Universidad del Valle de Guatemala

Facultad de Ingeniería

Departamento de Ciencias de la Computación y Tecnologías de la Información

Clasificador de música en base a estados anímicos basándonos en análisis de señales y el contexto cultural de Guatemala

Trabajo de graduación presentado por

Pablo José Estrada Cordón

para optar al grado académico de Licenciado en Ingeniería en Ciencias de la Computación y Tecnologías de la Información.

GUATEMALA

2017

Clasificador de música en base a estados anímicos basándonos en análisis de señales y el contexto cultural de Guatemala

UNIVERSIDAD DEL VALLE DE GUATEMALA

Facultad de Ingeniería

Clasificador de música en base a estados anímicos basándonos en análisis de señales y el contexto cultural de Guatemala

Trabajo de graduación presentado por

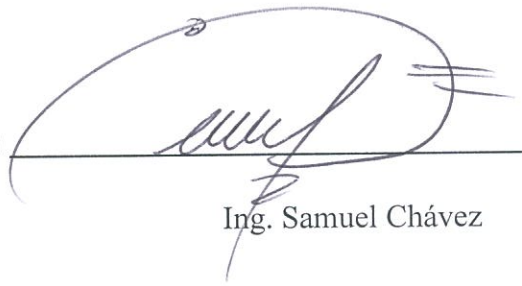
Pablo José Estrada Cordón

para optar al grado académico de Licenciado en Ingeniería en Ciencias de la Computación y Tecnologías de la Información

Guatemala,

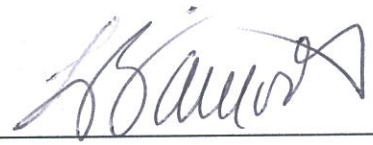
2017

Vo.Bo.:

(f) 

Ing. Samuel Chávez

Tribunal Examinador:

(f) 

Msc. Douglas Barrios

(f) 

Lic. Isabel Ciudad Real

(f) 

Ing. Samuel Chávez

Fecha de Aprobación: Guatemala, 6 de diciembre del 2017

PREFACIO

Con el desarrollo de los medios digitales y el fácil acceso a internet, las personas están expuestas a librerías de música cada vez más grandes que se vuelven más complejas de consultar. Hoy más que nunca es común ver a personas produciendo música desde la comodidad de sus casas, compartiéndola en redes sociales y colaborando con otros artistas en creación de obras musicales. La facilidad de acceso a librerías enormes de música trae nuevos retos para la creación de sistemas que faciliten a los usuarios un acceso rápido y con resultados satisfactorios. Esta necesidad creó una nueva rama de investigación que ha crecido en complejidad año con año llamada EIM (Extracción de Información Musical). Esta combina a expertos de áreas como musicología, computación, analistas de datos, antropólogos, psicólogos e incluso expertos en letras y procesamiento de señales. Año con año la investigación en EIM ha progresado mostrando nuevos retos que se relacionan con la búsqueda, comparación y clasificación de música. Los resultados de estas investigaciones han traído avances en la búsqueda para plataformas de música, procesamiento de señales de audio, herramientas de producción musical e incluso en el desarrollo de teorías psicológicas.

Además de las complejidades técnicas que trae el procesamiento de señales de audio, se pueden añadir más variables como el contexto cultural, los estados anímicos de las personas, recuerdos implantados sobre una pieza específica, el procesamiento e interpretación de la letra de una canción e incluso el lugar físico en donde la pieza está siendo escuchada. Esta gran complejidad requiere que poco a poco, se logren descifrar las características y variables más importantes para representar y comparar de manera satisfactoria y eficiente la similitud y el contenido de un conjunto de obras musicales. Este trabajo busca enfocarse en el contenido de la pieza, relacionando características encontradas en las señales de audio con las emociones percibidas en una pieza específica, dejando en un plano secundario los aspectos contextuales, culturales y psicológicos, aunque sin descuidarlos por completo.

El procesamiento de señales y extracción de características de una pieza de audio tienen una fuerte relación con las emociones con las que sus oyentes la clasifican. Por lo que se construyó un clasificador de música con ayuda de algoritmos de aprendizaje de máquina para categorizar de manera automática un conjunto de canciones dentro de cuatro emociones básicas, bajo un conjunto de datos de entrenamiento generado exclusivamente en Guatemala. Todo este trabajo aporta datos sobre las características sonoras más relevantes al momento de clasificar música automáticamente, y sus resultados pueden ser útiles en aplicaciones reales tales como la creación de sistemas de búsqueda, recomendación y ordenamiento de librerías masivas de música, relaciones entre emociones y sonidos para aplicaciones en industrias de entretenimiento como lo son videojuegos y cine, e incluso para ramas de investigación en psicología como musicoterapia.

Este trabajo introduce los conceptos relacionados con EIM y muestra los resultados obtenidos al crear varios modelos de clasificación automática con el conjunto de piezas bajo estudio. Con esto se pretende aportar al desarrollo de esta rama de estudio y explorar las relaciones que se encuentren entre los contenidos de una pieza musical y la emociones que las personas perciben en la misma. Finalmente se presenta un análisis de la eficiencia de distintos algoritmos de aprendizaje de máquina que pueden ser utilizados para este tipo de problemas y los distintos resultados que cada uno de estos proveen en la clasificación de música basada en emociones.

ÍNDICE

LISTA DE CUADROS

LISTA DE FIGURAS

RESUMEN

Los algoritmos de clasificación de música han sido cada vez más relevantes durante la última década, ya que las grandes librerías de música han creado la necesidad de una navegación eficiente a través de estos conjuntos grandes de música. Esta investigación analiza las características relevantes para la creación de un clasificador de música en base a la emoción expresada en una canción, entrenado con data generada únicamente por guatemaltecos para eliminar algún sesgo cultural. El trabajo toma 34 características con sus desviaciones estándar para la construcción de varios modelos utilizando los algoritmos de K Vecinos Más Cercanos, Máquinas de Vectores de Soporte y Bosques Aleatorios. Los resultados finales presentan el mejor modelo con un desempeño de 80% y un 73% de las canciones de prueba clasificadas correctamente, y recomiendan los Coeficientes Espectrales de Mel y los Vectores Cromáticos como una de las características más relevantes para la clasificación de música en base a emociones.

I. INTRODUCCIÓN

La necesidad de los usuarios de poder explorar de manera eficiente grandes librerías y poder adaptar sistemas de recomendación inteligentes que sepan reconocer y diferenciar entre distintos tipos de música ha creado la necesidad de explorar las características principales que pueden ayudar a la diferenciación de conjuntos de música. Este trabajo se enfoca en desarrollar un clasificador de música orientado a diferenciar la música según cuatro de las emociones básicas utilizadas en psicología. Para esto, se buscará encontrar las características extraídas de la señal de audio que sean más relevantes para clasificar la música en distintas emociones utilizando procesamiento de señales de audio y los algoritmos de K Vecinos Más Cercanos, Máquinas de Vectores de Soporte y Bosques Aleatorios. Además, se construyó un set de datos con la ayuda de participantes de nacionalidad guatemalteca, quienes clasificaron 860 canciones bajo las emociones de alegría, tristeza, enojo y miedo.

Dentro del campo de extracción de información de la música, encontramos varios modelos de clasificación, como el contenido, contexto musical, propiedades del usuario y contexto del usuario. En este trabajo se enfocaron esfuerzo única y exclusivamente en el contenido musical de la pieza, dejando fuera elementos como el contexto musical, las propiedades y características de quien escucha la pieza y el contexto en donde la pieza fue reproducida. Del mismo modo, queda fuera del alcance de este trabajo la creación de algún modelo de emociones o cualquier análisis cultural o antropológico relacionado a los resultados obtenidos de las clasificaciones. El trabajo pretende encontrar los mejores algoritmos y características de la señal de audio que puedan servir como base para diferenciar piezas en base a la emoción que expresan en base a la opinión de los oyentes guatemaltecos.

Para la creación de los modelos se obtuvieron 34 características de nivel bajo y medio extraídas de las 860 piezas. Posteriormente se construyeron modelos utilizando los tres algoritmos previamente mencionados y se realizaron modificaciones a los datos tomados en cuenta para poder analizar qué características y acercamientos eran los más efectivos para elevar el desempeño de los modelos. Al finalizar el análisis, se concluye que el modelo de Bosques Aleatorios utilizando las 34 características junto con su desviación estándar son los más efectivos para esta clasificación, dando un desempeño de 80% con un 73% de las clasificaciones realizadas correctamente

II. OBJETIVOS

A. Objetivo general

Desarrollar un algoritmo de clasificación de música con una precisión por encima de 50% utilizando técnicas de aprendizaje de máquina y procesamiento de señales de audio, basando el criterio de separación de la emoción en un conjunto de datos de entrenamiento creado únicamente por personas guatemaltecas.

B. Objetivos específicos

1. Aplicar modelos de aprendizaje supervisado para la clasificación de música con base en los estados anímicos que estas expresan.
2. Aplicar algoritmos de extracción de características de bajo nivel y nivel medio en señales de audio para poder obtener aspectos sonoros relevantes para la deducción de las emociones expresadas en una pieza musical.
3. Crear un conjunto de datos con canciones etiquetadas por emoción clasificados únicamente por personas guatemaltecas para uso en este y futuros estudios relacionados a este campo.
4. Determinar las características más relevantes para la clasificación de música con base en emociones

III. JUSTIFICACIÓN

Con la inclusión de casi toda la música existente dentro de la internet, nuevos problemas han surgido para los usuarios al tratar de navegar en librerías de música de gran tamaño. Los usuarios están buscando la música correcta para el momento correcto sin la necesidad de tener que esforzarse mucho en la búsqueda. Por esta razón ha surgido la necesidad de crear listas de reproducción automáticas en base a ciertos parámetros como género, emociones o estilos musicales. Muchísimas empresas millonarias como Last.fm, Spotify, Itunes, Google Play o Xbox Music están en búsqueda de los modelos adecuados para generar estas listas de reproducción de manera automáticas y satisfacer las necesidades de sus usuarios. Esta necesidad llevo a la creación de la rama de investigación de Extracción de Información Musical.

Aunque a simple vista la extracción de información musical puede verse con un campo limitado de aplicaciones, la necesidad del desarrollo de este tipo de tecnologías se encuentra en muchísimas áreas dentro de la industria de la música y el entretenimiento. Problemas como la detección de instrumentos, el proceso de tomar una señal de audio polifónica y separarla en señales individuales correspondientes a cada instrumento pueden ser de gran utilidad para sistemas de Karaoke o de aprendizaje de instrumentos como Yousician. Otra aplicación se ve en la transcripción de música de manera automática. Investigadores están intentando separar la señal de audio a través de las técnicas de extracción de información musical para poder obtener partituras de manera automática para una pieza. Esto enriquecería aún más las bibliotecas de música y permitiría a los estudiantes de música tener acceso a partituras de música que nunca fue transcrita en la historia. Otros problemas como la segmentación estructural automática de música, podría permitir identificar de manera automática los versos, coros, puentes, motivos y frases de una canción, enriqueciendo aún más la metadata de una pieza y permitiendo crear los bloques básicos para la creación de música de manera automática.

Este trabajo pretender explorar las principales características de una señal de audio. De esta se podrá proveer al lector de una guía inicial para construir modelos computacionales adecuados que permitan diferenciar de manera automática las emociones en conjuntos de música que son de gran tamaño.

IV. MARCO TEÓRICO

A. Música y emociones dentro del marco de la clasificación automática

La música es una forma de expresión artística que tiene una relación directa con las emociones humanas. La tonalidad, armonías y letras de una pieza pueden determinar en gran parte la emoción que un oyente percibe en esta. Sin embargo, existen muchísimas más variables que deben de considerarse para realmente determinar qué elementos hacen que una persona etiquete una pieza con una emoción. Aspectos como recuerdos, cultura e incluso el contexto físico e histórico en el que la persona se encuentra pueden influir en la decisión de una persona con respecto a la etiqueta emocional que le coloca a una pieza de música determinada.

Es importante primero diferenciar que los estudios apuntan a que dentro de la música se pueden encontrar dos tipos de emociones, intrínsecas y extrínsecas. Las emociones intrínsecas en una pieza de música hacen referencia a todos los elementos relacionados a los contenidos auditivos de la pieza (instrumentos, voces, efectos de sonido, etc.). Las emociones extrínsecas en cambio se refieren a todos los elementos fuera del contenido de la música, como recuerdo implantados, el lugar de escucha o características culturales. Además, existen ciertas teorías como la esperanza musical, que se refiere a la costumbre sobre la escucha de ciertos instrumentos. En ciertos casos, las personas tienden a cambiar sus opiniones sobre las emociones cuando los instrumentos utilizados en la pieza no son conocidos para ellas. A esto le podemos añadir la familiaridad que pudiese existir con la pieza, los recuerdos implantados previamente e incluso las asociaciones con la letra a ciertas experiencias previas. Estas variables usualmente determinan lo que se considera la parte extrínseca de la emoción de una canción y usualmente tienden a determinarse con estudios de carácter más cualitativo y orientado a ramas como la sociología, antropología y psicología. (Lee, 2012)

Como se puede observar, existen muchas variables que influyen en la decisión de una persona sobre la emoción que percibe dentro de una pieza musical. Por esta razón es importante estudiar previamente los conjuntos de canciones a analizar y tomar en cuenta el contexto social, cultural

para controlar la mayor cantidad de variables posibles. Toda esta complejidad nos da una idea del gran campo de trabajo que existe para las distintas ramas de la ciencia en el estudio de la clasificación, comprensión y búsqueda automática de música

1. Cultura y percepción de la música. Un elemento de gran importancia para el reconocimiento de emociones en la música es el contexto cultural del cual cada individuo proviene. Se sabe que, históricamente, distintas civilizaciones han empleado diferentes sistemas tonales, sistemas armónicas y conjuntos de instrumentos para la creación de sus obras musicales. Toda esta herencia cultural ha llevado a que las distintas culturas del mundo desarrollen diferentes colores y características estructurales en su música. A pesar de esta gran diferencia en el desarrollo musical, las emociones parecen no haberse desviado tanto. Varios estudios indican que en general, las emociones básicas (alegría, tristeza, miedo, enojo, disgusto, sorpresa) tienden a ser interpretadas de igual forma a través de las distintas culturas. A pesar de esta homogeneidad en las emociones, la interpretación del contenido acústico de una pieza puede presentar diferencias sutiles a considerar al momento de crear clasificadores automáticos de música. (Balkwill, 1999) (Laukka, 2013)

Con el desarrollo de distintas investigaciones se ha logrado determinar que las clasificaciones de música tienden a converger más fácilmente cuando se delimita culturalmente a los individuos. Esto se debe principalmente a que la familiaridad con instrumentos y la cercanía en la interpretación de los distintos sistemas tonales hacen que la interpretación del contenido acústico de una pieza sea mucho más similar, y por tanto la forma en que se designa una emoción a la pieza música tenga bastante parecido entre los distintos individuos dentro del mismo contexto cultural. Por otro lado, los aspectos académicos, como conocimientos sobre teoría musical, armonía, contrapunto o la interpretación de uno o más instrumentos musicales hace que los análisis para emitir una opinión sean más profundos y los relacionen las experiencias e interpretaciones musicales previas de la persona. Por ejemplo, un intérprete violinista podrá detectar muchas más sutilezas en la intensidad y las melodías de obras para violín que una persona común, esto puede provocar cambios en la emoción intrínseca de la pieza para cada individuo, por lo que es otro aspecto que se debe tomar en cuenta. Además de lo anterior se ha encontrado que usualmente las personas con entrenamiento musical tienden a tener clasificaciones mucho más homogéneas a través de las culturas, principalmente debido a la alta exposición a distintos géneros, instrumentos y corrientes musicales en general. (Argstatterm, 2012)

Los efectos que tienen las diferencias culturales en la clasificación de música por emociones, ha provocado que los investigadores de la rama de extracción de información de la música empiecen a reclutar investigadores de cada una de las distintas culturas alrededor del mundo para trabajar en este tipo de proyectos. Esto se debe principalmente a que las personas que están sumergidas dentro de la cultura podrán describir mejor las sutilezas y ajustar detalles en la clasificación de música para obtener resultados mucho más satisfactorios. Estos proyectos proponen que de alguna manera (geolocalización, perfiles de usuarios) se pueda segmentar al individuo dentro de una cultura o área geográfica específica para utilizar el clasificador que tenga los ajustes pertinentes a dicha cultura. Este trabajo puede ser bastante extenso, pero seguramente será de los que traerá resultados más prometedores para sistemas de clasificación de música globales, ya que

se podrá tomar en cuenta los rasgos culturales específicos de cada persona y ajustar los resultados de búsqueda y clasificación automática a estos parámetros previamente dados.

2. Emoción inducida vs emoción percibida en la música. La música es consumida por sus oyentes principalmente por las funciones emocionales que tienen en quién la escucha. Con esto nos referimos tanto a las emociones que puede inducir en el oyente, como a las emociones que expresa la pieza por el contenido en sí misma. Varios estudios a lo largo de la historia han mostrado que la capacidad para identificar emociones en una pieza musical se puede desarrollar desde una edad muy temprana y además puede ir transformándose al agregar las experiencias de vida, educación y procesos de enculturación y socialización. (Bella *et al*, 2001). Por otro lado, la música se consume en distintos contextos y con diferentes propósitos, estas diferencias en el contexto y el propósito con el cual se escucha puede inducir diferentes emociones. En general, se han encontrado algunas relaciones entre la emoción inducida y la percibida, pero y aunque en algunos casos estas dos emociones pueden ser las mismas, hay muchos casos en donde son distintas e incluso opuestas.

Podemos definir las emociones inducidas como las emociones "sentidas" o aquellas que son experimentadas por el oyente durante la reproducción. Estas emociones son internas del individuo y en general no presentan resultados homogéneos bajo una misma pieza musical. Esta variación en resultados se debe principalmente a que las emociones inducidas se ven en gran parte afectadas por las experiencias previas, los recuerdos asociados a la pieza, la letra o los elementos contextuales que rodean al individuo en el momento en el que está reproduciendo la pieza musical. (Vuoskoski, 2012)

La emoción percibida, en cambio, se refiere a la emoción que la canción expresa en base a su contenido acústico y las interpretaciones psicoacústicas del individuo. La emoción percibida en una pieza usualmente se convierte en un proceso más cognitivo, por lo que el hecho que una persona clasifique una pieza como triste, no necesariamente implica que esta persona está sintiendo tristeza durante el tiempo en el que la pieza está reproduciéndose. Con esto en mente, la emoción percibida se caracteriza entonces, por el contenido de la pieza y es el tipo de emoción de interés dentro de este trabajo ya que a través del procesamiento y análisis de señales será posible obtener las características más relevantes en el audio que llevan a la persona a elegir cierta emoción en una pieza musical determinada. (Vuoskoski, 2012)

La existencia de las emoción inducidas y percibidas dentro de la clasificación de música por emociones ha llevado a que investigadores descubran varias relaciones entre la emoción percibida y la emoción inducida:

- Relación positiva: esta se crea cuando la emoción que se induce en el oyente concuerda con la que la pieza expresa en términos musicales. (Halpern, 2016)
- Relación negativa: la relación se presenta cuando el oyente reacciona de manera opuesta a la emoción que la música expresa. Por ejemplo, sintiendo tristeza en una canción que evidentemente expresa alegría. (Halpern, 2016)
- Sin relación sistemática: ocurre cuando el oyente se mantiene neutral durante la pieza sin importar la emoción que está expresando la pieza. (Halpern, 2016)

- Sin relación: ocurre cuando la persona siente una emoción que no puede ser de ninguna manera expresada por la pieza. (Halpern, 2016)

Esta divergencia entre los dos tipos de emociones ha creado que los investigadores de estas ramas se dividan en dos corrientes: los cognitivistas musicales y emotivistas musicales. Los musicólogos de la corriente cognitivista aseguran que la descripción de una canción se basa únicamente en su emoción expresada, mientras que los emotivistas se basan únicamente en la emoción que siente el individuo con la pieza, es decir, la emoción inducida. A pesar de esta divergencia, la línea divisoria entre las emociones inducidas y percibidas aún es difusa y la relación entre estos tipos de emociones depende mucho de los mecanismos a través de los cuales una pieza particular pueda inducir una respuesta emocional en el individuo. Finalmente, también existen discusiones sobre cuáles son los posibles modelos de emociones sobre los cuales se pueden relacionar las emociones percibidas e inducidas para el área de clasificación de música. Para comprender estas divergencias, es necesario comprender sobre algunos modelos de clasificación de emociones que han sido utilizados en estudios de clasificación de música, y como estos modelos pueden servir en distintos tipos de investigaciones destinadas a encontrar esa relación entre las emociones inducidas, emociones percibidas y piezas musicales.

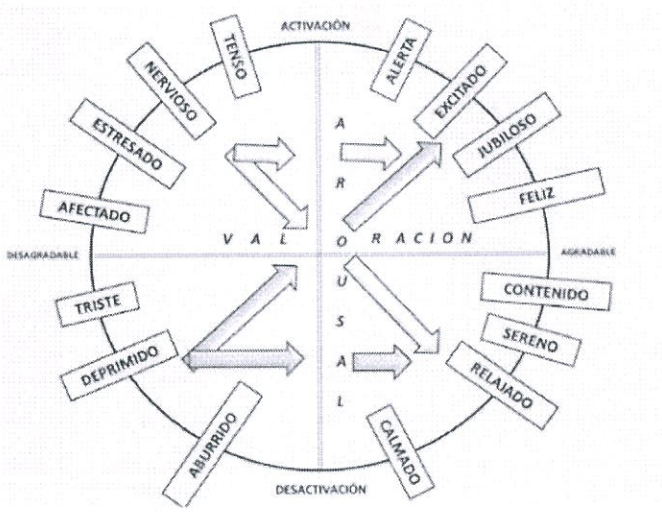
3. Modelos de clasificación de emociones. Dentro de las distintas investigaciones y estudios que se han realizado con respecto a la clasificación de emociones y estados de ánimo se han llegado a varios modelos que se detallarán a continuación y que tienen aplicaciones en distintas situaciones relacionadas con el análisis de la relación entre las emociones y la música. Estos modelos pueden ser del tipo categórico, dimensional y evaluativo. El modelo categórico clasifica las emociones y/o estados de ánimo con etiquetas discretas que son diferentes entre ellas. Los modelos dimensionales se basan en la noción de que las emociones se describen en términos de dimensiones afectivas, lo cual convierte una emoción en un punto en un plano multidimensional. La teoría no limita el número de dimensiones, pero usualmente los modelos van desde una hasta cuatro dimensiones afectivas. El modelo evaluativo es el más complejo de todos y busca imitar la evaluación cognitiva para poder distinguir de manera cualitativa entre distintas emociones. El modelo se compone de cinco componentes cognitivo, motivacional, motor, expresivo, subjetivo y eferente. (Vaiva, 2016)

a.Modelo circumplejo del afecto de Russel: El modelo circumplejo del afecto se basa en la teoría desarrollada por James Russel que expresa que las experiencias conscientes de los seres humanos pueden verse reflejadas en una mezcla de dos dimensiones centrales, el placer (agradable vs desagradable) y la intensidad con la que se sienten las emociones (activación vs desactivación).

