CIMAT

Centro de Investigación en Matemáticas, A.C.

Análisis de Datos Multimodales en Colecciones Musicales

T E S I S

Que para obtener el grado de

Maestro en Cómputo Estadístico

Presenta

Marco Tulio Pérez Ortega

Director de Tesis:

Dr. Víctor Hugo Muñiz Sánchez

Autorización de la versión final

Monterrey, Nuevo León, 2021

Espero que este trabajo llegue hasta el cielo.
Para mis abuelos: Juan Ortega, Jovita Ávila y Justino Pérez.
Para mi tío: Justino Pérez y su esposa Araceli García.
Vivo como aprendí de ustedes.

Resumen

En la actualidad, la mayor parte de los datos usados para abordar distintas tareas de aprendizaje máquina, tanto supervisado como no supervisado, son multimodales, es decir, se componen de distintos elementos de información que pueden provenir de distintas fuentes. Aunque tradicionalmente se han analizado las diferentes modalidades por separado, han surgido diferentes metodologías de inteligencia artificial que permiten incorporar todas las modalidades de información de los datos, y por ésta razón, los modelos con datos multimodales se ha convertido en un campo de investigación muy activo actualmente. Para el análisis de colecciones musicales, el enfoque multimodal constituye un reto, debido a la compleja naturaleza de cada modalidad de información, sumado a la poca disponibilidad de recursos de uso libre. Para hacer frente a esta problemática, en ésta tesis creamos y analizamos un conjunto de datos multimodales compuesto por señales de audio, texto, imágenes y una matriz pista-tag con el fin de crear representaciones vectoriales que mejor representen cada pista en nuestro conjunto de datos. Para esto, usamos técnicas de aprendizaje profundo, procesamiento de señales y procesamiento de lenguaje natural. Finalmente verificamos el desempeño de éstas representaciones con tareas de clasificación de género, y recuperación de información, mostrando buenos resultados al compararlos con métodos reportados en la literatura.

Palabras clave: Datos Multimodales, Colecciones musicales, Procesamiento de señales, Espectrogramas, Procesamiento de imágenes, Procesamiento de texto, Datos no estructurados, Perfiles de usuario, Espacio métrico, Sistema de recomendación, Aprendizaje máquina, Aprendizaje profundo, Aprendizaje métrico.

Agradecimientos

Este trabajo sintetiza una larga cantidad de horas de estudio, lluvias de ideas y múltiples líneas de código. Por otro lado, representa una etapa de mi vida que sin duda recordaré con mucho cariño. Y es por eso que quiero comenzar agradeciendo al Centro de Investigación en Matemáticas por darme la oportunidad de ser parte de su institución y permitirme vivir momentos invaluables .

Agradezco a mis compañeros de generación: Aleidali, Ameyalli, Benito, Enrique, Eva, Javier, Judith, Mariano, Orlando, Rafael, Stephania, Tatiana y Víctor por dar un toque mágico a esta historia llamada maestría.

También quiero agradecer a cada uno de los Doctores de la unidad Monterrey por el esfuerzo impreso en cada clase y especialmente al Doctor Víctor Muñiz quien me asesoró durante los dos años y un poco más del programa de maestría.

Agradezco mucho a mí familia. A mi madre Concepción Ortega por su apoyo y amor incondicional, también por repetirme constantemente "esfuerzate y se valiente" , a mí padre Marco Tulio Pérez por sus constantes y valiosos consejos, por enseñarme a siempre dar lo mejor de mí, a mis hermanos Itzel y Uriel por motivarme solo con una sonrisa o un mal chiste. Y a mi abuela Clara Butrón por su hospitalidad y su deliciosa comida las tardes en esas tardes abrumadoras de escribir tesis. A mí pequeño sobrino David Ramirez, por animarme con sus pequeñas travesuras.

Con cariño agradezco a todos mis amigos y familiares por creer en mí, por su aliento, compañía y apoyo. A mi primo Javier Reyes por escucharme hablar repetidamente sobre mi tesis.

Finalmente quiero agradecer al Consejo Nacional de Ciencia y Tecnología por la beca proporcionada durante el programa de posgrado.

...

Índice general

Índice de figuras

Capítulo 1

Introducción

Los datos que se generan actualmente y se comparten en medios digitales son diversos y complejos, entre otras cosas, por las diferentes modalidades de información que los componen. Por ejemplo, un vídeo pude ser identificado por el audio o las imágenes que contiene, además de una descripción textual o simplemente un título. Dicho elemento puede tener distintas apreciaciones en función del entorno del dato relacionado. Los elementos musicales son una muestra muy clara de esto, por lo que en este trabajo exploramos distintos modos de información relacionados a un elemento musical al cual nos referimos como pista o track.

Realizar análisis sobre colecciones musicales con un enfoque de datos multimodales nos da la oportunidad de explorar diversas técnicas de manejo de información para realizar extracción de características. Por otro lado, en la actualidad, el análisis de información multimodal es un concepto de gran importancia, dado que se trabaja con datos de este tipo en otras áreas por ejemplo vease Suris, Duarte, Salvador, Torres, y Giro-i Nieto (2018), por lo que consideramos importante abordar un trabajo que involucre el análisis de colecciones de datos que puedan aprovechar propiedades multimodales.

La importancia de éste trabajo también puede considerarse también desde una perspectiva comercial, dada la gran penetración que actualmente tienen los sistemas de streaming musical, derivando en una competencia por ofrecer el mejor servicio al usuario, creando sistemas de recomendación, listas de reproducción, y sistemas de

busqueda que sean óptimos y satisfagan de mejor manera las necesidades del usuario. Está competencia es principalmente impulsada por el impacto ecónomico que significa estar a la vanguardia en este tipo de sistemas, ya que son un factor determinante en el éxito comercial de este tipo de plataformas tales como Spotify, Deezer o youtube music, por mencionar algunos, dado que en muchos casos el ususario tiende a elegir la plataforma con un sistema de recomendación y busqueda que más comodo o útil le parezca. Lo anterior nos lleva a mencionar otro aspecto importante relacionado a estas plataformas y sus sistemas de recomendación y recuperación de información: el social. Existe una influencia bilateral entre usuarios y plataformas de streaming. Mientras la plataforma, por medio de este tipo de sistemas, puede crear perfiles de usuario de acuerdo a su historial de preferencia, al mismo tiempo tiene la capacidad de marcar tendencia en el usuario con sus sistemas de recomendación. Inclusive, tienen el potencial de impulsar las carreras de artistas emergentes. Así mismo, Los aspectos mencionados anteriormente nos motivan a realizar un proyecto de tesis sobre análisis de datos multimodales para colecciones musicales.

1.1. Objetivos

La música puede ser caracterizada de multiples formas, algo que representa una gran riqueza y flexibilidad de análisis, pero también representa un gran reto. En primer lugar, encontrar una representación adecuada que capture propiedades útiles de datos provenientes de colecciones musicales resulta ser una tarea de alto grado de complejidad. Lo anterior se complica más cuando se asumen diferentes modalidades o fuentes de información, donde es necesario también definir un espacio de representación que combine todas las modalidades. Dicho lo anterior, nuestro objetivo principal en éste trabajo de tesis es encontar un subconjunto adecuado de información así como un espacio vectorial compartido de tal forma que se puedan realizar tareas de aprendizaje supervisado y no supervisado.

Los objetivos particulares que se desprenden de nuestro objetivo principal son:

1. Identificar un subconjunto de información que mejor caracterice los items mu-

sicales de una colección.

2. Usando el subconjunto de información del punto anterior crear represenatciones y modelos que puedan usarse tareas específicas como recuperación de información y clasificación, de tal forma que sean al menos comparables con los modelos existentes en la literatura.

1.2. Contribuciones

Este trabajo contribuye al análisis musical en distintos sentidos, y se pueden enumerar de la siguiente manera:

1. **Modelos creados con archivos de audio de uso libre y de fácil acceso**. Los archivos de audio utilizados para crear los modelos provienen del conjunto de datos **fma** Defferrard, Benzi, Vandergheynst, y Bresson (2017), el cual es un conjunto de archivos de audio junto a sus datos editoriales de uso libre. Hasta donde hemos revisado en la literatura, éste es el primer análisis que se realiza para éste conjunto de datos desde una perspectiva multimodal.

2. **Un nuevo subconjunto de datos para análisis multimodal**. Además de utilizar el conjunto de archivos de audio **fma** enriquecimos el mismo con portadas de álbums y tags culturales obtenidos a través de las apis de spotify y last.fm. Esta información se agrega a un subconjunto de $11,517$ archivos de audio y sus respectivos datos editoriales pertenecientes al conjunto **fma**.

3. **Comparación de espacios métricos.** La construcción de espacios métricos puede realizarse de distintas formas y utilizando distintas combinaciones de modalidades de información, en este trabajo se presenta una comparación de espacios métricos construidos bajo diferentes combinaciones de modos de información. La comparación se realiza bajo dos métricas: precisión media promedio y precisión en k.

4. **Representaciones vectoriales multimodales.** A través de la creación de espacio métricos se obtuvieron distintas representaciones multimodales para cada pista, mismas que pueden utilizarse para tareas de aprendizaje supervisado y no supervisado.

5. **Sistema de Recuperación Multimodal.** Sabiendo que no siempre es posible contar con toda la información relativa a una pista, se propone una arquitectura para un modelo de recuperación multimodal aprovechando características unimodales, ya que el sistema puede recibir información en tres modalidades diferentes: audio, imagen y texto, para después regresar los resultados obtenidos a través del espacio multimodal.

1.3. Organización de la tesis

El resto de éste trabajo de tesis tiene la siguiente estructura:

- **Capítulo 2**. En las últimas dos décadas han surgido una cantidad significativa de trabajos referentes al análisis musical desde una perspectiva computacional, mismos que cimentan las bases del presente trabajo. Es por esta razón que en este capítulo realizamos un repaso por los trabajos que preceden el nuestro, así como una clasificación y descripción de los distintos tipos de información que se relacionan a un ítem musical.

- **Capítulo 3**. Las modalidades en las que se presenta la información relativa a una pista suelen ser diversas, en este capítulo se describen los distintos modos de información que se considerarán para realizar nuestro análisis.

- **Capítulo 4**. El conjunto de datos considerado para nuestro trabajo es un componente elemental, es por eso que en este capítulo se describe a detalle el conjunto de datos utilizado en nuestros experimentos y creación de los espacios métricos.

- **Capítulo 5**. En este capítulo se presenta la metodología seguida para construir los distintos espacios métricos, desde la extracción de características, la concatenación de los distintos modos de información, hasta la función de pérdida utilizada para construir estos espacios y las métricas de evaluación.

- **Capítulo 6**. Después de haber descrito nuestro conjunto de datos y la metodología empleada, en este capítulo pasamos a la experimentación y análisis de resultados obtenidos a través de los distintos espacios creados.

- **Capítulo 7**. Para cerrar nuestro trabajo, dedicamos este capítulo para expresar las conclusiones obtenidas en los distintos sentidos que este trabajo ha tomado, además se abre una senda a futuros trabajos que pueden surgir a partir de este.

Capítulo 2

Antecedentes

La llegada de la música digital junto a la era de la información abrió una senda de investigación bastante amplia, ya que la información que se puede obtener relativa a un ítem musical (pista, álbum, artista, etc.) es bastante amplia y diversa, como puede comprobarse en el catálogo de Spotify, una de las plataformas más populares de streaming musical, que para 2021 ya supera los 70 millones de pistas y sigue creciendo constantemente. La información de este catálogo no se refiere sólo a señales de audio sino a toda la información relacionada a la pista la cual puede presentarse en texto, indicadores numéricos o imágenes, entre otros. En la actualidad los servicios de streaming musical han evolucionado al grado de que podemos encontrar una pista entre un catálogo bastante amplio, mediante consultas de texto, y en algunos casos mediante consultas de melodías; también recibimos sugerencias basadas en nuestro historial de pistas escuchadas. Detrás de esta evolución hay una cantidad considerable de trabajos relativos al análisis musical desde una perspectiva computacional y estadística, comenzando con el problema de definir qué es lo que consideramos música y cómo podemos describirlo de una manera tal que una computadora pueda procesarlo, pasando después a construir indicadores musicales que nos ayuden a caracterizar nuestro universo de pistas y hacer cada vez más eficaces los mototres de busqueda y recomendación de los sistemas de streaming musical. A continuación repasamos algunos trabajos que ha contribuido a la evolución del análisis musical.

2.1. Tipos de información relacionados a un item musical

En 2005 François Pachet, director del laboratorio de Investigación Tecnológica Spotify Creator escribe en Pachet (2005) sobre los tres principales tipos de información que componen a un ítem musical y su relevancia en el análisis de datos, los cuales se describen a continuación.

- **Información editorial**. Esta información se refiere a los datos que se conceden a un ítem musical al momento de su producción, se puede pensar en ellos como "datos de nacimiento " y se le llama así porque son literalmente dados por su editor. Por ejemplo la canción " And I love her " de los Beatles aparece en el álbum "A Hard Day's Night" en el Reino Unido en 1964, en este caso los datos editoriales de la pista son el año, el país, el artista y el álbum, sin embargo puede haber más datos editoriales de la pista, ya que esta información cubre un amplio rango, desde elementos administrativos hasta históricos, por ejemplo festivales en los que se ha tocado, covers, inclusive participaciones en bandas sonoras de películas.

- **Información cultural**. Esta información es determinada por el ambiente o cultura con los que interactúa el ítem musical a diferencia de la información editorial esta no se encuentra explícitamente en algún sistema de información ya que se deriva de una colección de distintas fuentes o documentos. Una de las formas más comunes de obtener esta información es mediante filtrado colaborativo Cohen y Fan (2000). En este caso la información proviene de una colección de perfiles de usuario.

  Además del filtrado colaborativo existen más esquemas de información cultural aplicables a la música, como programas de radio en las que estas pistas son transmitidas o texto encontrado en la web referente a reseñas o descripciones de los ítems musicales.

- **Información Acústica**. Esta información está compuesta por las característi-

cas que componen el audio de los ítems musicales. Esta información pretende ser puramente objetiva ya que representa por el " contenido" de la pista. Desde la perspectiva musical existen bastantes indicadores acústicos que describen una señal de audio, tales como ritmo, tempo, tono, energía y volumen, obtener estos indicadores no es una tarea sencilla y se han propuesto distintos algoritmos para estimar los valores de estos indicadores en una señal de audio. Un elemento que nos otorga un panorama más general de las características acústicas de la pista (y en general de una señal de audio) son los espectrogramas de los que hablaremos con más detalle en la Subsección 3.1.2 y corresponden a una representación en tres dimensiones del tiempo, la frecuencia y la energía de la señal de audio.

Uno de los mayores desafíos en el análisis musical es la subjetividad a la que se encuentra ligada, ya que notaremos que puede haber discrepancia entre cada tipo de información por ejemplo supongamos que el editor de una pista le asigna el género Pop, sin embargo, con el tiempo esta pista se convierte en una precursora del Rock ahora y culturalmente es percibida bajo este género, sin embargo un análisis acústico podría ponerla más cerca de ser Folk. Es por esto que es bastante importante definir que tipo de información usaremos como referencia en un análisis musical.

Particionar de esta forma la información relacionada a los ítems musicales nos es de gran utilidad para explicar las características de los trabajos y aportaciones que se han realizado en este ámbito.

2.2. Conjuntos de datos

La piedra angular de estos trabajos son los datos disponibles para la experimentación, en los últimos años han surgido distintos conjuntos de datos con este objetivo y de los cuales han surgido distintos trabajos de esta índole. Algunos de los conjuntos de datos más populares en la materia son los siguientes:

- **GTZAN.** Este conjunto de datos aparece en 2002 en Tzanetakis y Cook (2002), está compuesto por 1000 pistas asociadas a 10 géneros musicales en este aspecto el conjunto está perfectamente balanceado ya que a cada género le corresponden 100 pistas. Cada pista está representada por archivos de audio mono de 16 bits a $22,050Hz$ en formato **.wav**. Este conjunto de datos tiene varias limitaciones, comenzando por el hecho de ser un conjunto no tan grande como se requiere para aplicar las técnicas de aprendizaje profundo que se utilizaran en este trabajo, sumado a esto la falta de datos editoriales como nombre de la pista, artista, álbum o año de lanzamiento.

- **TagATune.** Este conjunto aparece en 2007 en Law, Ahn, Dannenberg, y Crawford (2007), conjunto compuesto por $5,405$ pistas de 230 artistas. Cada pista está representada por archivos de audio mono a $16,000Hz$ en formato **.mp3** y cuenta con información editorial como artista, nombre de la pista, género e inclusive tags, sin embargo dejó de estar públicamente disponible en 2009.

- **Million Song Dataset**. Publicado en 2011 por Bertin-Mahieux, P.W. Ellis, Whitman, y Lamere (2011) este conjunto es una colección de un millón de pistas de uso libre acompañadas de su respectiva información editorial. El tamaño de este conjunto de datos es ideal para trabajos de aprendizaje profundo, sin embargo tiene un gran inconveniente, ya que la representación de cada pista está compuesta por características de audio de medio y alto nivel dadas por Echonest y no directamente por archivos de audio que es lo deseable para la mayoría de experimentos con propiedades acústicas.

- **FMA** El *Dataset For Music Analysis* presentado en 2017 en Defferrard y cols. (2017) está compuesto a la fecha de este trabajo (julio 2021) por $106,574$ archivos de audio junto a distintos datos editoriales como artista, nombre, álbum y fecha de lanzamiento. Las pistas están representadas por archivos de audio stereo de $128kb/s$ a $44,100Hz$ en formato **.mp3**. Este conjunto es el que más se adecua a nuestros propósitos y hablaremos a detalle de él en el Capítulo 4.

En las siguientes secciones se describirán algunos trabajos realizados con los distintos tipos de información y sus resultados más relevantes. En general todos utilizan información editorial de algún modo, sin embargo haremos más énfasis en la información utilizada para caracterizar la pista.

2.3. Trabajos con información acústica

Los trabajos con información acústica aprovechan las señales de audio para extraer características que puedan representar adecuadamente una pista por ejemplo en Tzanetakis y Cook (2002) se realiza una clasificación de género para el conjunto de datos **GTZAN** utilizando características numéricas de la señal de audio que median indicadores como ritmo, tono o timbre, está clasificación alcanzó una exactitud del 61 %.

Más tarde en Grzywczak y Gwardys (2014) publicado en 2014, se realiza un trabajo de clasificación de género utilizando el mismo conjunto de datos **GTZAN** la metodología consiste en tomar dos tipo de espectrogramas por pista, uno melódico y otro rítmico para después realizar una extracción de características usando redes convolucionales, una vez extraídas las características pasan por un clasificador **SVM** para predecir el género, la exactitud alcanzado con esta metodología fue del 78 %, mejorando de gran forma a su predecesor.

En 2018 en Murauer y Specht (2018) se presentan distintos modelos para clasificar canciones del conjunto de datos **fma** en 16 géneros musicales. Se utilizan dos representaciones acústicas en este trabajo, la primera son los llamados MEL espectrogramas y la segunda compuesta por indicadores numéricos representando distintas características acústicas, referentes al ritmo, tono, energía o volumen de la pista, cabe mencionar que esta clasificación no es balanceada ya que algunos géneros aparecen con mayor frecuencia que otros. El mayor f_1 score (Pastor-Pellicer, Zamora-Martínez, España-Boquera, y Castro-Bleda (2013)) obtenido en este trabajo es de 78 %.

2.4. Trabajos con información cultural

Los trabajos con información cultural tratan de aprovechar los datos que existen entorno a una pista como por ejemplo el trabajo presentado en 2014 en Choi, Lee, y Downie (2014) donde se realiza una clasificación del tema sobre el que habla la canción, teniendo diez temas posibles (guerra, religión, drogas,sexo, madre, lugares, engaño, tisteza, soledad y noviazgo) caracterizando cada canción mediante reviews y opiniones en distintos foros además de la letra de la canción. El mejor resultado alcanzado es una exactitud de 54.11 %.

Dos años más tarde en Oramas, Espinosa-Anke, Lawlor, y et al. (2016), se utiliza el mismo tipo de información, exceptuando la letra, para realizar una clasificación de 1300 álbums en 13 géneros diferentes (100 álbumes por género) logrando una exactitud de 69.1 % .

En 2020 en Korzeniowski y cols. (2020) se utilizan filtros colaborativos para crear matrices usuario-pista y obtener embeddings que captaran la relación social que guardan las pistas basado en las preferencias de los usuarios y con esto crear un modelo que clasifique las pistas según el "estado anímico" que estas transmitieran, . Las pistas en cuestión fueron tomadas del **Million song Dataset** y es necesario resaltar que cada pista podía corresponder a más de un estado anímico, es decir se trató de un problema multietiqueta. El número de estado anímicos considerado fue de 188 para un total de 66993 pistas. Finalmente el modelo que caracterizaba las pistas mediante la matriz pista usuario obtuvo una precisión promedio de 47 % y fue comparado con uno que usaba información acústica extraída con mel espectrogramas y redes convolucionales el cual obtuvo una precisión promedio de 32 %. Superandolo en este sentido.

2.5. Trabajos con información multimodal

Finalmente, los trabajos de información multimodal combinan distintos modos de información para obtener representaciones vectoriales que ayuden a alcanzar los objetivos planteados.

En Oramas, Barbieri, Nieto Caballero, y Serra (2018) se realiza una clasificación multi etiqueta de los géneros a los que pertenece un álbum musical, utilizando tres fuentes de información distintas:

1. La portada del album.
2. El espectrograma combinado de las pistas que componen el álbum.
3. Reviews y opiniones sobre el álbum.

La extracción de características se realizó usando redes convolucionales en los dos primeros casos y métodos de procesamiento de lenguaje natural para las reviews y opiniones. Las pistas utilizadas en este trabajo provienen del **Million Song Dataset** y los archivos de audio fueron recuperados de 7digital.com[1]. El experimento contó con $31,471$ álbumes y 250 géneros musicales obteniendo un AUC-ROC de $93.6\,\%$

También en este trabajo se realiza una clasificación sencilla no balanceada de género para $30,713$ pistas pertenecientes a 15 géneros diferentes utilizando las siguientes fuentes de información:

1. La portada del álbum donde se incluye la pista.
2. El espectrograma de la pista.

Logrando obtener un f_1-score de $42.7\,\%$.

En 2020 en Won, Oramas, Nieto, Gouyon, y Serra (2020) se presenta un modelo de recuperación de información basado en tags utilizando tres fuentes de información:

1. Una matriz pista-usuario.
2. Espectrogramas de las pistas.
3. Los tags asociados a las pistas.

[1] https://www.7digital.com/

La idea principal en este trabajo es construir un espacio compartido en el que se puedan representar pistas y tags, con el fin de recuperar la pista más cercana a un tag otorgado. Las pistas y la matriz pista-usuario con las que se construye este modelo pertenecen al **Million Song DataSet** y sus archivos de audio fueron recuperados de 7digital.com, mientras que los tags asociados fueron obtenidos de last.fm[2], en total se consideran los 100 tags más populares para construir la base del espacio métrico y $115,000$ pistas, el modelo alcanza una precisión en 10 de $31.2\,\%$ cuando se considera la matriz pista-usuario junto a los espectrogramas de las pistas, $32\,\%$ cuando solo se considera a la matriz pista-usuario y $35\,\%$ cuando solo se consideran los espectrogramas de las pistas .Este trabajo es nuestro referente principal y será constantemente citado a lo largo del texto.

[2]https://www.last.fm/

Capítulo 3

Modalidades de información

En este capítulo hablaremos de los distintos modos de información que se considerarán en este trabajo, clasificados en tres tipos de información: Acústica, editorial y cultural.

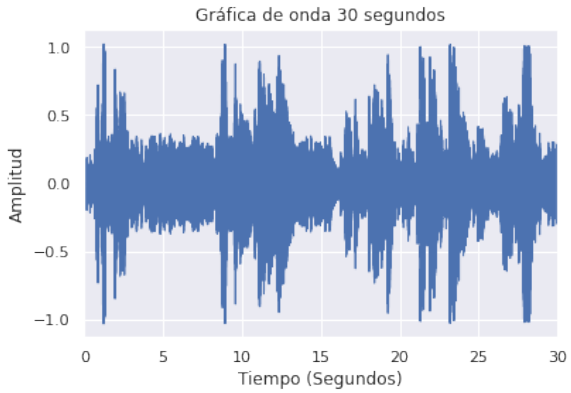
3.1. Información Acústica[1]

La información acústica de una pista es aquella que se encuentra en las señales auditivas producidas por los diferentes sonidos que la componen, caracterizar está información es una tarea de bastante importancia además de compleja. En este trabajo nos enfocaremos en dos caracterizaciones de la información acústica, una visual usando espectrogramas y una numérica mediante el cálculo de indicadores que resumen el comportamiento de las señales de audio. Para poder entender mejor cómo se realiza la caracterización de la información acústica repasaremos algunos conceptos relacionados al sonido y a la música en concreto.

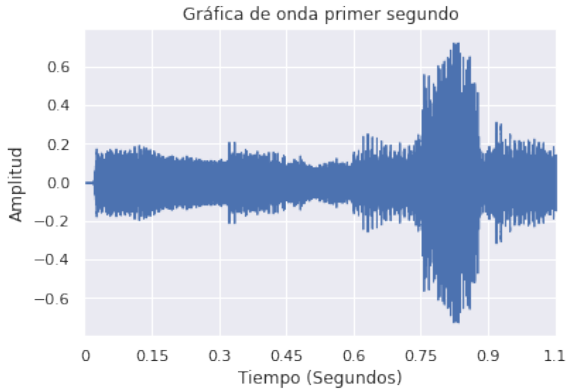
3.1.1. Principales características del sonido

El sonido es generado por la vibración de un objeto, estas vibraciones causan desplazamientos y oscilaciones de moléculas de aire, resultando en regiones locales

[1]El contenido de esta sección retoma conceptos de Müller (2015)



(a) Primeros 30 segundos



(b) Primer segundo

Figura 3.1: Gráfica de onda para una pista compuesta por piano y voz

de compresión y rarefacción generando una presión alternante que viaja en forma de ondas a través del aire, desde la fuente hasta el oyente. En otras palabras, el sonido puede ser visto como variaciones en la presión del aire. En la figura Figura 3.1 podemos observar gráficas de onda del sonido para una pista compuesta principalmente por piano y voz, estas gráficas muestran cómo la presión en el aire se desvía de la presión promedio la cuál está fijada en cero. Si los puntos de máxima y mínima presión se alcanzan alternadamente y de forma regular la gráfica de onda resultante es llamada periódica. En este caso el período de una onda se define como el tiempo requerido para completar un ciclo, es decir el tiempo que tarda la onda desde que pasa por un punto de presión p_0 hasta volver a pasar por ese punto de presión. La frecuencia, medida en Hertz es lo recíproco del período, el número de ciclos que se completan por unidad de tiempo. La relación entre el sonido y la frecuencia es positiva entre más alta es la

frecuencia más alto será el sonido, el rango audible para los humanos oscila entre los $20Hz$ y $20,000Hz$. Podemos definir entonces la gráfica de ondas como una función del tiempo que mapea la presión en el aire a está función se le conoce como **función de señal audio**, una de las funciones más sencillas y que mejor representan este comportamiento son las sinusoidales, estás funciones pueden ser consideradas como el prototipo de una nota musical, ya que en ocasiones el sonido resultante de una función sinusoidal es llamado sonido armónico o tono puro. Por ejemplo, una función sinusoidal con frecuencia de 440 Hz corresponde a la nota A4. Dos frecuencias se perciben como iguales si difieren en una potencia de dos, por ejemplo A3(220Hz), A3(440Hz) y A4(880Hz). Además la distancia percibida entre A3 y A4 es la misma que la percibida entre A4 y A5, por tanto decimos que la percepción humana de las notas es de naturaleza logarítmica. La frecuencia de las notas de la escala musical está dada por la siguiente expresión

$$F(p) = (440)2^{(p-69)/12} \qquad p \in [0:127] \tag{3.1}$$

note que el número 69 corresponde a la nota A4, el 70 a A4#, 71 a B4, 72 a C5 y así sucesivamente.

En el mundo real los sonidos están lejos de ser un simple tono puro con una frecuencia bien definida, una simple nota tocada por algún instrumento puede resultar en un sonido complejo que contiene una mezcla de frecuencias cambiando en el tiempo (como puede apreciarse en la figura Figura 3.1), este tono musical puede ser descrito como una superposición de tonos puros o funciones sinusoidales, cada uno con sus propias características. Cada una de las funciones sinusoidales que describen un tono musical es llamada parcial, la frecuencia de la parcial más baja es llamada la frecuencia fundamental del sonido.

