5$), obteniendo el siguiente resultado:

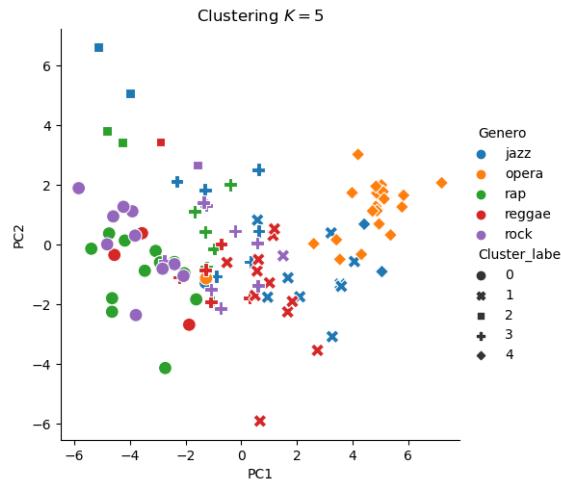


Figura 26: Clustering $K = 5$

Donde se observa únicamente un agrupamiento destacado en el género de ópera, siendo el resto de grupos una mezcla de géneros. Si analizamos el comportamiento del género ópera con las

diferentes variables y géneros, se observa claramente como es el grupo que más se dispersa del resto, manteniéndose unidos.

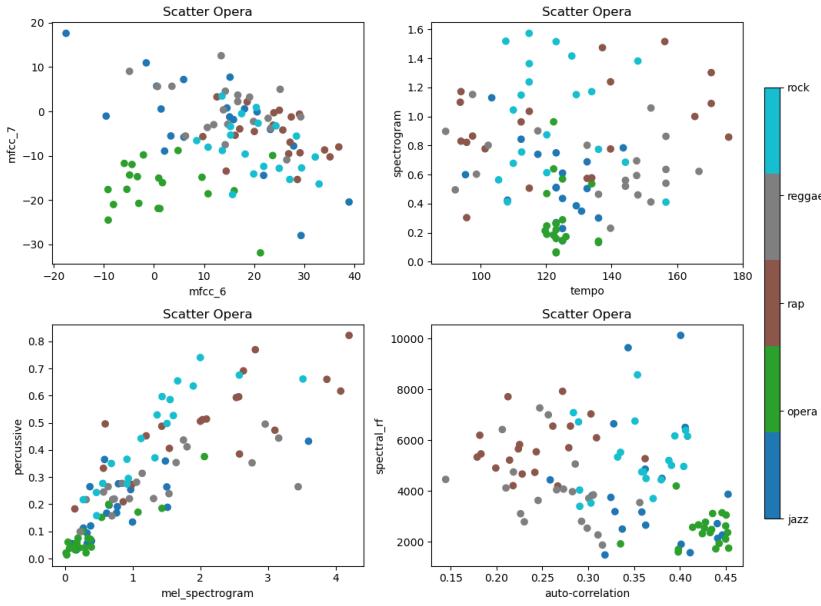


Figura 27: Gráficos de dispersión de diferentes variables

3.2. Supervisado (Random Forest Classifier)

Como método alternativo para comprobar la variabilidad de las características por género, se entrena un Clasificador, esta vez basado en la entropía en lugar de distancias. Para paliar el sobreajuste de los árboles de decisión se decide usar un bosque, en este caso se usa el clasificador de scikit-learn Random Forest Classifier.

Para ello se divide el conjunto de datos en un conjunto de entrenamiento y otro de test de manera balanceada. Debido a que solo se disponen de 20 canciones por género, se usan 80 de ellas para entrenar y 20 para evaluar el modelo.

El entrenamiento se realiza en dos fases, uno inicial para comprobar las puntuaciones iniciales, y posteriormente se hace una búsqueda de variables que optimicen el modelo y una búsqueda básica de hiperparámetros con los siguientes resultados:

	Accuracy Test	Accuracy CV
Inicial	0.63	0.675 (sd 0.086)
Ajustes	0.66	0.71 (sd 0.083)

Tabla 1: Precisión

Una vez establecidos los parámetros, se vuelve a entrenar el modelo con todas las observaciones (100 canciones).

Para implementar el modelo y comprobar con otras canciones que no se han visto implicadas en el proceso se decide descargar 20 canciones más (4 por género) y comprobar el género que devuelve, obteniéndose los siguientes resultados:

	Jazz	Opera	Rap	Reggae	Rock
Accuracy	0.5	0.75	0.5	0.5	1

Tabla 2: Precisión

Analizado los resultados, se comprueba que, pese al poco número de observaciones de entrenamiento, las características extraídas podrían suponer un buen punto de partida para poder clasificar el género en función del análisis temporal, frecuencial, cepstral y temporal-frecuencial.

4. Conclusión

A modo de conclusión, nos gustaría expresar lo que ha supuesto la realización de este trabajo. Haber podido investigar y conocer los diferentes fundamentos teóricos que hay detrás de cada una de las características, documentadas en este proyecto, que se pueden extraer de una señal de audio ha hecho posible el análisis y estudio de diferentes géneros musicales. No solo eso sino que, además, ha sido posible el conocimiento de la programación en Python de las diferentes extracciones. Así como la implementación de algoritmos de machine learning para la clasificación en diferentes géneros a partir la señal de audio de una canción. Donde, mediante la implementación de dichos algoritmos hemos comprobado que con un dataset de apenas 100 muestras es posible entrenar un clasificador con el cual poder obtener resultados positivos.

Referencias

- [1] AGRAWAL, N. Decoding the symphony of sound: Audio signal processing for musical engineering. *Medium* (2023).
- [2] GROUPA. Github repository. <https://github.com/nican2/ProyectoAS2023.git>. [Consulta: 23 de noviembre de 2023].
- [3] LARA, A. I. P. Clasificación de música a través del análisis de señales de audio. *Medium* (2020).
- [4] MARTÍN, F. A. Desarrollo y análisis de clasificadores de señales de audio.
- [5] VELARDO, V. Mel-frequency cepstral coefficients explained easily. <https://www.youtube.com/watch?v=4SH2nfbQZ8list> = *RDCMUCZPFjMe1uRSirmSpznqvJfQstart_radio = 1rv = 4SH2nfbQZ8t = 56*(2020).