La propuesta de Russel afirma que las emociones están mejor representadas a través de un círculo de dimensiones bipolares, en lugar de utilizar dimensiones independientes para cada emoción. Esto debido a que los grados de intensidad o de la magnitud de los estados afectivos varían con respecto a las situaciones particulares de cada individuo. Interpretando este modelo (Figura 3), se puede notar que todas aquellas estrategias, situaciones o vivencias que impulsen el vector de valoración a la derecha, izquierda arriba o abajo serán actividades mentales, comportamientos o movimientos físicos que regulen los estados afectivos. Uno

de estos es claramente la música, sin embargo, es importante entender las relaciones entre las emociones y como este modelo puede ir moviéndose a través de los distintos vectores posibles. Por ejemplo, una persona tensa o molesta debe reducir su activación y volver crear sus valoraciones de una manera más positiva. Las personas tristes o deprimidas deberán activar sus emociones, pero al mismo tiempo deberán crear una nueva valoración positiva de su situación o vivencia. Por otro lado, si la persona deprimida quisiera sentirse relajada, la activación no sería tan necesaria, sino que la importancia estaría en cambiar la valoración de los pensamientos depresivos que existen dentro de la persona. (González, 2012)

Figura 1. Representación gráfica del Modelo Circumplejo Afectivo con el esquema funcional de la regulación de las emociones

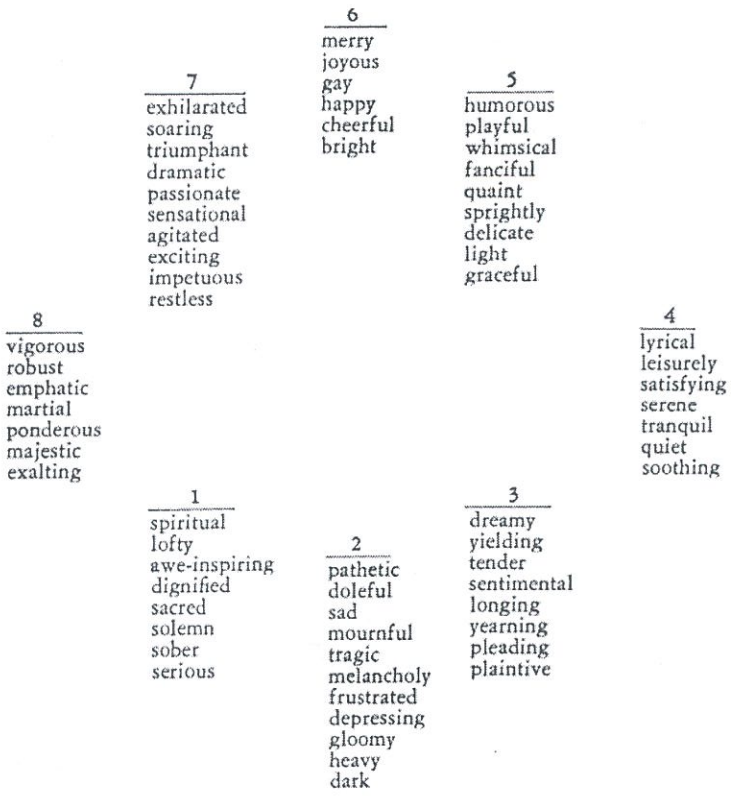


b. Modelo de Hevner. Hevner es uno de los primeros autores en publica literatura analizando la relación entre las emociones expresadas a través de la música. En una de sus publicaciones más populares, desarrolla el modelo de adjetivos para la descripción de los elementos expresado a través de la música. El estudio es desarrollado con un conjunto de 66 adjetivos, los cuales son presentados a un número de sujetos experimentales junto con un conjunto de elementos musicales para que sean clasificados de la manera más simple y objetiva posible. Luego de la tabulación de los datos, Hevner obtiene ocho grupos principales de adjetivos que describen música con las mismas características sonoras. Estos grupos han sido utilizados de base en muchos estudios de clasificación de música automática por la gran variedad de adjetivos que pueden usarse para describir un mismo grupo y las palabras tan comunes que resultan fácil de clasificar para personas con pocos conocimientos en psicología.

El experimento fue desarrollado con las siguientes composiciones: Debussy, Reflections of Water; Mendelssohn Midsummer Night´s Dream, Scherzo; Paganini, Etude on Eb Major; Sinfonía de Tschaikowsky No.6 en Si Menor; Wagner Lobengrin, Preludio al acto III. Todas estas composiciones fueron dadas a 52 individuos quienes marcaron los adjetivos que consideraban adecuados para la pieza que estuvieran escuchando. Los resultados del estudio demostraron bastante consistencia en las clasificaciones de los

individuos por lo que el estudio sentó bases para continuar el análisis de las emociones expresadas por la música y la percepción de las mismas en distintos individuos.

Figura 2. Modelo de adjetivos clasificados en los ocho grupos encontrados por Hevner.



c.Modelo de las emociones básicas de Paul Ekman. El Dr. Paul Ekman ha sido uno de los pioneros en la investigación de las emociones y las micro expresiones faciales, desarrollando uno de los modelos básicos de emociones más aceptado y utilizado en la comunidad de investigación de psicología. Ekman afirma que las emociones son innatas y están fuertemente atadas a la evolución biológica del ser humano. Además, estableció varios criterios para poder determinar una emoción como básica, estos incluyen las señales universales distintivas en la cara, un conjunto de características fisiológicas para cada emoción básica, eventos específicos que activan ciertas emociones, corta duración, así como experiencias, recuerdo e imágenes mentales subjetivas distintivas para cada emoción básica. Ekman también encontró que las emociones básicas son derivadas de elementos biológicos y por tanto deben ser iguales a través de distintas culturas y sociedades.

A continuación, se describen a detalle cada una de las características que definen a una emoción básica:

1) Señales universales distintivas: estas características se refieren a todas aquellas expresiones faciales que son características durante la experimentación de una emoción básica. Esta es la única

característica que se puede encontrar en todas las emociones básicas y es considerada la piedra angular para la identificación de emociones. (Ekman, 1999)

2) *Fisiología específica de la emoción:* cada emoción básica lleva consigo ciertas reacciones específicas en el sistema nervioso autónomo. Los estudios indican que estas características fisiológicas no varían tanto entre culturas, por lo que se puede decir que las emociones básicas son respaldadas por elementos biológicos que se han adquirido de manera evolutiva. (Ekman, 1999)
3) Mecanismos automáticos de estimación: este concepto se refiere a los mecanismos internos del ser humano que analizan un estímulo y de alguna manera lo evalúan y clasifican en una de las emociones básicas. Las emociones básicas cuentan con un mecanismo de estimación mucho más rápido que los mecanismos de estimación secundarios, que van más ligados a la cognición y razonamiento de situaciones. Es por esta razón que los mecanismos de estimación de emociones básicas son respuestas biológicas automáticas que ocurren de manera inconsciente en el individuo. (Ekman, 1999)
4) Eventos antecedentes universales: Otra característica importante en la definición de una emoción básica es que puedan relacionarse eventos que provoquen la emoción. Tras varios estudios se ha descubierto que existen cierta tendencia a que eventos específicos provoquen emociones básicas específicas. Por ejemplo, una situación donde existen potencial daño físico o psicológico es usualmente relacionado a la emoción del miedo. A pesar de la existencia estos eventos universales, es aún difuso el origen de estos como fue el proceso que origino la relación con el evento y las reacciones fisiológicas y psicológicas de la emoción básica. (Ekman, 1999)
5) Además de las antes mencionadas podemos encontrar otras características que se pueden encontrar en algunas de las emociones básicas como:
    a) Presencia en otros primates.
    b) Duración rápida.
    c) Pensamientos memorias e imágenes distintivas.
    d) Experiencias subjetivas distintivas. (Ekman, 1999)

En los primeros resultados de las investigaciones de Ekman, se determinaron 6 emociones principales basadas específicamente en las características mencionadas anteriormente. Estas son: alegría, tristeza, miedo, sorpresa, disgusto y enojo. Este modelo inicial es uno de los modelos más utilizados en el estudio de las emociones y ha sido adaptado en partes para estudios de análisis de emociones en música.

B. Extracción de información de la música

La rama de investigación de extracción de información de la música es bastante reciente, y poco a poco se ha expandido a nuevos tipos de problemas que involucran conocimiento de varias ramas de estudio que antes no tenían una relación tan estrecha. La tarea principal de esta rama de investigación es desarrollar métodos para la extracción de descriptores con significados musicalmente relevantes ya sea a través de la señal de audio directamente o a través las fuentes de información contextuales que rodean a la pieza musical.

Todo este trabajo tiene el objetivo de facilitar la búsqueda, clasificación, análisis y exploración de librerías musicales de gran tamaño, para satisfacer las necesidades del usuario final de una manera mucho más rápida.

Dentro del dominio de la extracción de información de la música, podemos encontrar 3 paradigmas básicos:

- Obtención: Este caso se presenta cuando el usuario tiene una necesidad de información musical específica, es decir, encontrar una pieza musical específica. El usuario activamente expresa esta necesidad a través de una consulta. Esta consulta se puede representar ya sea como un texto, como un conjunto de notas musicales o como un fragmento de audio. El resultado pueden ser una o varias canciones, partituras o meta-data de las piezas encontradas.
- Exploración: El usuario posee una necesidad de información sin una dirección específica, y solo desea explorar la colección de música a su disposición. Este paradigma se centra en la construcción de interfaces de usuario efectivas y fáciles de usar.
- Recomendación: Este paradigma busca filtrar la colección de música para mostrar únicamente piezas que sean del potencial interés del usuario. Este interés se puede deducir de manera implícita observando las acciones del usuario como sus listas de reproducción, piezas favoritas y patrones de búsqueda, así como de manera explícita a través de la exploración de categorías existentes en la colección de música.

El paradigma principal de la clasificación de música por emociones es el de recomendación ya que es usual que en sistemas de reproducción de música de gran tamaño (Spotify, Deezer, SoundCloud) se presente la posibilidad de generar listas de reproducción automáticas en base a las preferencias y parámetros de filtrado obtenidos del usuario. Algunos de estos parámetros pueden incluir las emociones en el audio, por lo que es importante encontrar los métodos más efectivos para poder clasificar música en base a emociones para mejorar los sistemas de recomendación de música automáticos.

1. Similitud entre piezas musicales. Para poder comprender correctamente el proceso de clasificación de música, primero se debe conocer de manera detallada cómo se mide la similitud entre dos piezas. El concepto de similitud entre música es de los más estudiados dentro de la rama de extracción de información musical y resulta ser bastante más complejo de lo que parece en primera instancia. Por ejemplo, al momento de medir similitud para identificar géneros similares o identificar "covers" automáticamente, el análisis de la melodía y el ritmo de las canciones puede ser una muy buena decisión. En otro contexto como la recomendación de música a un usuario específico, esta podría ser una mala decisión, y en cambio la decisión correcta podría ser el análisis de sus patrones de búsqueda y exploración de música. Con esto se puede observar que la similitud de piezas puede no depender únicamente de la señal de audio como tal, sino que existen elementos contextuales que afectan la percepción de la similitud entre dos piezas musicales para un usuario. (Wiggins, 2009)

Varios investigadores en la rama de psicología se han enfocado en encontrar cual es la naturaleza de la percepción de similitud entre dos piezas musicales. Los resultados han dado tres principales formas de codificar la música para poder determinar su similitud:

a. Codificación icónica: se refiere a las diferencias formales entre las señales de audio. Incluye características como el tempo, intensidad, volumen, frecuencias, etc. (Juslin, 2013)
b. Codificación intrínseca: abarca todas las relaciones sintácticas y de estructura dentro de la misma pieza, como motivos, frases y subfrases. Las ramas de estudio musicales como el análisis de la forma nos pueden dar características importantes dentro de este tipo de codificación. (Juslin, 2013)
c. Codificación asociativa: se enfoca en las relaciones entre la música y otros elementos arbitrarios fuera de la música. Por ejemplo, ¿la música es de carácter religioso? ¿se utiliza en alguna celebración o ritual específico? Todas estas preguntas influyen de alguna manera en como una pieza es similar a otra. (Juslin, 2013)

Adicionalmente, se han discutido las categorías más relevantes en cuanto a la información que se puede extraer al momento de determinar la similitud entre dos piezas. Todas estas características proporcionan factores que pueden ser medidos de manera automática y que colaboran en el proceso de medir la similitud entre dos piezas.

a. Contenido musical: Se refiere a todos aquellos elementos y características contenidos en la señal de audio. Cualesquiera de los aspectos que permitan medir características musicales como el timbre, el ritmo, la melodía y la armonía.
b. Contexto musical: Todos aquellos aspectos que me pueden inferirse directamente de la pieza musical, sino que son extraídos meta data social, cultural o histórica relacionada a la pieza. Incluye aspecto como el artista, género, cultura, período histórico, compañía disquera, etc.
c. Propiedades del usuario: Engloba todos los elementos de personalidad del usuario, como sus gustos musicales, conocimientos musicales y experiencia con cierto tipo de música. Esta categoría usualmente trae objetos de codificación asociativa e intrínseca.
d. Contexto del usuario: se refiere a todos los elementos y factores que rodean al usuario al momento de interactuar con la pieza musical. Incluye aspectos como la ubicación, época, contexto social y cultural y estados de ánimo propios del usuario.

C. Percepción de la intensidad y la frecuencia:

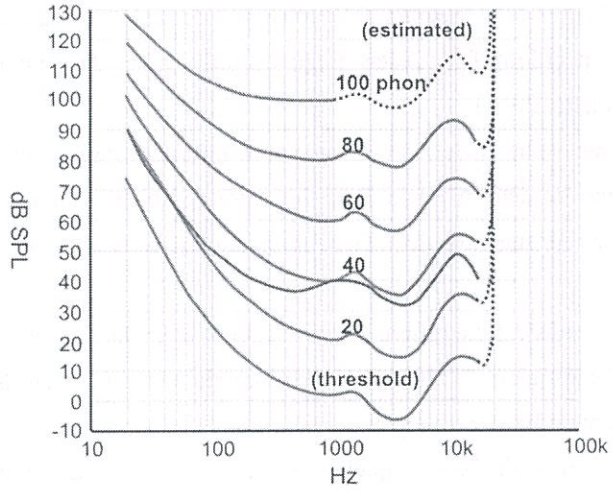
Para poder empezar a construir características relevantes para algoritmos de aprendizaje de máquina de la señal de audio, es importante conocer cuál es la relación real entre las escalas de frecuencia y la percepción del oído humano. Esto ayudará a determinar los rangos y delimitar la data en intensidad y frecuencia a valores que sean relevantes a la percepción humana, para así poder construir características más cercanas a los conceptos musicales que se conocen hoy en día y no quedarnos en características de bajo nivel, más cercanas a conceptos de electrónica y matemática.

1. Percepción de la intensidad y fuerza del sonido. Existen bastantes discrepancias entre las propiedades físicas del sonido y la manera en cómo el ser humano lo percibe. La intensidad del sonido, a un

nivel físico, es descrita como la presión del sonido sobre el medio que la rodea y usualmente se describe en decibeles (dB). Esta es una escala logarítmica con base 10, por lo que un incremento de 10 dB corresponde a un incremento de 10 en la presión del sonido.

$$i_{db} = 10 * \log_{10}(\frac{i}{i_0})$$

Figura 3: Curva de intensidad del sonido



(Knees, 2016)

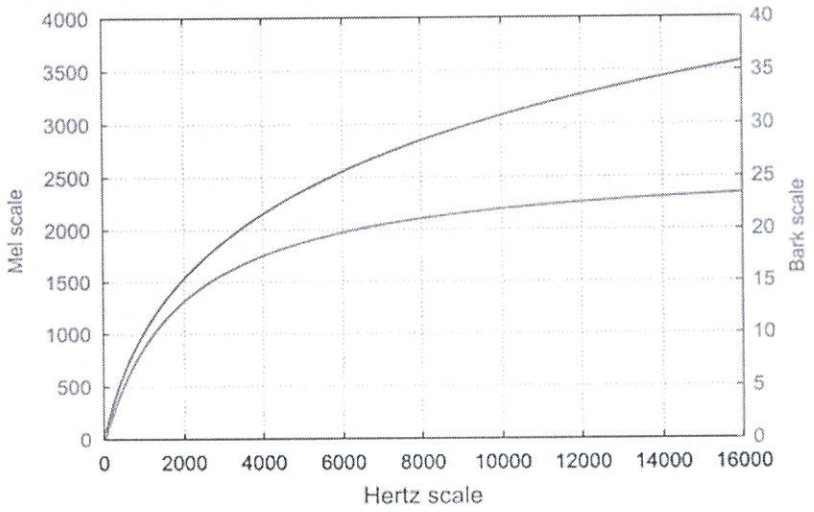
Aunque esta escala es utilizada en muchos casos para el análisis de intensidad del sonido, la percepción de intensidad del sonido varía dependiendo de la sensibilidad del oído, la cual también varía entre las distintas personas. El oído humano también percibe las frecuencias a distintas intensidades siendo las más sensibles entre 2000 y 5000 Hz. Para llegar a determinar esta relación entre la percepción humana y la intensidad física, se han llevado a cabo varios experimentos con los cuales se ha llegado a crear una relación descrita por la curva de la Figura 1. Cada curva relaciona la presión del sonido en dB con la percepción de intensidad humana a distintas frecuencias. Para poder medir la percepción del sonido en el ser humano se creó la unidad de medida llamada Phon. Un Phon se define como el mismo valor en dB a una frecuencia de 1000 Hz. Por lo que la percepción de una señal de 1000 Hz en su estado más puro es igual a un Phon. (Knees, 2016)

2. Percepción de la Frecuencia: En el estudio de la percepción humana de la frecuencia, se ha llegado a determinar que todas las frecuencias mayores a 20,000 Hz y menores a 20 Hz no deberían de tomarse en cuenta para el análisis piezas musicales ya que son prácticamente imperceptibles por los seres humanos. Además de la percepción en intensidad, también existe la percepción de tono, es decir, sonidos agudos o graves. Esta percepción de tonos no tiene una relación lineal con la frecuencia, por lo que también se han llevado a cabo experimentos para determinar una escala que modele esta relación entre frecuencia y percepción de tonalidad.

Las investigaciones llevaron a la creación de las escalas como las de Mel y de Bar. De estas haremos énfasis en la escala de Mel, ya que tiene aplicaciones directas en la creación de clasificadores de música. Esta escala busca definir los rangos de frecuencias que son percibidos como iguales a nivel de tonos por una persona. Esta escala permitió la creación de fórmulas, como la que se presenta a continuación, para describir la relación entre Hz y la escala de Mel. Sin embargo, debido a que las escalas fueron construida a través de experimentos de audición, estas fórmulas son sólo aproximaciones de la escala real. (Knees, 2016)

$$m = 1127 * \log(1 + \frac{f}{700})$$

Figura 4: Escalas de Mel y de Bark



Si observamos la Figura 2, la relación entre Hertz y Mel es casi lineal cuando la frecuencia está por debajo de los 500 Hz. De los 500 Hz en adelante, los intervalos de frecuencias que se perciben como iguales son cada vez más grandes. Otro aspecto importante en el estudio de la percepción del sonido es lo que se conoce como máscaras espectrales. Este fenómeno ocurre cuando se presentan dos sonidos en el cual uno es mucho más fuerte que el otro y por tanto no llega a ser perceptible por la persona. Compresores de audio como el MP3 se aprovechan de este tipo de fenómenos para reducir el tamaño de los archivos, aunque no siempre puede ser útil remover estas frecuencias, en muchas de las aplicaciones reales de la música se tienden a eliminar. (Knees, 2016)

D. Características musicales y procesamiento de señales de audio:

Una de las tareas más importantes de la extracción de información de la música es extraer características relevantes en la señal de audio que nos puedan permitir discernir entre dos o más piezas. Se considera una característica de la música cualquier valor numérico de alguna pieza musical bajo estudio que resulta ser un buen descriptor de algún aspecto en específico del sonido, ya sea ritmo, melodías, armonías, instrumentos o incluso artistas o grupos que la interpretan. Dentro de la extracción de características relevantes del sonido

podemos encontrar varios tipos de categorizaciones, ya sea por niveles de abstracción, el aspecto musical que describen, el dominio de la señal o su ámbito temporal.

En cuanto a niveles de abstracción, podemos encontrar que las características usualmente se describen en tres niveles. Las características de alto nivel que se refieren a conceptos musicales que son comprendidos por las personas comunes como instrumentos, acordes, melodías, ritmo, tempo, letras o género. Un nivel más abajo de estas encontramos las características de nivel medio, las cuales nos dan ideas sobre ciertas características de alto nivel, y al mismo tiempo son mucho más fáciles de comprender por computadoras. Entre estas entran los patrones de fluctuación de frecuencias, coeficientes espectrales de Mel, detección de impulsos y descriptores del beat y altura. Finalmente, encontramos las características de bajo nivel las cuales se refieren a descriptores de la señal de audio cercanas a propiedades físicas del sonido. Entre los descriptores de bajo nivel encontramos la amplitud, energía, centroides espectral, flujo espectral y la tasa de cruce de ceros.