Volumen, intensidad y poder

Otro factor que debe considerarse en la información acústica es el volúmen con el que se tocan las notas musicales, por ejemplo una nota de piano puede ser tocada

fuerte o suavemente, creando una diferencia de volúmen en el sonido emitido. Para el volúmen los sonidos son ordenados desde silenciosos hasta ruidosos. El poder de sonido, medido en Watts (W), expresa que tanta energía emite un sonido por unidad de tiempo a través del aire. El término intensidad del sonido se usa para denotar el poder de sonido por unidad de área, la intensidad mínima de sonido de un tono puro que un humano puede percibir (threshold of hearing) es tan pequeño como

$$I_{TOH} = 10^{-12} W/m^2 \tag{3.2}$$

Mientras que la intensidad de sonido máximo (threshold of pain) es de

$$I_{TOP} = 10 W/m^2 \tag{3.3}$$

Por razones prácticas suele medirse la intensidad del sonido en escala logarítmica, más precisamente en escala de decibeles, la relación entre la intensaidad I y los decibeles es la siguiente:

$$dB(I) = (10) log_{10}(\frac{I}{I_{TOH}}) \tag{3.4}$$

De esta forma $dB(I_{TOH}) = 0$ y duplicar la intensidad resulta en incrementar aproximadamente tres decibeles.

$$dB(2I) = (10) log_{10}(2) + dB(I) \approx dB(I) + 3 \tag{3.5}$$

Timbre

Además de la nota, el volumen y la duración, existe otro aspecto fundamental del sonido conocido como timbre o color del tono. El timbre nos permite distinguir el tono musical de un violín, una guitarra o una trompeta aunque se haya tocado la misma nota al mismo volúmen. Tener una medida de timbre es una tarea poco sencilla por lo que suele ser descrito de una forma indirecta. Investigadores han tratado de aproximar el timbre observando las correlaciones con características de sonido más objetivas como la evolución temporal y espectral la ausencia o presencia de tonos

y componentes de sonido o la energía distribuida a través de las parciales de un tono musical. La composición de un sonido en términos de sus parciales puede ser visualizado en un **espectrograma** tema que abordaremos en la sección 3.1.2.

3.1.2. Espectrogramas

Como se mencionó anteriormente, el espectrograma es una representación visual del sonido, en el cuál podemos observar la intensidad de las frecuencias a través del tiempo. Un espectrograma puede también puede ser visto como un arreglo matricial s_{ij} en el que para cada $i-$ ésima partición de rango de frecuencia y para cada $j-$ ésima partición de tiempo se obtiene la intensidad de la frecuencia mediante la transformada de Fourirer discreta de corto plazo (Short-Time Fourier Transform. STFT). La idea princicpal de la transformada es comparar las señales de audio con funciones sinusoidales con diferente frecuencia ω. Cada función sinusoidal puede ser pensada como un prototipo de oscilación. Como resultado, para cada valor de ω considerado obtenemos un coeficiente de magnitud: $d_\omega \geq 0$. Para comparar funciones la transformada de fourier utiliza el producto punto para funciones definido como:

$$d = \int_{t \in R} f(t)g(t)dt \tag{3.6}$$

donde $f(x)$ es la función de la señal de sonido y $g(x)$ su aproximación sinusoidal. expresando la función sinusoidal en su representación exponencial se tiene que:

$$g(t) = exp(-2\pi i \omega t) \tag{3.7}$$

de esta forma la transformada de Fourier para un ω dado queda definida como :

$$\widehat{f}(w) = \int_{t \in R} f(t)exp(-2\pi i \omega t)dt \tag{3.8}$$

De esta forma si nuestra función de señal de audio se aproxima a la función sinusoidal para un ω la transformada de Fourier obtendrá una valor grande, en caso contrario la transformada de Fourier obtendrá un valor pequeño.

Transformada de Fourier discreta

Aunque el oído humano puede percivir las señales de audio como una función continua en el tiempo, debemos tener en cuenta que para una máquina es más bien discreta, ya que para poder guardar la información de señal particiona la escala de tiempo en intervalo muy pequeños equidistantes y almacena en un arreglo numérico los valores de amplitud para esos valores de tiempo pertenecientes a la partición. Teniendo esto en cuenta necesitamos pasar de la definición de la transformada de Fourier continua a una expresión discreta y esto se logra reemplazando la integral por su equivalente para funciones discretas: la sumatoria. Formalmente si definimos $f : \mathbb{R} \longrightarrow \mathbb{R}$ como la función de señal de audio y sea $T > 0$ un valor real definimos $x : \mathbb{Z} \longrightarrow \mathbb{R}$

$$x(n) = f(nT) \tag{3.9}$$

La función $x(n)$ es llamada la muestra tomada en el tiempo $t = nT$ de la señal original f. Este proceso es conocido como $muestreoT$ donde el número T es llamado **período de muestra** y su inversa

$$F_s = \frac{1}{T} \tag{3.10}$$

es llamada **tasa de muestreo** De esta forma la transformada de Fourier discreta queda definida como

$$\sum_{n\in\mathbb{Z}} Tx(n)exp(-2\pi i\omega nT) \approx \widehat{f}(\omega) \tag{3.11}$$

si hacemos $T = 1$ obtenemos

$$\widehat{x}(\omega) = \sum_{n\in\mathbb{Z}} x(n)exp(-2\pi i\omega n) \tag{3.12}$$

Y podemos observar que

$$\widehat{x}(\omega) \approx \frac{1}{T}\widehat{f}\left(\frac{\omega}{T}\right) \tag{3.13}$$

Para realizar cálculos computacionales aún tenemos un par de inconvenientes:

1. La sumatoria de la Ecuación 3.12 involucra un número infinito de sumandos

2. El parámetro ω es continúo.

Para el primer problema supondremos que la información más importante de f está limitada a cierta duración en el tiempo. Es decir suponemos que el valor de la señal es cero para un tiempo que esté fuera de un intervalo definido con límite inferior en cero. De está forma sólo necesitamos considerar un número finito de muestras $x(0), x(1), ..., x(N-1)$ para $N \in \mathbb{N}$ con esto la suma de la Ecuación 3.12 se vuelve finita. Para el segundo problema, al igual que se hace con el tiempo se calculará la transformada de Fourier para un número finito de frecuencias considerando muestras sobre los valores de frecuencias $\omega = \frac{k}{M}$ para $M \in \mathbb{N}$ y $k \in [0 : M-1]$. En la práctica suele hacerse $M = N$ de está forma podemos hacer que la transformación sea invertible y además nos resulta un algoritmo computacionalmente eficiente para calcular la transformada. Con estas consideraciones transformamos la Ecuación 3.12 en:

$$X(k) = \widehat{x}(k/N) = \sum_{n=0}^{N-1} x(n) exp(-2\pi i k n/N) \tag{3.14}$$

Es importante resaltar que solamente es necesario considerar los coeficiente de $X(k)$ para $k \in [0 : \lfloor N/2 \rfloor]$ debido a que la transformada de Fourier tiene propiedades de simetría y considerar los valores de la mitad superior sería redundante.

Transformada de Fourier Local Discreta (STFT)

La transformada de Fourier obtiene información de la frecuencia para un dominio de tiempo entero, por lo que la información del momento en que ocurren estás frecuencias queda oculta en la transformada. Es por eso que en 1946 Dennis Gabor introduce la Transformada de Fourier Local Discreta o **STFT** por sus siglas en inglés, donde en lugar de considerar la señal entera solo se considera una fracción de ella. Con este fin se establece una función llamada **función ventana** la cuál toma valores diferentes de cero solo para un período de tiempo corto. De esta forma la señal se multiplica por la función ventana w para obtener información de un período de tiempo específico, está ventana se va moviendo en el eje del tiempo y se calcula la transformada de Fourier discreta en cada fracción de tiempo considerado. El tamaño

de la ventana determina cuánto dura la una sección de señal determinada, es decir si se han realizado un muestreo F_s Hertz por segundo, y siendo N el tamaño de la ventana considerada, cada sección de señal durará $\frac{N}{F_s}$ segundos. Para la STFT discreta es necesario considerar un parámetro más: el tamaño de paso H, el cuál determina la diferencia temporal entre dos ventanas subsecuentes. De esta forma la STFT discreta X de la señal x está dada por:

$$X(m,k) = \sum_{n=0}^{N-1} x(n+mH)w(n)exp(-2\pi ikn/N) \tag{3.15}$$

Donde $m \in \mathbb{N}$ y $k \in [0:K]$ con $K = N/2$ y el número complejo $X(m,k)$ denota el $k-$ésimo coeficiente de Fourier para la m-ésima ventana de tiempo, observemos que para cada valor de m obtendremos un vector de tamaño $K+1$, note que si hacemos $H = 1$ obtendremos vectores con valores muy parecidos debido a que ventanas subsecuentes compartirán bastante información, por lo que suele usarse un tamaño de paso $H = N/2$ con lo que, además de reducir esta redundancia, logramos un equilibrio entre resolución temporal y el volúmen de vectores calculados. Cada coeficiente $X(m,k)$ estará relacionado con la posición temporal de la siguiente forma:

$$T_{coef}(m) = \frac{mH}{F_s} \tag{3.16}$$

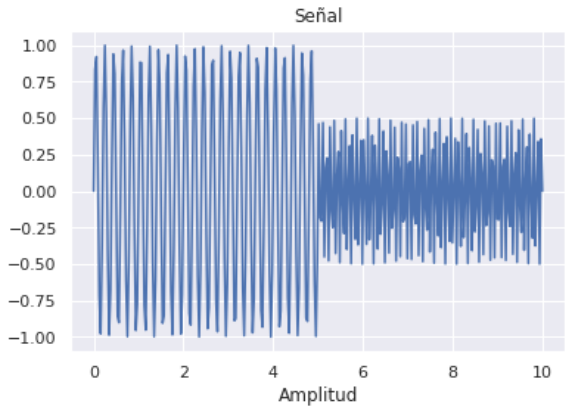
y con un valor de frecuencia determinado por:

$$F_{coef}(k) = \frac{kF_s}{N} \tag{3.17}$$

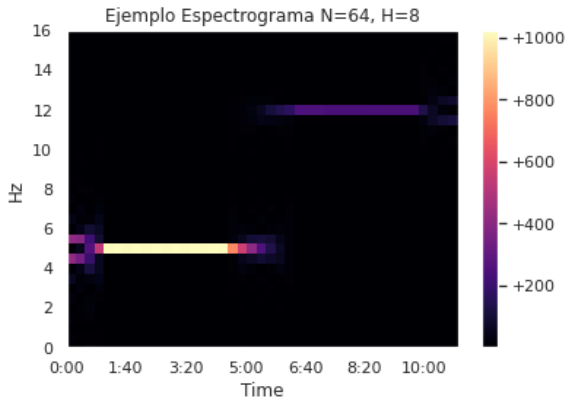
Finalmente el espectrograma de una señal de audio estará dado por:

$$Y(m,k) = |X(m,k)|^2 \tag{3.18}$$

Por ejemplo, considerando una señal de audio de 10 segundos, como la que se muestra en la imágen 3.2a si consideramos una tasa de muestreo de $32HZ$ obtendremos un vector con 320 muestras. Además tomando un tamaño de ventana de $N = 64$ muestras



(a) Gráfica de la señal



(b) Espectrograma

Figura 3.2: Gráfica de onda y espectrograma para una señal de 10 segunos, podemos apreciar dos claros comportamientos de frecuencia y amplitud, los cuáles se ven reflejados en el espectrograma

y un tamaño de paso $H = 8$ muestras obtendremos un espectrograma de dimensión 33×41 como el que se muestra en 3.2b. La señal de 3.2 es bastante sencilla ya que consta de solo dos comportamientos constantes, los primeros 5 segundos tiene una amplitud mayor que los últimos 5 segundos lo cual se ve reflejado en el mapa de calor del espectrograma, por otro lado, la frecuencia de los últimos 5 segundos de la señal es mayor que la de los primeros 5, lo cuál también se ve reflejado en el eje de la frecuencia del espectrograma.

.

Función ventana

La función ventana juega un papel importante en la obtención de los espectrogramas, en la literatura podemos encontrar distintas funciones que han sido definidas para ciertas tareas en específico relacionadas al procesamiento de señales, el espectrograma de la figura 3.2 se usó una **función rectangular**, la cuál está definida como:

$$w_{rectangular}(n) = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & \text{en otro caso} \end{cases} \tag{3.19}$$

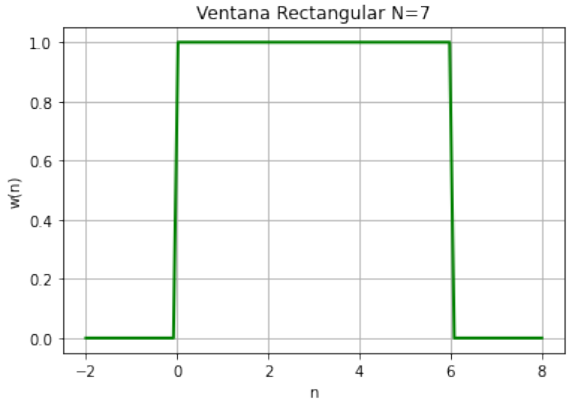
Donde M es el conjunto de índices dentro de la ventana temporal, de esta forma los valores que están fuera de la ventana son iguales a cero y los que están dentro tendrán el mismo peso. En bastantes ocasiones es preferible que los valores de la ventana tengo menor peso conforme se acercan a su frontera, una de las funciones que logra este comportamiento es la **función triangular** definida como:

$$w_{triangular}(n) = \begin{cases} 1 - |\frac{n-\frac{N}{2}}{\frac{N}{2}}| & 0 \leq n \leq N-1 \\ 0 & \text{en otro caso} \end{cases} \tag{3.20}$$

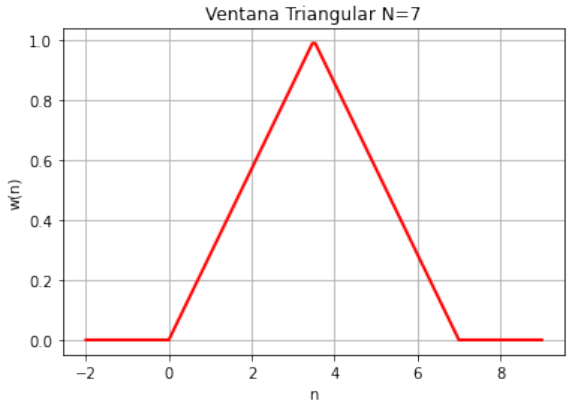
Sin embargo para señales de audio suele usarse una versión suavizada, como la **función de Hann** definida por:

$$w_{Hann}(n) = \begin{cases} sen^2(\frac{\pi n}{N}) & 0 \leq n \leq N-1 \\ 0 & \text{en otro caso} \end{cases} \tag{3.21}$$

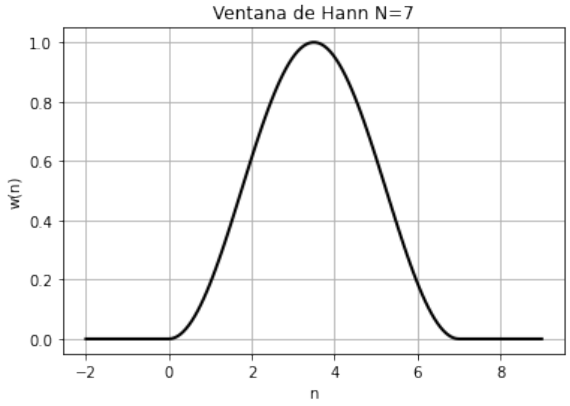
En la figura 3.3 podemos observar el comportamiento de cada una de las funciones ventana mencionadas anteriormente, en este trabajo usaremos la ventana de Hann, ya que da menor peso a los valores dentro de la ventana que se aproximan a la frontera además de ser una función suave en comparación a la ventana con función triangular adecuandose de mejor manera a la percepción humana de la frecuencia de la onda.



(a) Ventana rectangular



(b) Ventana triangular



(c) Ventana de Hann

Figura 3.3: a) Función ventana rectangular, b) Función ventana triangular, c) función ventana da Hann

Mel-Espectrograma

En la Subsección 3.1.1 se mencionó que la percepción humana de las notas es de naturaleza logarítmica ya que percibimos la misma diferencia de tonalidad entre una frecuencia de 110Hz a un de 220Hz y una de 220Hz a una de 440Hz, es por esto que en la práctica se suelen usar espectrogramas con escalas de frecuencia logarítmicas, con lo que se logra obtener una representación de notas musicales equidistantes en el eje de frecuencias. Dada una nota musical p en el intervalo $[0:127]$ podemos obtener su frecuencia mediante la Ecuación 3.1, con el fin de optimizar la cantidad de información extraída de una señal de audio ligada a una pista musical es común calcular sólo los coeficientes de la transformada de Fourier $X(m,k)$ solo para las frecuencias de estas 128 notas musicales, definiendo el siguiente conjunto:

$$P(p) = k : F_{nota}(p - 0.5) \leq F_{coef}(k) < F_{nota}(p + 0.5) \tag{3.22}$$

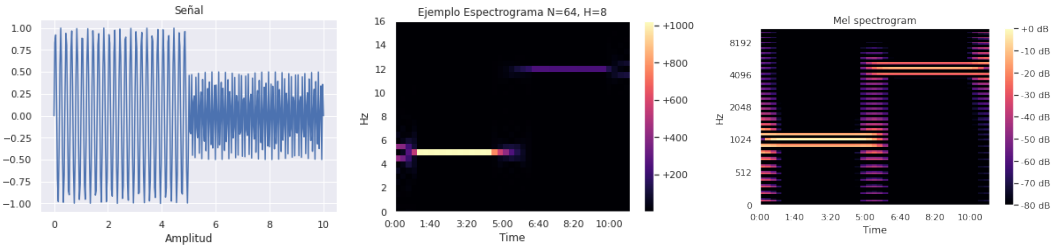
es decir agrupamos las frecuencias más cercanas a las frecuencias correspondientes a las notas musicales y calculamos los valores del espectrograma mediante:

$$Y_{LF}(m,p) = \sum_{k \in P(p)} |X(m,k)|^2 \tag{3.23}$$

Finalmente, transformamos estas magnitudes a una escala de decibeles mediante la Ecuación 3.4 para obtener lo que se conoce como mel- espectrograma o espectrograma en escala de mel. En 3.4 podemos observar la señal de audio junto a sus respectivos espectrograma y mel-espectrograma, podemos observar que el espectrograma cuenta con mayor resolución en comparación con su similar en escala de mel, sin embargo la información que ofrece es en esencia la misma.

3.1.3. Indicadores acústicos

Otro modo de caracterizar la información acústica es el derivado de las propiedades estadísticas que guardan las señales de audio, en este trabajo utilizaremos la librería de python Essentia Bogdanov y cols. (2013) para obtener indicadores acústicos referentes



(a) Grafica de la señal **(b)** Espectrograma **(c)** Mel-Espectrograma

Figura 3.4: a) Señal de audio, b) Espetrograma, c) Mel espectrograma

al ritmo, energía, volumen y tonalidad de la pista.

Características de ritmo

El ritmo se define como la forma de sucederse y alternar una serie de sonidos que se repiten periódicamente en un determinado intervalo de tiempo. Un concepto fuertemente relacionado al ritmo es el **tempo** el cual hace referencia a la velocidad con la que se ejecuta una pieza musical. Comúnmente se suele indicar en pulsaciones por minuto entre más pulsaciones por segundo mayor será la velocidad, **essentia** utiliza el algoritmo **BeatTrackerDegara** presentado en Degara y cols. (2012) para estimar el tempo de una pista dada su señal de audio, en la Figura 3.5 se muestran 30 segundos de tres señales de audio correspondientes a tres pistas, la primera del género rock exhibe el tempo más rápido con un estimado de 144.24 pulsaciones por minuto, la segunda del género electronic con un tempo más lento estimado de 117.75 pulsaciones por minuto y la última con el tempo más lento estimado de 88.96 correspondiente al género Classical. Las marcas rojas indican la posición estimada de los pulsos en la pista observemos que el número de marcas son proporcionales al tempo estimado en cada pista.

Otro indicador referente al ritmo es la danzabilidad o bailabilidad, que mide que tan bailable es la pista dada la señal de audio, el algoritmo de essentia obtiene un indicador entre 0 y 3 donde la pista es más danzable conforme el indicador se acerca a 3. El algoritmo para calcular este indicador se presenta en Streich y Herrera (2012) y se basa en el análisis de correlaciones dentro de series temporales con diferentes escalas de tiempo.

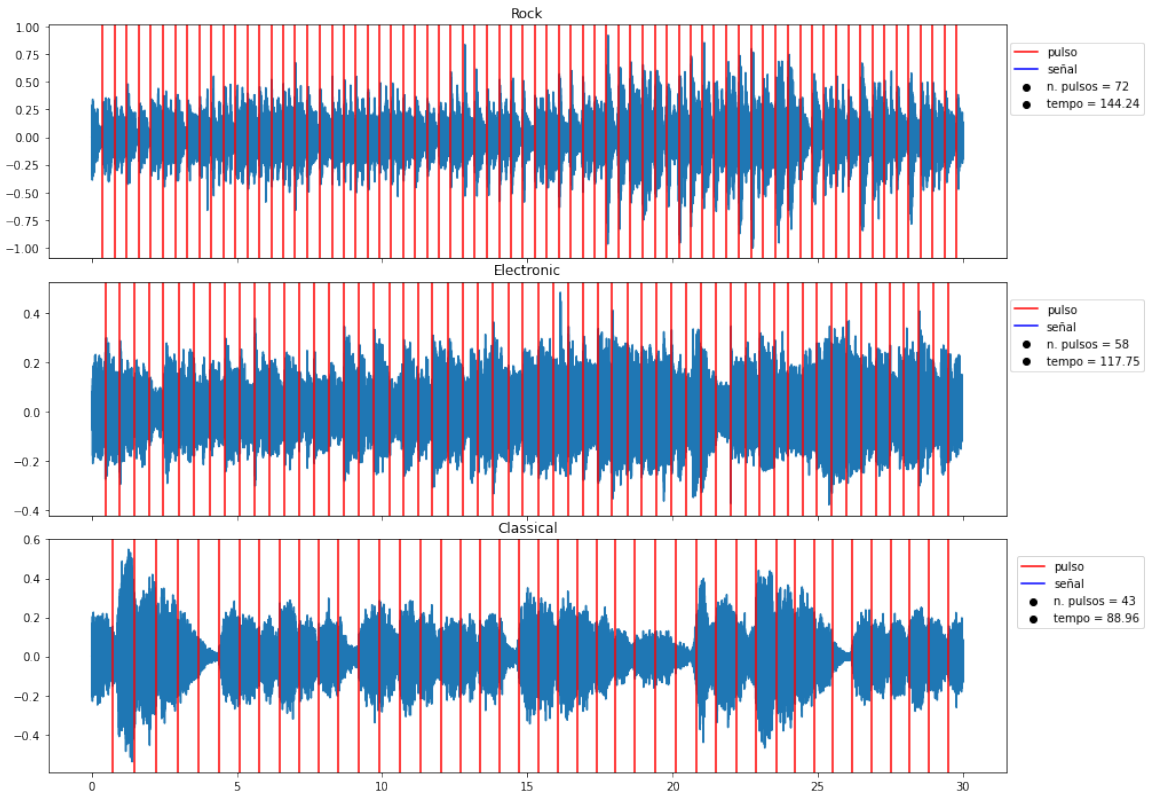


Figura 3.5: Pulsaciones detectadas para tres pistas con géneros musicales diferentes.

3.1.4. Energía

La energía se define como la capacidad de generar cambio, en el caso de las señales de audio este cambio se presenta en la presión de audio. Dada una señal de audio $\mathbf{x}$, el indicador de energía se calcula de la siguiente forma:

$$E = \sum_{i=0}^{n} |x(i)|^2 \tag{3.24}$$

3.1.5. Indicadores de tono

Podemos usar algunos indicadores relacionados con el tono de la pista dada una señal de audio :

- Nota estimada.
- Escala de la nota.

- fuerza de la nota.
- Velocidad de cambio de acordes.
- Acorde más frecuente.
- Proporción del número de acordes detectados respecto al total de acordes.
- Escala de la progresión de acordes.
- Fuerza en cada acorde detectado, tomando los siguientes descriptores estadísticos:
  - Número de acordes en la progresión.
  - Media de la fuerza en los acordes.
  - Desviación estándar de la fuerza en los acordes.
  - Asimetría de la fuerza en los acordes.
  - Curtosis de la fuerza en los acordes.

Estos indicadores se calculan mediante una serie de algoritmos propuestos en Temperley (1999), Fujishima (1999) y Gómez (2006).

3.1.6. Indicadores de duración y silencio

Estos indicadores se calculan mediante una secuencia del poder de sonido de la señal de audio, la idea principal para detectar un desvanecimiento del sonido se basa en el valor del poder de sonido x, si este valor disminuye, entonces se infiere que comienza un desvanecimiento, si el valor aumenta, entonces podemos inferir que el desvanecimiento ha terminado.

Con ayuda de Essentia detectamos los desvanecimientos del sonido en una pista a través de su señal de audio, con lo que obtuvimos los intervalos de tiempo que pasaban para que la pista entrará en una etapa de desvanecimiento de sonido para después obtener medidas descriptivas para estos intervalos, siendo estas la media, la desviación estándar, coeficiente de asimetría y curtosis. De igual forma se realizaron los mismos

cálculos para obtener los intervalos de tiempo que pasaban para que la pista saliera de estas etapas de desvanecimiento para después obtener las cuatro medidas descriptivas mencionadas anteriormente.

3.1.7. Indicadores de volumen

Para este componente se estimó el volumen global en decibeles de la pista dada su señal de audio, además de la complejidad dinámica, definida como la desviación absoluta promedio de la estimación global del volumen. También se calculó un indicador de la intensidad de la pista que toma tres valores:

- -1 Para pistas relajadas.
- 0 Para pistas moderadas.
- 1 Para pistas agresivas.

3.2. Información Editorial

La información editorial de una pista es aquella que no es acústica y está dada por los creadores, productores o editores, directa o indirectamente a la pista, como ejemplos podemos mencionar el nombre de la pista, el nombre del álbum en el que se incluye, el número que ocupa dentro del álbum, portada y contraportada del álbum, lugar donde fue creada la pista, nacionalidad del creador, género de la pista, entre otras.

3.2.1. Portadas de álbum

En este trabajo consideramos que portadas de álbums similares podrían apuntar a grupos de público similares, por ejemplo, se esperaría que la portada de un disco de música infantil fuera colorida y amigable, por otro lado los discos de algunos subgéneros del Heavy Metal que regularmente son dirigidos a cierta subcultura en específico suelen tener portadas con matices más oscuros. En este trabajo buscamos

Figura 3.6: Portadas de álbums de diferente década

encontrar los patrones existentes en portadas de álbums que nos ayuden a identificar el subconjunto al que pertenece cada pista, estos subconjuntos pueden hacer referencia a distintos conceptos tales como la década de la canción (Figura 3.6) o bien, su género musical (Figura 3.7). De esta forma podemos agregar una caracterización de las pistas desde la perspectiva de los autores.