Si nos enfocamos en la categorización por el ámbito temporal, encontramos tres tipos de categorías, características instantáneas, segmentadas y globales. Las características instantáneas van en tiempos de 10 ms aproximadamente, este es el tiempo elegido debido a que es la resolución de tiempo del oído humano. Posteriormente, las características segmentadas se construyen a través de ventanas de tiempo fijas a través de toda la pieza o bien con fragmentos definidos de una manera más semántica como a través de versos, coros, frases o motivos musicales. Finalmente, las características globales se enfocan en la pieza en su totalidad, es decir una canción completa, un movimiento completo o un concierto completo. (Knees, 2016)

La categorización por aspecto musical se refiere simplemente al concepto musical de alto nivel que describen. Aspecto como el tempo, ritmo, timbre, melodía, contrapunto, armonías, acordes y dinámicas son parte de esta categorización. Finalmente, la categorización de dominio de la señal básicamente se divide en el dominio de la frecuencia y el dominio del tiempo. Cuando una señal se encuentra en el dominio del tiempo se representa como un conjunto de puntos indicando la amplitud de la señal a través del tiempo. Para el dominio de la frecuencia, representamos la señal como magnitudes a distintas frecuencias en todo el espectro de frecuencias disponible. El dominio de la frecuencia ha sido de gran utilidad para la creación de clasificadores de música, y no sería posible realizar todos estos análisis sin la existencia de la Transformada de Fourier. (Knees, 2016)

E. Transformada discreta de Fourier

La transformada discreta de Fourier es un tipo de transformación discreta que toma una función en el dominio de frecuencia como parámetro y la transforma a una nueva función en el dominio del tiempo. Al utilizar esta transformada estamos asumiendo que se analiza una señal periódica que se extiende de manera finita (Ali, 2010). La entrada de la transformada discreta de Fourier es una secuencia finita de números reales o complejos.

A continuación se presenta la fórmula a utilizar para la transformación de una secuencia de N números complejos.

$$X_k \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}kn} \; k = 0, \ldots, N-1$$

Algunas propiedades de la transformada de Fourier son bastante útiles al momento de computar la transformación en una máquina, ya que permiten realizar optimizaciones a los algoritmos que la calculan y reducir el tiempo de cálculo. Entre las propiedades de la transformada discreta de Fourier encontramos:

Linealidad: se puede demostrar que la transformada de Fourier es lineal. Esto significa que cualquier combinación lineal de las señales, entonces en el dominio de la frecuencia estas dos señales transformadas corresponderán a la combinación lineal de las señales iniciales (Serra, 2016).

$$DFT(ax_1[n] + bx_2[n]) = \sum_{n=0}^{N-1} (ax_1[n] + bx_2[n]) e^{-i2\pi kn/N}$$
$$= a\, X_1[n] + bX_2[n]$$

Traslación: la propiedad de traslación de la transformada de Fourier nos dice que si trasladamos la señal de entrada una cantidad n_0 de muestras corresponde a la transformada de Fourier de la señal original multiplicada por un exponencial complejo (Serra, 2016).

$$DFT(x[n - n_0]) = \sum_{m=0}^{N-1} x[m] e^{-i2\pi km/N}$$
$$= e^{-i2\pi kn_0/N} X[k]$$

Simetría: las propiedades de simetría de la transformada de Fourier afirman que, si tenemos una señal real y tomamos la transformada discreta de Fourier, la parte real tendrá una simetría par mientras que la parte compleja tendrá una simetría impar. De manera similar, si la señal de entrar es real y compleja entonces la parte real de la transformada de Fourier de esta señal será par y la parte compleja será cero (Serra, 2016).

$$x[n]\, real \leftrightarrow \Re\{X[k]\}\, par \wedge \Im\{X[k]\}\, impar$$

$$x[n]\, real\, y\, par \leftrightarrow \Re\{X[k]\}\, par \wedge \Im\{X[k]\} = 0$$

Convolución: esta propiedad nos indica que, si aplicamos la convolución dos señales, en el dominio espectral, corresponderá a la multiplicación de las transformadas discretas de Fourier de estas dos señales (Serra, 2016).

$$DFT(x_1 * x_2) = X_1[k]\, x\, X_2[k]$$

F. Pre procesamiento para la extracción de características de la música

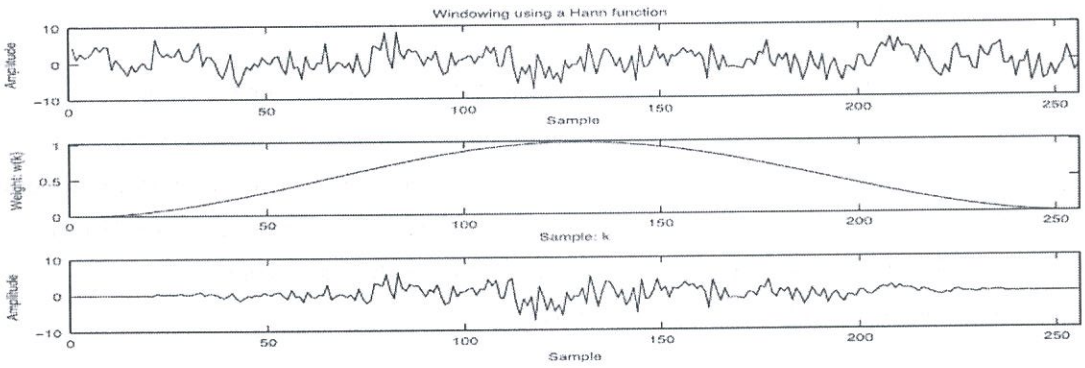
Con el objetivo de eliminar ruido y mejorar los resultados de los algoritmos de extracción de datos de la señal de audio, es necesario realizar algunos procedimientos de limpieza y fragmentación de la señal de audio. En primera instancia, se debe hacer un muestreo de la señal analógica para poder convertirla a un formato digital que pueda utilizarse dentro de algún lenguaje de programación. Para esto se debe tomar en cuenta el teorema de Nyquist, el cual nos indica que las señales deben muestrearse al menos al doble de la frecuencia para poder obtener un resultado de una calidad aceptable. Posteriormente debe tomarse en cuenta la cuantificación de la señal analógica y definir la cantidad de bits a utilizar para representar los valores de amplitud leídos de la señal analógica.

1. Ventanas y marcos de datos. Una vez tenemos una señal correctamente muestreada, se tendrán fracciones de datos que corresponden a la amplitud durante 0.0227ms si se utiliza una frecuencia de muestreo de 44,100 Hz (una de las más utilizadas). Sin embargo, un solo dato no será suficiente para poder computar características relevantes al oído humano, lo que implica que se deben agrupar varios datos para crear fragmentos más largos llamados marcos. Los tamaños de marcos más comunes van desde 256 hasta 8192 muestras. Es importante que, al momento de elegir el tamaño del marco, se tome en cuenta la frecuencia de muestreo ya que distintas frecuencias darán distintos tamaños de marcos en tiempo a pesar de tener las mismas frecuencias y el marco no deberá ser menor a 10ms, ya que esté es el mínimo tiempo de percepción del oído humano.

Con un conjunto de marcos creados, ya es posible empezar a computar algunas características en el dominio del tiempo. Pero se pueden lograr aún mejores resultados si a estos marcos les aplicamos una ventana antes de trasladarlos al dominio del tiempo con la transformada de Fourier. Una ventana es una función que se aplica a la señal de audio con el objetivo de limpiar la señal y evitar que aparezcan artefactos inexistentes en la señal de audio tras la transformada de Fourier (fenómeno conocido como drenaje espectral). Una de las ventanas más utilizadas en el procesamiento de audio es la ventana Hann:

$$w(k) = 0.5 * (1 - \cos\left(\frac{2 * \pi * k}{K - 1}\right))$$

Figura 5: Aplicación de la ventana Hann a una señal de audio.



En la Figura 3 se puede observar el efecto que tiene en una señal la aplicación de la ventana Hann. Se puede notar que la ventana mantiene las frecuencias del centro y elimina las frecuencias de ambos bordes de la señal. Para evitar perdida de datos, se debe aplicar un tamaño de salto. Este tamaño de salto permitirá aplicar la ventana entre traslapes de distintos marcos para evitar que la ventana provoque pérdida de datos en los bordes de cada marco. Por ejemplo, si se utiliza un tamaño de marco de 2048 muestras y un tamaño de salto de 1024, estaremos traslapando las ventanas en un 50%, esta elección es bastante común en procesamiento de audio y es la que se utilizará para el procesamiento del set de datos en este trabajo.

G. Características del audio

Con una señal debidamente tratada, es posible empezar a realizar la extracción de características para la construcción de un conjunto de datos de entrenamiento en un clasificador de música. Todas estas características se computarán en cada marco y posteriormente pueden agregarse en estadísticas a nivel de bloques (varios marcos en conjunto) A continuación, se mencionan varias características que han sido elegidas como posibles candidatos para la creación del clasificador:

1. Amplitud máxima. Una característica de bajo nivel. Nos indica la amplitud máxima de todas las muestras de un marco. Por el hecho de utilizar la amplitud, es una característica que se computa en el dominio del tiempo y se define a través de la siguiente ecuación:

$$AE = max_{k=t*K}^{(t+1)*K-1} s(k)$$

Donde t es un marco de tamaño K. Esta característica puede ser utilizada en conjunto con otras para detección del tempo en características de nivel medio o alto. (Caetano, 2010)

2. Raíz de la energía cuadrática media. Esta es otra característica de bajo niveles en el dominio del tiempo. Nos da un indicador de la intensidad del sonido percibida por el oyente:

$$\text{RMS} = \sqrt{\frac{1}{K} * \sum_{k=t*k}^{(t+1)*(K-1)} s(k)^2}$$

(Caetano, 2010)

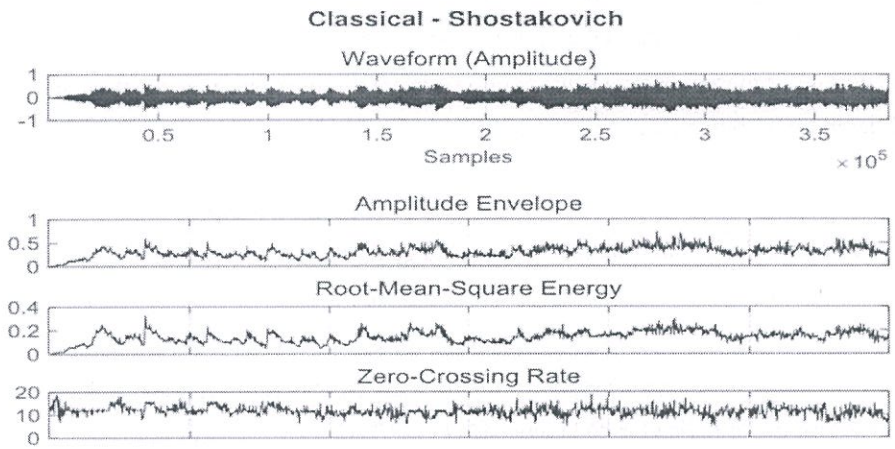
3. Tasa de cruzamiento de ceros. La velocidad de cruzamiento de ceros nos indica la cantidad de vez que la señal cambia de signo en su amplitud durante un marco de audio específico. Es también una característica de bajo nivel en el dominio del tiempo Se utiliza principalmente para la detección de ruido y sonidos de percusión. Además, puede ser un indicador de notas musicales cuando hablamos de señales monofónicas. Es común que los afinadores electrónicos de instrumentos utilicen esta característica para determinar la frecuencia del sonido del instrumento.

$$ZCR = \frac{1}{2} \sum_{k=t*K}^{(t+1)*(K-1)} |sgn(s(k)) - sgn(s(k+1))|$$

(Gouyon, 2000)

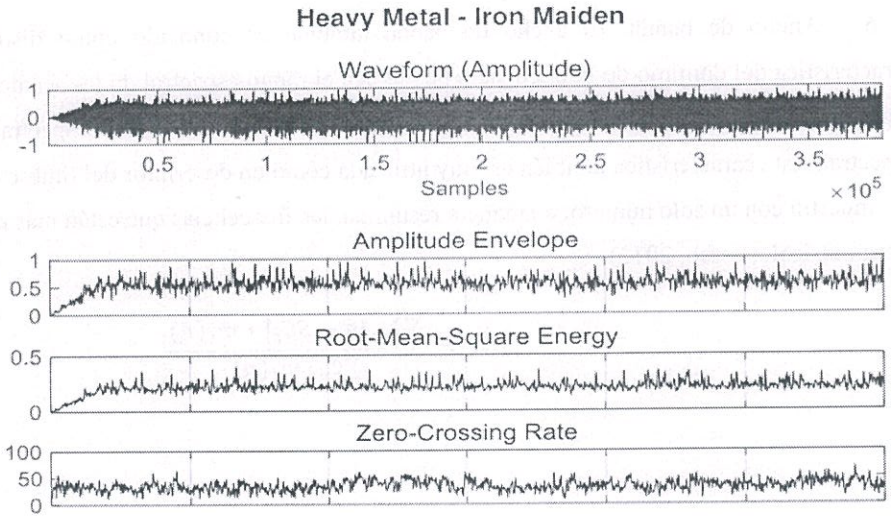
Nótese como en las Figuras 4 y 5 se pueden empezar a observar diferencias notables entre dos tipos de música distintos solo con estas características de bajo nivel. La amplitud máxima es mucho más baja en la pieza clásica, de igual manera que la energía cuadrática media, además la tasa de cruzamiento de ceros es bastante más estable que en la canción de metal. Por el contrario, la canción de heavy metal, tiene amplitudes máximas mucho mayores y una tasa de cruzamiento de ceros mucho más variada, lo cual puede indicar mayor cantidad de percusiones y sonidos distorsionados.

Figura 6: Características de bajo nivel para una pieza de música clásica.



(Knees, 2016).

Figura 7: Características de bajo nivel para una pieza de Heavy Metal.



(Knees, 2016)

4. Proporción de energía en las bandas. Esta es una característica de bajo nivel en el dominio del tiempo. Se utiliza principalmente en el reconocimiento de voz y nos da una idea de que tan dominantes son las frecuencias bajas con respecto a las frecuencias altas de una señal:

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

En esta fórmula, F denota la frecuencia máxima para distinguir las frecuencias altas y las bajas. Una elección correcta de esta F puede mejorar bastante los resultados de esta característica. En algunos casos se invierte el numerador y denominador o se aplican frecuencias de corte para los casos en que las frecuencias bajas son demasiado predominantes. (Knees, 2016).

5. Centroide espectral. El centroide espectral es una característica muy utilizada en la construcción de clasificadores de música a pesar de ser una característica de bajo nivel. Representa de alguna manera el centro de gravedad del espectro de frecuencias dentro de un marco y por tanto nos puede dar ideas de características musicales como la brillantez del sonido o característica del timbre de un instrumento o pieza musical. Es recomendable no filtrar los sonidos bajos para este tipo de características ya que los resultados que podría traer pueden verse distorsionados por eliminar estas frecuencias.

$$SC_t = \frac{\sum_{n=1}^{N} m_t(n) * n}{\sum_{n=1}^{N} m_t(n)}$$

(Knees, 2016).

6. Ancho de banda. El ancho de banda también es conocido como dispersión espectral. Esta característica del dominio de la frecuencia nos indica el rango espectral de las señales cercanas al centroide espectral. Se puede interpretar como una especie de varianza sobre la media espectral, es decir, el centroide espectral. Esta característica también es muy utilizada como un descriptor del timbre de una canción, ya que nos muestra con un solo número, de manera resumida, las frecuencias que están más presentes dentro de una pieza musical. (Lerch, 2012)

$$BW = \frac{\sum_{n=1}^{N} |n - SC_t| * m_t(n)}{\sum_{n=1}^{N} m_t(n)}$$

7. Flujo espectral. El flujo espectral es una característica bastante utilizada para el análisis de timbre y detección de voz. Nos da una idea de los cambios de intensidad en todo el espectro de frecuencias entre dos marcos de audio consecutivos:

$$SF_t = \sum_{n=1}^{N} \left(D_t(n) - D_{t-1}(n)\right)^2$$

En esta fórmula, D_t es la distribución normalizada de todas las frecuencias del marco de audio t. Por lo que el flujo espectral es la suma de todas las magnitudes en el espectro de las frecuencias de todos los marcos de audio consecutivos. (Lerch, 2012)

8. Entropía de la energía. La entropía de la energía es también una característica de bajo nivel en el dominio de la frecuencia. Esta característica nos indica la cantidad de cambios abruptos que podemos encontrar en los niveles de energía de una señal de audio. Para poder computarla dividimos un marco de audio en K submarcos de la misma duración. Luego calculamos la energía de cada submarco y la dividimos dentro de la energía total del marco:

$$e_j = \frac{E_{submarco_j}}{E_{marco_i}}$$

Donde $E_{marco_i} = \sum_{k=1}^{K} E_{submarco_k}$. (Giannakopoulos, 2014)

El paso final para calcular la entropía es sumar las energías de cada submarco multiplicadas por su logaritmo en base 2:

$$H(i) = \sum_{j=1}^{K} e_j * \log_2(e_j)$$

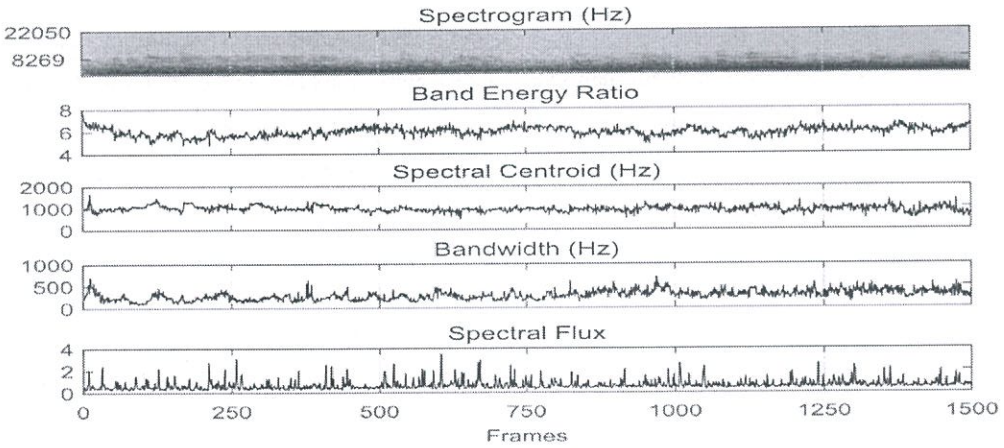
Mientras más pequeño sea el valor, mayor será el cambio de energía. Por lo que valores bastantes pequeños puede utilizarse para detección de sonidos fuertes como percusiones o efectos de sonido como armas o ruidos estridentes. (Giannakopoulos, 2014)

9. Rodamiento espectral. Esta característica del dominio del tiempo nos indica la frecuencia más baja que posee debajo de sí misma, al 80% al 90% de la magnitud espectral de la señal de audio. Esta característica es bastante útil para la detección de ruidos en una señal, ya que el valor será más alto mientras más dispersas estén las magnitudes de todas las frecuencias en el espectro. Música con melodías bastante claras y armonías sencillas tendrán valores bajos de rodamiento espectral, mientras que sonidos más estridentes y armónicamente más complejos puede tener rodamientos espectrales más altos. (Lerch, 2012)

$$vsr(n) = i \longrightarrow \sum_{k=0}^{i} |X(k,n)| = \tau * \sum_{k=0}^{\frac{\tau}{2}-1} |X(k,n)|$$

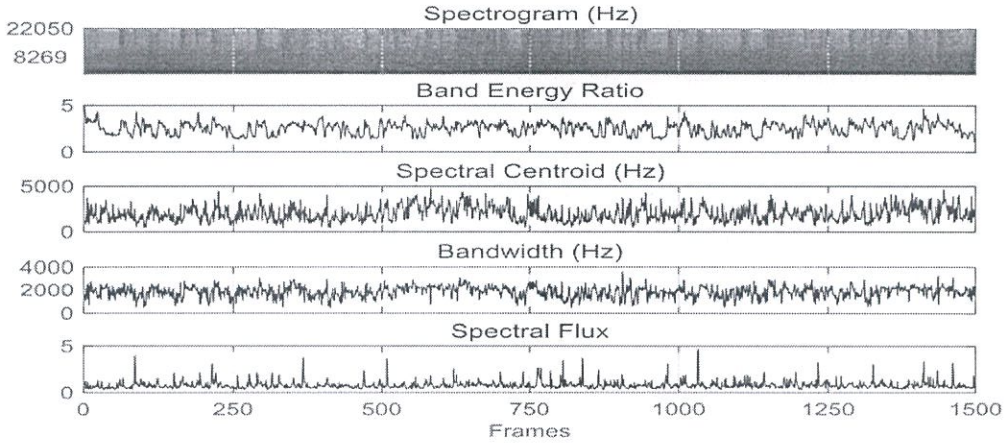
La ecuación nos indica que se regresará el valor de frecuencia en el cual la suma de las magnitudes de la frecuencia sean al menos τ. Donde τ es un valor entre 0.80 y 0.90. (Lerch, 2012)

Figura 8: Características en el dominio de la frecuencia para una pieza de música clásica (Shostakovich)



(Knees, 2016)

Figura 9: Características en el dominio de la frecuencia para una pieza de música de heavy metal (Iron Maiden).



(Knees, 2016)

Como se puede observar en las Figuras 6 y 7, las características en el dominio de la frecuencia nos dan aún más indicios de las diferencias en el contenido de música con características de timbre, armónicas y melódicas bastante diferentes. Un claro ejemplo es el centroide espectral, en done la pieza clásica tiene valores mucho más estables que la pieza de heavy metal. Esto nos puede dar una idea de la comparación del timbre brillante y estable de un piano versus los timbres más estridentes de las guitarras eléctricas y baterías en el heavy metal. De la misma manera el ancho de banda nos muestra que en la música clásica se respeta mucho más una estructura armónica y melódica ya que los anchos de banda son considerablemente menores a los anchos de banda en la pieza de heavy metal.