3.2.2. Tags editoriales

Un tag es una etiqueta o una palabra clave asociada a un ítem, en particular en este trabajo nuestro item es una pista musical. Un tag editorial es entonces una palabra clave asignada, directa o indirectamente, a la pista musical por sus creadores, esta palabra bien puede tratarse de un género musical, el nombre de alguna región del mundo, el nombre de algún instrumento musical, el nombre de un sentimiento o hasta de una época del año, también puede tratarse de una fecha específica como el año

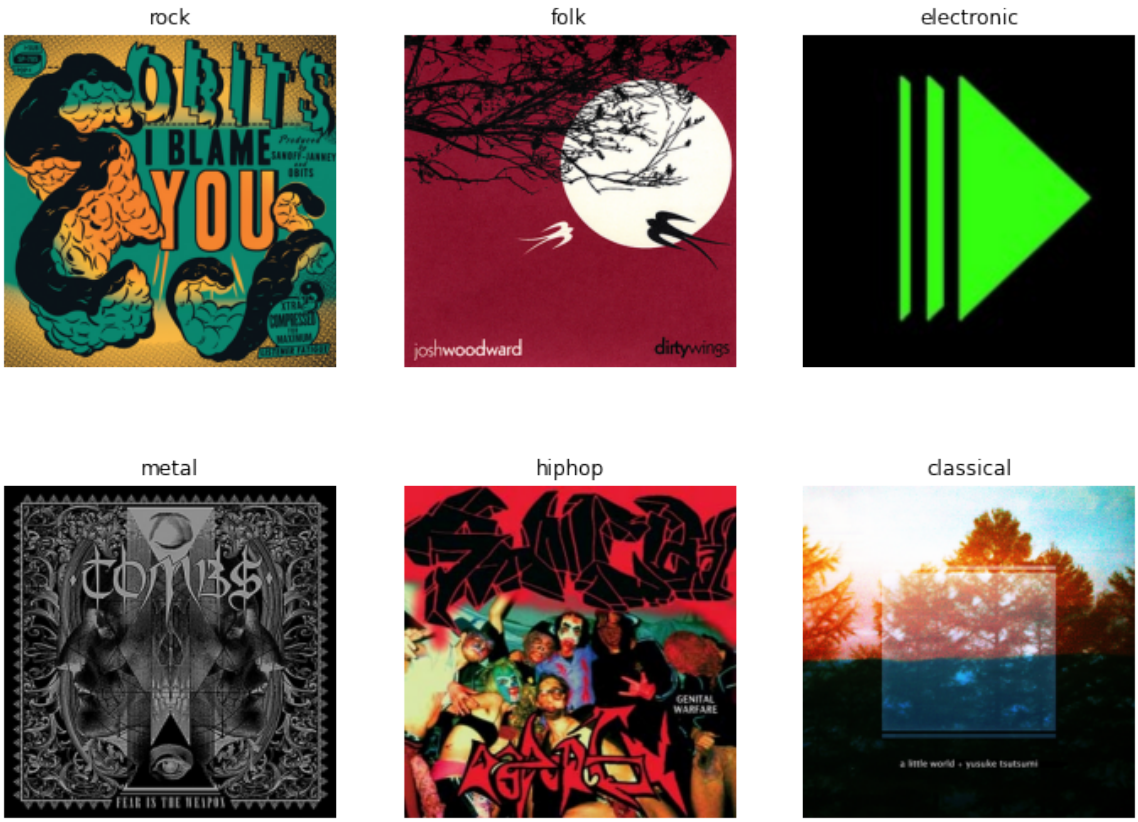


Figura 3.7: Portadas de álbums de diferente género

de su creación, década a la que pertenece e inclusive puede ser el nombre del artista que creó la pista, de tal forma que además del título de la pista existen palabras relacionadas a esta que le dan una esencia única a la pista.

3.3. Información cultural

La información cultural de una pista es entendida como la forma en que la pista es percibida en un entorno social, por ejemplo, un grupo de personas puede asociar una pista a cierto género musical independientemente de si su creador la clasifica en ese género. En ocasiones capturar y representar la información obtenida desde un enfoque cultural puede convertirse en una tarea bastante compleja.

3.3.1. Perfiles de usuario

Un perfil de usuario es un conjunto de características o preferencias que la persona tiene al interactuar con el sistema de búsqueda de los sitios web que frecuenta. Dicho perfil se estructura mediante el ingreso del usuario a estos sitios web. En nuestro caso particular estamos interesados en perfiles de usuarios originados por sitios web dedicados al streaming musical. En ocasiones obtener información de estos perfiles suele ser bastante complicado, debido a que la información de un usuario es bastante sensible y por tanto privada, es por eso, que en este trabajo se trabaja con tags culturales para capturar la información que los perfiles de usuario otorgan a las plataformas de streaming musical.

3.3.2. Tags Culturales

La idea de tag cultural es similar a la de un tag editorial del cuál se habló en la Subsección 3.2.2 la diferencia radica en quién o quienes asignan el tag a la pista musical, en el caso editorial son los creadores, mientras que en el caso cultural son los usuarios quienes asignan estas palabras clave a la pista de acuerdo a la percepción que tengan de la misma. Por ejemplo habrá perfiles que asignen el tag “electronic”

a una pista y otros que asignen el tag “ dancing” a la misma, lo que nos otorga dos características de dicha pista, extraídas indirectamente de los usuarios que en el caso ideal conocen la pista. Por otro lado, si dos pistas comparten un tag cultural por ejemplo “ rainig” podremos esperar una similaridad alta desde esta perspectiva.
Otra característica importante de los tags culturales es que cuentan con un peso para cada pista, es decir cuentan el número de veces que cierto tag fue asignado a una pista, de esta forma si el tag “guitar ” fue asignado a una pista el 500 veces mientras el tag “piano” solo fue asignado 10 veces podremos inferir que dicha pista es más conocida por incluir un sonido de guitarra que por incluir un sonido de piano y esta pista tendrá más similaridad con pistas conocidas por incluir guitarras en su sonido.

De tal forma que se puede pensar en los los tags culturales como una agrupación de la percepción de los usuarios a través de sus perfiles, y así cada tag puede actuar como un macro -perfil de usuario.

Un aspecto que debe tenerse en cuenta es que los tags culturales son palabras asignadas libremente por usuarios y una pista puede desde enriquecerse hasta sobrecargarse de información, por ejemplo, si a un usuario cierta canción le recuerda a un evento sucedido en el otoño, es posible que este usuario asigne un tag con la palabra “otoño” a esta pista, sin embargo podría haber otro usuario con una percepción diferente y asigne la palabra “invierno” a la misma pista, de esta forma cada pista puede adquirir tags que recuerden eventos específicos al usuario generando una cantidad bastante amplia de tags culturales que en muchas ocasiones puede resultar en información redundante o en información poco relevante e incluso confusa, es por eso que consideramos prudente cuidar el equilibrio de información cultural que rodea a una pista musical, quitando tags que podrían no ser muy representativos de la pista, de esto se hablará más a detalle en la Sección 5.2.3 .

En la Figura 3.8 se muestran los tags que con mayor frecuencia son asociados a nuestro conjunto de pistas.

Figura 3.8: Tags culturales más frecuentes en nuestro conjunto de pistas

Capítulo 4

Conjunto de Datos

Un elemento de bastante relevancia en este trabajo, es el conjunto de datos utilizado para crear los distintos espacios métricos, ya que caracteriza la estructura de dichos espacios y la representaciones de cada modo de información así como las representaciones multimodales.
Nuestro conjunto de datos está compuesto en gran medida por datos del conjunto de datos **FMA**, de donde tomamos los archivos de audio junto a sus correspondientes datos editoriales, en la siguiente sección describimos dicho conjunto de datos.

4.1. FMA

Este conjunto se presenta en Defferrard y cols. (2017) y es un conjunto de archivos de audio junto a una base de datos editoriales. La recolección de archivos de audio de audio stereo de $128kb/s$ a $44,100Hz$ en formato **.mp3** comenzó en abril de 2016 con un total de 89,912 pistas y a la fecha de este trabajo (2021) está compuesto por $106,574$ archivos de audio. El sitio web del conjunto de datos FMA es gestionado por WFMU la estación de radio libre de más larga duración en los Estados Unidos, este sitio provee un extenso catálogo de artistas y pistas de alta calidad en formato mp3. Cada canción puede ser descargada de forma gratuita, ya que cada artista ha decidido trabajar bajo licencia tipo creative commons. El objetivo del sitio web **FMA** es ofrecer una plataforma legal y tecnológica que pueda ser aprovechada por artistas y

escuchas. Entre los datos editoriales que otorga la base de datos adjunta a los archivos de audio destacamos los siguientes:

- Título de la pista.
- Un total de $14,299$ álbumes diferentes y 1024 pistas sin esta información.
- Un total de $16,294$ artistas diferentes .
- Un total de 163 géneros diferentes, jerarquizados en 4 niveles. El primer nivel está compuesto por 16 géneros musicales que serán considerados como *géneros principales* y se muestran en la Figura 4.1, de los cuales se desprenden 106 subgéneros que constituyen el segundo nivel, los subgéneros que componen el tercer nivel son 37 y se derivan de los subgéneros del segundo nivel , finalmente el último nivel está compuesto solo por tres subgéneros derivados del tercer nivel. En promedio cada pista tiene asociado 2.38 géneros diferentes. Además se tienen $2,231$ pistas sin un género asociado. En la Figura 4.2 podemos observar la distribución del número de géneros asociados a una pista, podemos destacar que en la mayoría de las veces el número de géneros asociados a una pista son 3, mientras que el mayor número de géneros asociados a una pista es de 25. Por otro lado en Figura 4.3 podemos observar los 20 géneros más recurrentes en el conjunto de datos.
- Un total de $4,926$ tags diferentes. Estos tags contienen a los 163 géneros musicales, sin embargo las anotaciones de los tags en las pistas no están completas ya que $83,549$ pistas no tienen ningún tag anotado, a pesar de que la mayoría sí tienen al menos un género asignado. No obstante, esta información puede complementarse con la de los géneros musicales.

Es importante resaltar que hay más artistas que álbumes debido a que varios álbumes son de formato colaborativo.

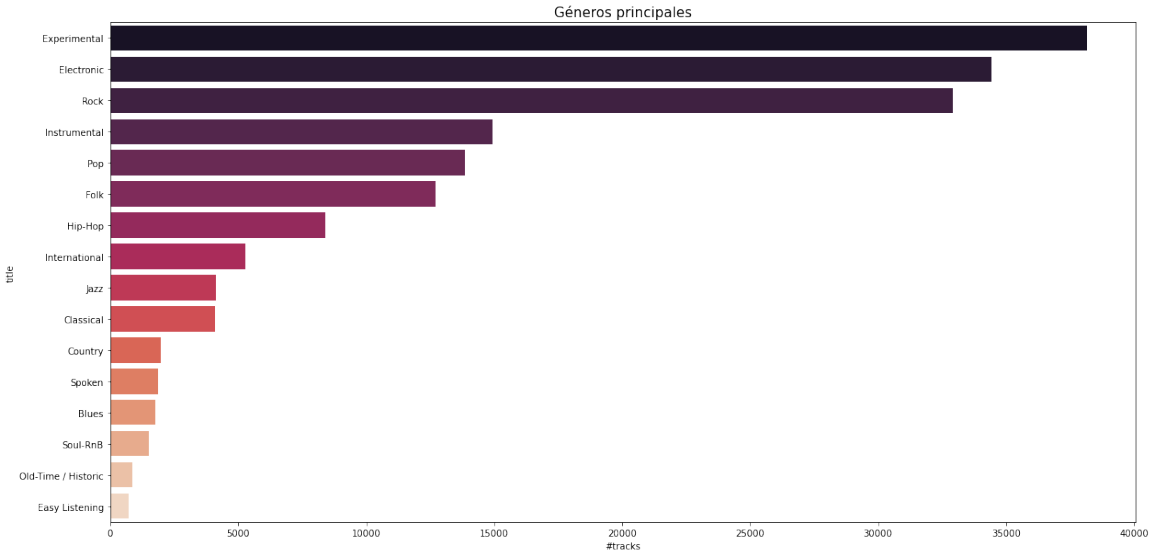


Figura 4.1: Recurrencia de los 16 géneros principales

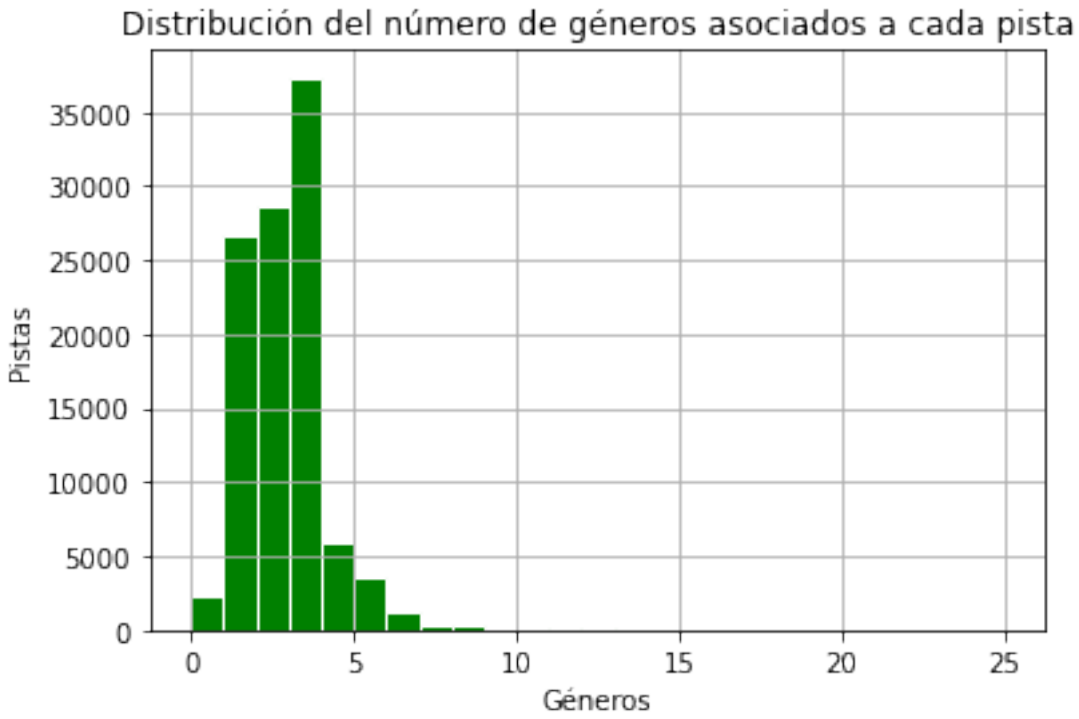


Figura 4.2: Distribución del número de géneros asociados a una pista

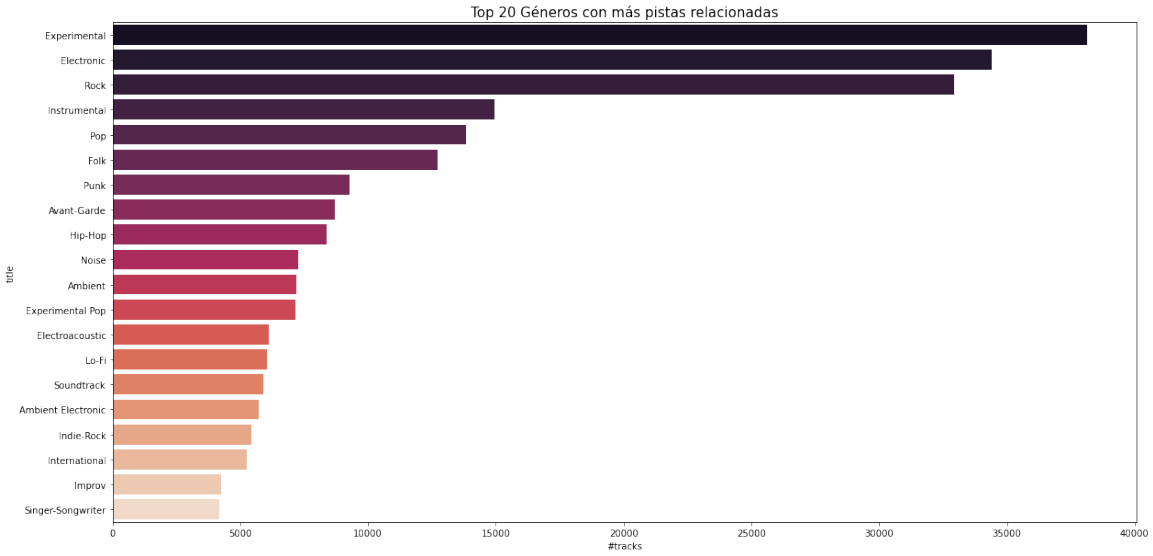


Figura 4.3: Top 20 géneros más recuerrentes en las pistas

4.2. Sub conjunto seleccionado

Mediante el nombre de la pista, el nombre del álbum y el nombre del artista se realizaron consultas para obtener tags culturales y portadas de los álbums asociadas a las pistas del conjunto de datos **FMA**, en total se logró relacionar $11,517$ pistas del conjunto de datos, a los tags culturales y portadas de álbum, de esta forma estas $11,517$ pistas pasan a conformar el subconjunto de datos con el que trabajaremos del cual resaltamos los siguiente:

- $11,517$ pistas. Un $10.8\,\%$ del conjunto original.
- $3,557$ álbumes. Un $24.8\,\%$ del conjunto original.
- $3,692$ artistas. Un $22.6\,\%$ del conjunto original.
- 149 géneros diferentes, 14 géneros menos que el conjunto original, esto es menos del $10\,\%$ del conjunto original. La distribución de los géneros tiene algunas diferencias con respecto a la distribución del conjunto original, en la Figura 4.4 se muestra la nueva distribución del número de géneros asociados a cada pista, puede observarse que en este caso no hay cambios significativos respecto a la Figura 4.2, en cuanto a la recurrencia de géneros hay un cambio importante ya

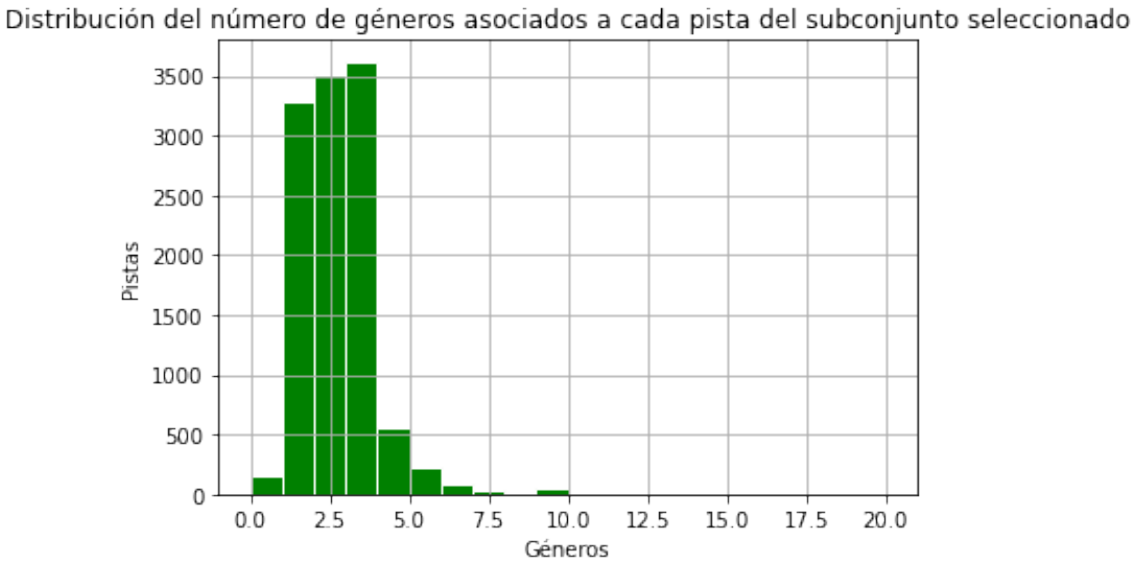


Figura 4.4: Distribución del número de géneros asociados a una pista para el subconjunto seleccionado.

que como podrá apreciarse en la Figura 4.5 y en la Figura 4.6 el género con más recurrencia pasa a ser **electronic** mientras que en el conjunto original el género con más recurrencia es **experimental**.

El conjunto de datos **FMA** proporciona una cantidad importante de archivos de audio junto a información editorial de la pista lo cual es de gran utilidad en nuestros objetivos.

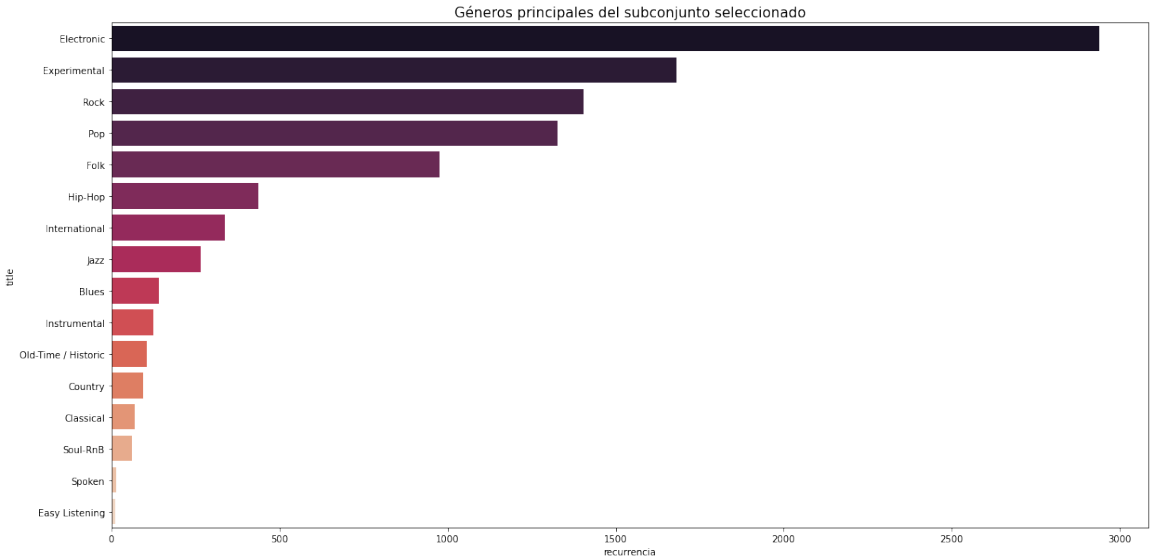


Figura 4.5: Recurrencia de los 16 géneros principales en el subconjunto de pistas seleccionado.

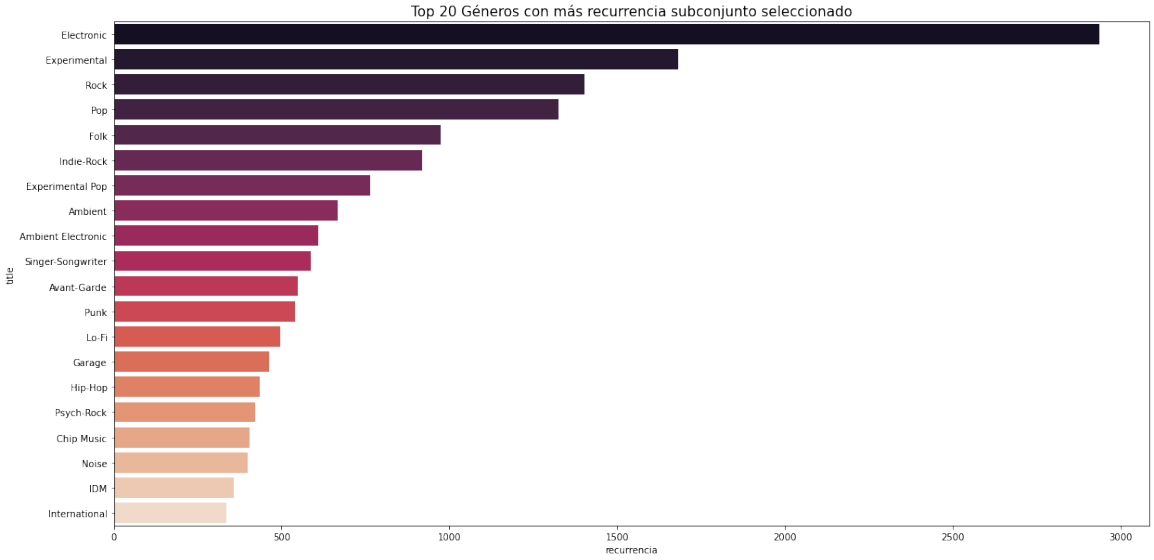


Figura 4.6: Top 20 géneros más recuerrentes en las pistas en el subconjunto de pistas seleccionado.

Capítulo 5

Metodología

En este capítulo se describe la metodología usada para construir un modelo que genere un espacio métrico, en el que cada pista tenga una representación vectorial que considere todos los modos de información disponible, así como las métricas utilizadas en la evaluación del desempeño de dicho modelo. Por simplicidad cada que se haga referencia a una representación vectorial utilizaremos el término **embedding**.

5.1. Descripción general del modelo

El objetivo de nuestro modelo es generar un espacio métrico en el que cada pista tenga una representación vectorial que conserve la esencia de cada modalidad de información disponible. Para lograr esto utilizaremos un modelo de aprendizaje métrico como el que se presenta en Xing, Jordan, Russell, y Ng (2002) y se compone de los siguientes elementos:

- **Base de referencia.** Un conjunto de vectores que sirven como referencia del espacio métrico, en el contexto de este trabajo esta base puede estar conformada por embeddings que representen tags, géneros, álbumes, artistas , pistas, etc. A los embeddings que componen la base los denotaremos como $\mathbf{b_i}$

- **Embeddings multimodales.** Estos embeddings corresponden a las pistas, están conformados por las características extraídas de los distintos modos de in-

formación y son los elementos que buscamos ubicar de mejor manera posible en nuestro espacio métrico **S**.

- **Muestreo.** Para generar **S** es necesario entrenar un modelo basado en las diferentes modalidades de las pistas de nuestro conjunto de datos. Para que dicho entrenamiento sea eficaz en términos de costo computacional es necesario definir un tipo de muestreo para elegir las pistas que se usarán para entrenar nuestro modelo en cada época, este muestreo suele ser aleatorio, la idea principal como se verá en la Subsección 5.3.1 es tomar un embedding de la base $\mathbf{b_i}$, un embedding que represente a una pista relacionado a $\mathbf{b_i}$ y un tercer embedding no relacionado a $\mathbf{b_i}$.

- **Función de pérdida.** Finalmente, la función de pérdida es fundamental en el proceso de entrenamiento, ya que se encarga de orientar el modelo hacía la construcción de un espacio métrico óptimo. Esta función toma como base alguna medida de distancia, el objetivo es acercar embeddings que estén asociados de alguna forma y alejar embeddings no asociados.

En la Figura 5.1 se presenta un esquema general del modelo utilizado en este trabajo. Consideramos cuatro modos de información:

1. Señales de audio.
2. Texto.
3. Imagen.
4. Filtros colaborativos.

Para cada modo de información se realizan distintas operaciones con el fin de extraer sus características, las modalidades que representan a las pistas son concatenadas para crear un solo embedding multimodal por pista, se realiza lo mismo para la modalidad o modalidades que representen a los embeddings de la base, posterior a esto mediante un método de muestreo se seleccionan dos embeddings que representen a la pista y

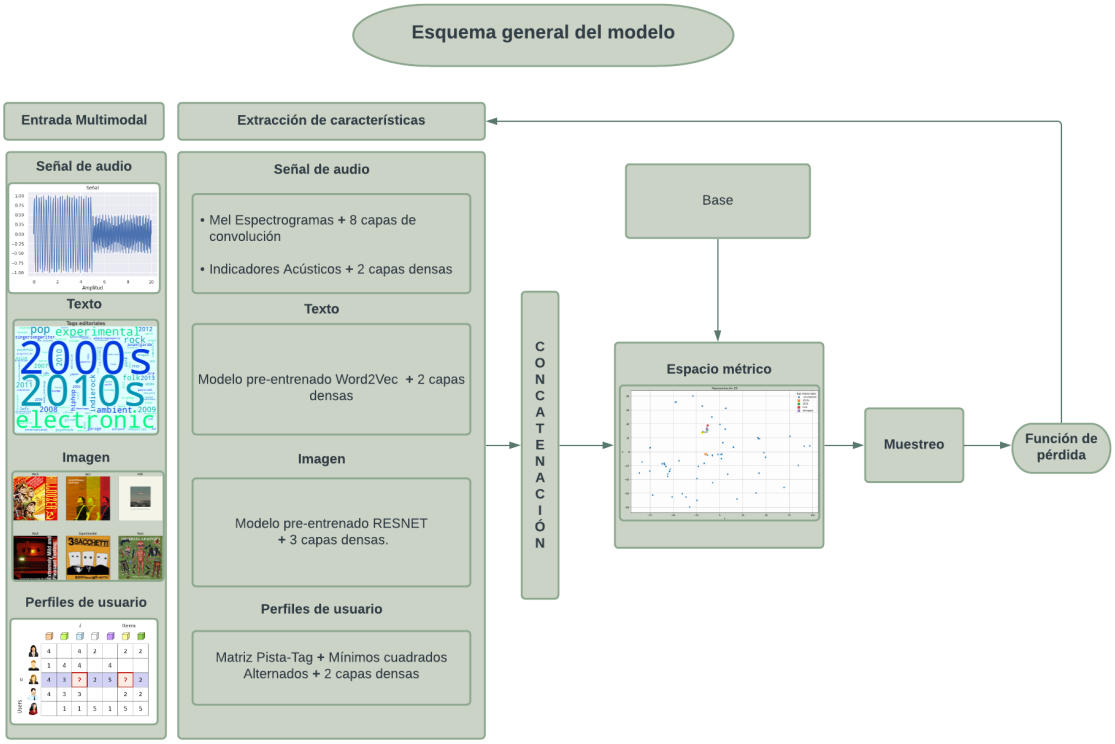


Figura 5.1: Esquema general del modelo

un embedding de la base para ser evaluado por la función de pérdida, a partir de esta función se realiza la actualización de los pesos del modelo.

5.2. Extracción de características

En esta sección explicaremos el proceso de extracción de características para cada modo de información considerado .