10. Coeficientes espectrales de Mel. Los coeficientes espectrales de Mel son una característica de nivel medio, más cercana a los conceptos de alto nivel de música. Es una característica que vino a revolucionar el campo de la extracción de información de música por su gran cantidad de aplicaciones y sus resultados tan satisfactorios en distintos tipos de problemas de análisis de señales de audio. Esta característica se deriva del espectrograma de la señal. Al tener las magnitudes de cada frecuencia, el primer paso es convertir la escala de Hertz tradicional a la escala de Mel. De esta representación en la escala de Mel, se toma el logaritmo y posteriormente se aplica la transformada discreta de Fourier (aunque también se utiliza la transformada discreta de cosenos). El resultado de esta transformada nos da el espectro sobre todas las frecuencias de Mel en lugar de sobre el tiempo. Este espectro es el que representa los coeficientes espectrales de Mel para el marco en consideración. Si hacemos lo mismo para cada marco de audio en la señal, obtenemos un conjunto de vectores ordenados en el tiempo donde cada vector es un conjunto de coeficientes espectrales de Mel, esto es bastante similar al espectrograma producido por la transformada discreta de Fourier. (Majeed, 2015)

Es importante mencionar que esta característica está fuertemente ligada a experimentos de el comportamiento de la percepción de la frecuencia en humanos. La cóclea presenta vibraciones en distintos lugares dependiendo de la frecuencia que está siendo escuchada. De manera similar, al tener el espectrograma y las magnitudes de las frecuencias podemos determinar que frecuencias son las que están más presentes en la señal. Luego, tomamos en cuenta la diferencia en cómo se perciben las frecuencias en la escucha humana aplicando el banco de filtros de Mel. Esta escala y filtros son una de las claves que ha hecho a esta característica muy útil tanto para la clasificación de música como para el procesamiento de voz.

En la Figura 8 se puede observar el banco de filtros de la escala de Mel. Este filtro da una mayor resolución a las frecuencias bajas y una menor resolución a las frecuencias altas, de manera similar al comportamiento de la percepción humana del sonido.

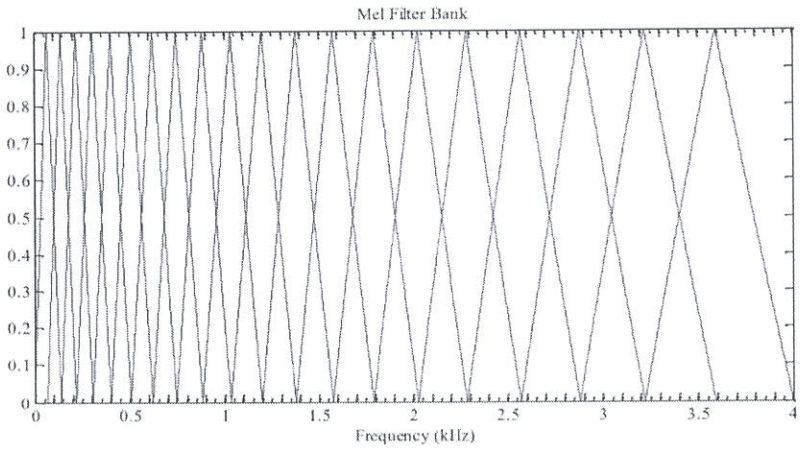
$$f_m = 2525 * \log\left(1 + \frac{f}{\tau}\right)$$

La energía de la señal transformada a la escala de Mel, se obtiene de la siguiente ecuación:

$$E_t^X = \sum_{k-1}^{N} |X(k)|^2 \varphi_i(k)$$

Donde $X(k)$ es el espectro de la amplitud de la señal, k es el índice de la frecuencia y φ_i es el i-ésimo filtro de la escala de Mel. Finalmente, con base en los resultados dentro del campo de la clasificación de música se vio la necesidad de aplicar la Transformada Discreta de Fourier para eliminar la alta correlación del banco de filtros, ya que, como se observa

Figura 10: Filtro de la escala de Mel



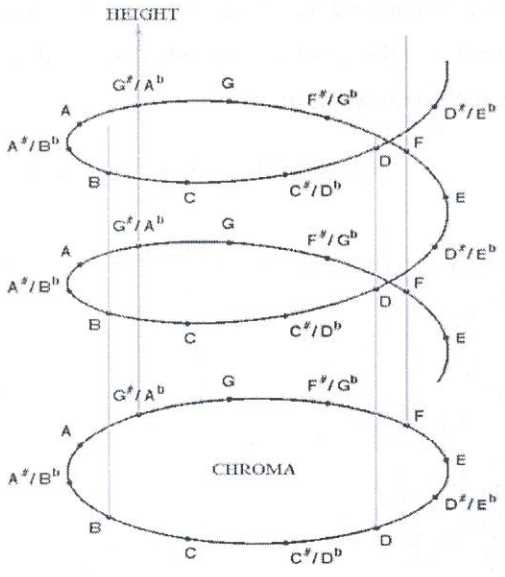
(Majeed, 2015)

11. Vectores cromáticos. La gran mayoría de música occidental se basa en un sistema tonal, este sistema ordena los diferentes tonos en base a una relación de espaciado de frecuencias específico. Los vectores cromáticos son una característica de nivel medio en el dominio de la frecuencia que busca explotar estas relaciones para poder determinar la presencia de los diferentes tonos del sistema tonal cromático occidental (12 tonos). Esta es una característica bastante utilizada para poder determinar la tonalidad de una pieza y empezar a realizar análisis armónicos y melódicos de manera automática.

Para poder comprender los vectores cromáticos, es necesario conocer un poco sobre la percepción de la altura del sonido en el ser humano. La percepción cromática nos indica que las personas no solo perciben las distintas frecuencias del sonido como altas y bajas, sino que además tienden a agruparlas bajo ciertos rangos en los que no notan mayor diferencia. Estos rangos de frecuencias están espaciados en potencias de 2. (Cho, 2010)

La Figura 9 nos muestra la hélice de percepción cromática. Esta se utiliza principalmente para modelar la relación entre los intervalos de octava en la teoría musical. Este modelo se representa con dos dimensiones, la primera nos indica la organización de sonidos de grave a agudos (altura). La segunda nos muestra la relación circular que existe en la organización de los tonos. En este caso, por ser música occidental, el croma se organiza en 12 tonos, cada uno de los que corresponden al sistema tonal occidental.

Figura 11: Hélice Cromática.



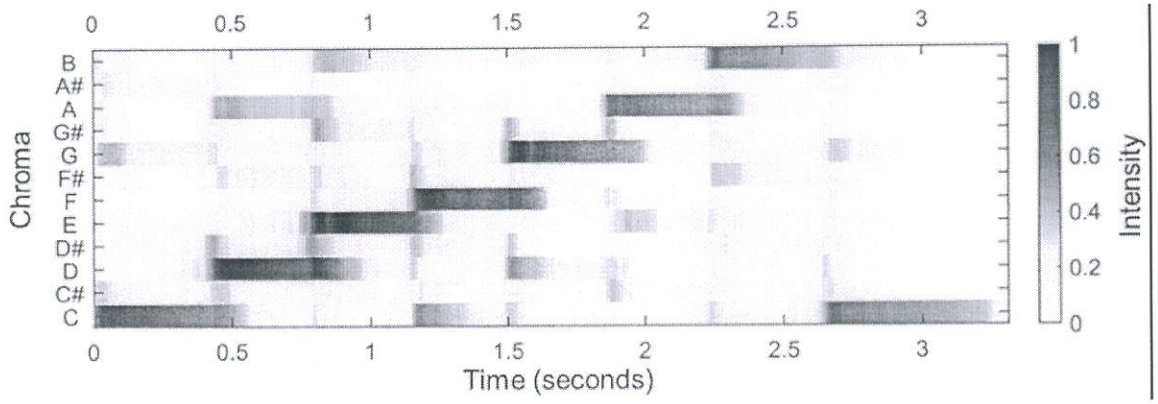
(Cho, 2010)

Los vectores cromáticos son calculados con la siguiente fórmula:

$$C_f(b) = \sum_{z=0}^{Z-1} |X_{lf}(b + z\beta)|$$

Donde X_{lf} es el espectro de la frecuencia en una escala logarítmica, b un entero que representa la nota del sistema tonal, Z el total de octavas y β el número de elementos por octava. Una vez obtenida esta sumatoria, obtenemos un conjunto de vectores cromático para un marco de audio en específico. El conjunto de todos los vectores cromático en una señal es conocido como cromagrama y se muestra un ejemplo en la Figura 10. Nótese cómo se puede evidenciar de manera más clara las notas presentes en la señal de audio con la ayuda de esta característica. Esto es de gran ayuda para determinar la tonalidad de una pieza y realizar análisis armónicos más complejos. En la clasificación automática de música puede ayudar a determinar las notas principales de una pieza y segun esto tomar decisiones sobre la clase a clasificarla al compararla con las notas existentes en los datos de entrenamiento. (Cho, 2010)

Figura 12. Cromagrama (Evidencia la escala de Do Mayor en la señal)



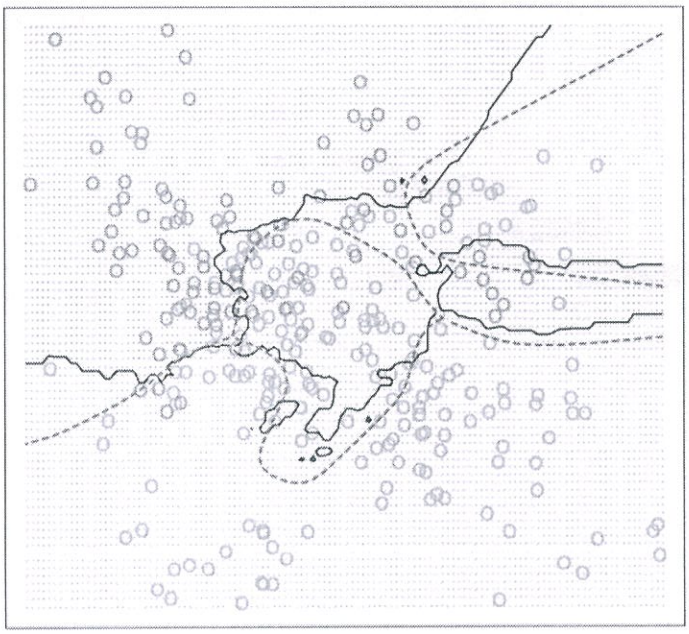
(Cho, 2010)

H. Algoritmos de Aprendizaje de Máquina

Los algoritmos de aprendizaje de máquina, son una rama de la inteligencia artificial que se dedica a la creación de modelos predictivos en base a análisis estadísticos sobre gran cantidad de datos. Todo esto para crear en la máquina un comportamiento inteligente sin programarlo de manera explícita. Existen muchísimos tipos de algoritmos de aprendizaje de máquina para realizar tareas de clasificación de datos. En este trabajo se centrará la atención en tres algoritmos, K Vecinos más Cercanos, Máquina de Vectores de Soporte y Bosques Aleatorios.

1. K vecinos más cercanos. Este es uno de los algoritmos más comunes y más sencillos dentro del aprendizaje de máquina. Se considera un algoritmo basado en memoria ya que no necesita de ningún modelo para ser entrenado. Dado un vector de entrada x, se encuentran k puntos de entrenamiento $x(r), r = 1,2 \dots k$ cuyas distancias sean las más cercanas a la entrada. Luego, para realizar la clasificación, se realiza una votación con las clases de cada uno de los puntos, siendo la clase más votada la que utilizará para la clasificación. Usualmente la distancia a utilizar es la distancia euclidiana, pero existen casos en donde se usa la distancia de manhattan u otros modelos de distancia.

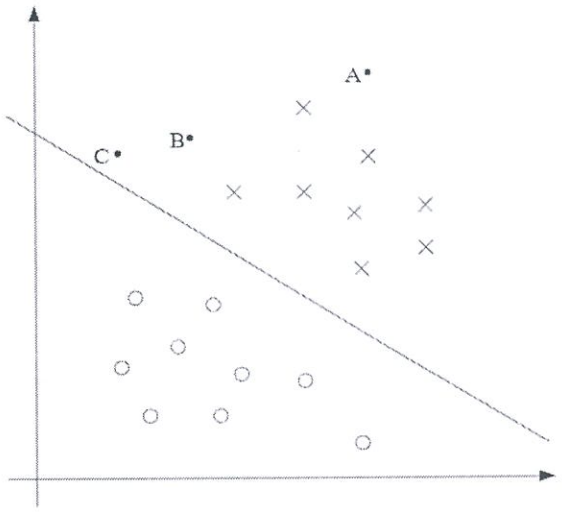
Figura 13: K vecinos más cercanos con K=11



2. Máquina de vectores de soporte. El algoritmo de máquina de vectores de soporte es uno de los algoritmos más utilizados para el aprendizaje supervisado y uno de los que mejores resultados ha mostrado en distintos campos de aplicación. La noción inicial del algoritmo se basa en los márgenes y la confianza en la predicción de un dato. Para esto se considera la Figura 14, en donde se muestra "X" y "O" como dos clases

distintas de resultados de un aprendizaje, la línea separadora llamada hiperplano separador y tres puntos etiquetados como A, B y C (Ng, 2016).

Figura 14. Hyperplano separador y puntos de entrenamiento.



Se puede observar que el punto A se encuentra bastante lejos del hiperplano separador, esto nos dice que seguramente la predicción para este punto será y=1. La confianza de esta predicción se reduce para C, ya que se encuentra mucho más cerca del hiperplano separador y un pequeño cambio en su valor haría que la predicción y=1 se vuelva incorrecta. La idea central de este algoritmo es buscar un hiperplano separador tal que la mayoría de nuestras decisiones sean lo más correctas y confiables posibles. (Ng, 2016)

Para encontrar el hiperplano óptimo que cumpla estas condiciones se comienza por una definición formal del hiperplano:

$$f(x) = \beta_o + \beta^T x$$

En esta función, β es el vector de peso y β_o es el sesgo. Con esta definición es posible definir el hiperplano óptimo de infinitas maneras distintas al escalar β_o y β de diferentes maneras. La representación del hiperplano a elegir de las maneras de representarlo será:

$$\beta_o + \beta^T x = 1$$

x simboliza los datos de entrenamiento que están más cercanos al hiperplano y es llamado el vector de soporte. La representación elegida se conoce como la representación canónica del hiperplano. Posteriormente, se define la distancia entre el punto x y el hiperplano canónico. (Ng, 2016)

$$D = \frac{|\beta + \beta^T x|}{||\beta||}$$

Definimos M como el máximo margen que hay entre el hiperplano y los datos más cercanos para las dos posibles clasificaciones en el plano de la Figura 14. Con esto en mente, notamos que:

$$D = \frac{1}{||\beta||}$$

$$M = \frac{2}{||\beta||}$$

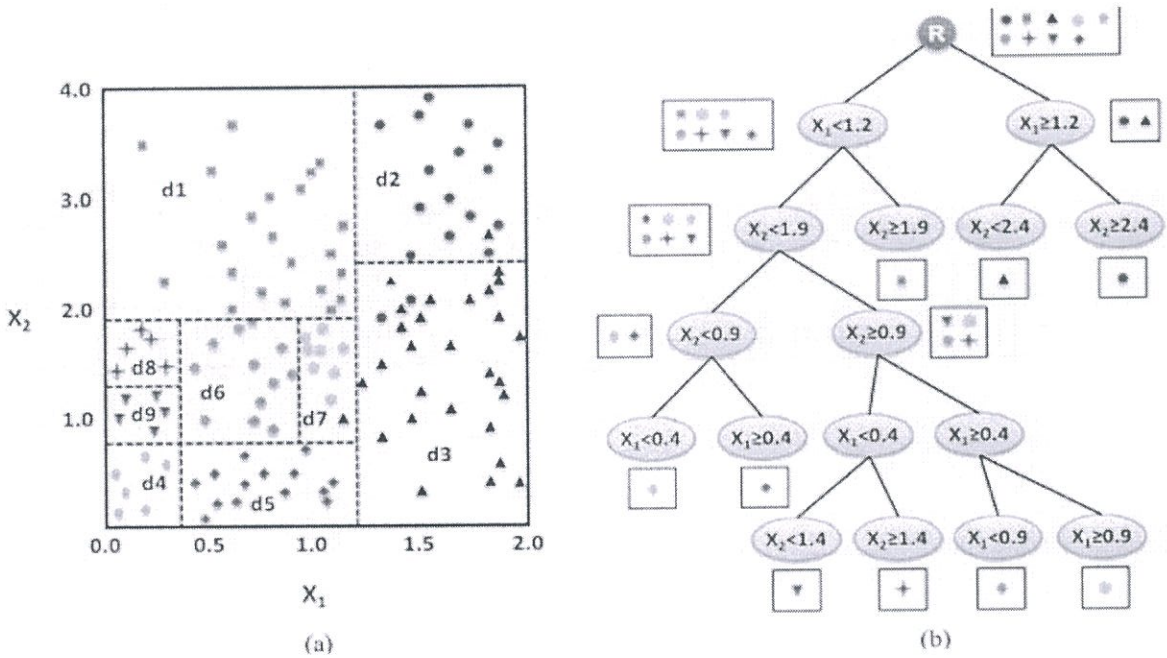
Las fórmulas anteriores nos dicen que el margen M es entonces el doble de la distancia entre las muestras más cercanas del conjunto de datos de entrenamiento. Lo que se busca entonces es maximizar M, y esto es equivalente a minimizar una función $L(\beta)$ sujeta a ciertas restricciones. Estas restricciones son las que modelan el requerimiento para que el hiperplano clasifique correctamente todos los datos de prueba x_i. Esto se puede escribir formalmente como sigue:

$$min_{\beta,\beta_0} L(\beta) = \frac{1}{2}|\beta|^2 \; sujeto \; a \; y_i(\beta^T x_i + \beta_0) \geq 1 \forall i$$

Este es un problema de optimización de LaGrange, que puede ser resulto con multiplicadores de LaGrange para obtener el vector de peso y el sesgo del hiperplano óptimo. (Ng, 2016)

3. Bosques aleatorios. El algoritmo de bosques aleatorios es una mejora al modelo de árboles de decisión y es muy utilizado para la clasificación de datos no lineales. Tiene como ventajas principales su excelente desempeño ante conjuntos de datos con ruido y su velocidad de creación. La idea principal de este algoritmo es crear árboles de decisión. Lo árboles de decisión son estructuras de datos que dividen el conjunto de datos en base a las características que presenten mayor cantidad de entropía. Cada uno de estos árboles nos dan la probabilidad de que un dato pertenezca a una clase en específico. Los bosques aleatorios simplemente crean una cantidad N de árboles de K datos del conjunto de datos de entrenamiento seleccionados aleatoriamente, y promedian las probabilidades para dar el veredicto final sobre la clase del dato bajo análisis (Hastie, 2009).

Figura 15. Partición de datos con base en un árbol de decisión.



(Hastie, 2009)

Para la creación cada uno de los árboles, se eligen una cantidad M de variables de manera aleatoria y luego se determina la mejor partición con m variables de las M originales donde $m \leq M$. El criterio para determinar la mejor partición es la fórmula de entropía definida a continuación:

$$H = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Esta fórmula nos da una idea de impureza de la información, donde p_i nos indica la probabilidad de que un punto del conjunto de datos sea de la clase i. Con esta fórmula podemos crear el concepto de ganancia de información. La ganancia de información resulta ser la diferencia en las entropías luego de haber determinado una forma de dividir el conjunto de datos. Si imaginamos un árbol que se divide en dos ramas A, B. La ganancia de información sería la suma de la entropía de ambas ramas menos la entropía que ya se tiene en el conjunto de datos. Esto se definiría de la siguiente manera: (Hastie, 2009)

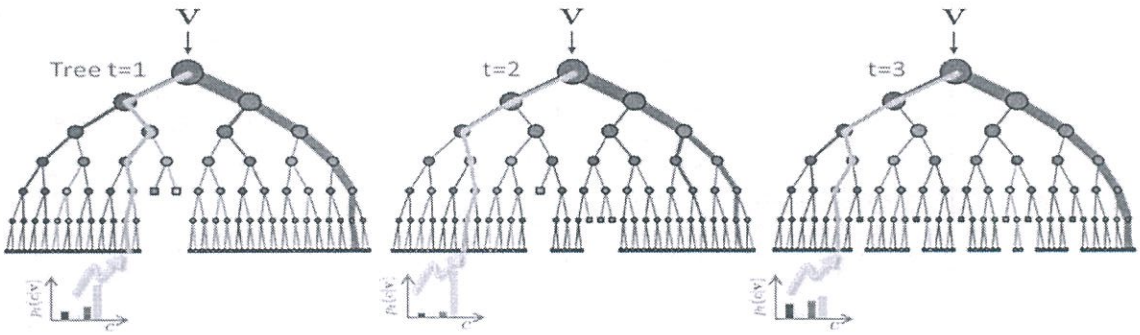
$$ganacia = H_{anteior} - H_{despues}$$

$$H_{despues} = H_A + H_B$$

Con estos criterios ya es posible construir una cantidad N de árboles en los cuáles los datos y las características serán seleccionadas de manera aleatoria. Para un dato, se tendrán resultados de la probabilidad de pertenencias a cada una de las clases en cada árbol. Para el voto final pueden utilizarse varios criterios, pero los más comunes son promediar la probabilidad o elegir la clase que más se repita en todos los árboles.

Este modelo se utiliza en aplicaciones donde la data puede llegar a ser ruidosa como análisis de video e imágenes. También se utiliza bastante en la detección facial y el procesamiento de señales en general.

Figura 16. Visualización de un bosque aleatorio.



(Hastie 2009)

V. Marco metodológico

A. Elección del modelo de clasificación de emociones

Tras analizar los distintos modelos de clasificación de emociones, se determinó que el tipo de modelo a utilizar debía se evidentemente categórico. Esto es principalmente debido a que podemos utilizar las clases del clasificador como las etiquetas de emociones básicas. Si se hubiese utilizado un modelo dimensional, la clasificación sería un punto en el plano afectivo y no de podría dar una categoría de emoción definitiva a la pieza musical bajo análisis. Aunque el plano dimensional si se ha utilizado para construir clasificadores de música en base a emociones, el objetivo final de estos clasificadores es proveer a investigadores más información sobre dinámicas en las emociones dentro de una misma canción. El objetivo del clasificador en este caso es poder categorizar música para facilitar su búsqueda, crear sistemas de recomendación automática y facilitar la exploración de grandes librerías al usuario.