5.2.1. Información Acústica

Como se mencionó en la Sección 2.1 esta información está compuesta por las características que componen el audio de los items musicales, específicamente la señal de audio. En este trabajo se utilizaron dos caracterizaciones de esta información: espectrogramas en escala mel e indicadores acústicos.

Espectrogramas en escala mel

Para cada pista se obtuvo un espectrograma en escala de mel considerando los siguientes hiper-parámetros:

- Tasa de muestreo $F_s = 22,050Hz$ (661,500 elementos por señal)
- Función de ventana: **Hann**
- Tamaño de ventana $N = 1024$
- Tamaño de paso $H = 512$

Esto dió como resultado representaciones matriciales de dimensión 128×1293, de las cuales, siguiendo la metodología de Won y cols. (2020), se seleccionó aleatoriamente una sub matriz A_s de dimensión 128×173, habiendo seleccionado A_s, que representa $\approx 4s$ de la pista, procedimos a aplicar 7 capas de convolución, con filtros de tamaño 3×3, zero padding $p = 1$, stride $s = 1$ un **Max pooling** de tamaño 2 y función de activación **Relu**. Previo a cada capa de convolución se aplica **batch normalization**. El tensor resultante es pasado por una octava capa de convolución sin **Max pooling** para finalmente obtener el vector que representará la información aportada por los espectrogramas el cual consta de 200 elementos. En la Figura 5.2 se muestra un esquema del proceso descrito anteriormente.

Indicadores acústicos de la señal de audio

En la Subsección 3.1.3 se describen distintos indicadores acústicos derivados de la señal de audio que describen las siguientes características de la pista :

- Ritmo. Compuesto por medidas de tempo y danzabilidad
- Energía. Calculada directamente de la onda producida por la señal de audio
- Tono. Compuesto por medidas relacionadas a los acordes y notas detectadas en la señal de audio.

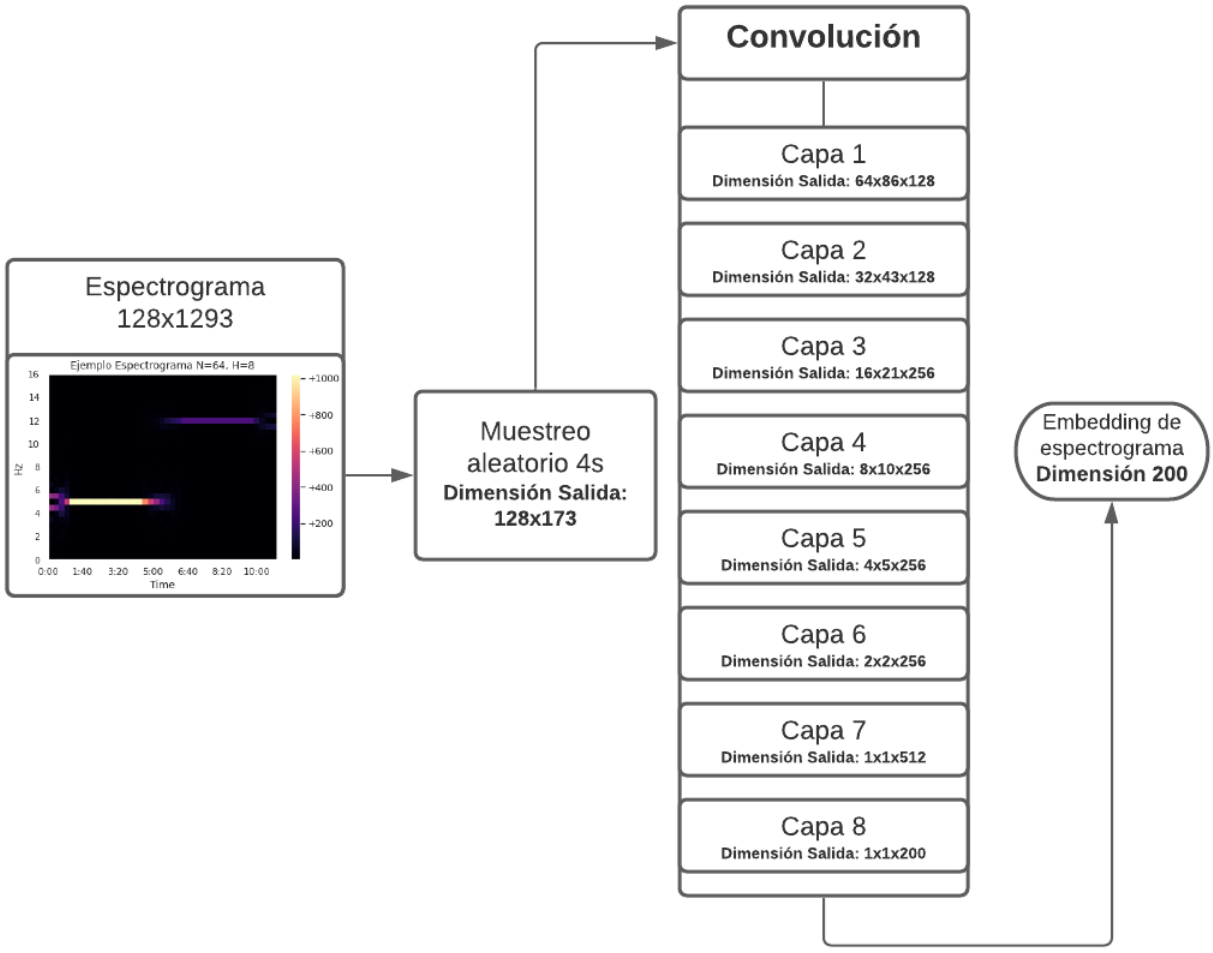


Figura 5.2: Esquema del proceso de extracción de características para señales de audio por medio de espectrogranas

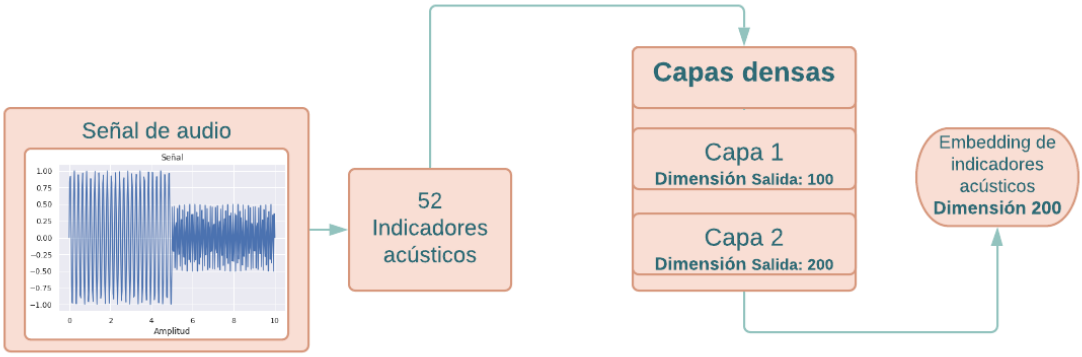


Figura 5.3: Esquema del proceso de extracción de características para señales de audio por medio de indicadores acústicos

- Duración y silencio. Medidas relativas al desvanecimiento del sonido y del silencio.

- Volumen . Medidas calculadas a través de la complejidad dinámica e intensidad del sonido.

En total se obtuvieron 52 indicadores por cada pista los cuales fueron almacenados en vectores, para después ser estandarizados y agregarse al modelo. Los vectores son pasados por una capa densa con 100 unidades ocultas, para después aplicar **batch normalization**, enseguida se aplica una función de activación **Relu** y un **dropout** con parámetro de probabilidad $p = 0.5$, finalmente este resultado es pasado por una segunda capa densa con 200 unidades ocultas, el vector resultante representará este modo de información en el modelo y será concatenado con los demás tipos de información. En la Figura 5.3 podemos observar el proceso descrito anteriormente.

5.2.2. Información Editorial

Respecto a la información editorial de las pistas consideramos dos elementos: los tags editoriales y las portadas de álbumes.

Tags editoriales

El conjunto de datos FMA[1], nos proporciona los archivos de audio de cada una de las pistas en nuestro conjunto de datos y además incluye una base con los metadatos referentes a estas pistas, entre los que se incluyen los tags editoriales. Se obtuvieron $4,926$ tags editoriales, a los que se les agregó 50 tags referentes al año y la década de las pistas que contarán con esta información disponible, sumando un total de $4,976$ tags editoriales a los que se les realizó el siguiente conjunto de acciones manuales:

1. Se convirtieron todos los caracteres a minúsculas.
2. Se eliminaron espacios y caracteres especiales.
3. Se agruparon palabras con significado similar, por ejemplo: electronic y electrónica.
4. Se eliminaron tags que pudiesen considerarse muy específicos, estos son por ejemplo tags que hace referencia al artista, título del álbum, o título de la canción. Esto se logró tomando solo tags relacionados con al menos 50 canciones.

Después de aplicar este pre-proceso se redujo el número de tags de $5,186$ a 317. El siguiente paso fue obtener una representación vectorial de los tags utilizando un modelo **Word2vec** como el que se presenta en Mikolov, Chen, Corrado, y Dean (2013) y pre-entrenado en Won y cols. (2020) con un textos musicales los cuales incluyen reseñas de Amazon, biografías musicales, y páginas de Wikipedia sobre teoría musical y géneros musicales. De esta forma, para cada uno de los tags se consiguió un embedding de dimensión 300.

Con estos embeddings se obtiene una representación para cada pista considerando un promedio de los embeddings correspondientes a los tags relacionados a esta, es decir que si p_i es el embedding de la i-ésima pista y T_i es el subconjunto de embeddings de los tags editoriales relacionadas a esta pista entonces

$$p_i = \frac{1}{|T_i|} \sum_{t_k \in T_i} t_k.$$

[1] Defferrard y cols. (2017)

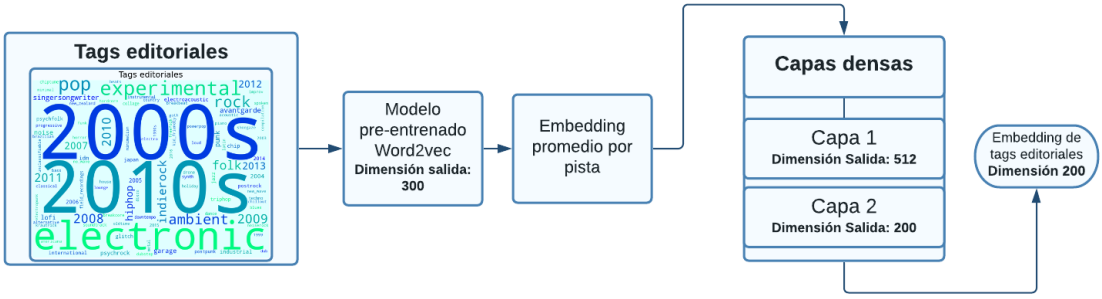


Figura 5.4: Esquema del proceso de extracción de características para tags editoriales

Posteriormente los embeddings que representan a cada pista pasan por una capa densa con 512 unidades ocultas, para después aplicar **batch normalization**, enseguida se aplica una función de activación **Relu** y un **dropout** con parámetro de probabilidad $p = 0.5$, seguida de una segunda capa densa con 200 unidades ocultas. En la Figura 5.4 se muestra un esquema del proceso descrito anteriormente.

Portadas de álbum

Para cada pista en nuestro subconjunto de datos se cuenta con la portada de álbum en formato de imagen en escala de grises de tamaño 300×300 pixeles. Al igual que en Oramas y cols. (2018) se utilizaron las capas y los pesos del modelo **ResNet 101** pre-entrenado con el conjunto de datos **ImageNet**, con el fin de extraer las características más relevantes de cada imagen, de esta forma para cada imagen se obtienen 1024 filtros de tamaño 10×10 al que aplicamos **max pooling** de tamaño 10 sobre las dos dimensiones, para obtener un vector de 1024 características por imagen, mismos que se incluyen al modelo y se procesan de forma muy similar a los vectores de indicadores numéricos de la Sección 5.2.1, los vectores son pasados por una capa densa con 512 unidades ocultas, para después aplicar **batch normalization**, enseguida se aplica una función de activación **Relu** y un **dropout** con parámetro de probabilidad $p = 0.5$, seguida de una segunda capa densa con 256 unidades ocultas, se aplica batch normalization, la función de activación Relu y dropout para finalmente aplicar una tercer capa densa con 200 unidades ocultas, el vector resultante representará este modo de información en el modelo, en la Figura 5.5 podemos observar un esquema

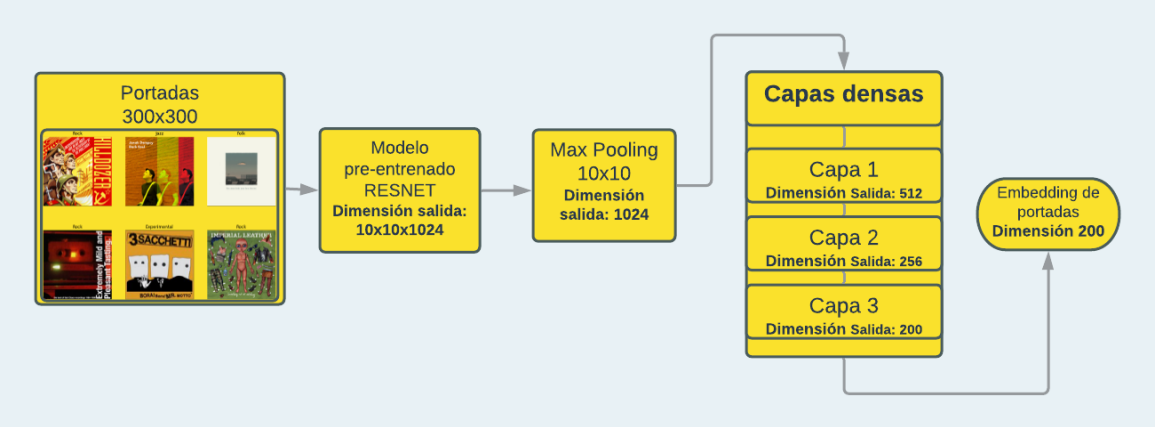


Figura 5.5: Esquema del proceso de extracción de características para las portadas de álbum

del proceso descrito anteriormente.

5.2.3. Información cultural

Para la información de origen cultural nos auxiliamos en la información de usuarios que indirectamente guardan los tags culturales, para lograr esto se realizaron los procesos que se explican a continuación.

Perfiles de usuario: Tags culturales

Para captar la información de los perfiles de usuario en plataformas de streaming musical se utilizó el concepto de tag cultural del cual se habla en la Subsección 3.3.2. Mediante la API de **last.fm** se obtuvieron 19,694 tags culturales diferentes, un número considerablemente alto si se toma en cuenta que cada tag representará una columna en nuestra matriz pist-tag que posteriormente será factorizada con el método de mínimos cuadrados alternantes, por lo que resulta importante realizar una limpieza de tags que podrían estar duplicando información. Resulta evidente que canciones ligadas a los mismos tags serán identificadas como culturalmente similares, sin embargo existen otros aspectos no tan explícitos que trataremos de capturar, por ejemplo, supongamos que una canción ha sido frecuentemente ligada al tag "electronica" y ocasionalmente ligada al tag "electronic", en esencia el tag representa lo mismo:

el género musical de la pista, sin embargo el primer tag nos índica que existe un alta probabilidad de que la canción sea escuchada principalmente en hispanoamérica y no en regiones anglosajonas, aspecto cultural que no sería prudente pasar por alto, por otro lado puede haber tags diferentes que tengan el mismo significado y su diferencia guarde poca información cultural por ejemplo un error al escribir la palabra esto es, puede haber un tag llamado " eperimental" y otro "experimental" que en esencia son lo mismo y su diferencia no nos indica nada más que la omisión de la letra " x", este tipo de tags deben ser agrupados bajo el mismo nombre. Por este tipo de detalles es que, comparado con los tags editoriales, resulta más complejo realizar una limpieza de los tags culturales. Las acciones llevadas a cabo para el reducir el número de tags culturales fueron las siguiente:

1. Se convirtieron todos los caracteres a minúsculas.

2. Se quitaron caracteres raros o poco usuales como: [@-_/́]

3. Se quitaron palabras que se consideran de término general, como: music, songs, song y tag

4. Cadenas con más de 2 caracteres iguales consecutivos fueron sustituidos por una cadena de un solo caracter.

5. Buscando cuidar el aspecto descrito anteriormente, se dividió el conjunto de tags en dos subconjuntos, utilizando el vector s como discriminante donde s queda determinado por la matriz pista - tag P de la siguiente forma $s = \mathbf{1}^T P$, es decir s es la suma por renglones de la matriz pista-tag, luego los subconjuntos quedan determinados de la siguiente forma :

   a) $s_t > 1000$. Con 766 elementos

   b) $s_t \leq 1000$. Con $18,824$ elementos

   Finalmente verificamos si cada tag del subconjunto 2 está contenido en algún tag del subconjunto 1 y viceversa, las coincidencias de tag se agregaron a los tags del subconjunto 1 y se eliminó el tag del subconjunto 2.

6. Por último se quitaron aquellos tags que tuviesen relación solo con una pista, pensando en que solo aportan dimensionalidad y no aporta información que nos ayude a relacionar pistas.

Matriz pista-Tag

Para captar la información cultural seguiremos la idea presentada en Hu, Koren, y Volinsky (2008). en la que se crea una matriz de ítems-usuario, con el fin de localizar items culturalmente cercanos y poder realizar un sistema de recomendación. En este trabajo las pistas musicales juegan el papel de ítem. Al no tener acceso libre a información que relaciona usuarios con nuestros archivos de audio usaremos tags culturales para suplir el elemento "usuario" en nuestra tarea de caracterización cultural. De tal forma que nuestra **matriz pista-tag** mapea en cada uno de sus elementos la relación entre un tag asignado por el usuario de una plataforma de streaming musical y una pista. La escala de relación entre una pista y un tag va de 0 a 100, donde cero significa que el tag nunca ha sido relacionado al tag y el 100 representa que el tag es asignado a la canción frecuentemente. De esta forma, no tan directa, intentamos capturar la percepción de los usuarios sobre cada pista, de esta manera la matriz pista-tag funcionará como una matriz de preferencia en la que cada tag "calificará" a cada pista de acuerdo al número de veces que el tag es relacionado a la pista. Nuestro siguiente paso fue buscar, a partir de la matriz pista-tag una configuración de vectores para cada pista, que guarden y representen de mejor manera la información almacenada en ella, para esto se empleó el método de mínimos cudrados alternantes, descrito a continuación.

Mínimos cuadrados alternados

El método de mínimos cuadrados alternados fue propuesto por Hu y cols. (2008) para encontrar configuraciones derivadas de matrices de preferencias entre ítems y usuarios. La idea de mínimos cuadrados alternados se describe a continuación. Sea R una matriz de preferencias con n items y m usuarios y sea r_{ui} un elemento de la matriz R correspondiente al nivel de preferencia que tiene el usuario u sobre el ítem

i, consideramos además que r_{ui} tiene como mínimo a cero y es un indicador de no preferencia, definimos la variable indicadora de preferencia p_{ui} como

$$p_{ui} = \begin{cases} 1 & r_{ui} > 0 \\ 0 & r_{ui} = 0 \end{cases} \tag{5.1}$$

y sea c_{ui} una variable de preferencia definida como

$$c_{ui} = 1 + \alpha r_{ui} \tag{5.2}$$

donde α es un meta parámetro que determina el nivel de diferencia entre dos valores de preferencia distintos. **MCA** intenta obtener m vectores $x_u \in \mathbb{R}^f$ y n vectores $y_i \in \mathbb{R}^f$ que minimicen:

$$\sum_{u=1}^{n}\sum_{i=1}^{m} c_{u,i}(p_{ui} - x_u^T y_i)^2 + \lambda(\sum_{u=1}^{m}||x_u||^2 + \sum_{i=1}^{n}||y_i||^2) \tag{5.3}$$

Donde f es un meta-parámetro que representa la dimensión de los vectores resultantes y se elige antes de aplicar el algorítmo. Es claro que conforme el producto $x_u^T y_i$ se acerque a p_{ui}, el valor de Ecuación 5.3 será menor. En este caso los vectores x_u representarán a los usuarios y los vectores y_i serán las representaciones de los ítems. Para resolver la Ecuación 5.3 supondremos que contamos con los vectores y_i y derivando encontramos que

$$x_u = (Y^T C^u Y + \lambda I)^{-1} Y^T C^u p(u) \tag{5.4}$$

donde C^u es la matriz diagonal de tamaño $n \times n$ con $C_{ii} = C_{ui}$ y $p(u) \in \mathbb{R}^n$ es el vector que contiene todas las preferencias del usuario u, es decir. los valores p_{ui}. Tras obtener los vectores x_u procedemos a actualizar los vectores y_i, que después de derivar respecto a y_i la Ecuación 5.3 e igualar el resultado a cero podemos obtenerlos mediante la siguiente ecuación :

$$y_i = (X^T C^i X + \lambda I)^{-1} X^T C^i p(i) \tag{5.5}$$

análogamente que en la Ecuación 5.4 C^i es la matriz diagonal de tamaño $m \times m$ con $C_{uu} = C_{ui}$ y $p(i) \in \mathbb{R}^m$ es el vector que contiene todas las preferencias para el item i es decir los valores p_{ui}. Tras obtener la configuración de vectores y_i comenzamos un proceso iterativos reconfigurando en cada paso los vectores x_u y y_i hasta que la Ecuación 5.3 se estabilice, empíricamente 10 iteraciones suele dar buenos resultados. Un punto importante que debemos resaltar en la Ecuación 5.4 es que la igualdad $Y^T C^u Y = Y^T Y + Y^T (C^u - I) Y$ donde $Y^T Y$ es independiente de u por lo que solo necesita ser calculada una vez en cada iteración y $C^u - I$ tiene n_u elementos distintos de cero con $n_u \leq n$ lo que economiza los cálculos computacionales significativamente, para la Ecuación 5.5 podemos realizar un proceso análogo reduciendo de igual forma los cálculos para obtener los vectores y_i. En nuestro trabajo los usuarios serán sustituidos por tags culturales, las pistas musicales tomarán el lugar de los items y los vectores resultantes y_i serán nuestras representaciones culturales de las pistas musicales, que usaremos más adelante en nuestros diferentes modelos. Por otro lado los vectores x_u servirán como base para obtener representaciones vectoriales y_{m+k} de pistas fuera de nuestro conjunto de prueba y entrenamiento, haciendo uso de Ecuación 5.5.

Inclusión al modelo

Tras realizar la búsqueda de pistas con al menos un tag cultural y aplicar el preproceso correspondiente, obtuvimos una matriz pista-tag de dimensiones $11,517 \times 4,898$. Posteriormente factorizamos la matriz pista-tag mediante mínimos cuadrados alternados obteniendo una configuración de vectores con $f = 200$ elementos para cada pista $(11,517)$. En la figura Figura 5.6 podemos apreciar una represenatación en dos dimensiones de la configuración obtenida identificada por género musical.

La técnica seguida para obtener los vectores que representan la información cultural de cada pista puede resumirse en los siguientes pasos:

1. Búsqueda de los tags culturales relacionados a las pistas de nuestro conjunto de datos mediante la api de **last.fm**. Se obtuvieron $19,694$ tags correspondientes a $11,517$ pistas.

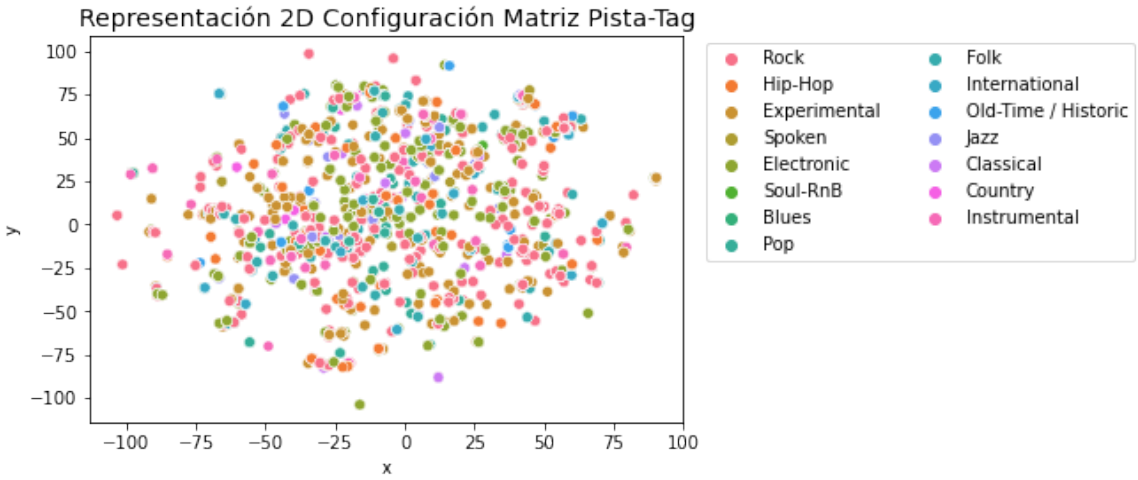


Figura 5.6: Representación en 2D de la configuración de las pistas obtenida mediante mínimos cuadrados alternantes. Identificadas por género musical

2. Preproceso de los tags obtenidos. Se redujo el número de tags de $19,694$ a $4,898$.

3. Construcción de la matriz pista-tag. Esta matriz de dimensiones $11,517 \times 4,898$ indica el nivel de relación que hay entre la pista i y el tag j.

4. Factorización de la matriz pista-tag, usando el método de mínimos cuadrados alternados visto en la Ecuación 5.3, obteniendo una representación vectorial para cada pista de tamaño 200.

Posteriormente, los vectores resultantes son pasados por una capa densa con 512 unidades ocultas, para después aplicar **batch normalization**, enseguida se aplica una función de activación **Relu** y un **dropout** con parámetro de probabilidad $p = 0.5$, seguida de una segunda capa densa con 200 unidades ocultas, el vector resultante representará este modo de información en el modelo. En la Figura 5.7 podemos observar un esquema con el proceso descrito anteriormente.

5.3. Espacio Métrico

Una vez obtenidos los embeddings que representan cada modo de información procedemos a construir el espacio métrico, para el cual utilizamos una función de pérdida

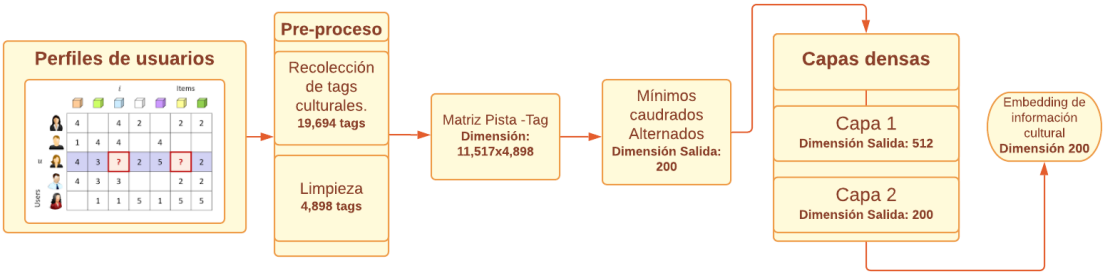


Figura 5.7: Esquema del proceso de extracción de características para perfiles de usuarios mediante tags culturales

basada en la distancia coseno que se presenta en la Subsección 5.3.2 junto a una estrategia de muestreo que se presenta en la Subsección 5.3.1. Se construyeron dos tipos de espacios métricos ambos toman como base los tags editoriales pero los embeddings de esta base se construyen de forma diferente en cada caso. Al primer espacio métrico lo llamaremos **espacio métrico semántico** y se presenta en la Subsección 5.3.3, mientras que al segundo, presentado en la Subsección 5.3.5 los llamaremos **espacio métrico acústico** .

5.3.1. Muestreo

La función de pérdida de nuestro modelo requiere la selección de tres elementos:

- Embedding de un elemento de la base del modelo.
- Embedding de una pista relacionada al tag en cuestión. Al que llamaremos embedding positivo.
- Embedding de una pista no relacionada al tag en cuestión. Al que llamaremos embedding negativo.

Para seleccionar los tres elementos mencionados anteriormente implementaremos el muestreo **muestreo aleatorio balanceado y ponderado** que se presenta en Won y cols. (2020) y tiene las siguientes características:

- **Balanceado**. En la base del modelo puede haber elementos que tienen más pistas relacionadas que otros, por tanto, si se selecciona primero una pista y

después el elemento de la base relacionado obtendremos un muestreo desbalanceado respecto a la base del modelo ya que algunos de sus elementos serían seleccionados más veces que otros, es por eso que el muestreo se realiza para cada elemento de la base, buscando así que el muestreo se balanceado para la base .

- **Aleatorio**. El muestreo es aleatorio al momento de seleccionar los embeddings de las pistas. En la selección del embedding positivo cada pista tiene la misma probabilidad de ser seleccionada, mientras que la selección del embedding negativo es ponderada, lo que nos lleva a la siguiente característica.

- **Ponderado**. Buscando mejores resultados, la selección del embedding negativo es aleatoria ponderada, siendo la probabilidad de selección directamente proporcional a la similaridad entre el embedding de la base y el embedding negativo. Con esta estrategia es más probable seleccionar embeddings negativos cercanos al embedding de la base en cuestión, para después tratar de alejarlos.

Entonces el muestreo se realiza para cada elemento de la base siguiendo las acciones que se muestran a continuación:

1. Se selecciona una pista aleatoriamente relacionada al elemento en cuestión. Todas las pistas relacionadas a este elemento tienen la misma probabilidad de ser seleccionadas.

2. Se obtiene la similitud coseno entre el embedding del elemento de la base y los embeddings de las pistas no relacionadas a este.

3. Se selecciona una pista aleatoriamente no relacionada al elemento en cuestión. La probabilidad de selección de cada pista es directamente proporcional a la similaridad que mantiene su embedding con el embedding del elemento en cuestión. Es decir, los embeddings más parecidos al embedding del elemento de la base tienen una probabilidad mayor de ser seleccionados.

Al haber seleccionado nuestro triplete de embeddings procedemos a evaluar la función de pérdida que se describe a continuación en la Subsección 5.3.2.

5.3.2. Función de pérdida

Nuestra función de pérdida recibe tres embeddings.

- **Embedding de un elemento de la base :** E_b.
- **Embedding de una pista relacionada al elemento en turno:** E_p.
- **Embedding de una pista no relacionada al elemento en turno:** E_n.

Nuestro modelo busca minimizar:

$$L = ReLU(D(E_b, E_p) - D(E_b, E_n) + \delta) \tag{5.6}$$

Donde la función $ReLU$ esta definida como:

$$ReLU(x) = min(0, x) \tag{5.7}$$

y $D(.)$ es la distancia coseno definida como:

$$D(x, y) = 1 - \frac{< x, y >}{\|x\|\|y\|} \tag{5.8}$$

Mientras que δ funciona como margen y evita que Ecuación 5.6 sea siempre cero. Lo que se busca con la función de pérdida dada en Ecuación 5.6 es que $D(E_b, E_n)$ sea mayor a $D(E_b, E_p)$ de tal forma que los embeddings de los elementos de la base se acerquen a los embeddings de las pistas relacionadas y se alejen de los embeddings de las pistas no relacionadas en el mejor de los casos se espera que $D(E_b, E_n) > D(E_b, E_p) + \delta$ y de esta forma L alcanza su mínimo en 0, por tanto δ determina que tan restrictivo será nuestro proceso de entrenamiento, ya que un δ muy bajo permitirá que L alcance su mínimo de manera muy sencilla y obtendremos un modelo bastante libre con el que podríamos no estar acercando lo suficiente los embeddings de los elementos de la base con los embeddings de sus correspondientes pistas relacionadas, mientras que para un δ muy alto será más difícil que L alcance su mínimo lo que puede desembocar en un modelo sobre ajustado que no funciona de manera adecuada

para datos fuera del conjunto de entrenamiento, por lo que este valor debe elegirse con cautela, en este trabajo haremos $\delta = 0.4$ que representa un 20 % de la distancia máxima que puede haber entre dos embeddings.

5.3.3. Espacio métrico semántico

El nombre de este espacio métrico se deriva del proceso que se siguió para obtener los embeddings que componen su base ya que fueron construidos bajo características semánticas de los tags editoriales.

Base de referencia

Este espacio métrico toma como base los 100 tags editoriales más recurrentes en nuestro subconjunto de datos cada embedding que representa a cada elemento de la base es tomado del modelo word2vec pre-entrenado en Won y cols. (2020). En la Figura 5.8 podemos apreciar una representación en dos dimensiones de esta base, la representación fue obtenida mediante la técnica de reducción de dimensión **T-SNE**, podemos observar que esta representación conserva características semánticas, el ejemplo más claro es el grupo de tags que se forma en la parte central inferior de la imagen, el cual corresponde a números que hacen referencia al año de la pista, aspecto que resalta el origen semántico de esta base.

Antes de ser evaluados por la función de pérdida estos embeddings pasan por una capa densa con 512 unidades ocultas, para después aplicar **batch normalization**, enseguida se aplica una función de activación **Relu** y un **dropout** con parámetro de probabilidad $p = 0.5$, seguida de una segunda capa densa con 256 unidades ocultas.

Embeddings Multimodales de las pistas

Cada embedding multimodal que representan a las pistas está conformado por embeddings de tamaño 200 que representan a las siguientes modalidades:

- Embedding que representa los tags editoriales de cada pista .

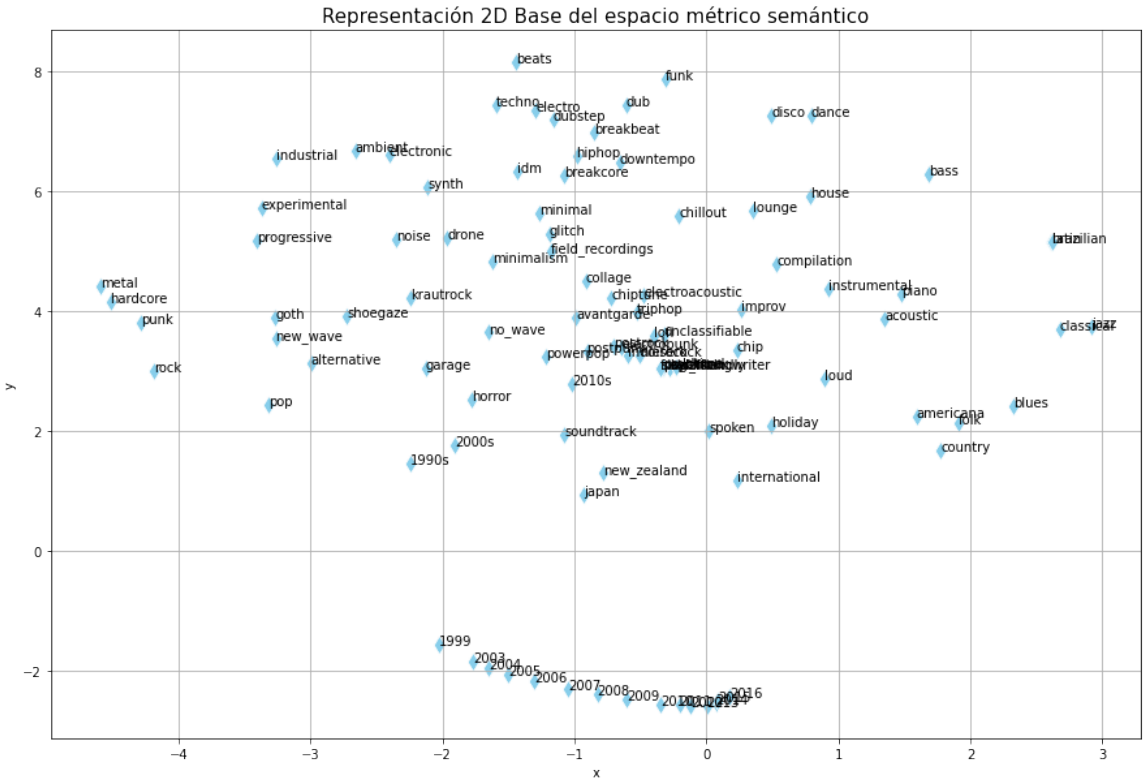


Figura 5.8: Base del espacio métrico semántico

- Embedding que representa los indicadores acústicos.

- Embedding que representa las portadas de álbum.

- Embedding que representa los perfiles de usuario (tags culturales).

Estos cuatro embeddings (que se obtuvieron siguiendo la metodología vista en Sección 5.2) son concatenados formando un embedding de 800 elementos que después pasa por una capa densa con 512 unidades ocultas, para después aplicar **batch normalization**, enseguida se aplica una función de activación **Relu** y un **dropout** con parámetro de probabilidad $p = 0.5$, seguida de una segunda capa densa con 256 unidades ocultas. De esta forma obtenemos un vector con 256 elementos que integra todos los modos de información de la pista.

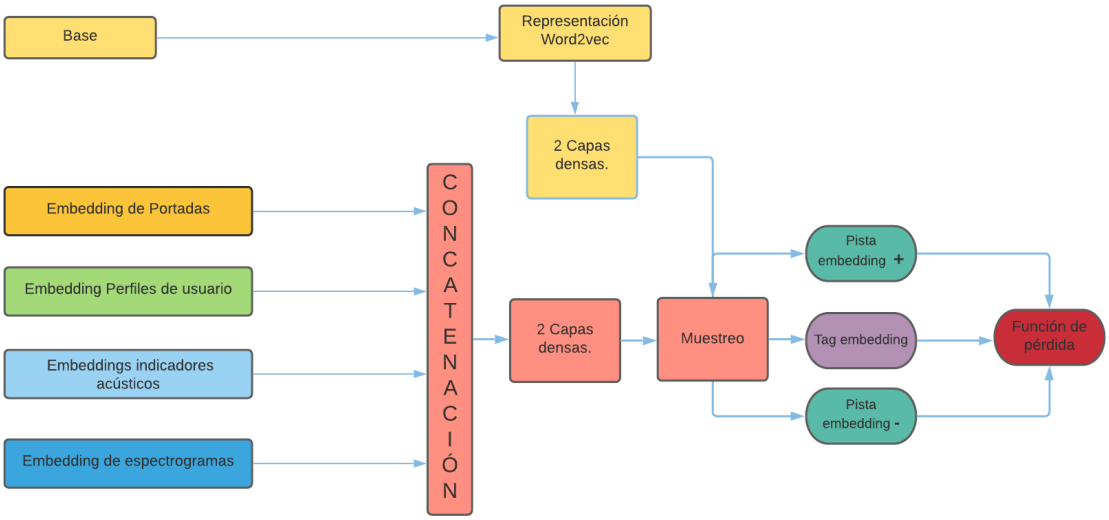


Figura 5.9: Diagrama. Construcción de espacio métrico semántico

5.3.4. Proceso de entrenamiento

Una vez obtenidos los embeddings que representan a los elementos base y los embeddings que representan a las pistas procedemos a realizar el muestreo descrito en Subsección 5.3.1 para seleccionar los tres embeddings que se usarán para evaluar la función de pérdida, este muestreo se realiza un total de $10,000$ veces por época. Una vez realizada la actualización se procede a actualizar los pesos de las distintas capas correspondientes a cada modo de información así como a las capas correspondientes a los embeddings de la base y de los embeddings multimodales. El diagrama del proceso de construcción de este espacio métrico se muestra en Figura 5.9. á

5.3.5. Espacio métrico acústico

Este espacio métrico recibe este nombre debido a que los embeddings que componen su base fueron construidos bajo características acústicas de los tags editoriales.

Base de referencia

Este espacio métrico toma como base los 100 tags editoriales más recurrentes en nuestro subconjunto de datos cada embedding que representa a cada elemento de

la base es construido a través de un modelo pre-entrenado con los espectrogramas obtenidos a mediante la señal de audio de las pistas y la relación que conecta a una pista con otra por medio de tags editoriales. La idea es crear un espacio métrico unimodal en el que pistas del mismo artista esten representadas por embeddings con similaridad alta, embeddings de pistas que compartan algún tag editorial tengan una similaridad media y embeddings de pistas que no compartan tags editoriales ni artistas obtengan una similaridad baja. La metodología para entrenar este modelo es el siguiente:

1. Se selecciona aleatoriamente un fragmento de pista de aproximadamente $4s$, del cuál tomamos la parte del espectrograma correspondiente equivalente a una matriz de dimensión 128×173 que funcionará como ancla.

2. Se selecciona aleatoriamente un segundo fragmento que al menos comparta artista con el primer fragmento de pista, es decir puede ser un fragmento de la misma pista o un fragmento de otra pista pero del mismo artista. Este segundo fragmento será nuestra muestra positiva.

3. Se selecciona un tercer fragmento que no comparta ni artista ni ningun tag editorial con el primer fragmento, esta pista se entenderá como una muestra negativa.

4. Cada fragmento pasa por las primeras 7 capas descritas en la Sección 5.2.1, mientras que la octava capa es una convolucional con 256 filtros, obteniendo así tres representaciones vectoriales.

5. Estas representaciones sirven como entradas a la función de pérdida descrita en Subsección 5.3.2, de esta forma buscamos un espacio en el que representaciones de pistas muy relacionadas estén cerca y pistas sin ninguna relación se mantengan a una distancia considerable.

Una vez entrenado el modelo lo usamos para obtener los embeddings que representen a las pistas.

Finalmente la representación de cada tag editorial se obtiene agrupando los embeddings de las pistas que se le relacionen. Formalmente se puede decir que si t_{ei} es el embedding del i-ésimo tag editorial y P_i el subconjunto de los embeddings de las pistas relacionadas al i-ésimo tag editorial entonces

$$t_{ei} = \frac{1}{|P_i|} \sum_{p_k \in P_i} p_k$$

En la Figura 5.10 podemos apreciar una representación en dos dimensiones de esta base, la representación fue obtenida mediante la técnica de reducción de dimensión **T-SNE**, podemos observar que a diferencia de la Figura 5.8 en esta representación los tags que hacen referencia al año de la pista no se agrupan entre sí, sino más bien se integran con los demás tags considerados. Esto se debe a que en este espacio no se toman en cuenta las características semánticas del tag sino las acústicas y por tanto este espacio acerca tags acústicamente parecidos.

Antes de ser evaluados por la función de pérdida estos embeddings pasan por una capa densa con 512 unidades ocultas, para después aplicar **batch normalization**, enseguida se aplica una función de activación **Relu** y un **dropout** con parámetro de probabilidad $p = 0.5$, seguida de una segunda capa densa con 256 unidades ocultas.

Embeddings Multimodales de las pistas

Cada embedding multimodal que representan a las pistas está conformado por embeddings de tamaño 200 que representan a las siguientes modalidades:

- Embedding que representa a los espectrogramas.
- Embedding que representa los indicadores acústicos.
- Embedding que representa las portadas de álbum.
- Embedding que representa la información de los perfiles de usuario (tags culturales).

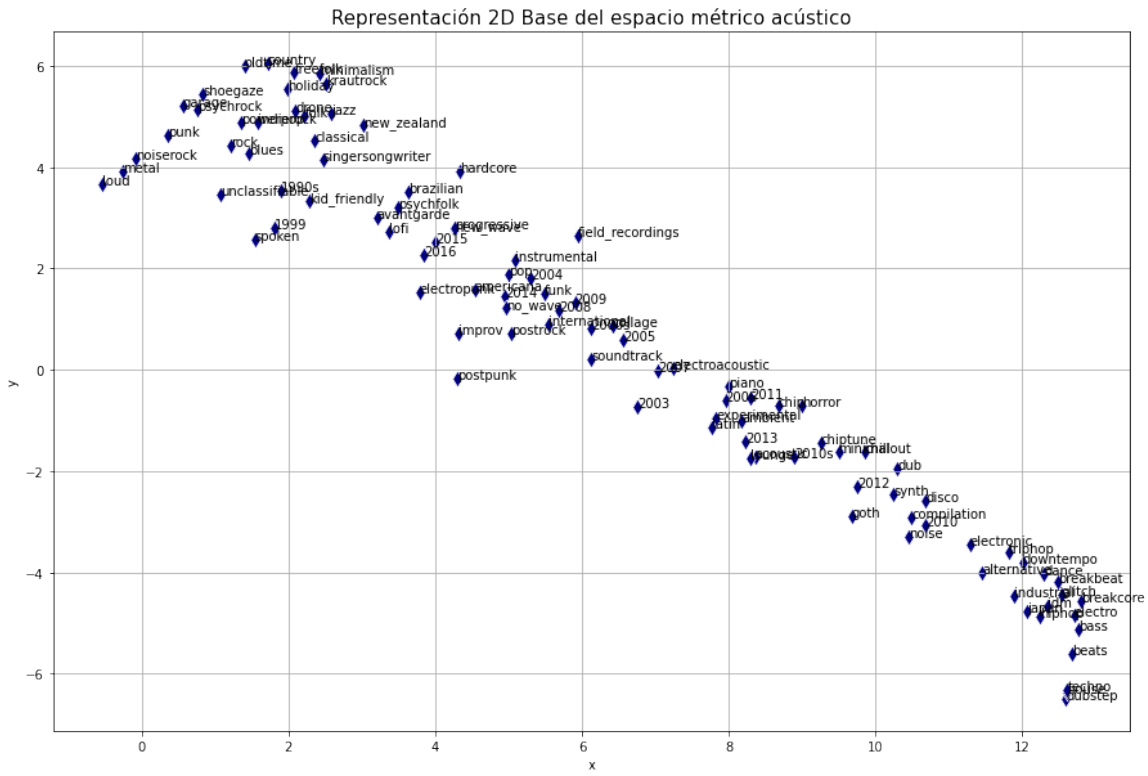


Figura 5.10: Base del espacio métrico acústico

Estos cuatro embeddings (que se obtuvieron siguiendo la metodología vista en Sección 5.2) son concatenados formando un embedding de 800 elementos que después pasa por una capa densa con 512 unidades ocultas, para después aplicar **batch normalization**, enseguida se aplica una función de activación **Relu** y un **dropout** con parámetro de probabilidad $p = 0.5$, seguida de una segunda capa densa con 256 unidades ocultas. De esta forma obtenemos un vector con 256 elementos que integra todos los modos de información de la pista.

5.3.6. Proceso de entrenamiento

Una vez obtenidos los embeddings que representan a los elementos base y los embeddings que representan a las pistas procedemos a realizar el muestreo descrito en Subsección 5.3.1 para seleccionar los tres embeddings que se evaluarán en la función de pérdida, este muestreo se realiza un total de $10,000$ veces por época. Una vez realizada la actualización se procede a actualizar los pesos de las distintas capas

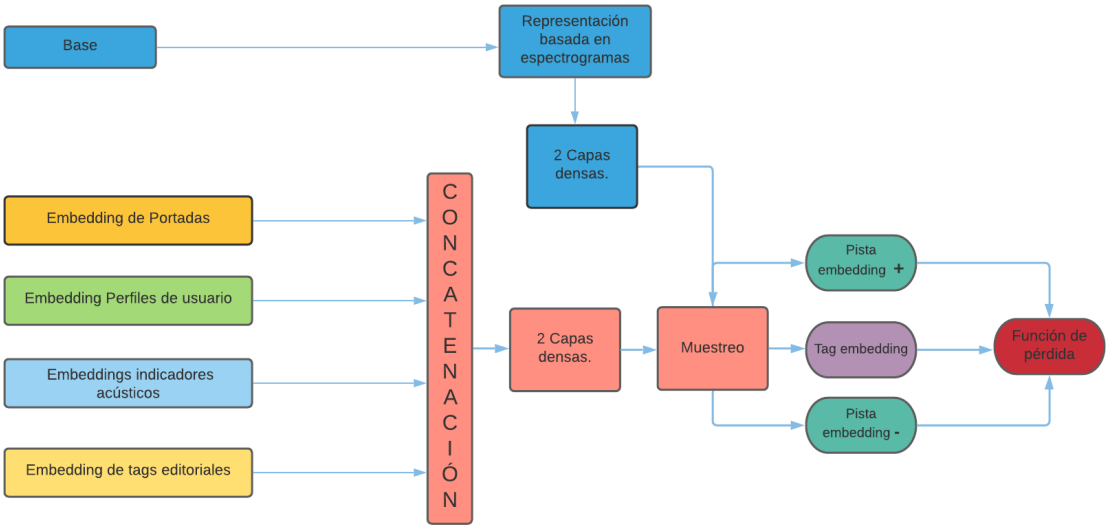


Figura 5.11: Diagrama. Construcción de espacio métrico acústico

correspondientes a cada modo de información así como a las capas correspondientes a los embeddings de la base y de los embeddings multimodales. El diagrama del proceso de construcción de este espacio métrico se muestra en Figura 5.11.

5.4. Métricas de evaluación

Para evaluar el desempeño del modelo se siguió la misma estrategia usada en Won y cols. (2020) donde se utilizan la precisión media promedio (**map** por sus siglas en inglés) y la precisión en k, para ambas métricas se utilizan la matriz de similaridades S y la matriz binaria B ambas de tamaño $m \times n$, donde m es el número de elementos en la base y n el número de pistas involucradas en la evaluación. La matriz de similaridades S está dada por

$$s_{ij} = \frac{< \mathbf{v}_i^{(t)}, \mathbf{v}_j^{(p)} >}{\|\mathbf{v}_i^{(t)}\| \|\mathbf{v}_j^{(p)}\|} \tag{5.9}$$

donde $\mathbf{v}_i^{(t)}$ es el embedding del i-ésimo elemento en la base y $\mathbf{v}_j^{(p)}$ el embedding que representa la j-ésima pista, es decir, cada elemento s_{ij} esta dado por la similaridad coseno entre los embedding del elemento de la base i y el embedding de la pista j. La matriz binaria B está definida por

$$b_{ij} = \begin{cases} 1 & \text{si el tag } i \text{ está relacionado a la pista } j \\ 0 & \text{en otro caso} \end{cases} \tag{5.10}$$

5.4.1. Precisión media promedio

Para el cálculo de esta métrica haremos uso de los conceptos de **precisión** y **exhaustividad** para lo que definimos la función indicadora de coincidencias como:

$$I_{coincidencia}(\mathbf{x}, i) = \begin{cases} 1 & \text{si } argmax(\mathbf{x}) = i \\ 0 & \text{en otro caso} \end{cases} \tag{5.11}$$

con $x \in \mathbb{R}^n$, $i \in \mathbb{N}$, $i \leq n$.

Luego para cada elemento de la base definimos el vector de coincidencias $\mathbf{c}_i$ con tamaño n donde cada elemento está dado por:

$$c_{ij} = I_{coincidencia}(\mathbf{s}_j^{(c)}, i) \tag{5.12}$$

donde $\mathbf{s}_j^{(c)}$ es la j-ésima columna de la matriz de similaridad. Es decir, el vector de coincidencias nos indica si el embedding de la j-ésima pisa es el más cercano al embedding del i-ésimo elemento de la base. Con lo anterior podemos obtener expresiones para calcular los **verdaderos positivos**, **falsos positivos**, y **falsos negativos** para cada tag como se muestra a continuación:

- **verdaderos positivos**: Nos indican el número de veces en el que el que el embedding del elemento i es el más cercana al embedding de la pista j estando relacionados y se calcula mediante

  $$vp_i = \mathbf{b}_i^{(r)} \mathbf{c}_i$$

  donde $\mathbf{b}_i^{(r)}$ es el i-ésimo renglón de la matriz binaria B.

- **falsos positivos**: Nos indican el número de veces en el que el embedding del elemento i es el más cercano al embedding de la pista j sin que estos embeddings

esten relacionados. mediante

$$fp_i = \sum_{j=1}^{n} c_{ij} - vp_i$$

- **falsos negativos**: Nos indican el número de veces en el que el embedding del elemento i no es el más cercano al embedding de la pista j estando relacionados y se calcula mediante

$$fn_i = \sum_{j=1}^{n} b_{ij} - vp_i$$

Finalmente la precisión queda definida como:

$$p_i = \frac{vp_i}{vp_i + fp_i} = \frac{vp_i}{\sum_{j=1}^{n} c_{ij}} \tag{5.13}$$

mientras que la exhaustividad se obtiene mediante:

$$r_i = \frac{vp_i}{vp_i + fn_i} = \frac{vp_i}{\sum_{j=1}^{n} b_{ij}} \tag{5.14}$$

luego definimos:

$$ap_i = \sum_{j=1}^{n} (r_{i,j} - r_{i,j-1}) p_{i,j}. \tag{5.15}$$

donde $p_{i,j}$ y $r_{i,j}$ son la precisión y la exhaustividad parciales calculadas para las primeras j pistas y $r_{i,0} = 0$. La idea de la Ecuación 5.15 es capturar el área bajo la curva que dibuja la relación entre la precisión y la exhaustividad. Note que $(r_{i,j} - r_{i,j-1})$ solo tiene dos valores posibles, 0 cuando la j-ésima pista es no es un verdadero positivo y $\frac{1}{\sum_{j=1}^{n} b_{ij}}$ cuando la j-ésima pista es un verdadero positivo, de tal forma que solo las precisiones con que obtengan verdaderos positivos se verán reflejados en la suma, por otro lado conforme aumente el número de falsos positivos aumente, el valor de la precisión disminuirá provocando que el valor de ap_i sea más pequeño, el valor máximo posible para ap_i es 1 y ocurre cuando $fp = fn = 0$. Finalmente la **Precisión media**

promedio se calcula a través del promedio de los ap_i, esto es:

$$map = \frac{1}{m}\sum_{i=1}^{m} ap_i \tag{5.16}$$

5.4.2. Precisión en k

La precisión en k es una medida que se calcula utilizando los k embeddings de las pistas más cercanas al embedding del elemento de la base i. El escenario ideal es que las k pistas correspondientes a los embeddings estén relacionadas al elemento i. Entonces, para obtener la precisión en k para el tag i definimos el conjunto de índices ι_i que contiene los índices de los embedding de las k pistas más cercanas al elemento de la base en cuestión, luego para todo $j \in \iota$ calculamos la precisión en k para el i-ésimo elemento como:

$$p@k_i = \frac{\sum_{j\in\iota} b_{ij}}{k}. \tag{5.17}$$

Finalmente la precisión en k del modelo está dada por el promedio del las precisiones en k de los m tags, esto es:

$$p@k = \frac{\sum_{i=1}^{m} p@k_i}{m}. \tag{5.18}$$

En este trabajo tomaremos $k = 10$ con el fin de tener de referencia los valores obtenidos en Won y cols. (2020) donde esta métrica se cálcula bajo ese valor de k.

Capítulo 6

Análisis y resultados

Tener 4 modalidades de información para representar a las pistas en cada espacio métrico nos da la oportunidad de experimentar con distintas combinaciones de cada uno de ellos, pudiendo entrenar desde un modelo considerando solo una modalidad de información hasta un modelo que integra las cuatro modalidades de información, obteniendo un total de 15 posibles combinaciones de información para cada tipo de espacio métrico, con fines comparativos los 15 modelos fueron entrenados utilizando los mismos parámetros que se describen en la siguiente sección.

6.1. Ajuste del modelo

Para entrenar nuestro modelo se utilizó el algorítmo **Adam** como optimizador con un **learning rate** de $lr = 10^{-4}$ considerando 200 épocas y un **batch** de 128 elementos, como se mencionó en la Sección 5.3 solo se consideraron los 100 tags culturales más relacionados a las pistas. Tomamos 70 % del conjunto de pistas para el proceso de entrenamiento, 15 % para la validación y el 15 % restante para el proceso de prueba. Las métricas que se utilizaron fueron:

- **MAP**: Mean Average Precision.
- **P@10**: Precision at 10.

los resultados obtenidos sobre el conjunto de prueba para **Espacio métrico semántico** se muestran en la Tabla 6.1, donde **E** hace referencia a los espectrogramas, **C** a la información cultural, **P** a las portadas de álbum e **I** a los indicadores acústicos. Mientras que los reultados obtenidos sobre el conjunto de prueba para el **Espacio métrico acústico** se muestran en la Tabla 6.2. En este caso **W** hace referencia la representación Word to vec, **C** a la información cultural, **P** a las portadas de álbum e **I** a los inidcadores acústicos. Para ambos casos el 1 indica que la información se consideró en el modelo y 0 si no se consideró. Podemos observar que en general los modelos entrenados bajo el **Espacio métrico semántico** tienen mejor desempeño aquellos que fueron entrenados bajo el **Espacio métrico acústico**. Además, para el **Espacio métrico semántico** los modelos que incluyen la información de las portadas de los álbums tienen un mejor desempeño, mientras que el modelo con el peor desempeño es el que solo considera los indicadores numéricos. Para el **Espacio métrico acústico**, los modelos que incluyen la información dada por las representaciones word2vec tienen un mejor desempeño, mientras que el modelo con el peor desempeño es el que solo considera los indicadores numéricos. En resumen, de los resultados obtenidos podemos concluir lo siguiente:

- **El Espacio métrico semántico tiene un mejor desempeño que el Espacio métrico acústico**
- **Para el Espacio métrico semántico los mejores resultados se obtienen cuando se considera la información de las portadas de álbum**.
- **Para el Espacio métrico acústico los mejores resultados se obtienen cuando se considera la información de la representación word to vec**.
- **En contraste, en ambos casos, los modelos con información acústica son los de peor desempeño. Por tanto, valdría la pena explorar otros parámetros y caracterizaciones de esta modalidad de información**.
- **Los resultados obtenidos son bastante competentes con los existentes en la literatura, por ejemplo ver Won y cols. (2020)**, donde se realiza

E	C	P	I	T. Entrenamiento	MAP	P@10
0	1	1	1	47:29	47.64 %	64.5 %
0	1	1	0	44:27	46.88 %	62.6 %
1	1	1	0	2:38:49	46.48 %	62.2 %
1	1	1	1	2:41:41	46.16 %	62 %
1	0	1	0	2:32:12	43.98 %	61 %
0	0	1	0	37:29	43.25 %	58.4 %
1	0	1	1	2:36:46	42.67 %	57.3 %
0	0	1	1	43:58	41.63 %	56.2 %
1	1	0	1	2:38:27	29.56 %	42.3 %
1	1	0	0	2:34:40	29.58 %	42.2 %
0	1	0	1	38:43	27.93 %	40.08 %
0	1	0	0	37:06	26.75 %	39.3 %
1	0	0	0	2:29:19	14.08 %	21.10 %
1	0	0	1	2:34:43	13.17 %	19.3 %
0	0	0	1	35:33	10.09 %	13.7 %

Tabla 6.1: Resultados en el conjunto de prueba en el **Espacio métrico semántico** ordenados por el desempeño de cada uno. **E**: Espectrograma,**C**: Información cultural, **P**: Portadas de álbum, **I**: indicadores acústicos.

un ejercicio muy similar con pistas del conjunto de datos **Million dataset** considerando dos modalidades de información **Espectrogramas** e **Información cultural** por medio de matrices pista-usuario, los resultados obtenidos en dicho trabajo se presentan en Tabla 6.3.