Una vez elegido el modelo categórico, se tomó la decisión de basarnos en el modelo de Paul Ekman de las emociones básicas, por lo que se descartó el uso del modelo de Hevner. Esta decisión se debe principalmente a que, aunque el modelo de Hevner se utilizó en los inicios de la construcción de clasificadores automáticos de música, ha sido criticado por tener demasiadas categorías que para muchos usuarios resultan ser muy similares. Este problema resulta en precisiones menores de los clasificadores debido a que existen clases bastante similares donde el contenido musical bajo estas categorías puede llegar a tener gran parecido. Adicionalmente, se sabe que en estudios recientes se ha reducido las emociones básicas de 6 a 4. Esto debido a que, en muchos casos, el enojo y disgusto eran categorizados de manera muy similar, de la misma manera que con el miedo y la sorpresa. Por estas razones se definieron inicialmente cinco emociones básicas: alegría, tristeza, miedo, enojo y neutral. Al realizar las clasificaciones iniciales, se observó que los participantes estaban utilizando la emoción neutral para canciones que se les dificultaba clasificar. Esto hizo cambiar el modelo a las emociones de alegría, tristeza, enojo y miedo.

Para tomar las clasificaciones, se construyó una pequeña aplicación web en donde los participantes ingresaban sus datos y luego se les presentaban diez canciones aleatorias de un total de

859 canciones. El participante debe elegir una emoción de las cuatro presentadas, a las cuales se les colocaron imágenes correspondientes a las expresiones faciales asociadas a la emoción. Las expresiones se presentan en la Figuras 12, 13, 14 y 15. Adicionalmente, en cada clasificación se tomó el tiempo de escucha de la canción antes de que el participante emitiera su respuesta. Este tiempo fue utilizado para obtener el tiempo promedio de escucha durante toda la toma de datos.

Figura 17: Alegría

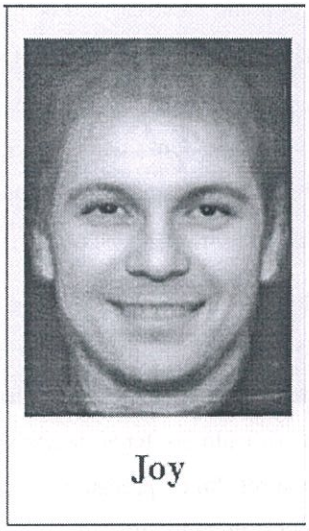


Figura 18: Tristeza

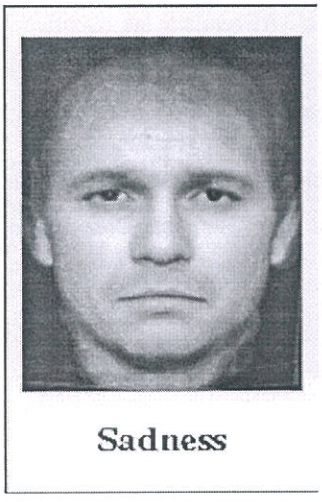


Figura 19: Miedo

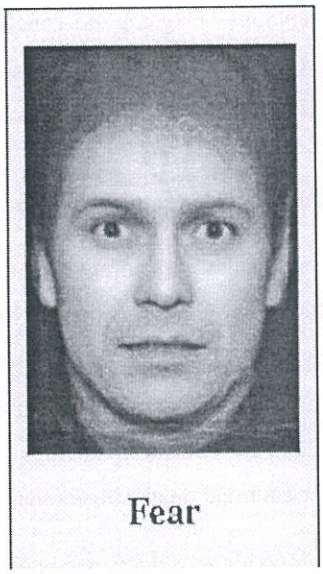
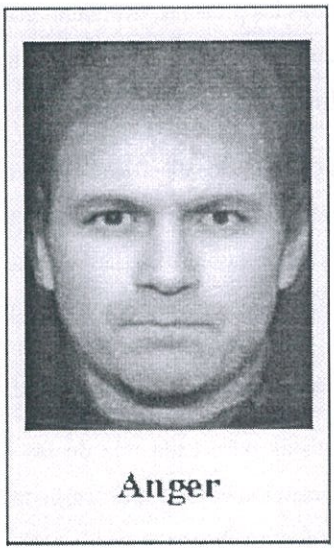


Figura 20: Enojo



B. Construcción del conjunto de datos de entrenamiento

Para la construcción de los datos de entrenamiento se obtuvo un conjunto de 859 canciones de los géneros de ambiente, clásica, jazz, pop, rock, hip-hop, blues r&b y country. Estas fueron clasificadas por tres personas distintas para poder determinar la emoción final sobre la cual se creará el modelo de clasificación. Es importante notar que las canciones elegidas era canciones netamente instrumentales, esto para poder reducir el sesgo en la elección de la emoción por la letra de la canción y limitarse únicamente al contenido sonoro de esta. Además, se eligieron canciones libres de licencias de distribución o licencias mecánicas, para respetar los derechos de autor de los artistas populares y para evitar que los participantes hayan tenido recuerdos y emociones asociadas a canciones que ya habían escuchado previamente.

Con estos criterios para la creación de conjunto de canciones, logramos reducir la mayor cantidad de variables que puedan afectar los datos de entrenamiento. Entre estos evitamos que los participantes tengan recuerdos asociados a la música al tener canciones sin letras y fuera del ámbito de la música popular. Además, nos aseguramos de tener una emoción válida al tener al menos tres clasificaciones de diferentes personas hacia la misma canción. Si en dado caso existiera un empate, se realizó una ronda más de clasificación hasta llegar a una emoción con mayor cantidad de clasificaciones.

Cada participante proporcionó algunos datos personales para poder obtener datos sobre la población en la cual se realizó el experimento. Estos datos incluyen edad, género, nivel de escolaridad, conocimientos sobre teoría musical y problemas auditivos y/o psicológicos. Estos datos pueden ser útiles para las personas que deseen utilizar el conjunto de datos para otros estudios o para continuar el presente trabajo. Finalmente se debe tomar en cuenta que todas las personas evaluadas eran guatemaltecas para poder crear un clasificador enfocado únicamente en una cultura y proporcionar clasificadores automáticos de emociones en la música entrenados en un contexto cultural bien definido.

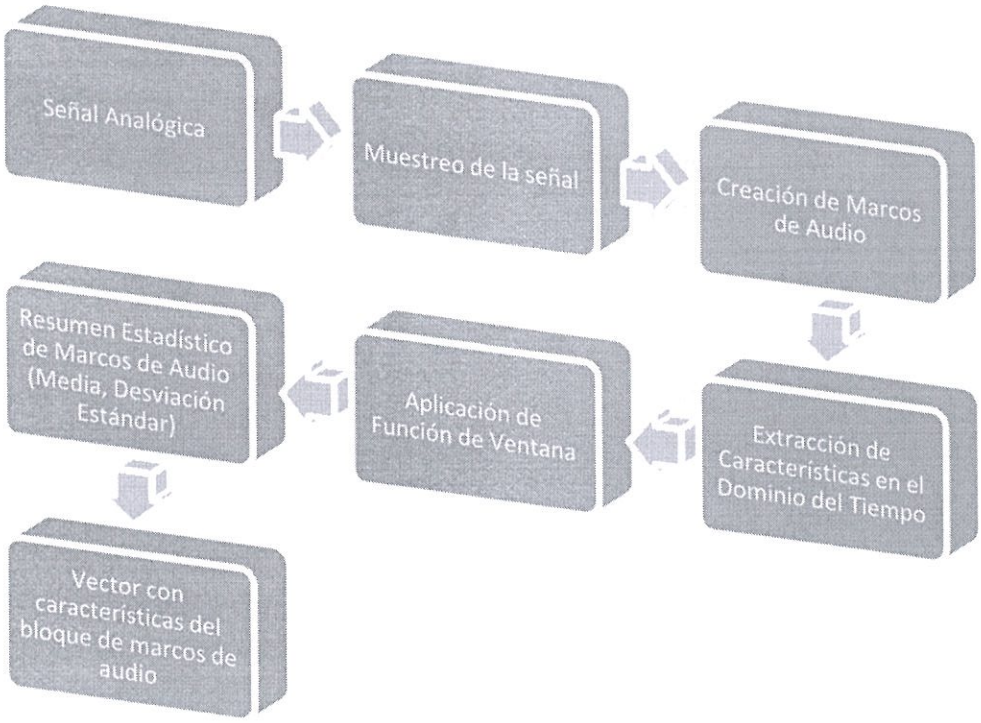
Al finalizar la obtención de todas las clasificaciones se obtuvieron 48 piezas con empates por lo que se procedió a hacer una clasificación más para cada una de estas piezas y así determinar la emoción final de la pieza. Con este trabajo concluido, fue posible iniciar la extracción de características de audio para poder crear los vectores de datos para los algoritmos de aprendizaje de máquina.

C. Análisis de audio, pre procesamiento y características del audio utilizadas

Con las clasificaciones listas para cada una de las 859 canciones, se procedió con el análisis del contenido de audio. Todas las canciones fueron descargadas en formato wav para no perder ningún tipo de dato por los filtrados de frecuencias que hacen otros formatos como el mp3. Las canciones fueron descargadas a una frecuencia de muestreo 44100 Hz. Para evitar procesamiento pesado en las máquinas locales y uso excesivo de memoria, se procedió a subir todas las piezas a utilizar al servicio S3 de Amazon Web Services para su acceso público. Este procedimiento facilitaría el manejo de datos en el proyecto y optimizaría espacio al analizar las piezas de música ya que únicamente se mantendrían dentro del servidor de aplicación los

resultados de las características de audio extraídas de cada pieza, dejando todos los archivos wav en la nube y a disposición de cualesquiera personas que desee accederlos para otros estudios.

Figura 21: Proceso de extracción de datos



El proceso general para la creación de las características de audio, se puede observar en la Figura 16, y da inicio con el muestreo de la señal de audio. Se realiza un muestreo de 44100Hz a 16 bits de resolución. Una vez digitalizada la señal, iniciamos el proceso de creación de marcos. En este trabajo se eligió un tamaño de marco de 50 ms con un tamaño de salto de 25 ms. Esto quiere decir que, en el proceso de aplicación de la función de ventana en la señal de audio, tenemos un traslape de 50% por cada marco de audio. Posteriormente, se procede a obtener la transformada discreta de Fourier para poder computar todas las características que se encuentran en el dominio del tiempo.

Una vez se obtienen las señales en ambos dominios, se calculan las características de audio en cada uno de los dominios y se procede a realizar algún tipo de análisis estadístico para reducir la cantidad de datos. Esto es debido a que la cantidad de datos por una canción es extremadamente grande si se deja tan solo a nivel de marco de audio. Por ejemplo, para una canción de 2:30 minutos, tendríamos 150,000 marcos de audio. A cada uno de estos 150,000 marcos debemos calcularles todas las características necesarias tanto en el dominio de tiempo como de frecuencia. En este trabajo se calcularon 34 características que se detallaran más adelante, por lo que tenemos aproximadamente 5,100,000 datos solo para una canción. Esta cantidad de

datos es demasiada para crear un modelo en tiempos cortos y puede llegar a ser bastante repetitiva por el tamaño tan pequeño de los marcos de audio. Por esta razón se toman bloques de 1 segundo en donde se toma la desviación estándar y la media para reducir la cantidad de datos y mejorar el rendimiento de la construcción de los modelos de clasificación de música por emociones.

Las características extraídas que se utilizaron para la construcción de los datos de entrenamiento de clasificador se extrajeron con la ayuda de las librerías Librosa y PyAudioAnalysis y se presentan a continuación:

a. Tasa de Cruzamiento de Ceros
b. Energía Cuadrática Media
c. Entropía de la Energía
d. Centroide Espectral
e. Ancho de Banda
f. Entropía Espectral
g. Rodamiento Espectral
h. Coeficientes Espectrales de Mel
i. Vectores Cromáticos (12 coeficientes basados en el sistema tonal occidental)
j. Desviación Estándar de los Vectores Cromáticos

Cada una de estas características fue almacenada en un archivo de numpy (.npy) para su fácil acceso a través del lenguaje de programación de Python. En este archivo se almacenó el promedio de cada una de las características por un bloque de marcos de audio de 1 segundo, junto con la desviación estándar del respectivo bloque. Dándonos una matriz de 64 elementos x número de bloques de 1 s. en la canción.

Es importante mencionar que los datos de entrenamiento no son por pieza completa, sino que son por bloque de audio. Es decir, si se clasifico una pieza con la emoción de alegría (por ejemplo) no damos al clasificador un solo vector con las características de toda la pieza, sino que damos K conjunto de vectores, donde K es el número de bloques de 1 segundo y cada vector representa las características resumidas de los marcos de audio que quepan en un segundo, es decir, 20 marcos de audio.

Al momento de entrenar los modelos de clasificación, se utilizaron dos acercamientos. El primero entregando al modelo todos los datos de la canción y el segundo entregando los datos correspondientes únicamente al tiempo de escucha promedio. Esto permitirá comparar las precisiones de los clasificadores en ambos casos para determinar si el tiempo de escucha es influyente en los resultados del clasificador. Finalmente, de esto, se entrenaron los modelos añadiendo las desviaciones estándar de cada característica para determinar si la adición de estos datos al vector mejora o emporan los resultados del clasificador.

Los algoritmos de aprendizaje de máquina utilizados son K Vecinos Más Cercanos, Máquina de Vectores de Soporte y Bosques Aleatorios. Para esto se realizó el proceso de validación cruzada, tomando el 80% de los datos para entrenamiento, un 10 % para la validación y un 10% para la prueba. Se entrenaron

varios modelos con las variantes anteriormente mencionadas y se obtuvieron las precisiones de cada uno de estos algoritmos para determinar cuál de estos trajo mejores resultados. Adicionalmente, es importante notar que el clasificador no clasifica la pieza como un todo, sino que clasificará en bloques de 1 segundo. Esto quiere decir que luego de construir el clasificador se debe elaborar un método para determinar la emoción final de toda la pieza. En este caso se optó por utilizar la moda de las clasificaciones en todos los bloques de audio. Después de tener la clasificación final de la canción se obtuvo un nuevo porcentaje de rendimiento final para determinar cuántas canciones del conjunto de canciones utilizadas para prueba fueron clasificadas exitosamente. Con estos dos porcentajes se puede determinar si las clasificaciones por marcos son similares a las clasificaciones globales de la canción e inferir algunos factores que pudiesen afectar en las diferencias de estos porcentajes.

VI. Resultados

A. Conjunto de datos de entrenamiento

Al finalizar la recolección de datos de entrenamiento para la creación del modelo se obtuvo los siguientes resultados de las características de las 254 personas que participaron en la clasificación de música:

Figura 22: Participantes por género

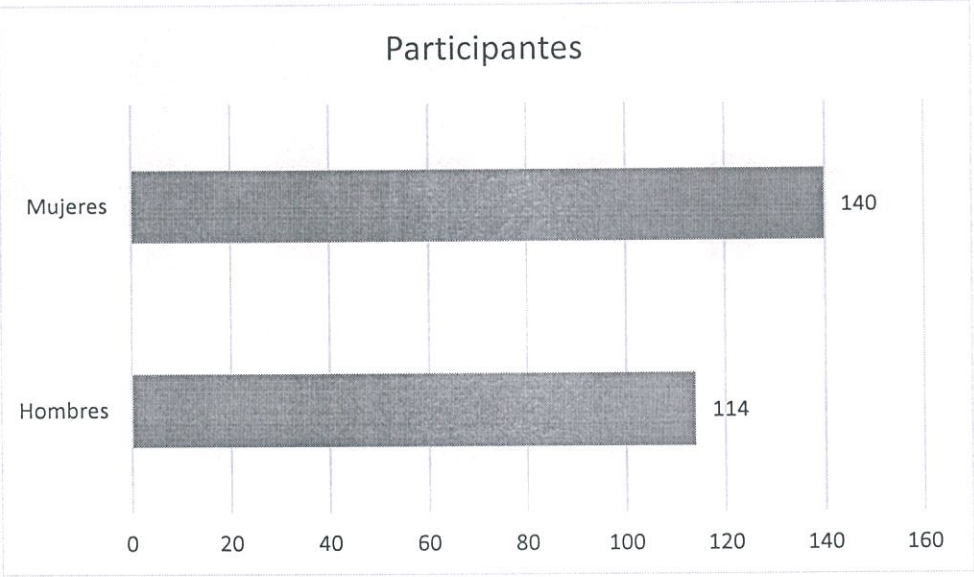


Figura 23: Participantes con problemas psicológicos, auditivos o físicos.

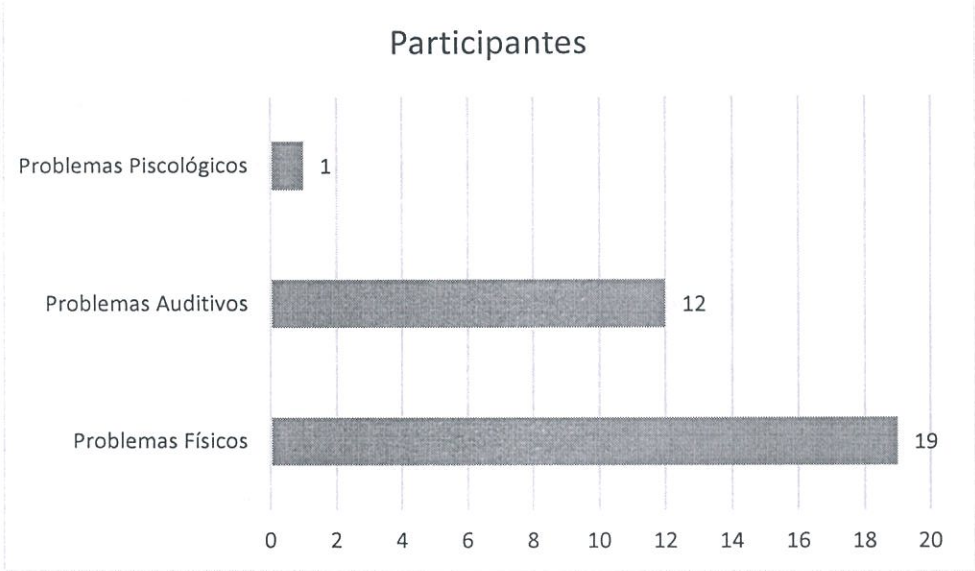


Figura 24: Participantes por edad

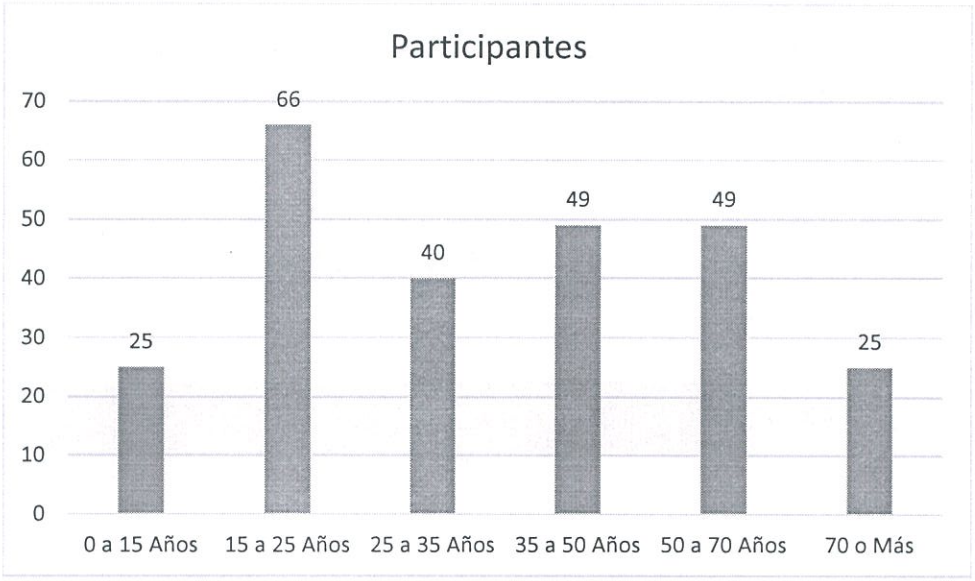
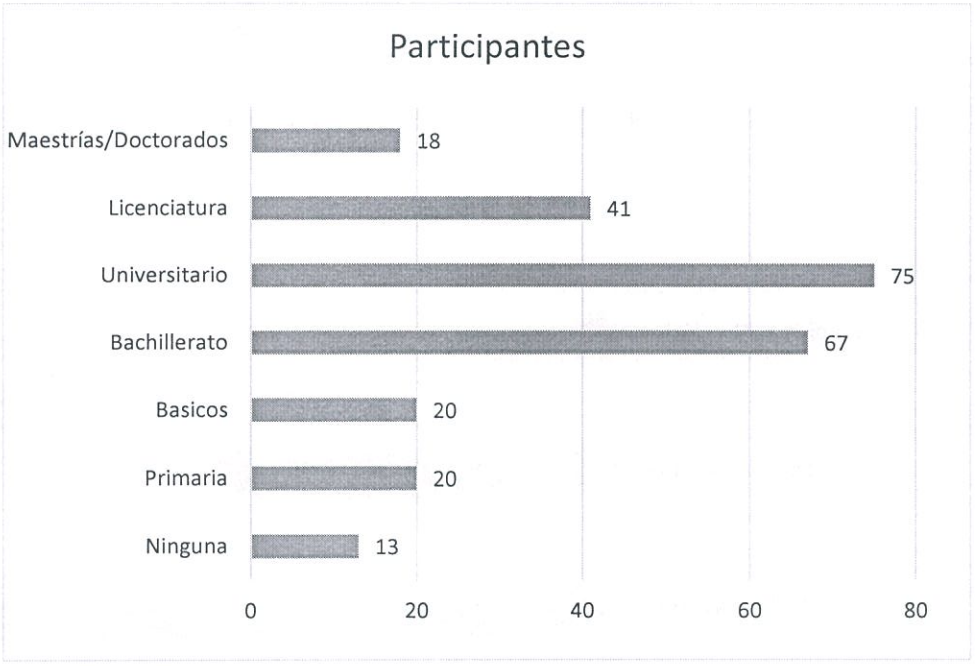


Figura 25: Participantes por nivel académico



B. Canciones utilizadas y resultados de las clasificaciones

Dentro del conjunto de datos de entrenamiento se utilizaron un total de 860 canciones distintas de diversos géneros. La distribución de los géneros del conjunto de canciones y los resultados de la clasificación de estas se presentan en las siguientes figuras. Además, se presenta el tiempo promedio de escucha y la distribución de géneros en cada etiqueta de emoción.