Para el **Espacio métrico semántico** el modelo que considera tres de las cuatro modalidades de información disponibles obtuvo el mejor desempeño en cuanto a las métricas consideradas, estas modalidades son: la información cultural, la información de portadas y los indicadores de audio numéricos. En la Figura 6.1 podemos observar una representación en dos dimensiones obtenida mediante la técnica **t-SNE** del **Espacio métrico semántico** generado usando esta información, en cada subimagen se muestra una pista diferente (punto amarillo) con sus respectivos tags editoriales relacionados, podemos destacar que en la mayoria de los casos la representación de la pista queda muy cerca de al menos la representación de un tag editorial relacionado, siendo solo las subimagenes de la derecha del segundo y tercer renglón las únicas en las que la representación de la pista no aparece suficientemente cerca de sus tags relacionados. Cabe mencionar que las pistas mostradas en Figura 6.1 fueron seleccionadas

W	C	P	I	T. Entrenamiento	MAP	P@10
1	1	1	1	0:43:9.0	43.75 %	51.4 %
1	0	0	1	0:43:16.0	41.37 %	49 %
1	1	0	1	0:44:21.0	41.21 %	46 %
1	1	1	0	0:44:19.0	37.96 %	43.4 %
1	1	0	0	0:43:48.0	36.55 %	41.5 %
1	0	1	0	0:43:17.0	37.84 %	39.5 %
1	0	0	0	0:42:23.0	32.79 %	39.3 %
0	0	1	1	0:43:41.0	28.5 %	38.1 %
0	1	1	1	0:45:11.0	26.68 %	35.5 %
1	0	1	1	0:44:43.0	30.95 %	34 %
0	1	1	0	0:45:43.0	24.23 %	33.1 %
0	0	1	0	0:42:29.0	21.82 %	29.6 %
0	1	0	0	0:42:59.0	15.89 %	20.8 %
0	1	0	1	0:44:5.0	10.51 %	14.9 %
0	0	0	1	0:42:28.0	7.5 %	8 %

Tabla 6.2: Resultados en el conjunto de prueba con el **Espacio métrico acústico** ordenados por el desempeño de cada uno. **W**: Representación Word2vec,**C**: Información cultural, **P**: Portadas de álbum, **I**: Indicadores acústicos.

Métrica	I. Cultural	Espectrogramas	I. Combinada
MAP	11.55 %	18.52 %	17.75 %
P@10	32 %	35 %	31.2 %

Tabla 6.3: Resultados obtenidos en Won y cols. (2020)

aleatoriamente.

La Figura 6.2 tiene la misma interpretación que la Figura 6.1 pero para el mejor modelo del **Espacio métrico acústico** el cuál considera todas las modalidades de información disponibles, las pistas seleccionadas son las mismas que las que se muestran en la Figura 6.1. Podemos observar que en este caso las veces que al menos una representación de los tags relacionados aparece cerca de su representación de pista correspondiente es menor que en la Figura 6.1, sin embargo las representaciones de las pistas siguen sin alejarse demasiado de las representaciones de sus tags editoriales asociados.

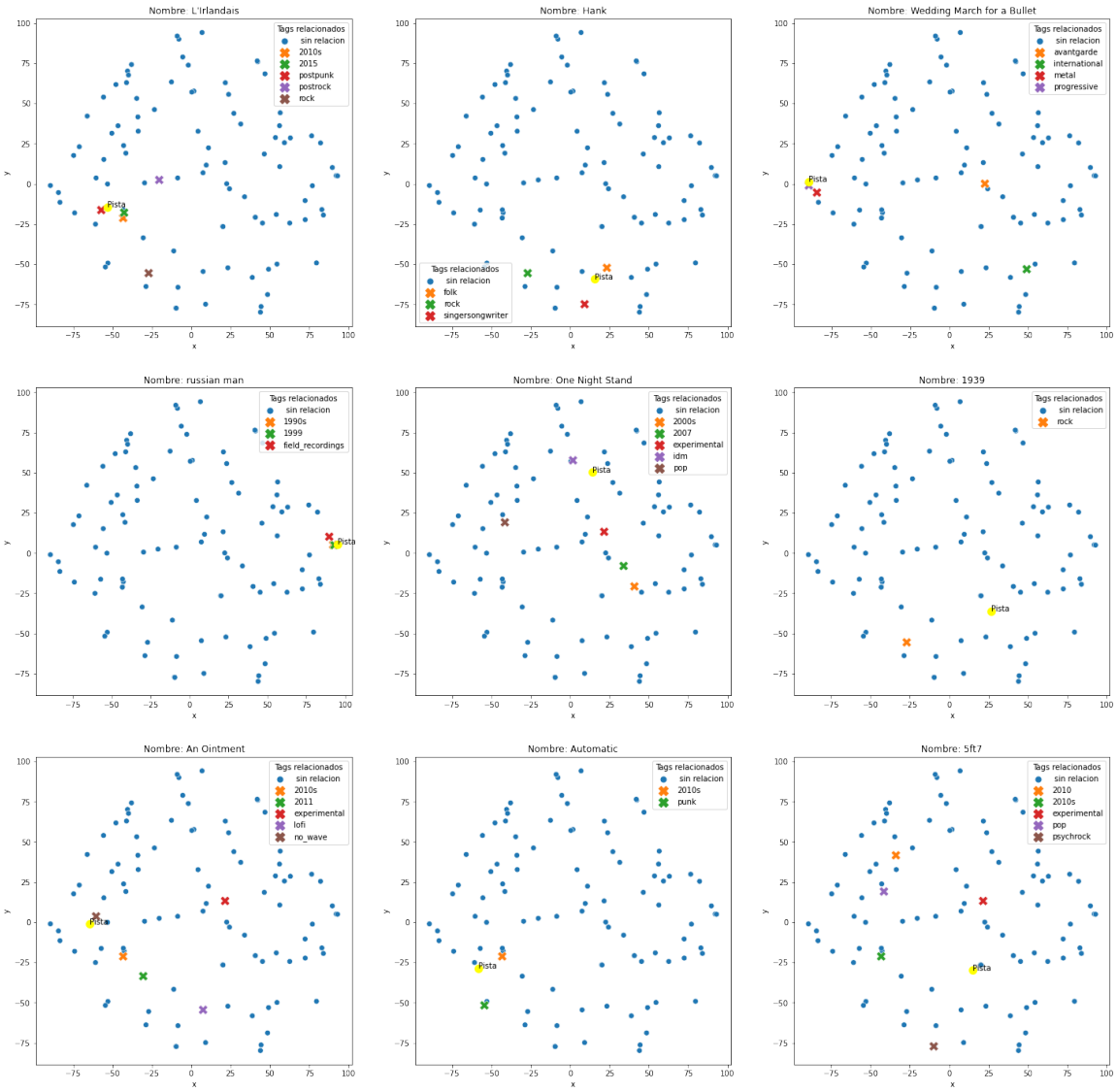


Figura 6.1: Representaciones 2D obtenidas mediante **t-SNE** del **Espacio métrico semántico** compartido entre tags y pistas. En cada sub imagen se muestran las representaciones de 100 tags editoriales y una pista seleccionada aleatoriamente

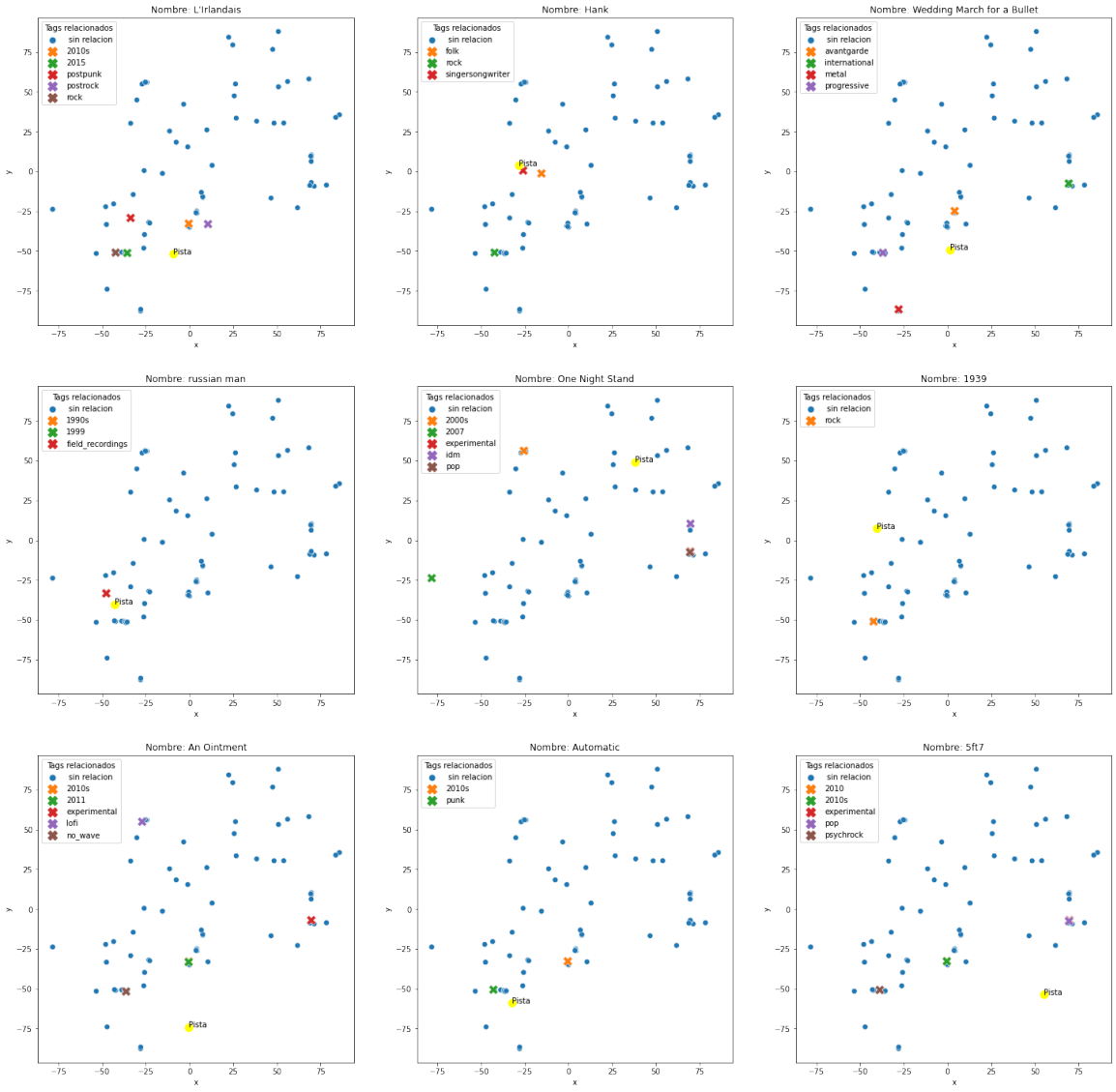


Figura 6.2: Representaciones 2D obtenidas mediante **t-SNE** del **Espacio métrico acústico** compartido entre tags y pistas. En cada sub imagen se muestran representaciones de 100 tags editoriales y una pista seleccionada aleatoriamente

6.2. Modelo de clasificación de género

El objetivo de construir cada uno de los modelos es obtener un espacio métrico y una representación multimodal para cada pista de nuestro subconjunto que resida en el espacio métrico obtenido. Con el fin de probar la calidad de las representaciones obtenidas en cada modelo creado, realizamos una tarea de clasificación en la que tratamos de identificar el género principal de cada pista dado por los metadatos del conjunto **fma** usando como entrada las representaciones obtenidas en cada uno de estos modelos.

El modelo que se ocupó para esta clasificación fue una red neuronal con 4 capas densas de 128, 64, 32 y 14 unidades respectivamente, a las tres primeras capas se les aplicó una función de activación $RELU$ y a la cuarta una función $softmax$. Además de la clasificación que se puede obtener de la red neuronal, se consideraron 6 clasificadores más:

- Regresión Logística
- Máquina de soporte vectorial con Kernel lineal
- Máquina de soporte vectorial con Kernel gaussiano
- Årboles de decisión
- Random Forest
- Extreme Gradient Boosting

Estos clasificadores fueron conectados a la penúltima capa del modelo de red neuronal, de esta forma cada clasificador recibe como entrada un vector de tamaño 32. El diagrama de la estrategia descrita anteriormente puede observarse en la Figura 6.3. Los meta-parámetros que se consideraron para cada clasificador son los fijados por deafault en la librería de **python sklearn**[1].

Dado que contamos con 15 representaciones vectoriales para cada tipo de espacio métrco, 7 clasificadores y utilizando validación cruzada considerando 5 particiones en

[1]La documentación puede encontrarse en https://scikit-learn.org/stable/

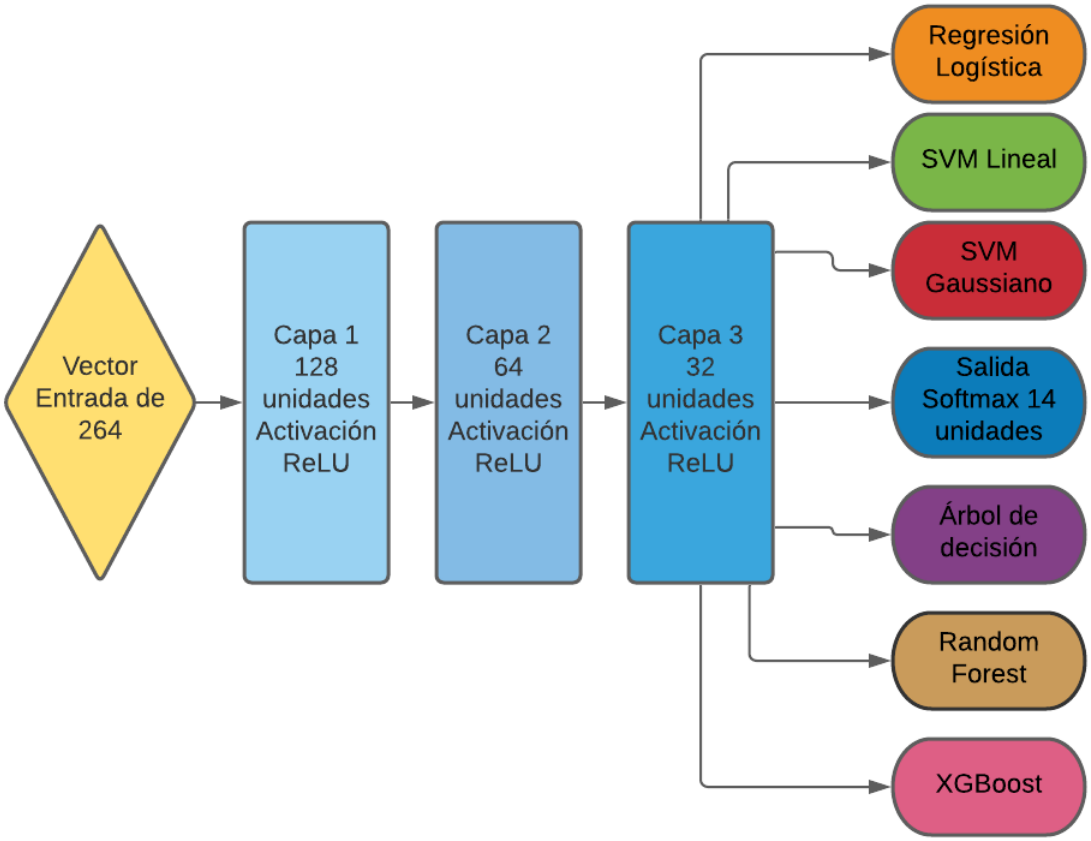


Figura 6.3: Diagrama de la estrategia de clasificación seguida

nuestro conjunto de datos, se repitió el proceso de entrenamiento y prueba del modelo un total de 525 veces por tipo de espacio métrico. Los resultados obtenidos se muestran en la Subsección 6.2.1 para el **Espacio métrico semántico** y Subsección 6.2.2 para el **Espacio métrico acústico**

6.2.1. Resultados: Clasificación bajo el Espacio métrico semántico

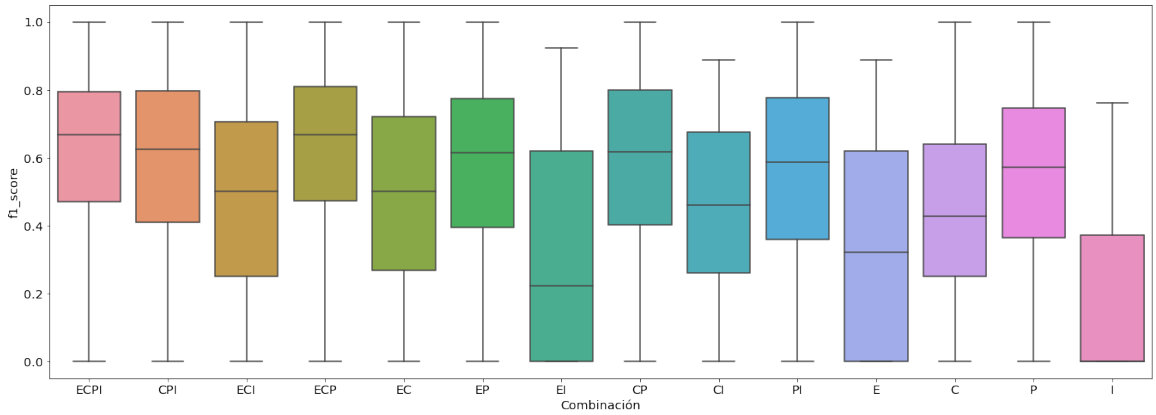
En la Figura 6.4a podemos observar el desempeño obtenido por cada combinación de modalidades usadas para crear los embeddings, cada combinación está codificada por cuatro dígitos donde cada posición representa a una modalidad diferente, la primer posición es para los espectrogramas, la segunda para la información cultural, la tercera para las portadas de álbum y la cuarta para los indicadores de audio, se asigna 1 cuando la modalidad fue considerada para crear el embedding y 0 cuando la modalidad no fue considerada. De las 15 combinaciones destacan principalmente tres :

1. La representación que se obtuvó al considerar espectrogramas, la información cultural y las portadas de álbums, obteniendo un f_1 score promedio de 0.618, mientras que su mediana se sitúa en 0.667.

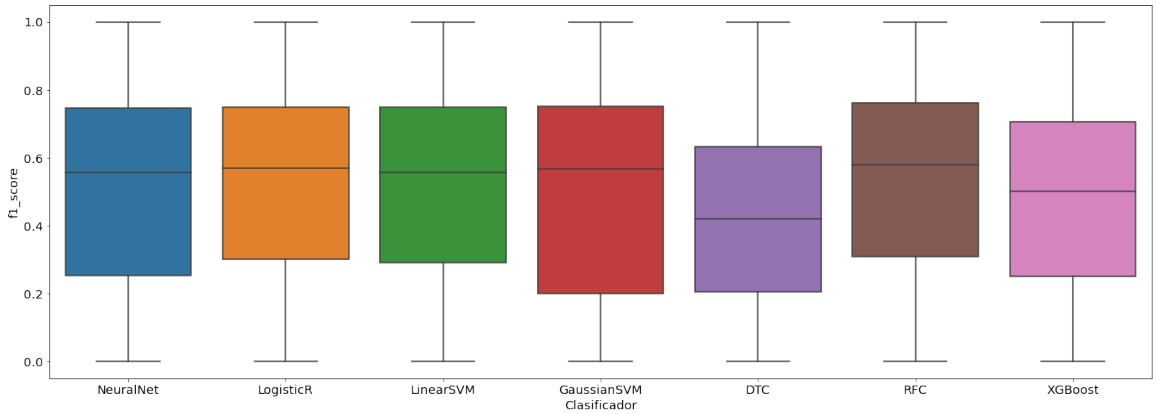
2. La representación que se obtuvó al considerar los cuatro modos de información, obteniendo un f_1 score promedio de 0.602 y con mediana de 0.667

3. La representación que se obtuvó al considerar la información cultural, las portadas de álbums y las características de audio numéricas obteniendo un f_1 score promedio de 0.596, con mediana de 0.657

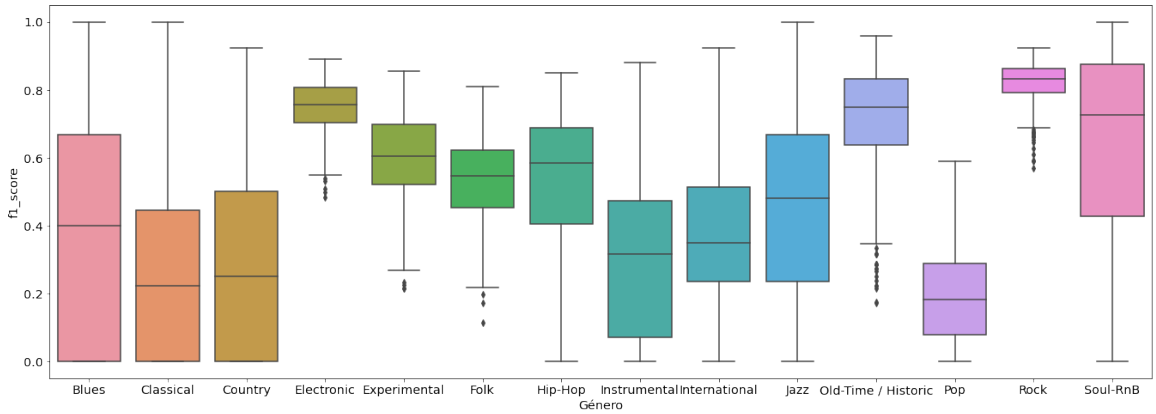
Por otro lado, las representaciones que mostraron el desempeño más bajo son las que involucran solo información acústica, como es el caso de la representación obtenida al considerar solo características de audio numéricas que obtiene un f_1 score promedio de 0.193 y el valor de la mediana es de 0, la representación dada al considerar espectrogramas y características de audio que obtuvo un f_1 score promedio de 0.32



(a) f1 score obtenido por combinación de modalidades



(b) f1 score obtenido por clasificador



(c) f1 score obtenido por género

Figura 6.4: Resultados clasificación para el **Espacio métrico semántico**

con mediana en 0.22 y la representación dada al considerar solo espectrogramas que obtuvo un f_1 score promedio de 0.346 con mediana en 0.32. Podemos ver que existe una clara relación entre los resultados mostrados en Tabla 6.1 y los obtenidos en estos experimentos de clasificación.

En Figura 6.4b observamos que en general los clasificadores tienen un desempeño muy similar, siendo **Random Forest** el de mejor desempeño con un f_1 score proemdio de 0.521 y mediana de 0.58, seguido por **Regresión Logística** con un f_1 score promedio de 0.51 con mediana de 0.569 y **Máquina de soporte vectorial con Kernel lineal** con un f_1 score promedio de 0.505, y mediana de 0.558.

Finalmente en la Figura 6.4c podemos ver como los géneros **rock** y **electronic** destacan de entre los demás con un f_1 score promedio de 0.821 y 0.751 respectivamente segudos por el género **Old-Time/ Historic** con un f_1 score promedio de 0.721.

Los géneros que resultaron más difíciles de clasificar fueron **Pop** que obtuvo un f_1 score promedio de solo 0.196, el género **Classical** con solo 0.251 de f_1 score promedio y el género **Country** con un f_1 score promedio de 0.292. En la Figura 6.5 podemos observar el top 10 de las combinaciones representación-clasificador con mejor desempeño, de donde destacamos las 5 primeras:

1. Conjunto: Espectrogramas, información cultural y portadas de álbum.
   Clasificador: Red neuronal.
   f_1 score promedio de 0.651.

2. Conjunto: Espectrogramas, información cultural y portadas de álbum.
   Clasificador: Random Forest.
   f_1 score promedio de 0.651.

3. Conjunto: Información cultural, portadas de álbum e indicadores de audio numéricos.
   Clasificador: Random Forest.
   f_1 score promedio de 0.65.

4. Conjunto: Los cuatro modos de información.

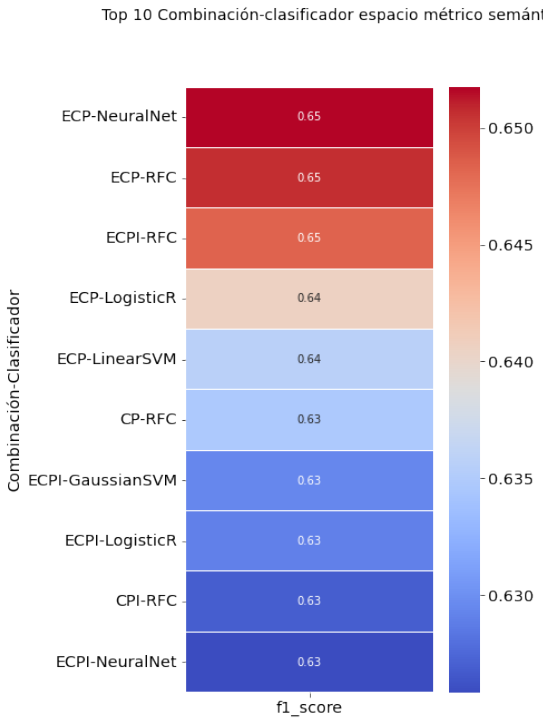


Figura 6.5: Top 10 combinaciones representación-clasificador con los mejores resultados para el **Espacio métrico semántico**

Clasificador: Random Forest.

f_1 score promedio de 0.648.

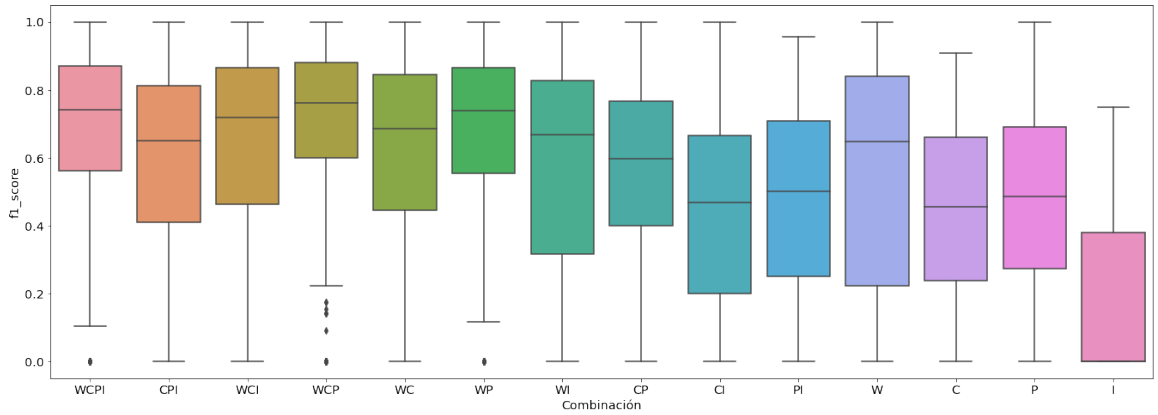
5. Conjunto: Espectrogramas, información cultural y portadas de álbum.

   Clasificador: Regresión Logística.

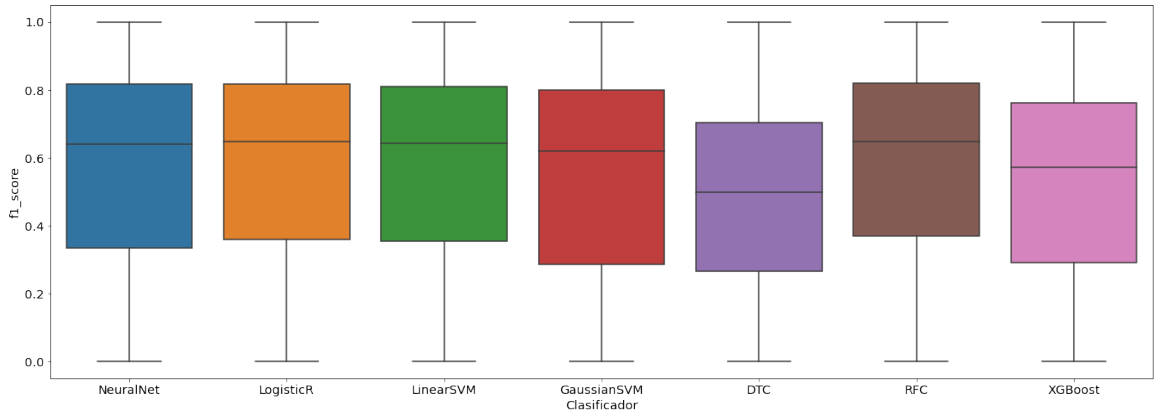
   f_1 score promedio de 0.641.

6.2.2. Resultados: Clasificación bajo el Espacio métrico acústico

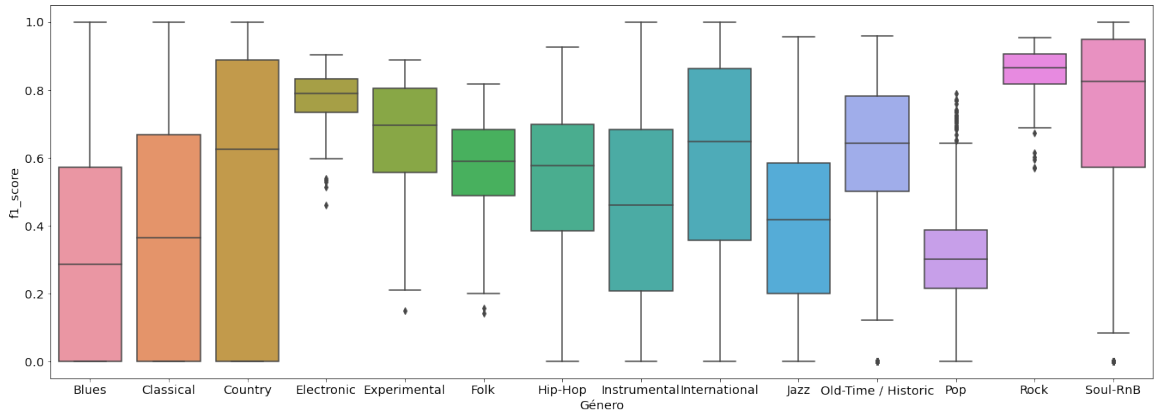
En la Figura 6.6a podemos observar el desempeño obtenido por cada combinación de modalidades usadas para crear los embeddings, cada combinación está codificada por cuatro dígitos donde cada posición representa a una modalidad diferente, la primer posición es para las representaciones word2vec, la segunda para la información cultural, la tercera para las portadas de álbum y la cuarta para los indicadores de audio,



(a) f1 score obtenido por combinación de modalidades



(b) f1 score obtenido por clasificador



(c) f1 score obtenido por género

Figura 6.6: Resultados clasificación para el **Espacio métrico acústico**

se asigna 1 cuando la modalidad fue considerada para crear el embedding y 0 cuando la modalidad no fue considerada. De las 15 combinaciones destacan principalmente tres:

1. La representación que se obtuvó al considerar las representaciones word2vec, la información cultural y las portadas de álbums, obteniendo un f_1 score promedio de 0.696 con mediana de 0.762.

2. La representación que se obtuvó al considerar los cuatro modos de información, obteniendo un f_1 score promedio de 0.741.

3. La representación que se obtuvó al considerar las representaciones word2vec, las poratdas de álbum y las características de audio numéricas obteniendo un f_1 score promedio de 0.68 y una mediana de 0.732.

Por otro lado las representaciones que mostraron el desempeño más bajo son las que solo involucran los indicadores de audio numéricos y la información cultural. Para la representación que solo considera indicadores de audio numéricos se obtuvo un f_1 score promedio de 0.191 y una mediana de 0, la representación dada al considerar solo información cultural obtuvo un f_1 score promedio de 0.431 con una mediana de 0.456 y la representación dada al considerar ambos tipos de información obtuvo un f_1 score promedio de 0.44 y mediana de 0.468. Observemos que existe una clara relación entre los resultados mostrados en Tabla 6.2 y los obtenidos en estos experimentos de clasificación.

En la Figura 6.6b observamos que en general los resultados son muy parecidos a los obtenidos con el **Espacio métrico semántico** donde los clasificadores tienen un desempeño muy similar, siendo en este caso la **Regresión Logística** la de mejor desempeño con un f_1 score proemdio de 0.578 y mediana de 0.647, seguido por **Random Forest** con un f_1 score promedio de 0.577 con mediana de 0.648 y **Máquina de soporte vectorial con Kernel lineal** con un f_1 score promedio de 0.57 y mediana de 0.643.

Finalmente en la Figura 6.6c podemos ver como los géneros **rock** y **electronic** destacan de entre los demás con un f_1 score promedio de 0.851 y 0.777 respectivamente

segudos por el género **Old-Time/ Historic** con un f_1 score promedio de 0.712. Los géneros que resultaron más difíciles de clasificar fueron **Blues** que obtuvo un f_1 score promedio de solo 0.3, el género **Pop** con solo 0.308 de f_1 score promedio y el género **Classical** con un f_1 score promedio de 0.388. En la Figura 6.7 podemos observar el top 10 de las combinaciones representación-clasificador con mejor desempeño, podemos observar que las represenatciones que consideraron tres de los cuatro modos de información existentes en esta arquitectura siendo estos las representaciones word to vec, la información cultural y las portadas de álbum ocupan los primeros cuatro puestos con diferentes clasificadores en cada caso los cuales describimos a continuación:

1. Conjunto: word to vec, información cultural y portadas de álbum.
   Clasificador: Máquina de soporte vectorial con kernel lineal.
   f_1 score promedio de 0.73.

2. Conjunto: word to vec, información cultural y portadas de álbum.
   Clasificador: Random Forest.
   f_1 score promedio de 0.73.

3. Conjunto: word to vec, información cultural y portadas de álbum.
   Clasificador: Máquina de soporte vectorial con kernel gausseano.
   f_1 score promedio de 0.73.

4. Conjunto: word to vec, información cultural y portadas de álbum.
   Clasificador: Regresión logística.
   f_1 score promedio de 0.72.

en general podemos ver que los resultados obtenidos bajo el **Espacio métrico acústico** superan a los obtenidos bajo el **Espacio métrico semántico** logrando un f_1 score promedio de 0.55 contra un f_1 score promedio de 0.49 obtenido con el **Espacio métrico semántico**, este resultado puede atribuirse a la contribución de las representaciones word to vec como modo de representación de las pistas, ya que esta

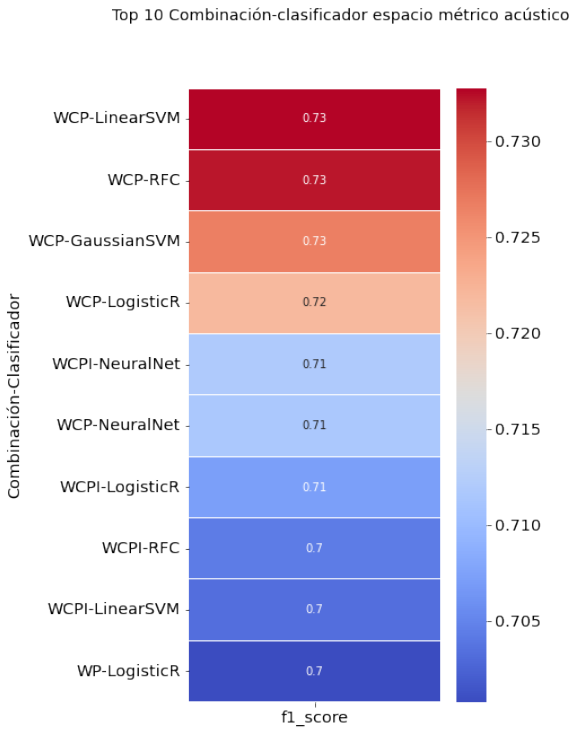


Figura 6.7: Top 10 combinaciones representación-clasificador con los mejores resultados para el **Espacio métrico acústico**

presente en los mejores resultados obtenidos por el **Espacio métrico acústico**.

6.3. Modelo de recuperación de información

Otra tarea que aprovecha el espacio construido con todos los modos de información disponible y cada modo de información de forma individual, es la recuperación de información multimodal, con el que se pueden realizar consultas bajo tres modos de información texto, imagen y audio.

1. **Texto**: Ingresando nombre de pista y artista. Con estos datos se realiza el siguiente proceso:

   a) buscamos los tags culturales mediante la api de **lastfm** para obtener nuestro vector pista-tag.

b) A nuestro vector pista-tag aplicamos el mismo proceso que se aplicó a nuestra matriz pista-tag para obtener un vector de tamaño 200 que represente a nuestra información cultural antes de ser concatenada, esto es: factorización con mínimos cuadrados alternantes, y dos capas densas.

2. **Imagen**: Ingresando la ruta de un archivo de imagen. Idealmente se espera que la imagen sea una portada de álbum. La imagen es preprocesada para que sus dimensiones sean 300×300, después aplicamos el mismo proceso que se aplicó a nuestro conjunto de entrenamiento para obtener un vector de tamaño 200 que represente a nuestra información editorial antes de ser concatenada, esto es: el modelo pre- entrenado RESNET101 y tres capas densas.

3. **Audio**: Ingresando la ruta de un archivo de audio. En este punto tenemos dos eventos de acuerdo a al espacio métrico considerado :

   - **Espacio métrico semántico**. Obtenemos el espectrograma del archivo de audio y sus características numéricas, después aplicamos el mismo proceso que se aplicó a nuestro conjunto de entrenamiento, 8 capas de convolución en el caso de los espectrogramas, y estandarización y dos capas densas en el caso de las características numéricas, obteniendo finalmente dos vectores de tamaño 200 cada uno.
   - **Espacio métrico acústico**. Solo obtenemos las características numéricas, estandarizamos y aplicamos las dos capas densas correspondientes a nuestro modelo previamente entrenado, obteniendo un vector de tamaño 200.

Una vez realizada la consulta y haber obtenido la representación vectorial correspondiente , buscamos entre nuestro subconjunto de datos la representación x más cercana en distancia coseno dada por la Ecuación 5.8, enseguida buscamos en el espacio métrcio compartido la representación de x y sus $k-1$ representaciones más cercanos también bajo la misma distancia. De esta forma podemos encontrar pistas similares considerando cuatro modos de información recibiendo solo uno, el diagrama

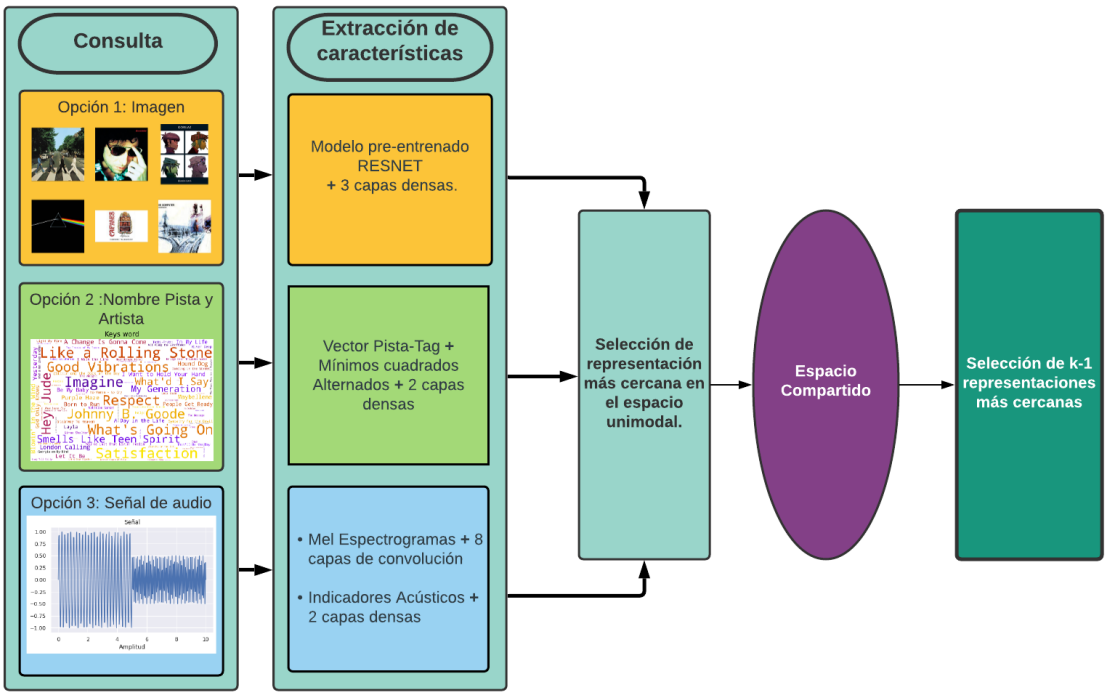


Figura 6.8: Diagrama del sistema de recuperación de información

de la Figura 6.8 describe a grandes rasgos el proceso de recuperación descrito anteriormente.

Medir el desempeño de nuestro modelo de recuperación de información resulta ser algo complejo debido a la subjetividad existente en la música. Con el fin de observar el desempeño del modelo de recuperación se realizaron distintas experimentaciones con archivos de audio, imágenes e información musical fuera del subconjunto que usamos para entrenarlo, los resultados se presentan a continuación.

6.3.1. Resultados: consultas con archivos de audio

Se seleccionó aleatoriamente una pista por cada género del conjunto de datos **fma** que no estuviese en el subconjunto que usamos para entrenar nuestro modelo principal, después se realizó una consulta para recuperar las 5 canciones más similares utilizando las representaciones vectoriales de espectrogramas y características numéricas.

Resultados: Espacio métrico semántico

Debido a que existen 16 géneros principales en la base **fma**, solo mostramos los resultados para los que consideramos podrían darnos un mayor contraste, como el **Rock** (género que obtuvo el mejor f_1 score en los modelos de clasificación para el **Espacio métrico semántico**) y el **Pop** (género que obtuvo el peor f_1 score en los modelos de clasificación para el **Espacio métrico semántico**). Esta primer arquitectura presenta una ventaja al realizar consultas con archivos de audio ya que puede tomar dos puntos como referencia (los espectrogramas y los indicadores numéricos) dando como resultado un un conjunto de pistas recuperadas más diverso como veremos a continuación.

En la Tabla 6.4 y Tabla 6.5 podemos observar los resultados obtenidos para una canción del género **Rock** de la década de los 2010's en la primera la recuperación se realizó usando el espectrograma de la pista y en la segunda sus características de audio numéricas. Podemos observar que cuando se usa el espectrograma se recuperan canciones de un solo artista coincidendo en el género principal y la década de la pista. Por otro lado, cuando se usan características de audio numéricas la recuperación resulta más variada en cuanto a artistas y en consecuencia en tags, vemos que todas las pistas a excepción de la cuarta más cercana coniciden en el género principal y ninguna logra coincidir en la década.

En Tabla 6.6 y Tabla 6.7 podemos observar los resultados obtenidos para una canción del género **Pop** de la década de los 2010's, relacionada a dos subgéneros, en la primera la recuperación se realizó usando el espectrograma de la pista y en la segunda sus características de audio numéricas. Podemos observar que cuando se usa el espectrograma ninguna canción logra atinar ni el género principal ni la la década, sino que parece inclinarse más hacia uno de sus subgeneros como lo es el experimental Pop que guarda relación con el género **Experimental** y el subgenero Drone los cuales aparecen en repetidas ocasiones en los tags de las canciones recuperadas . Por

Pista	Título	Artista	Tags
Consulta	Let The Kid Come Out	Glenn Morrow's Cry For Help	Rock, 2010's, 2017
Top 1	Waiting Room	Terry Malts	Rock, Punk, Lo-fi, 2010's, 2012
Top 2	Tumble Down	Terry Malts	Rock, Punk, Lo-fi, 2010's, 2012
Top 3	I Do	Terry Malts	Rock, Punk, Lo-fi, 2010's, 2012
Top 4	Something About Yout	Terry Malts	Rock, Punk, Lo-fi, 2010's, 2012
Top 5	Mall Dreams	Terry Malts	Rock, Punk, Lo-fi, 2010's, 2012

Tabla 6.4: Resultados con género Rock. Criterio: Espectrogramas. **Espacio métrico semántico**

Pista	Título	Artista	Tags
Consulta	Let The Kid Come Out	Glenn Morrow's Cry For Help	Rock, 2010's, 2017
Top 1	1,000,000 Kisses	Half Japanese	Rock, 2000's, 2008
Top 2	Punk Rock vs. Swiss Modernism	Double Dagger	Rock, 2000's, 2009
Top 3	Surfing With My 2 Little Brothers	Party People in a Can	Rock, Pop, Indie-Rock, Portugal, Surf, 2000's, 2006
Top 4	SMisen Gymnastics	Oorutaichi	Electronic, 2000's, 2009
Top 5	Wave Goodbye	Kelley Stoltz	Rock, Pop, Indie Rock, 2000's, 2006

Tabla 6.5: Resultados con género Rock. Criterio: Características de audio numéricas. **Espacio métrico semántico**

Pista	Título	Artista	Tags
Consulta	Dialogg of the Unknown	Phemale	Pop, Experimental Pop, Synth Pop, 2010's, 2011
Top 1	Divine Flesh	Pocahaunted	Drone, 2000's, 2009
Top 2	Kvallsol	Lithis	Drone, Experimental, Ambient, 2000's, 2006
Top 3	svartedauden	Rngmnn	Drone, Experimental, Ambient, 2000's, 2006
Top 4	Ljuset	Lithis	Drone, Experimental, Ambient electronic, 2000's, 2005
Top 5	Cr	Jari Pitkanen	Electronic, Audio Collage, 2000's, 2008

Tabla 6.6: Resultados con género Pop. Criterio: Espectrogramas. **Espacio métrico semántico**

otro lado, cuando se usan características de audio numéricas la recuperación resulta tener más coincidencias, específicamente en cuanto al subgénero Experimental Pop y la década de la pista, por otro lado no es variada en cuanto a artistas ya que solo se recuperaron pistas de un solo artista.

Resultados: Espacio métrico acústico

El segundo espacio métrico tiene una diferencia importante cons respecto al primero: solo toma como referencia los indicadores numéricos. Con el objetivo de comparar ambas arquitecturas se consideraron las mismas pistas usadas para la consulta con archivo de audio del **Espacio métrico semántico**, la información de las 5 pistas recuperadas con el audio de la pista del género **Rock** se muestran en la Tabla 6.8, podemos observar que se recuperaron pistas de solo un artista, en cuanto a tags las 5 comparten el correspondiente al género **Rock** lo que se traduce en un resultado positivo, los mienstras que los tags correspondientes a la época de la pista no son del todo acertados.

La información de las 5 pistas recuperadas con el audio de la pista del género

Pista	Título	Artista	Tags
Consulta	Dialogg of the Unknown	Phemale	Pop, Experimental Pop, Synth Pop, 2010's, 2011
Top 1	The Gift	Ergo Phizmiz	Avant-Garde, Electroacoustic, Experimental Pop, 2010's, 2010
Top 2	Lady Godiva's Operation	Ergo Phizmiz	Avant-Garde, Electroacoustic, Experimental Pop, 2010's, 2010
Top 3	I Heard Her Call My Name	Ergo Phizmiz	Avant-Garde, Electroacoustic, Experimental Pop, 2010's, 2010
Top 4	Here She Comes Now	Ergo Phizmiz	Avant-Garde, Electroacoustic, Experimental Pop, 2010's, 2010
Top 5	Sister Ray	Ergo Phizmiz	Pop, Rock, Experimental Pop

Tabla 6.7: Resultados con género Pop. Criterio: Características numéricas. **Espacio métrico semántico**

Pop se muestran en la Tabla 6.9, debemos resaltar que las pistas recuperadas son las mismas que se recuperaron con el **Espacio métrico semántico** usando esta pista y teniendo como referente los indicadores de audio numéricos, lo único que cambia es el orden de las pistas recuperadas, concluyendo que para este género tanto el espacio compartido como el espacio de las representaciones de inidcadores de audio podrían verse muy similares en ambas arquitecturas.

Considerando las coincidencias en los tags para ambas arquitecturas podemos ver que el sistema obtiene resultados aceptables.

Pista	Título	Artista	Tags
Consulta	Let The Kid Come Out	Glenn Morrow's Cry For Help	Rock, 2010's, 2017
Top 1	Cindy	Gary Wilson	Rock, 2000's, 2002
Top 2	When You Walk Into My Dreams	Gary Wilson	Rock, 2000's, 2002
Top 3	Chromium Bitch	Gary Wilson	Rock, 2000's, 2002
Top 4	And Then I Kissed Your Lips	Gary Wilson	Rock, 2000's, 2002
Top 5	You Think You Really Know Me	Gary Wilson	Rock, 2000's, 2002

Tabla 6.8: Resultados con género Rock. **Espacio métrico acústico**

Pista	Título	Artista	Tags
Consulta	Dialogg of the Unknown	Phemale	Pop, Experimental Pop, Synth Pop, 2010's, 2011
Top 1	The Gift	Ergo Phizmiz	Avant-Garde, Electroacoustic, 2010's, 2010, Experimental pop
Top 2	Here She Comes Now	Ergo Phizmiz	Avant-Garde, Electroacoustic, 2010's, 2010, Experimental pop
Top 3	Lady Godiva's Operation	Ergo Phizmiz	Avant-Garde, Electroacoustic, 2010's, 2010.0, Experimental pop
Top 4	I Heard Her Call My Name	Ergo Phizmiz	, Avant-Garde, Electroacoustic, 2010's, 2010, Experimental pop
Top 5	Sister Ray	Ergo Phizmiz	, Avant-Garde, Electroacoustic, 2010's, 2010, Experimental pop

Tabla 6.9: Resultados con género Pop. **Espacio métrico acústico**

6.3.2. Resultados: consultas con imágenes

El modelo permite realizar consultas mediante imagenes, en esencia se espera que estas imagenes sean portadas de álbum, sin embargo, se puede experimentar con cualquier tipo de imagen, por ejemplo la que se muestra en la Figura 6.9a, en este caso, la idea es que dada una imagen se recuperen pistas con portadas de álbum similares a ella y que estas pistas se relacionen de una manera a ella.

Resultados: Imagen que no pertenece a la portada de un álbum

Una vez más la calidad del resultado es subjetiva, sin embargo con el fin de dar una idea del desmepeño del sistema se realizó la una consulta considerando la imagen de la Figura 6.9a como entrada, las imágenes recuperadas utilizando el **Espacio métrico semántico** pueden observarse en la Figura 6.9b mientras que la recuperadas usando el **Espacio métrico acústico** se muestran en la Figura 6.9c, si tomamos en cuenta que lo que se observa en la imagen de consulta es a grandes rasgos un paisaje debemos favorecer los resultados regresados por el **Espacio métrico acústico**, sin embargo debemos recordadar que esta apreciacion es totalmente subjetiva. No basta con ver las imágenes recuperadas para darnos una idea de la calidad de recuperación de cada espacio métrico, por esta razón en la Tabla 6.10 se muestran los datos de las primeras 5 pistas recuperadas usando el **Espacio métrico semántico** mientras que en la Tabla 6.11 se muestran los datos de las pistas recuperadas utilizando el **Espacio métrico acústico**.

Las primeras cuatro pistas recuperadas con el **Espacio métrico semántico** son del género electronic, la quinta se realciona con la música electrónica mediante el subgénero **minimal electronic**, este tag caracteriza a la consulta ya que lo comparten las 5 pistas y es referente a música de corte experimental con una extructura repetitiva y que utiliza como elementos base el género electrónico. Además del subgénero estas pistas son contemporaneas siendo todas de la década del 2000 oscilando entre los años 2006 y 2008.

Las 5 pistas recuperadas con el **Espacio métrico acústico** pertenecen al mismo álbum y artista, por lo que comparten los mismos tags, este conjunto de tags podemos dividirlos en tres subconjunto:

- Dubstep, Electronic. Estos tags hacen referencia a la base del sonido de las pistas. Lo que nos indica que se usa como base los elementos de la música electránica y el dupstep.

- Chill-out, Instrumental, piano, Uplifting,atmospheric. Estos tags hacen referencia al tipo de música al que pertenecen las pistas, en etse caso se puede pensar como una música de de relajación, meditación o inspiración, ya que todos estos tags están relacionados a este tipo de sensaciones.

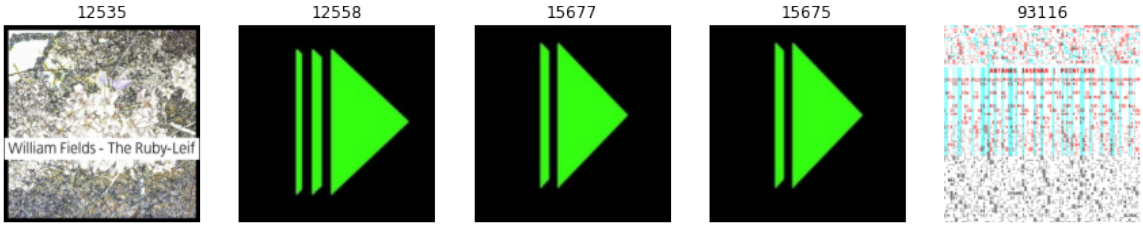
- 2010's, 2013. Estos tags solo hacen referencia a la época de las pistas.

Por lo anterior podemos concluir que las pistas recuperadas correponden a una mezcla de música considerada tranquila basada en sonidos electrónicos.
Determinar si estas pistas están o no relacionadas a la imagen de consulta esta totalmente condicionado al juicio del oyente.

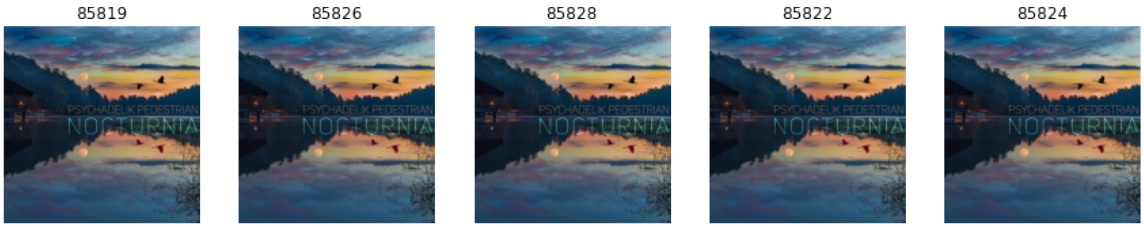
Resultados: Imagen de un álbum

La segunda prueba de consultas con imagen que se presenta en este trabajo es la realizada con la portada del álbum **Dark Side of The Moon** de la banda de rock progresivo **Pink Floyd**, la razón de escojer este álbum es que es bastante famoso por lo que análizar el resultado de las consultas será menos complejo que usando un álbum menos conocido. En la Figura 6.10 podemos observar la imagen usada como consulta y las imágenes de las pistas recuperadas (Figura 6.10b para el **Espacio métrico semántico** y Figura 6.10c para el **Espacio métrico acústico**), la primer arquitectura recupera pistas de tres álbums diferentes el primero tiene un triángulo en su portada, característica que comparte con la imagen consulta, el segundo comparte el fonfo oscuro y el tercero parece no compartir ninguna característica destacable.