Figura 26: Géneros de canciones

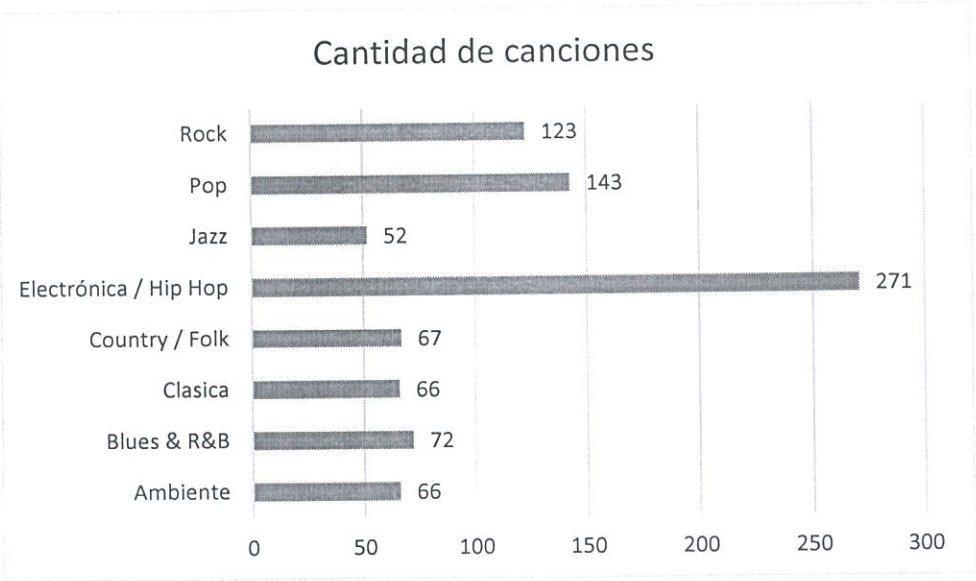


Figura 27: Emociones de canciones clasificadas

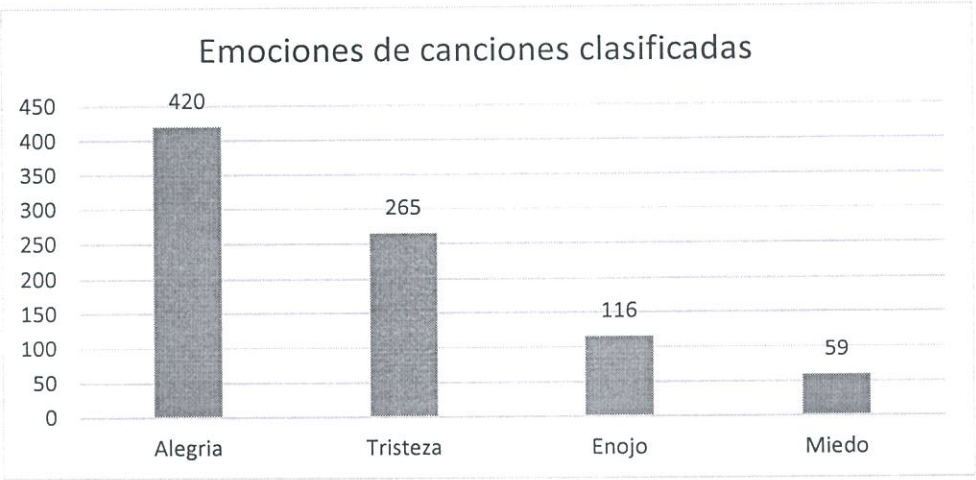
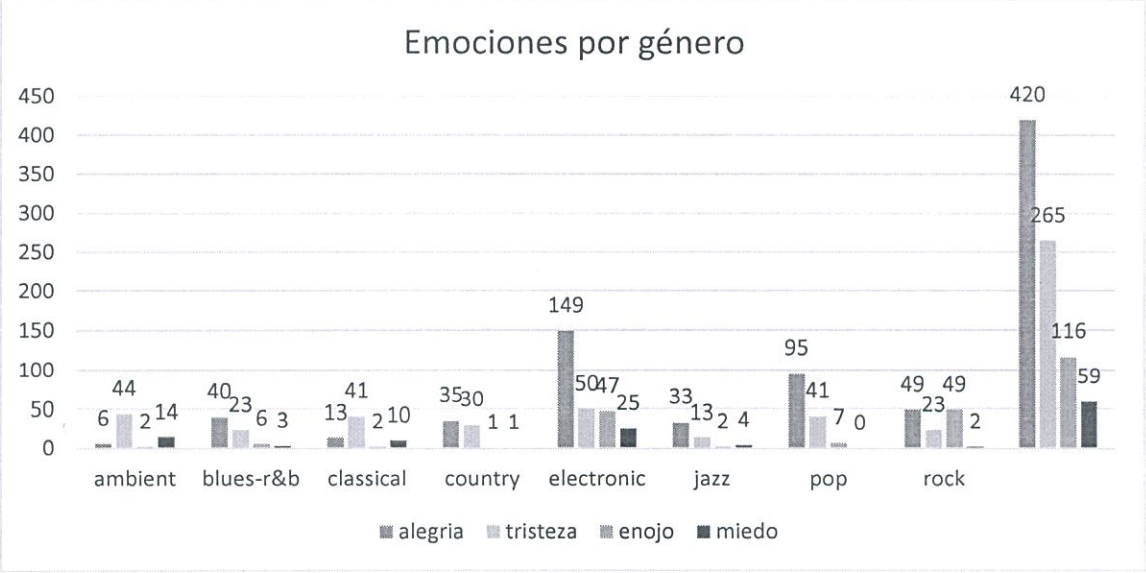
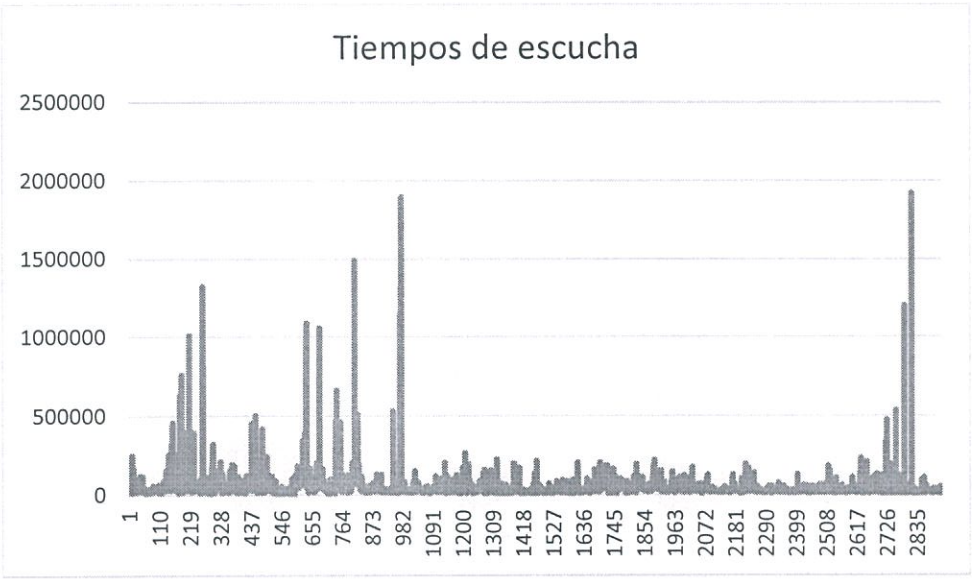


Figura 28: Emociones de canciones clasificadas por género



La figura a continuación nos muestra todos los tiempos de escucha en las clasificaciones de las canciones. De estos datos se obtuvo un promedio de escucha de 40.7 segundos.

Figura 29: Tiempos de escucha de cada clasificación



C. Características extraídas

A continuación, se presentan gráficas de las características extraídas para una canción seleccionada de cada una de las cuatro emociones. Estas canciones fueron seleccionadas de tal manera que se pudieran ejemplificar de mejor manera las ideas desarrolladas en el análisis de resultados Tasa de cruzamiento de ceros

Figura 30: Tasa de cruzamiento de ceros para canción de alegría

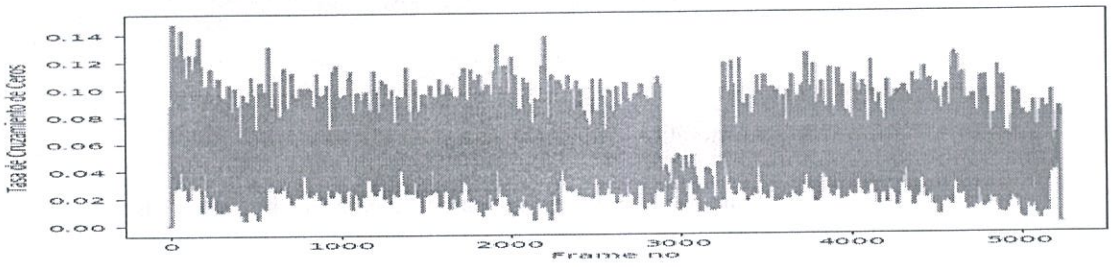


Figura 31: Tasa de cruzamiento de ceros para canción de tristeza

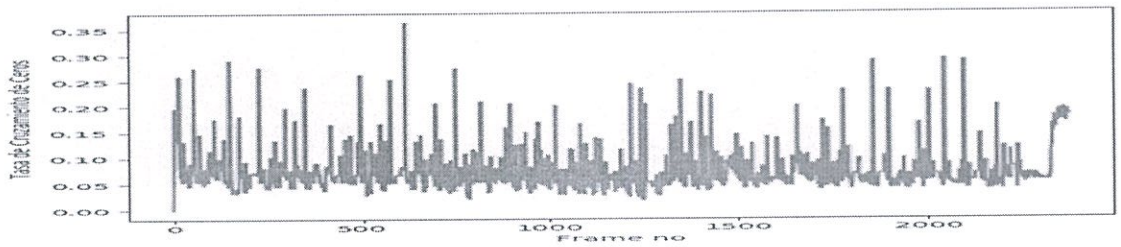


Figura 32: Tasa de cruzamiento de ceros para canción de enojo

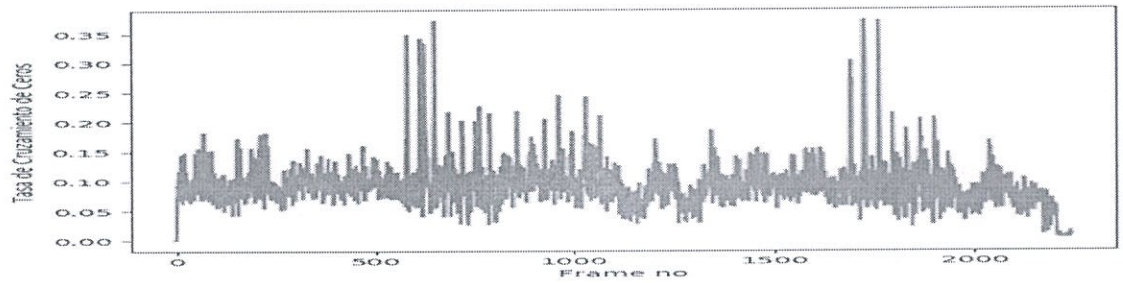
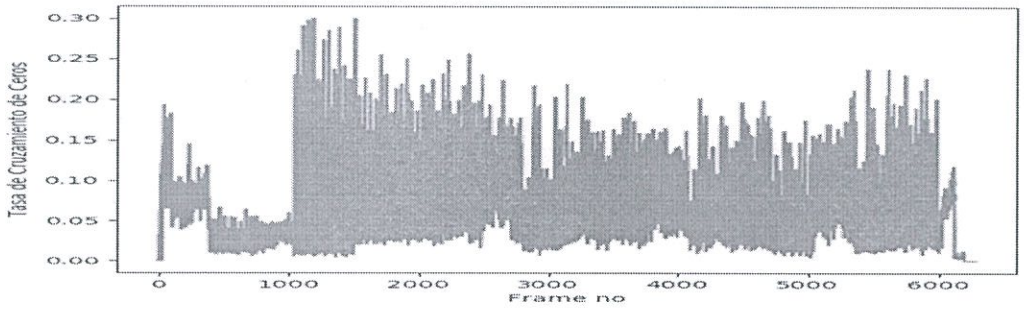


Figura 33: Tasa de cruzamiento de ceros para canción de miedo



1. Energía cuadrática media

Figura 34: Energía cuadrática media para canción de alegría.

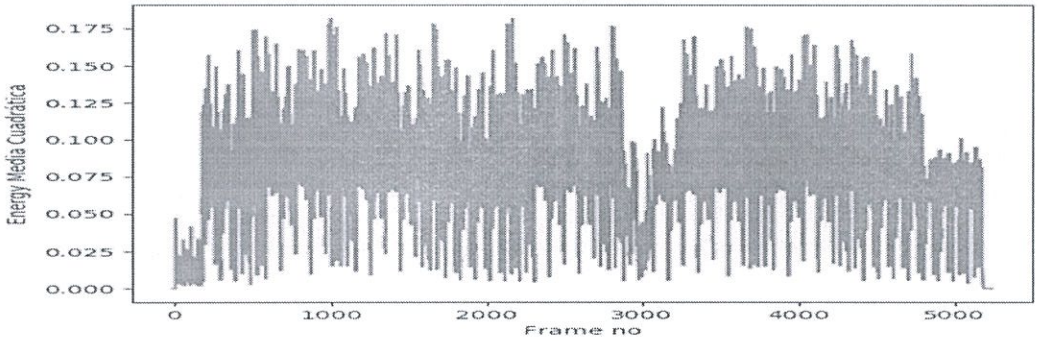


Figura 35: Energía cuadrática media para canción de tristeza.

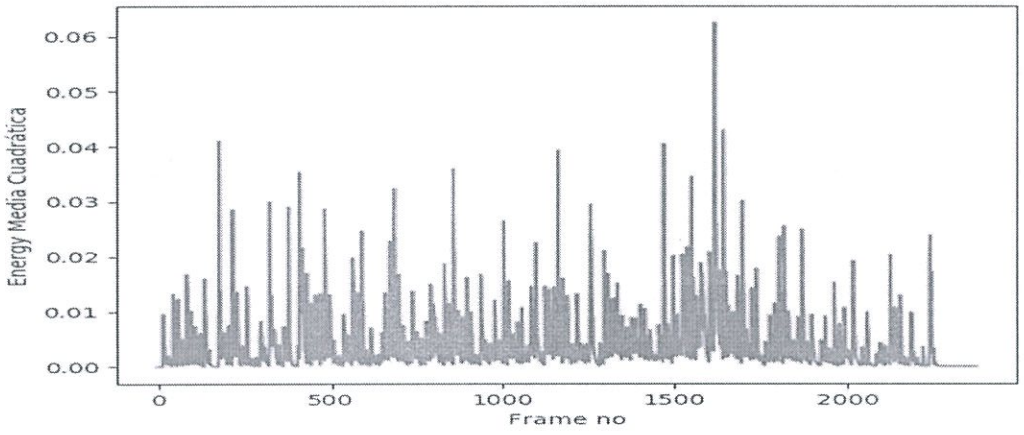


Figura 36: Energía cuadrática media para canción de enojo

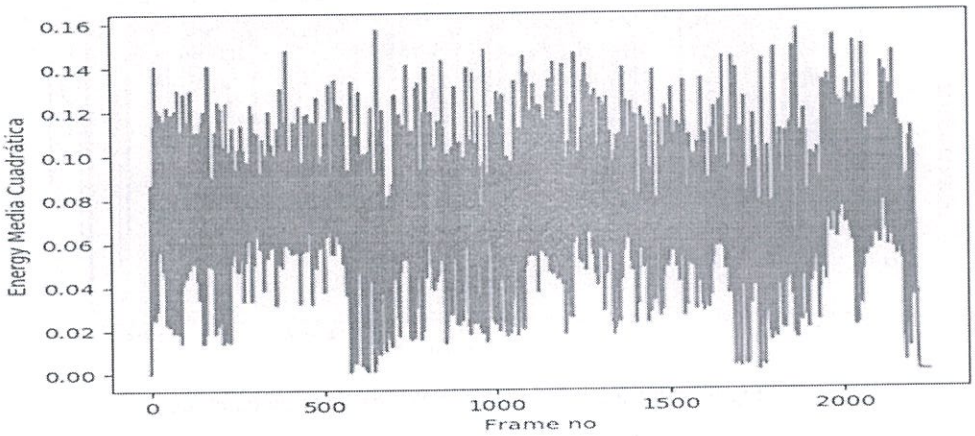
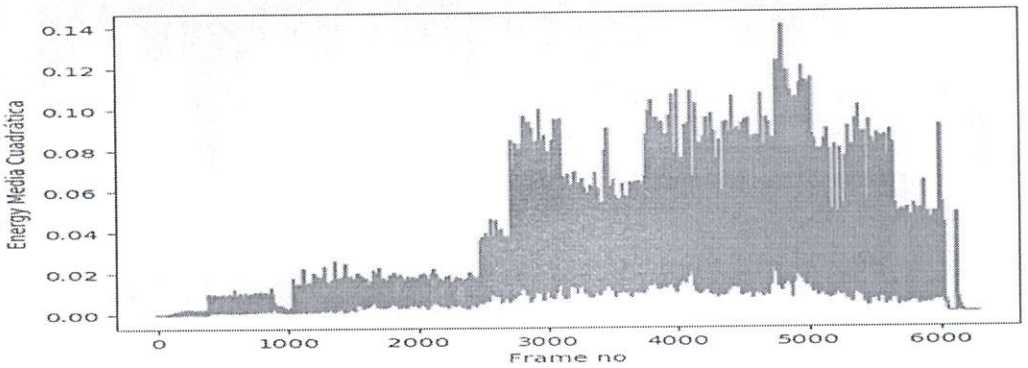


Figura 37: Energía cuadrática media para canción de miedo.



2. Entropía de la energía

Figura 38: Entropía de la energía para canción de alegría

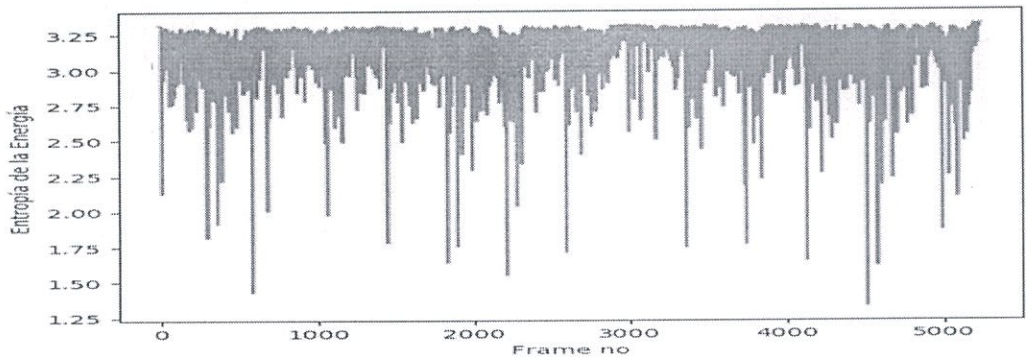


Figura 39: Entropía de la energía para canción de tristeza

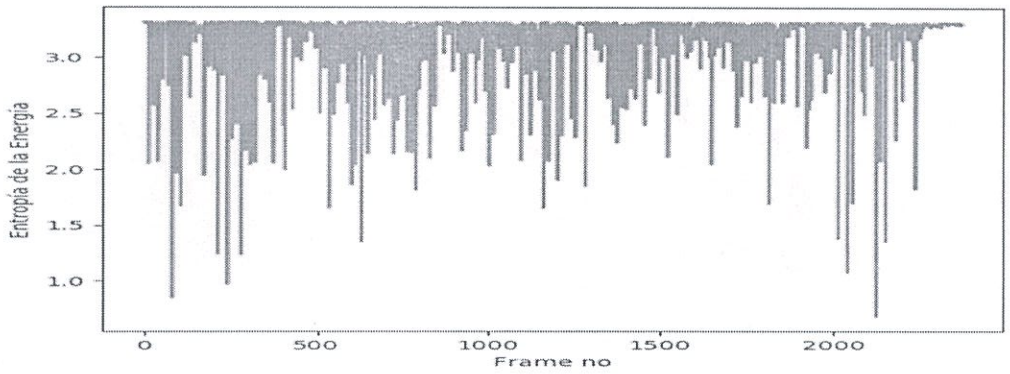


Figura 40: Entropía de la energía para canción de enojo

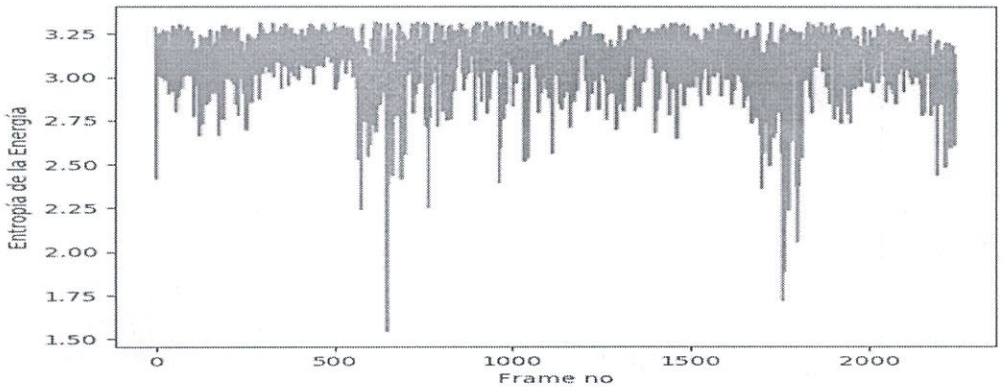
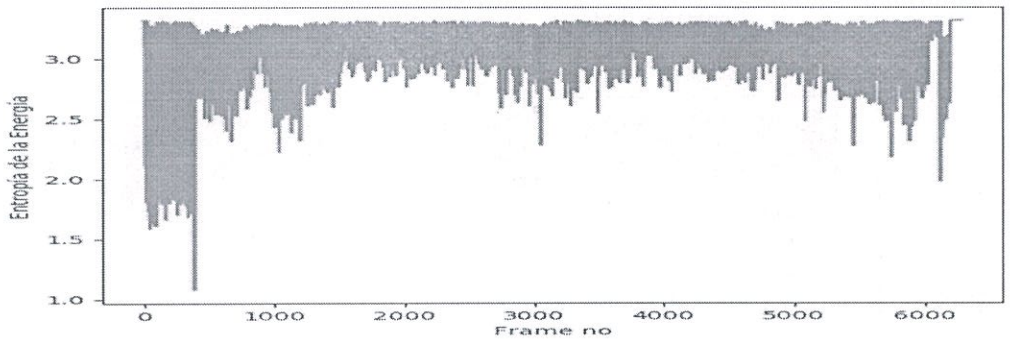