(a) Imagen consulta



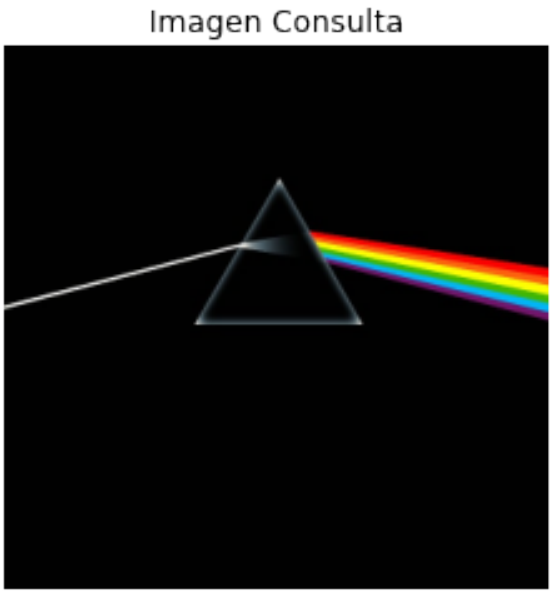
(b) Portadas de las pistas recuperadas. **Espacio métrico semántico**



(c) Portadas de las pistas recuperadas. **Espacio métrico acústico**

Figura 6.9: Resultados de la consulta con imagenes. Imagen que no pertenece a una portada de álbum en concreto

En cuanto a las imágenes de las pistas recuperadas con la segunda arquitectura se detectan dos álbums diferentes y ambos comparten el fondo oscuro. La información de las pistas recuperadas con el **Espacio métrico semántico** se resumen en la Tabla 6.12 donde se destacan los tags **Synth pop** y **2010's**. La década de las pistas no coincide con la década del álbum que se usó como consulta (1970), mientras que el subgénero **Synth pop** sí escontemporáneo del álbum ya que comenzó a aparecer en la década de los 70's pero toma elementos de la música disco, el glam rock y el post-punk que poco están relacionados al álbum Dark Side of the Moon, a juzgar por la información obtenida se puede inferir que la consulta no fue de una calidad aceptable, sin embargo aún existe la posibilidad de que exista una similaridad sonora, que una vez más es completamente subjetiva. La información de la consulta realizada con el **Espacio métrico acústico** puede apreciarse en la Tabla 6.13, en este caso aparece una gama muy amplía de tags que se repiten en las 5 pistas y que hacen referencia a música experimental y vanguardista creada a través del Jazz, la música electrónica y la improvisación, si consideramos que Dark Side of the Moon fue un álbum vanguardista en su época podemos encontrar un punto de similitud con las pistas recuperadas, recordemos que además de ser un disco principalmente de rock progresivo, en este álbum se experimentó con la fusión de distintos efectos de sonido lo que puede indicarnos que la calidad de la consulta es aceptable.



(a) Imagen consulta

(b) Portadas de las pistas recuperadas. **Espacio métrico semántico**

(c) Portadas de las pistas recuperadas. **Espacio métrico acústico**

Figura 6.10: Resultados de la consulta con imagenes. Álbum Dark Side of the Moon

Pista	Título	Artista	Tags
Top 1	Caim	William Fields	Electronic, minimal electronic, 2000's, 2008
Top 2	Orchestral Tortures	Lee Rosevere	Electronic, Experimental, ambient, minimal electronic, 2000's, 2007
Top 3	Nobody Goes To Heaven	Lee Rosevere	Electronic, Experimental, ambient, minimal electronic, 2000's, 2007
Top 4	There Is No Music On This Side	Lee Rosevere	Electronic, Experimental, ambient, minimal electronic, 2000's, 2007
Top 5	em_mxspSF-8	Antanas Jasenka	Experimental, Minimalism, minimal electronic, 2000's, 2006

Tabla 6.10: Resultados de consulta con imagen que no pertenece a una portada de álbum. **Espacio métrico semántico**

Pista	Título	Artista	Tags
Top 1	Cyberstalker	Psychadelik Pedestrian	Dubstep, Chill-out, Electronic, Instrumental, piano, Uplifting, atmospheric, 2010's, 2013
Top 2	In Suspense	Psychadelik Pedestrian	Dubstep, Chill-out, Electronic, Instrumental, piano, Uplifting, atmospheric, 2010's, 2013
Top 3	Cosmos (universal mix)	Psychadelik Pedestrian	Dubstep, Chill-out, Electronic, Instrumental, piano, Uplifting, atmospheric, 2010's, 2013
Top 4	Nocturnia	Psychadelik Pedestrian	Dubstep, Chill-out, Electronic, Instrumental, piano, Uplifting, atmospheric, 2010's, 2013
Top 5	Pacific	Psychadelik Pedestrian	Dubstep, Chill-out, Electronic, Instrumental, piano, Uplifting, atmospheric, 2010's, 2013

Tabla 6.11: Resultados de consulta con imagen que no pertenece a una portada de álbum. **Espacio métrico acústico**

Pista	Título	Artista	Tags
Top 1	Far Apart	Airglow	Electronic, 2010's, 2014, synth pop
Top 2	Electrifying Landscape	Airglow	Electronic, 2010's, 2014, synth pop
Top 3	Laugh Tracks	Garmisch	Pop, 2010's, 2010, synth pop
Top 4	I Was Awake (dustmotes Remix)	Stewart and Scarfe	Pop, Electronic, 2010's, 2013, synth pop
Top 5	I Was Awake (Don't Sleep)	Stewart and Scarfe	Pop, Electronic, 2010's, 2013.0, synth pop

Tabla 6.12: Resultados de consulta con poratda del álbum Dark Dide of the Moon. **Espacio métrico semántico**

Pista	Título	Artista	Tags
Top 1	The High Priestess	Goat	Experimental, Avant-Garde, Free-Jazz, Jazz, Electronic, clinical archives, improvisation, various, 2000's, 2008
Top 2	There's a Hole at the End of the Tunnel	The Black Hakawati	Experimental, Avant-Garde, Free-Jazz, Jazz, Electronic, clinical archives, improvisation, various, 2000's, 2008
Top 3	Tango Argentino	Joan Silver Pin	Experimental, Avant-Garde, Free-Jazz, Jazz, Electronic, clinical archives, improvisation, various, 2000's, 2008
Top 4	The Pink Shoes Of Marie Antoinette are sailing on a Ship of Thoughts	Infinitus Ensemble	Experimental, Avant-Garde, Free-Jazz, Jazz, Electronic, clinical archives, improvisation, various, 2000's, 2008
Top 5	Captain B (Track 1)	Populaere Mechanik	Experimental, Avant-Garde, Free-Jazz, Jazz, Electronic, clinical archives, improvisation, various, 2000's, 2008

Tabla 6.13: Resultados de consulta con portada del álbum Dark Dide of the Moon. **Espacio métrico acústico**

6.3.3. Resultados: consultas con texto

Siguiendo con la línea de la Sección 6.3.2 para este tipo de consulta se utilizó la canción **Time** de **Pink Floyd**[2] debido a que es una canción que pertence al álbum **Dark Side of the Moon** del cuál usamos su portada para realizar consultas con imágenes. Esta pista es posiblemente la más conocida de este álbum. Los resultados obtenidos se muestran en la Tabla 6.14 para el **Espacio métrico semántico** y en la Tabla 6.15 para el **Espacio métrico acústico**.

Las pistas recuperadas con el **Espacio métrico semántico** presentan una variedad importante en cuanto a sus tags asociados (ver la Tabla 6.14), los tags predominantes son el **internacional** y el **Folk** los que nos indica que las pistas recuperadas pertenecen a pistas fuera de los Estados Unidos o Gran Bretaña así como de sus géneros tradicionales, esta característica se ve reforzada por la tercer, cuarta y quinta pistas recuperada la cuál están acompañadas por los tag **Brazil** y **Balkan**, por lo que podemos inferir que estas pistas son o bien de Brasil o de los Balcanes situados en Europa y además puede persivirse como música tradicional de estos lugares del planeta. A juzgar por estos dos tags predominantes podemos inferir que no hay mucha relación con **time** en este sentido ya que justamente **Pink Floyd** es una banda proveniente de la Gran Bretaña sin embargo se pueden rescatar algunos tags que podrían relacionar las pistas con la pista usada como consulta, los cuales son **Experimental Pop**, **Singer-Songwriter** y **Avant-Garde**. Con estos resultados podemos inferir que la consulta mediante texto usando el **Espacio métrico semántico** no es del todo aceptable.

Las información de las pistas recuperadas con el **Espacio métrico acústico** puede consultarse en Tabla 6.15 en la que podemos ver que las 5 pistas son del año 2011, un punto en contra en cuanto a la calidad de la consulta, sin embargo tres de ellas están asociadas al tag **Psych-Rock** subgénero estrechamente relacionado a la banda **Pink Floyd** lo que anota un punto a favor de la calidad de esta consulta, el otro tag

[2]Los tags de esta pista fueron tomados de wikipedia

protagonista de esta consulta es el **Indie-Rock** el cual es un subgénero del rock que engloba la música Rock creada de una forma ïndependientez que hace referencia a músicos que no se apoyaron de grandes compañías discográficas para crear música y realizaron esta actividad a través de sus propios medios. En esta persepectiva no se puede hablar de Time como una pista perteneciente al **Inide-Rock** sin embargo no se descarta la influencia que pudo haber tenido la pista en el subgénero. Con base los puntos comentados anteriormente podemos concluir que la consulta medianete texto con el **Espacio métrico acústico** es aceptable.

En genral es necesario evaluar las pistas recuperadas de forma cualitativa para tener certeza sobre la calidad de cada una de las consultas realizadas, por lo que un trabajo interesante es tomar una muestra de oyentes que se dediquen a calificar las similitudes entre los elementos de consulta y los audios de las pistas recuperadas, de tal forma que se pueda construir un indicador de la calidad de cada consulta, actividad que está fuera del alcance del objetivo de este trabajo, por lo que se dejará como trabajo a futuro.

Pista	Título	Artista	Tags
Consulta	Time	Pink Floyd	Progressive rock, art rock, Funk rock, Psych-Rock, 1973, 1970's
Top 1	Village Mentality	Vialka	International, Folk, Experimental Pop, 2010's, 2010
Top 2	Bitter Heart	Zee Avi	International, Singer-Songwriter
Top 3	Peregum	Axial	International, Singer-Songwriter, Avant-Garde, Electroacoustic, Folk, sound poetry, Brazil, female vocals, 2000's, 2007
Top 4	Kissing By The Lake	Bonifrate	Folk, Singer-Songwriter, Brazil, 2000's, 2007
Top 5	Kandes Tsirkec	Paniks	International, Folk, Balkan

Tabla 6.14: Resultados con Time de Pink Floyd. Criterio: Nombre de pista y artista. **Espacio métrico semántico**

Pista	Título	Artista	Tags
Consulta	Time	Pink Floyd	Progressive rock, art rock, Funk rock, Psych-Rock, 1973, 1970's
Top 1	We Ask You To Ride	Wooden Shjips	Psych-Rock, 2010's, 2011
Top 2	Lazy Bones	Wooden Shjips	Psych-Rock, 2010's, 2011
Top 3	Raymond Play	Dumbo Gets Mad	Indie-Rock, 2010's, 2011
Top 4	Crossing	Wooden Shjips	Psych-Rock, 2010's, 2011
Top 5	New Age Dinosaur	Radical Dads	Indie-Rock, 2010's, 2011

Tabla 6.15: Resultados con Time de Pink Floyd. Criterio: Nombre de pista y artista. **Espacio métrico acústico**

Capítulo 7

Conclusiones y trabajo futuro

7.1. Conclusiones

Con el trabajo realizado en este proyecto podemos concluir lo siguiente:

- **La información relativa a un elemento puede encontrarse en diferentes modalidades**. Esto da pauta a la creatividad del investigador para identificar los modos de información relacionados a un elemento que mejor nos ayuden a caracterizarlo.

- **Los métodos de caracterización son bastos y pueden ser determinantes en el desempeño de un modelo**. Este punto se puede ejemplificar con el desempeño que tuvieron los modelos que incluian las portadas de albums y las representaciones word to vec, ya que ambos involucraban modelos pre-entrenados.

- **Trabajar con información de audio puede llegra a ser bastante complejo y costoso computacionalmente**. Esto se ve reflejado en los modelos que involucraban solo características de sonido ya que fueron los que tuvieron un menor desempeño y el tiempo que se tomó en entrenarlos fue mayor, sin embargo no pueden dejarse de lado ya que la persepectiva acústica es importante al momento de caracterizar pistas. Lo que supone un gran margen de

mejora en cuanto a preprocesamiento, caracterización y modelado de este tipo de información.

- **interpretar la información cultural puede ser bastante complejo** El método de caracterización cultural fue satisfactorio sin embargo puede mejorarse de gran manera ya sea pasando de una matrz pista-tag a una matriz pista-usuario o explorando más métodos siempre cuidando la perspectiva cultura que se puede tener del elemento en estudio.

- **Conseguir representaciones vectoriales puede facilitar tareas de aprendizaje.** Como se vió en el Capítulo 6 obtener vectores representativos es bastante práctico. Como ejemplo los tiempos de entrenamiento de los 15 modelos del **Espacio métrico acústico**, los modelos de clasificación de la Sección 6.2 y los de recuperación de información de la Sección 6.3.

- **Los modelos multimodales son bastante utiles para realizar consultas de información cruzada** Como se vió en Sección 6.3 los odelos que contemplan información multimodal nos sirven para realizar consultas que dada un tipo de información como entrada nos regrese una salida en otro modo de información pero relacionado a la información de entrada, ejemplos de estas consultas son: texto-audio, audio-texto, texto-imagen, imagen-texto, etc.

- **Los sistemas de recuperación de información multimodal pueden enriquecer la diversidad de los resultados**. Los sistemas de recuperación que consideran distintos modos de información pueden aprovecharlos para tomar distintos puntos de referencia, en concreto considerando el sistema que se implementó en este problema se pudieron haber tomado hasta cuatro modos de información como referencia (espectrogramas, indicadores de audio numéricos, imagenes y texto) cada uno regresando una pista cercana a cada representación, lo que puede ser bastante util en sistemas de recomendación donde se tiene abundante información de las pistas que suele escuchar un usuario.

En general este ha sido un proyecto bastante ambicioso con resultados muy in-

teresantes que dan pauta a una investigación más profunda en en distintos puntos del trabajo realizado.

7.2. Trabajo futuro

Este trabajo ofrece bastantes vertientes que seguir para expandir los alcances del mismo, entre los cuales se encuentran los siguientes:

- **Espectrogramas**. Se pueden explorar más métodos, modelos y arquitecturas para extraer información de los espectrogramas, con el fin de que obtengan un mejor desempeño al crear un espacio métrico.

- **Indicadores acústicos**. Se pueden agregar más indicadores acústicos y realizar un análisis bastante profundo que nos ayude a identificar aquellos que sean de más utilidad para caracterizar una pista, además de encontrar modelos que nos ayuden a mejorar su desempeño.

- **Información Cultural** La matriz pista-tag ofrece resultados aceptables en cuanto a información cultural se refiere, sin embargo, se puede considerar una matriz más amplia (pista- usuario) e inclusive encontrar más información de este tipo en distintas fuentes, como reseñas y opiniones de las pistas, explorar como obtenter y preprocesar esta información es una tarea bastante atractiva.

- **Información editorial** Otro tipo de información editorial que puede ser relevante al momento de caracterizar una pista es la letra que incluye, por lo que también resulta interesante explorar este tipo de información en un futuro.

- **Modalidad de referencia (base) en el espacio métrico**. En este trabajo la base del modelo métrico fueron tags editoriales con representación word to vec y representaciones obtenidas mediante espectrogramas, sin embargo aún quedan muchas modalidades por explorar, tales como embeddings de tags editoriales construidos a través de información cultural o de portadas de álbumes. Además el conjunto de referencia se cierra solo a tags editoriales sino que se pueden

tomar otros elementos como base, por ejemplo el género, año, artista, álbum o la misma pista.

- **Evaluación de los modelos de recuperación de información.** Estos modelos tienen la característica de que la calidad de sus resultados puede ser bastante subjetiva por lo que es necesario explorar formas de evaluarlos y de construir un indicador que nos ayude a decidir si los resultado obtenidos son aceptables o no.

- **Correlación entre tipos de información** Estes otro rubro bastante interesante que no se explora en este trabajo, por ejemplo se puede explorar los patrones que relaciona la información cultural con la acústica.

Aún hay bastante por hacer y mucha metodología por explorar sin embargo este trabajo sienta las bases de los siguientes proyectos relacionados.

Referencias

Bertin-Mahieux, T., P.W. Ellis, D., Whitman, B., y Lamere, P. (2011). The million song dataset. *Proc. of the 12th International Society for Music Information Retrieval Conference*. doi: https://doi.org/10.7916/D8NZ8J07

Bogdanov, D., Wack, N., Gómez, S., E.and Gulati, Herrera, P., Mayor, O., Roma, G., ... Serra, X. (2013). Essentia: an audio analysis library for music information retrieval.. Descargado de https://essentia.upf.edu

Choi, K., Lee, J. H., y Downie, J. S. (2014). What is this song about anyway?: Automatic classification of subject using user interpretations and lyrics. En *Ieee/acm joint conference on digital libraries* (p. 453-454). doi: 10.1109/JCDL .2014.6970221

Cohen, W., y Fan, W. (2000, 06). Web-collaborative filtering: recommending music by crawling the web. *Computer Networks*, *33*, 685-698. doi: 10.1016/S1389 -1286(00)00057-8

Defferrard, M., Benzi, K., Vandergheynst, P., y Bresson, X. (2017). FMA: A dataset for music analysis. En *18th international society for music information retrieval conference (ismir)*. Descargado de https://arxiv.org/abs/1612.01840

Degara, N., Rua, E. A., Pena, A., Torres-Guijarro, S., Davies, M. E. P., y Plumbley, M. D. (2012). Reliability-informed beat tracking of musical signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 290-301. doi: 10.1109/TASL.2011.2160854

Fujishima, T. (1999). Realtime chord recognition of musical sound: a system using common lisp music. En *Icmc*.

Grzywczak, D., y Gwardys, G. (2014, 08). Deep image features in music information

retrieval. En (Vol. 60, p. 187-199). doi: 10.1007/978-3-319-09912-5_16

Gómez, E. (2006, 08). Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing*, *18*, 294-304. doi: 10.1287/ijoc.1040.0126

Hu, Y., Koren, Y., y Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. En *2008 eighth ieee international conference on data mining* (p. 263-272). doi: 10.1109/ICDM.2008.22

Korzeniowski, F., Nieto, O., McCallum, M., Won, M., Oramas, S., y Schmidt, E. M. (2020). Mood classification using listening data. *CoRR*, *abs/2010.11512*. Descargado de https://arxiv.org/abs/2010.11512

Law, E., Ahn, L., Dannenberg, R., y Crawford, M. (2007, 01). Tagatune: A game for music and sound annotation. En (p. 361-364).

Mikolov, T., Chen, K., Corrado, G., y Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Murauer, B., y Specht, G. (2018). Detecting music genre using extreme gradient boosting. En *Companion proceedings of the the web conference 2018* (p. 1923–1927). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. Descargado de https://doi.org/10.1145/3184558.3191822 doi: 10.1145/3184558.3191822

Müller, M. (2015). *Fundamentals of music processing* (1.ª ed.). springer International Publishing. doi: 10.1007/978-3-319-21945-5

Oramas, S., Barbieri, F., Nieto Caballero, O., y Serra, X. (2018). Multimodal deep learning for music genre classification. *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.*.

Oramas, S., Espinosa-Anke, L., Lawlor, A., y et al. (2016). *Exploring customer reviews for music genre classification and evolutionary studies.*

Pachet, F. (2005, 09). Knowledge management and musical metadata. En (p. 672-). doi: 10.4018/978-1-59904-931-1.ch114

Pastor-Pellicer, J., Zamora-Martínez, F., España-Boquera, S., y Castro-Bleda, M. J. (2013). F-measure as the error function to train neural networks. En I. Rojas,

G. Joya, y J. Gabestany (Eds.), *Advances in computational intelligence* (pp. 376–384). Berlin, Heidelberg: Springer Berlin Heidelberg.

Streich, S., y Herrera, P. (2012, 01). Detrended fluctuation analysis of music signals: Danceability estimation and further semantic characterization.

Suris, D., Duarte, A., Salvador, A., Torres, J., y Giro-i Nieto, X. (2018, September). Cross-modal embeddings for video and audio retrieval. En *Proceedings of the european conference on computer vision (eccv) workshops.*

Temperley, D. (1999). What's key for key? the krumhansl-schmuckler key-finding algorithm reconsidered. *Music Perception: An Interdisciplinary Journal*, *17*(1), 65–100. Descargado de http://www.jstor.org/stable/40285812

Tzanetakis, G., y Cook, P. R. (2002). Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.*, *10*, 293-302.

Won, M., Oramas, S., Nieto, O., Gouyon, F., y Serra, X. (2020). Multimodal metric learning for tag-based music retrieval. *CoRR*, *abs/2010.16030*. Descargado de https://arxiv.org/abs/2010.16030

Xing, E., Jordan, M., Russell, S. J., y Ng, A. (2002). Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, *15*, 521–528.

Apéndice A

Taxonomía de géneros

	Nivel 1	Nivel 2	Nivel 3	Nivel 4	Nivel 5
0	Blues	-	-	-	-
1	Blues	Gospel	-	-	-
2	Classical	-	-	-	-
3	Classical	20th Century Classical	-	-	-
4	Classical	Chamber Music	-	-	-
5	Classical	Choral Music	-	-	-
6	Classical	Composed Music	-	-	-
7	Classical	Contemporary Classical	-	-	-
8	Classical	Opera	-	-	-
9	Classical	Symphony	-	-	-
10	Country	-	-	-	-
11	Country	Americana	-	-	-
12	Country	Bluegrass	-	-	-
13	Country	Country and Western	-	-	-
14	Country	Country and Western	Western Swing	-	-
15	Country	Rockabilly	-	-	-
16	Easy Listening	-	-	-	-
17	Easy Listening	Easy Listening: Vocal	-	-	-

Tabla A.1: Litsa de géneros parte 1

	Nivel 1	Nivel 2	Nivel 3	Nivel 4	Nivel 5
18	Easy Listening	Lounge	-	-	-
19	Easy Listening	Nu-Jazz	-	-	-
20	Electronic	-	-	-	-
21	Electronic	Ambient Electronic	-	-	-
22	Electronic	Breakcore - Hard	-	-	-
23	Electronic	Chip Music	-	-	-
24	Electronic	Chip Music	Chiptune	-	-
25	Electronic	Dance	-	-	-
26	Electronic	Downtempo	-	-	-
27	Electronic	Drum and Bass	-	-	-
28	Electronic	Dubstep	-	-	-
29	Electronic	Dubstep	Skweee	-	-
30	Electronic	Glitch	-	-	-
31	Electronic	House	-	-	-
32	Electronic	House	Chill-out	-	-
33	Electronic	IDM	-	-	-
34	Electronic	Jungle	-	-	-
35	Electronic	Minimal Electronic	-	-	-
36	Electronic	Techno	-	-	-
37	Electronic	Techno	Bigbeat	-	-
38	Electronic	Trip-Hop	-	-	-
39	Experimental	-	-	-	-
40	Experimental	Audio Collage	-	-	-
41	Experimental	Avant-Garde	-	-	-
42	Experimental	Drone	-	-	-
43	Experimental	Electroacoustic	-	-	-
44	Experimental	Field Recordings	-	-	-
45	Experimental	Improv	-	-	-
46	Experimental	Minimalism	-	-	-
47	Experimental	Musique Concrete	-	-	-
48	Experimental	Noise	-	-	-
49	Experimental	Novelty	-	-	-
50	Experimental	Novelty	Kid-Friendly	-	-

Tabla A.2: Litsa de géneros parte 2

	Nivel 1	Nivel 2	Nivel 3	Nivel 4	Nivel 5
51	Experimental	Novelty	Sound Effects	-	-
52	Experimental	Novelty	Sound Effects	Holiday	-
53	Experimental	Novelty	Sound Effects	Holiday	Christmas
54	Experimental	Sound Art	-	-	-
55	Experimental	Sound Collage	-	-	-
56	Experimental	Sound Poetry	-	-	-
57	Experimental	Unclassifiable	-	-	-
58	Folk	-	-	-	-
59	Folk	British Folk	-	-	-
60	Folk	Freak-Folk	-	-	-
61	Folk	Free-Folk	-	-	-
62	Folk	Psych-Folk	-	-	-
63	Folk	Singer-Songwriter	-	-	-
64	Hip-Hop	-	-	-	-
65	Hip-Hop	Abstract Hip-Hop	-	-	-
66	Hip-Hop	Alternative Hip-Hop	-	-	-
67	Hip-Hop	Breakbeat	-	-	-
68	Hip-Hop	Hip-Hop Beats	-	-	-
69	Hip-Hop	Nerdcore	-	-	-
70	Hip-Hop	Rap	-	-	-
71	Hip-Hop	Wonky	-	-	-
72	Instrumental	-	-	-	-
73	Instrumental	Ambient	-	-	-
74	Instrumental	New Age	-	-	-
75	Instrumental	Soundtrack	-	-	-
76	Instrumental	Soundtrack	Compilation	-	-
77	International	-	-	-	-
78	International	African	-	-	-
79	International	African	Afrobeat	-	-
80	International	African	North African	-	-
81	International	Asia-Far East	-	-	-
82	International	Balkan	-	-	-
83	International	Brazilian	-	-	-
84	International	Celtic	-	-	-
85	International	Europe	-	-	-
86	International	Europe	Fado	-	-
87	International	Europe	Klezmer	-	-

Tabla A.3: Litsa de géneros parte 3

	Nivel 1	Nivel 2	Nivel 3	Nivel 4	Nivel 5
88	International	Europe	Romany (Gypsy)	-	-
89	International	Europe	Spanish	-	-
90	International	Flamenco	-	-	-
91	International	French	-	-	-
92	International	Indian	-	-	-
93	International	Indian	Bollywood	-	-
94	International	Indian	N. Indian Traditional	-	-
95	International	Indian	South Indian Traditional	-	-
96	International	Latin	-	-	-
97	International	Latin America	-	-	-
98	International	Latin America	Cumbia	-	-
99	International	Latin America	Salsa	-	-
100	International	Latin America	Tango	-	-
101	International	Middle East	-	-	-
102	International	Middle East	Turkish	-	-
103	International	Pacific	-	-	-
104	International	Polka	-	-	-
105	International	Reggae - Dub	-	-	-
106	International	Reggae - Dub	Reggae - Dancehall	-	-
107	Jazz	-	-	-	-
108	Jazz	Be-Bop	-	-	-
109	Jazz	Big Band/Swing	-	-	-
110	Jazz	Free-Jazz	-	-	-
111	Jazz	Jazz: Out	-	-	-
112	Jazz	Jazz: Vocal	-	-	-
113	Jazz	Modern Jazz	-	-	-
114	Old-Time / Historic	-	-	-	-
115	Pop	-	-	-	-
116	Pop	Experimental Pop	-	-	-
117	Pop	Synth Pop	-	-	-
118	Rock	-	-	-	-
119	Rock	Garage	-	-	-
120	Rock	Garage	Surf	-	-
121	Rock	Goth	-	-	-

Tabla A.4: Litsa de géneros parte 4

	Nivel 1	Nivel 2	Nivel 3	Nivel 4	Nivel 5
122	Rock	Indie-Rock	-	-	-
123	Rock	Industrial	-	-	-
124	Rock	Krautrock	-	-	-
125	Rock	Lo-Fi	-	-	-
126	Rock	Loud-Rock	-	-	-
127	Rock	Loud-Rock	Noise-Rock	-	-
128	Rock	Loud-Rock	Noise-Rock	Sludge	-
129	Rock	Metal	-	-	-
130	Rock	Metal	Black-Metal	-	-
131	Rock	Metal	Death-Metal	-	-
132	Rock	Metal	Grindcore	-	-
133	Rock	New Wave	-	-	-
134	Rock	Post-Rock	-	-	-
135	Rock	Post-Rock	Space-Rock	-	-
136	Rock	Progressive	-	-	-
137	Rock	Psych-Rock	-	-	-
138	Rock	Punk	-	-	-
139	Rock	Punk	Electro-Punk	-	-
140	Rock	Punk	Hardcore	-	-
141	Rock	Punk	Hardcore	Thrash	-
142	Rock	Punk	No Wave	-	-
143	Rock	Punk	Post-Punk	-	-
144	Rock	Punk	Power-Pop	-	-
145	Rock	Rock Opera	-	-	-
146	Rock	Shoegaze	-	-	-
147	Soul-RnB	-	-	-	-
148	Soul-RnB	Disco	-	-	-
149	Soul-RnB	Funk	-	-	-
150	Soul-RnB	Funk	Deep Funk	-	-
151	Spoken	-	-	-	-
152	Spoken	Banter	-	-	-
153	Spoken	Comedy	-	-	-
154	Spoken	Musical Theater	-	-	-
155	Spoken	Poetry	-	-	-
156	Spoken	Radio	-	-	-
157	Spoken	Radio	Interview	-	-
158	Spoken	Radio	Radio Art	-	-
159	Spoken	Radio	Talk Radio	-	-
160	Spoken	Radio Theater	-	-	-
161	Spoken	Spoken Weird	-	-	-
162	Spoken	Spoken Word	-	-	-

Tabla A.5: Litsa de géneros parte 5