Figura 41: Entropía de la energía para canción de miedo



3. Centroide espectral

Figura 42: Centroide espectral para canción de alegría

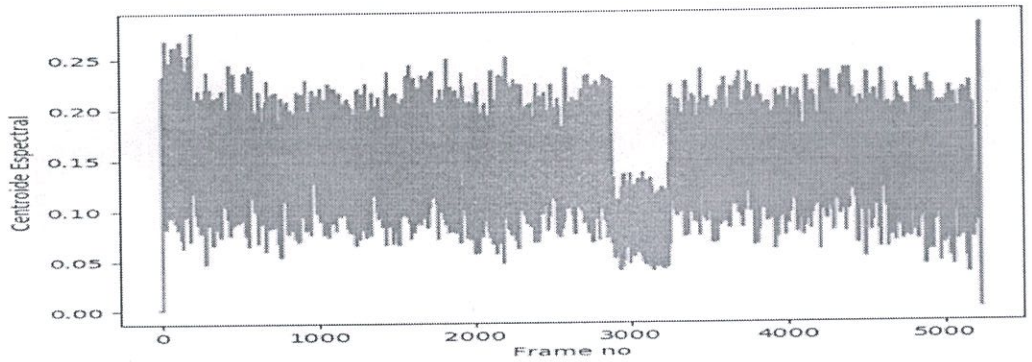


Figura 43: Centroide espectral para canción de tristeza

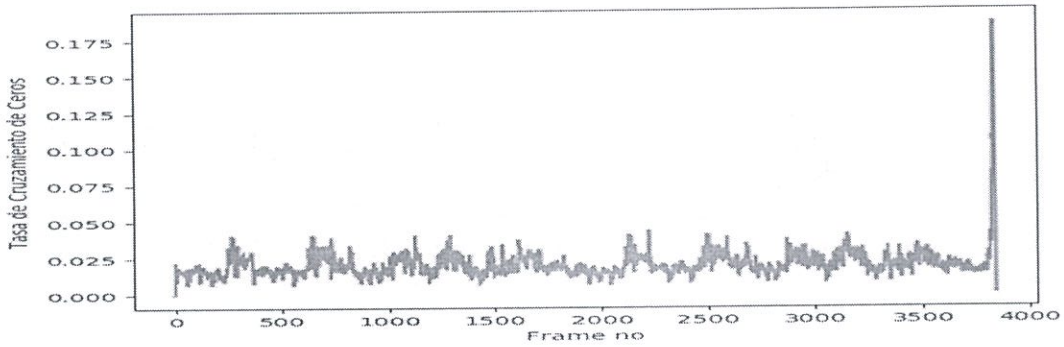


Figura 44: Centroide espectral para canción de enojo

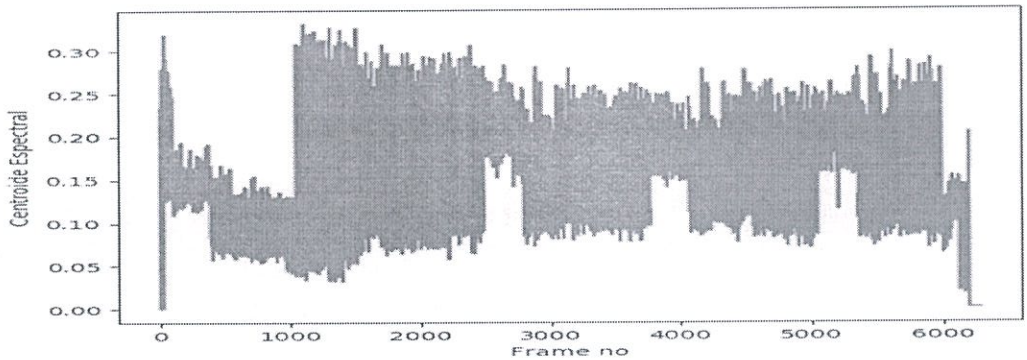
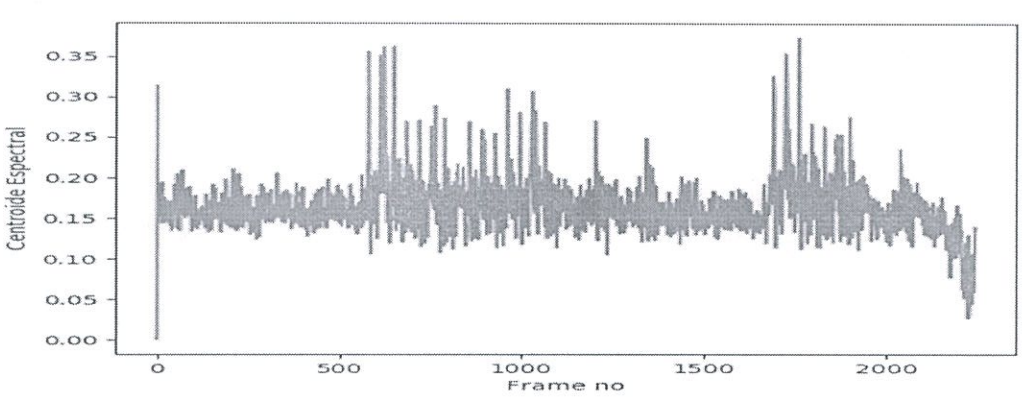


Figura 45: Centroide espectral para canción de miedo



4. Ancho de banda

Figura 46: Ancho de banda para canción de alegría

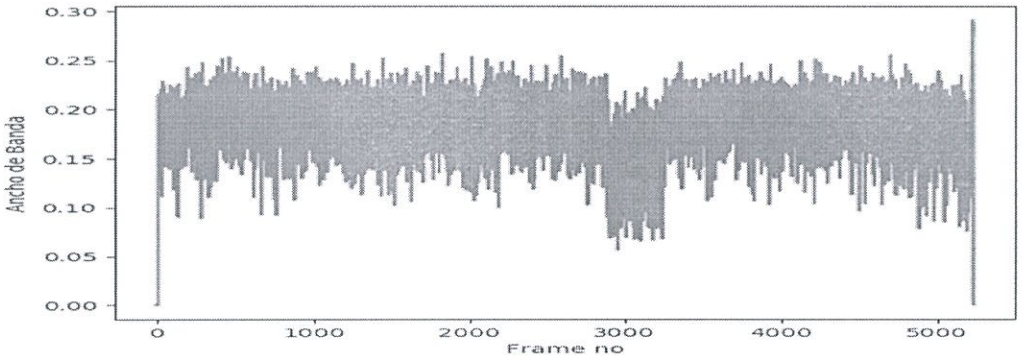


Figura 47: Ancho de banda para canción de tristeza

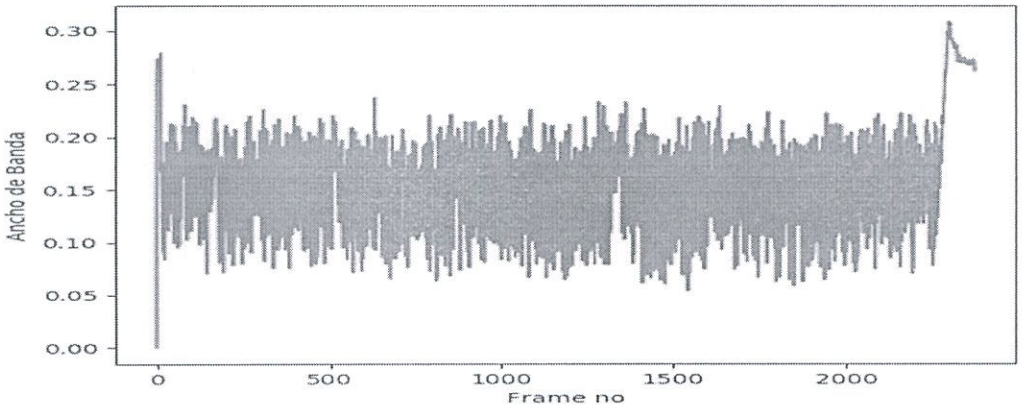


Figura 48: Ancho de banda para canción de enojo

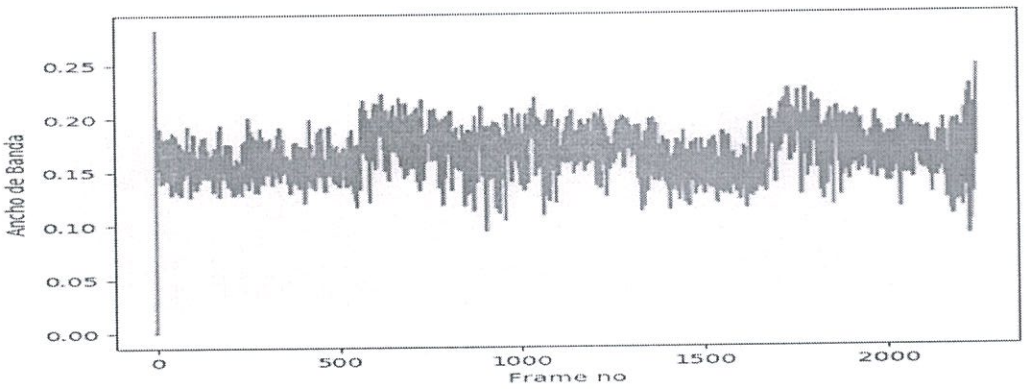
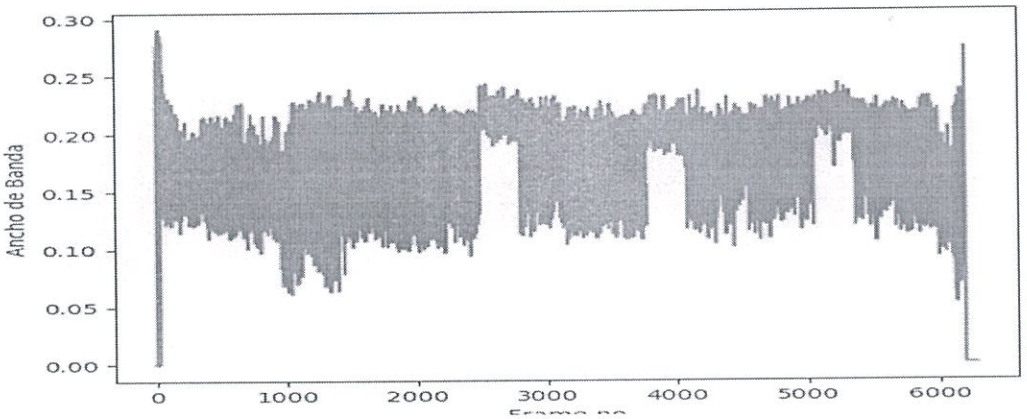


Figura 49: Ancho de banda para canción de miedo



5. Entropía espectral

Figura 50: Entropía espectral para canción de alegría

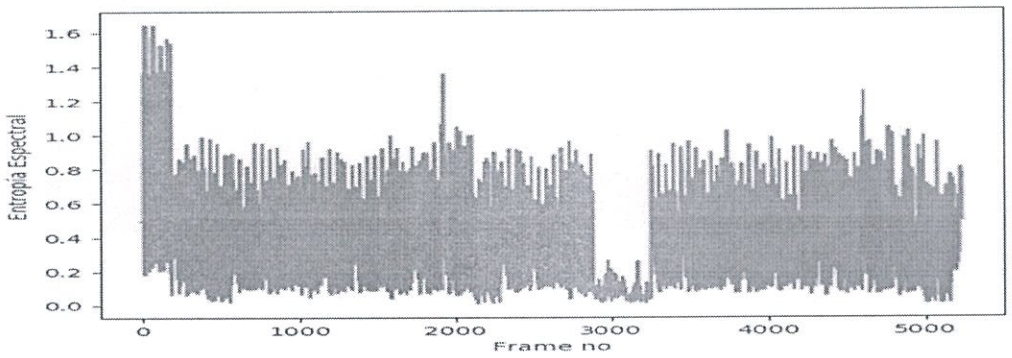


Figura 51: Entropía espectral para canción de tristeza

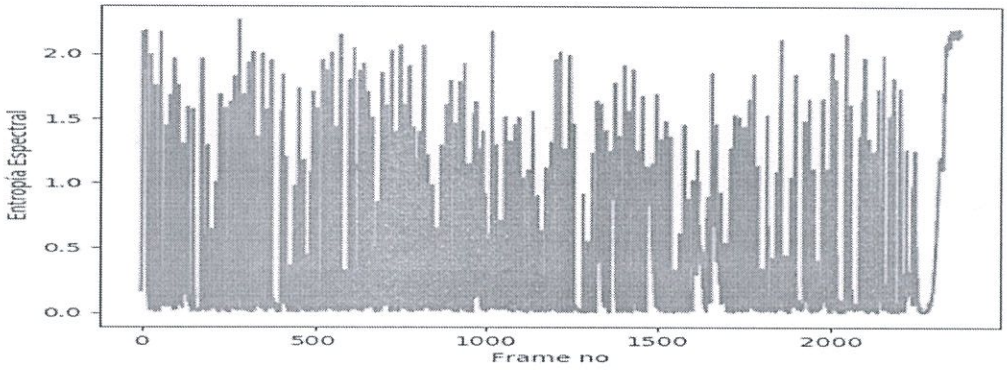


Figura 52: Entropía espectral para canción de enojo

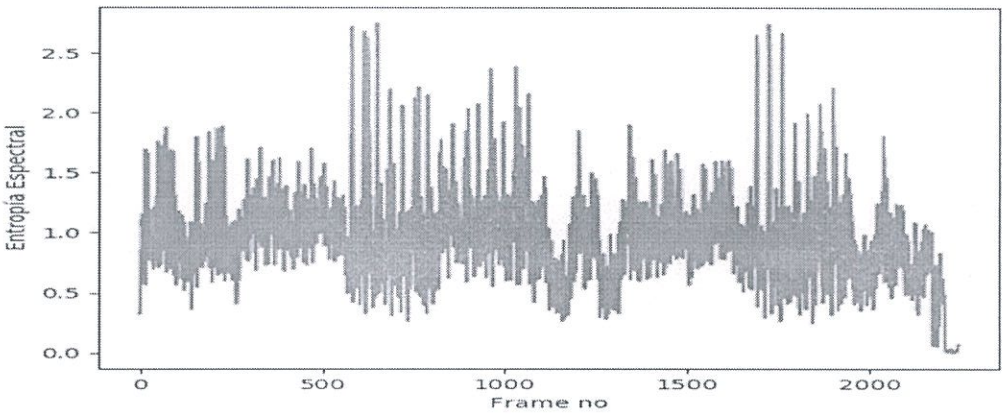
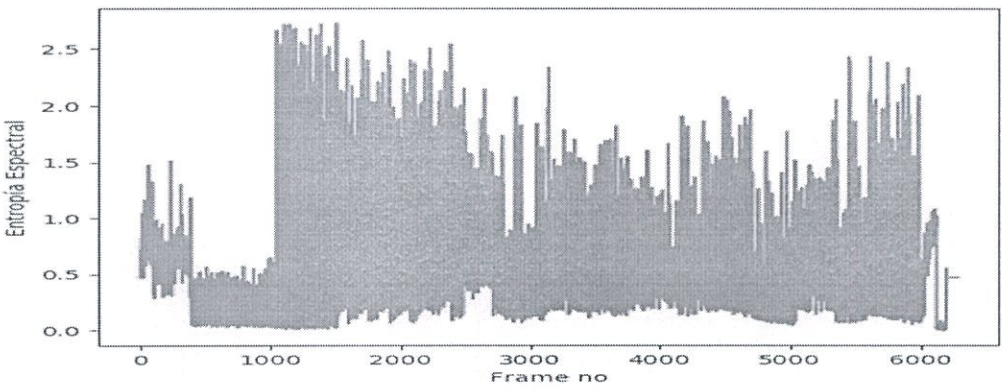


Figura 53: Entropía espectral para canción de miedo



6. Flujo espectral

Figura 54: Flujo espectral para canción de alegría

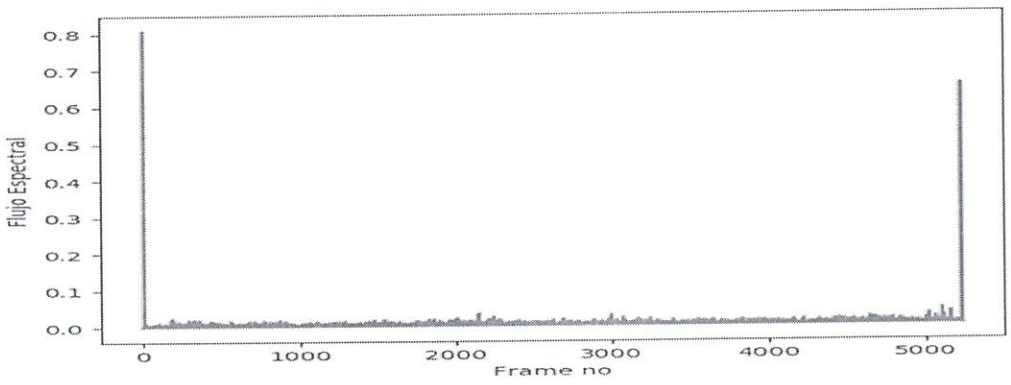


Figura 55: Flujo espectral para canción de tristeza

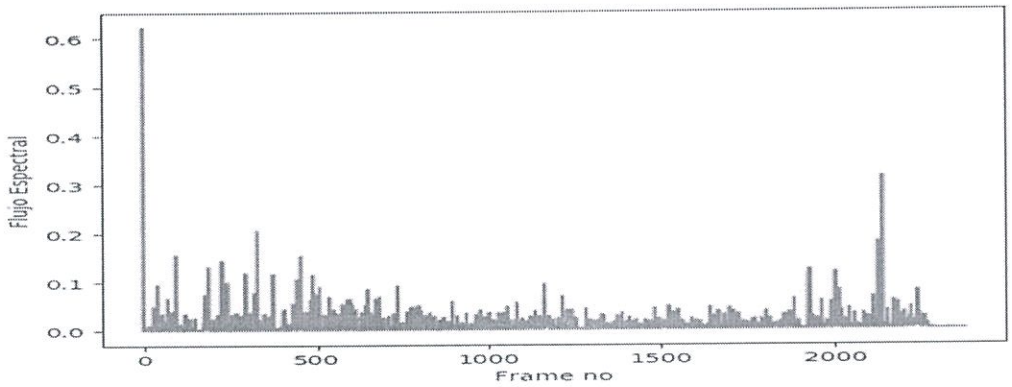


Figura 56: Flujo espectral para canción de enojo

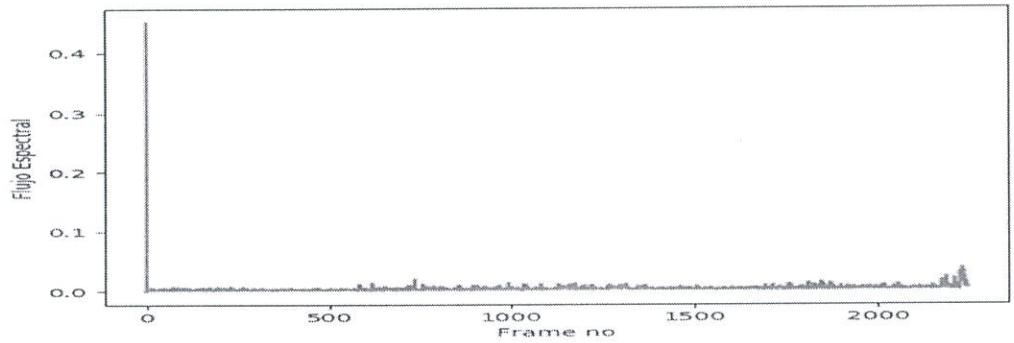
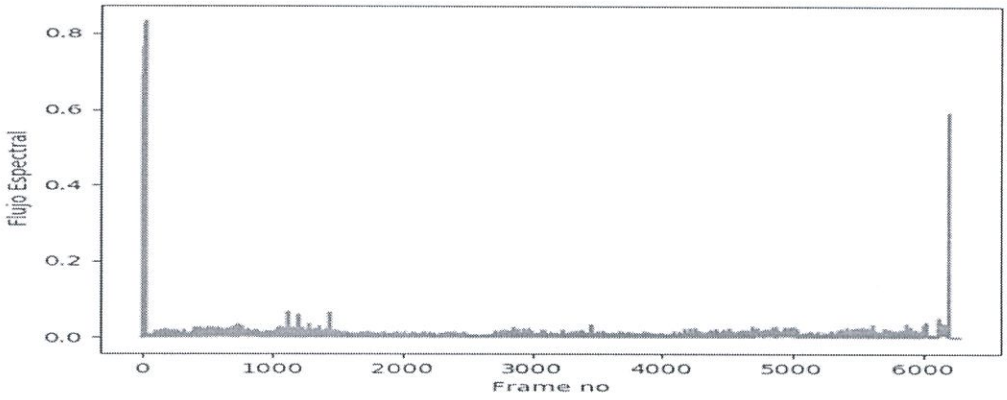


Figura 57: Flujo espectral para canción de miedo



7. Rodamiento espectral

Figura 58: Rodamiento espectral para canción de alegría

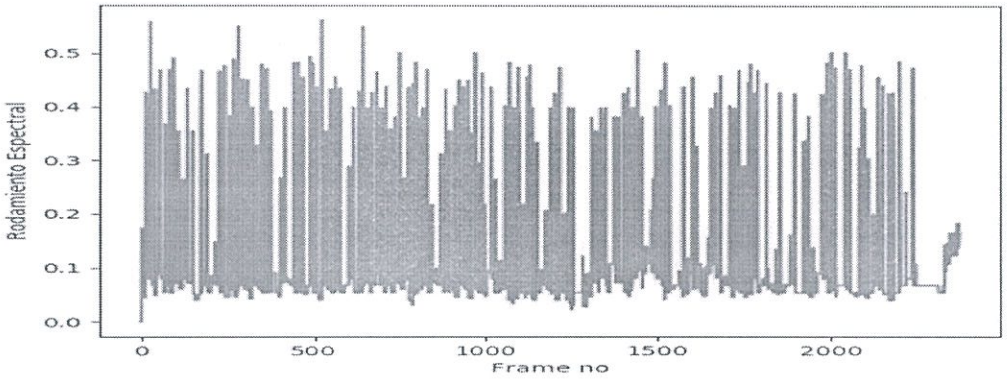


Figura 59: Rodamiento espectral para canción de tristeza

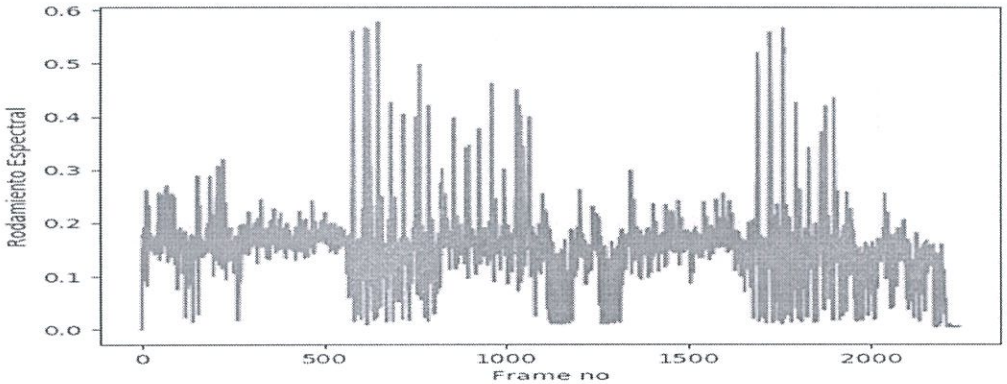


Figura 60: Rodamiento espectral para canción de enojo

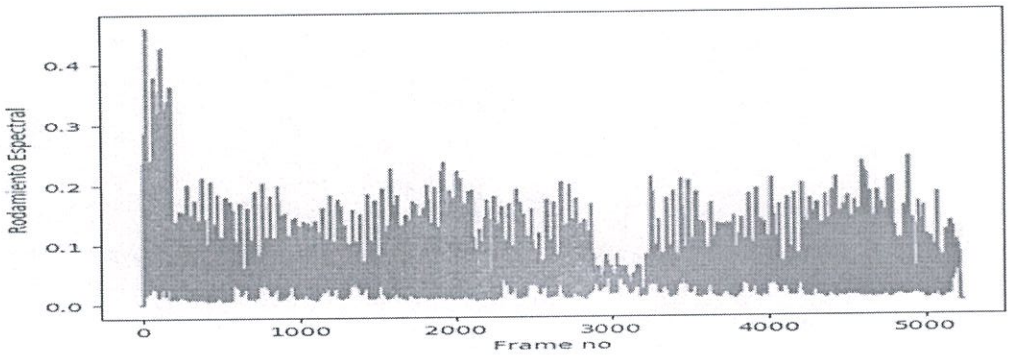
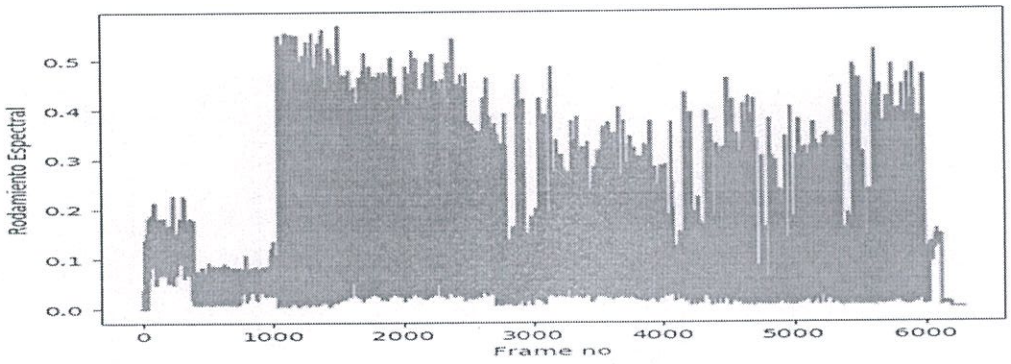


Figura 61: Rodamiento espectral para canción de miedo



8. Coeficientes espectrales de Mel

Figura 62: Coeficientes espectrales de Mel para canción de alegría

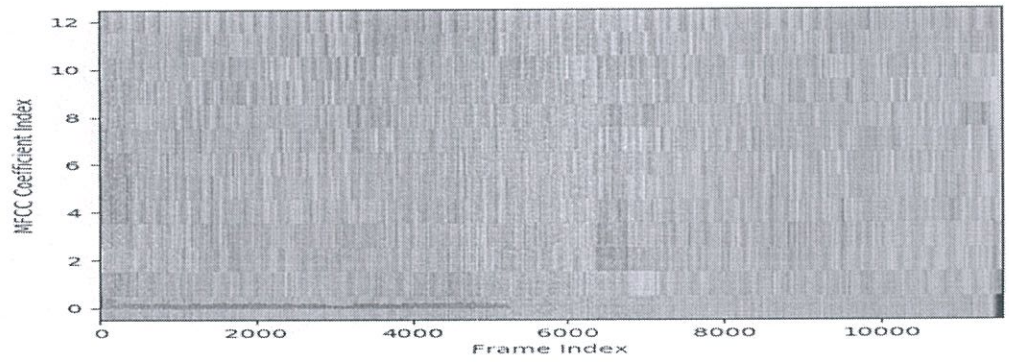


Figura 63: Coeficientes espectrales de Mel para canción de tristeza

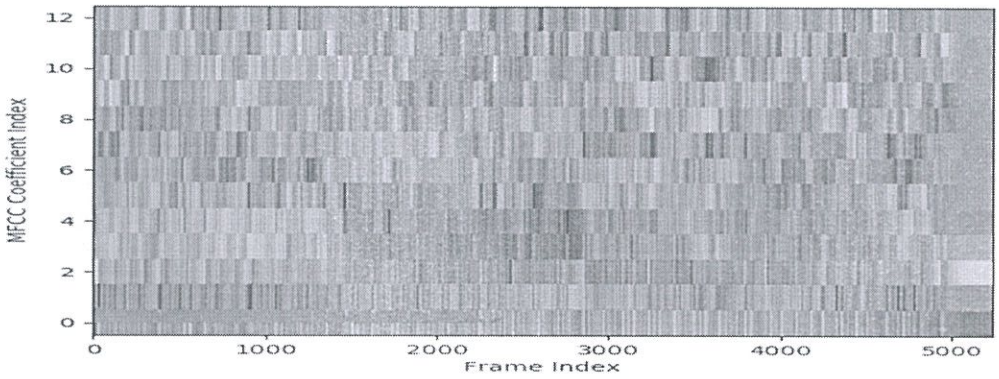


Figura 64: Coeficientes espectrales de Mel para canción de enojo

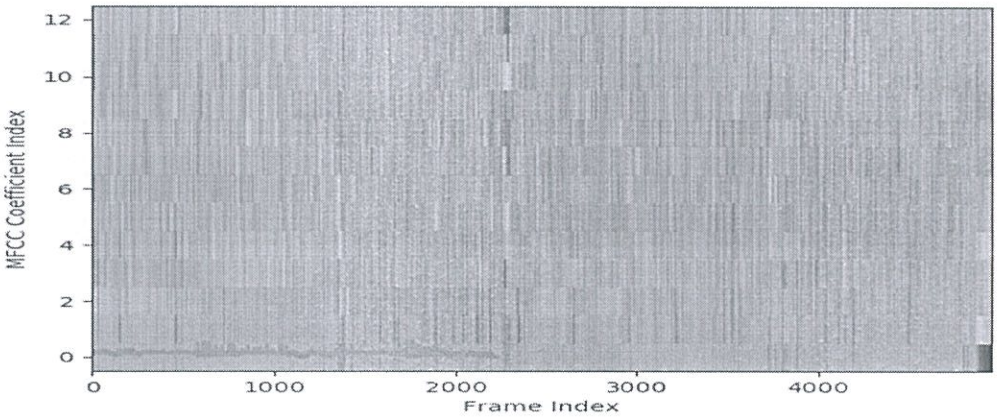
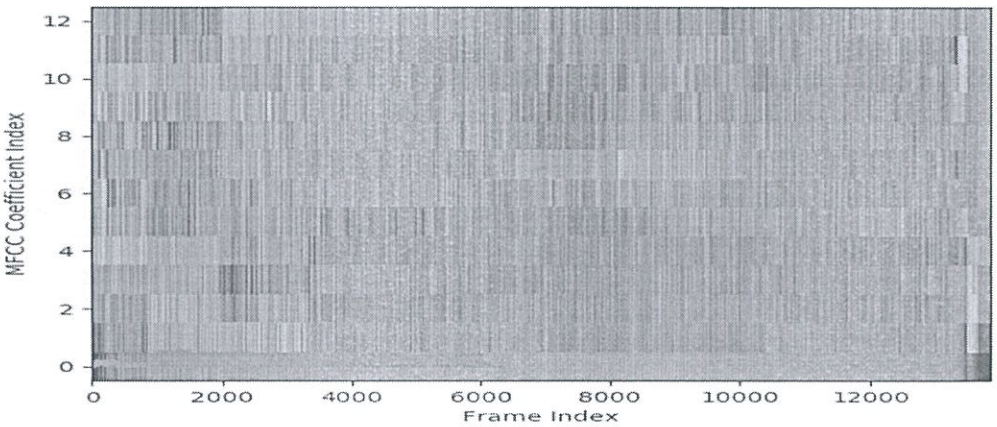


Figura 65: Coeficientes espectrales de Mel para canción de miedo



9. Vectores cromáticos (12 coeficientes basados en el sistema tonal occidental)

Figura 66: Vectores cromáticos para canción de alegría

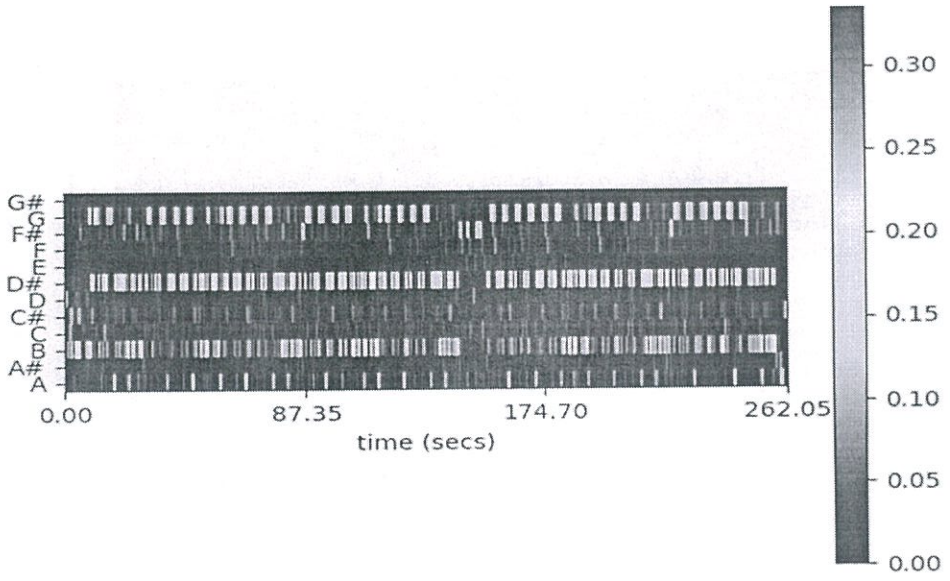


Figura 67: Vectores cromáticos para canción de tristeza

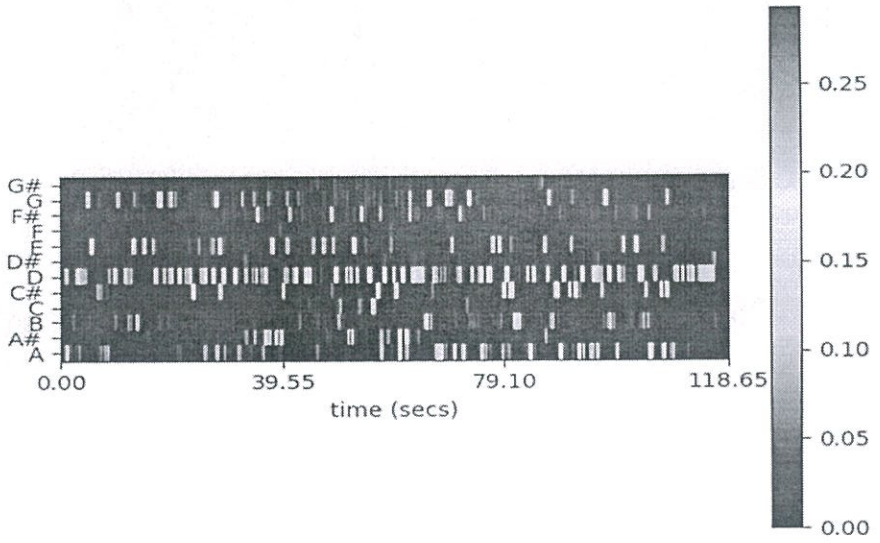


Figura 68: Vectores cromáticos para canción de enojo

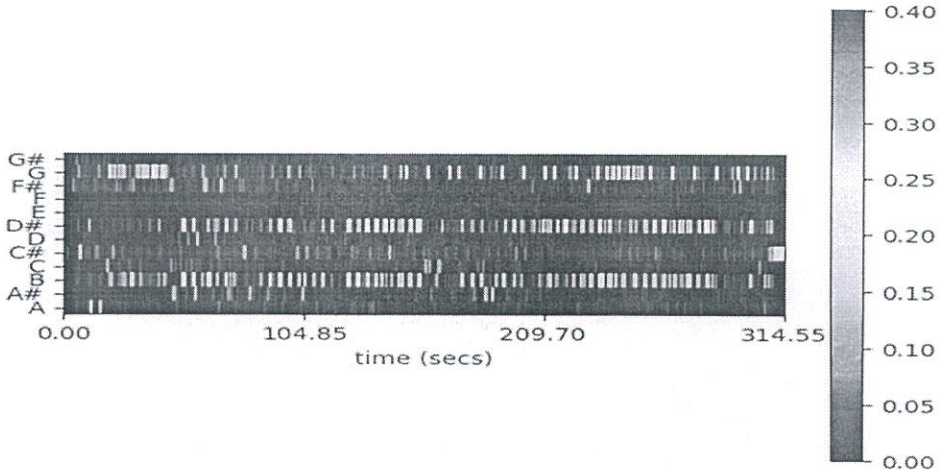
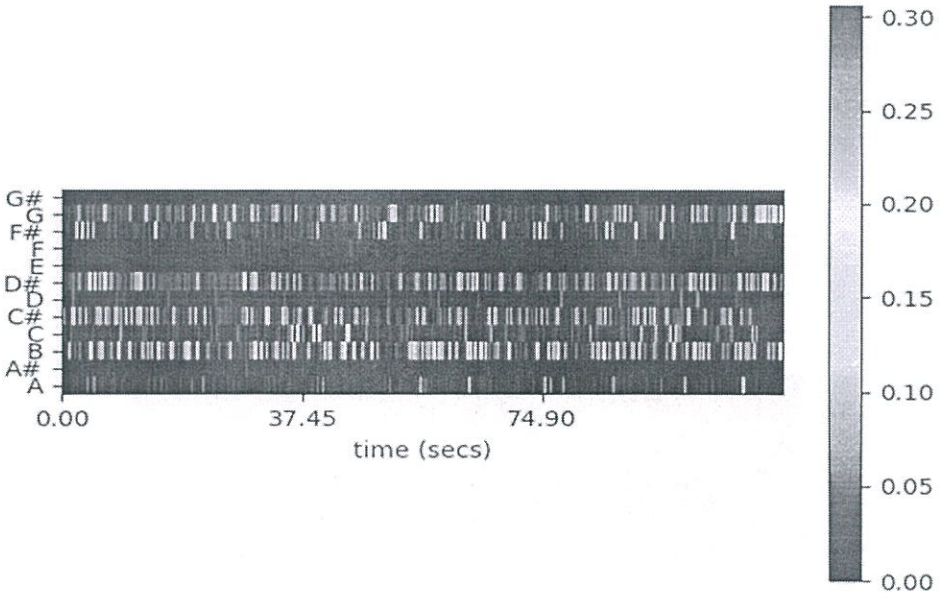


Figura 69: Vectores cromáticos para canción de miedo



D. Resultados de algoritmos de clasificación

Cuadro 1: Desempeños con K vecinos más cercanos

Características Extraídas	Cantidad de marcos de audio	Desviación Estándar	Parámetros adicionales del clasificador	Precisión	Clasificaciones correctas
Todas	Todos los Marcos de audio	Sí	K=5, algorithm='Auto'	76%	66%
Todas	Promedio del tiempo de escucha	Sí	K=5, algorithm='Auto'	76%	69%
Todas	Promedio del tiempo de escucha	Sí	K=5, algorithm='Auto' + Clases Balanceadas	55%	78%
Todas	Todos los marcos de audio	Sí	K=5, algorithm='Auto' + Clases Balanceadas	58%	70%
MFCC+ Vectores Cromáticos	Promedio del tiempo de escucha	Sí	K=5, algorithm='Auto'	75%	64%
MFCC+ Vectores Cromáticos	Todos los marcos de audio	Sí	K=5, algorithm='Auto'	73%	68%
Todas	Promedio del tiempo de escucha	No	K=2, algorithm='Auto'	79%	70%
Todas	Promedio del tiempo de escucha	NO	K=3, algorithm='Auto'	81%	68%
MFCC	Promedio del tiempo de escucha	NO	K=3, algorithm='Auto'	76%	66%

Cuadro 2: Desempeños con máquina de vectores de soporte

Características extraídas	Cantidad de marcos de audio	Desviación Estándar	Parámetros adicionales del clasificador	Precisión	Clasificaciones correctas
Todas	Todos los marcos de audio	Sí	kernel='rbf'	57%	64%
Todas	Todos los marcos de audio	No	kernel='rbf'	58%	66%
Todas	Promedio de escucha	No	kernel='sigmoid'	45%	50%
Todos características de bajo nivel	Promedio de escucha	No	kernel='polynomial' grado=3	58%	68%
MFCC + vectores cromáticos	Promedio de escucha	No	kernel='rbf'	67%	56%
MFCC	Promedio de escucha	No	kernel='rbf'	65%	55%
				64%	56%

Cuadro 3: Desempeños con bosques aleatorios

Características extraídas	Cantidad de marcos de audio	Desviación Estándar	Parámetros adicionales del clasificador	Precisión	Clasificaciones correctas
TODAS	Promedio de escucha	Sí	40 Árboles	80%	73%
TODAS	Promedio de escucha	Sí	40 Árboles	72%	66%
TODAS	Promedio de escucha	No	40 Árboles	77%	66%
TODAS	Todos los barcos de audio	Sí	40 Árboles	71%	71%
TODAS	Todos los marcos de audio	No	40 Árboles	73%	66%
Bajo Nivel	Todos los marcos de audio	Sí	40 Árboles	63%	70%
Bajo Nivel	Todos los marcos de audio	No	40 Árboles	63%	70%
Bajo Nivel	Promedio de escucha	No	40 Árboles	67%	63%
MFCC + Vectores Cromáticos	Todos los marcos de audio	Sí	40 Árboles	71%	66%
MFCC + Vectores Cromáticos	Promedio de escucha	Sí	40 Árboles	61%	69%
MFCC	Todos los marcos de audio	Sí	40 Árboles	66%	63%
MFCC	Promedio de escucha	Sí	40 Árboles	66%	63%

VII. Análisis de resultados

Tras completar las clasificaciones de las canciones para la creación del conjunto de datos de entrenamiento, se obtuvieron varias estadísticas sobre los participantes que se involucraron en el estudio. En primer lugar, se puede observar que, aunque hubo una cantidad mayor de participantes mujeres, en general el balance entre mujeres y hombres fue adecuado por lo que se puede descartar un género ligado específicamente a género en la clasificación de datos (Figura 21). Con base en la Figura 23, también es notable que las edades se concentraron en personas jóvenes menores a los 35 años y la menor cantidad de personas se encuentran en los extremos, es decir, adultos mayores y niños. Estos rangos de edades podrían darle un sesgo al clasificador a las percepciones de jóvenes y adultos jóvenes, aunque el hecho de pertenecer al mismo país y estar inmersos en la misma cultura reduce los problemas que podrían causar estos aspectos. Otro aspecto importante dentro de las estadísticas de los participantes es el hecho que los participantes con problemas psicológicos y/o auditivos no superan a los 20 participantes, por lo que las clasificaciones de emociones que pudiesen estar sesgadas por este aspecto son mínimas y se descartan al asegurarnos que una misma canción es clasificada por al menos tres personas diferentes.

Es importante mencionar que todas las estadísticas y datos de contacto tomados de los participantes son parte del conjunto de datos y son proporcionados en la base de datos construida para futuros estudios y realizar análisis psicológicos o antropológicos más profundos del conjunto de participantes. Por otro lado, se debe notar que la mayoría de participantes tenían un nivel de estudio de Bachillerato o superior a este por lo que la mayoría de clasificaciones no fueron realizadas por personas con poca preparación académica. Este aspecto podría sesgar un poco el clasificador al dejar de un lado las personas sin preparación académica. A pesar de eso, como ya se ha mencionado, en la mayoría de casos los estudios apuntan a que la percepción de emociones en la música es afectada principalmente por diferencias culturales y no se menciona nada acerca de aspectos académicos del individuo, aunque no se descarta por completo el este hecho. Finalmente, es importante notar que el tiempo de escucha promedio denota que las personas no necesitan escuchar una canción completa para poder emitir un juicio sobre la emoción que la canción presenta.

Pasando al análisis de las canciones en sí mismas, podemos observar que gran parte de las canciones obtenidas con derechos de libre distribución y uso son de género electrónico, como lo muestra la Figura 24. Esta cantidad se debe sencillamente a la disponibilidad de canciones encontradas en YouTube que poseían libre distribución y marca claramente la tendencia que existe en la actualidad sobre la producción de música electrónica y la gran variedad de música que se puede encontrar dentro de este mismo género. El resto de géneros se ven bastante balanceados, y el hecho de conseguir ocho géneros distintos para poder ser analizados fortalece la generalidad de los resultados de los modelos entrenados para la cultura guatemalteca. Esto es debido a que la variedad sonora entre estos géneros nos proporciona muchísimos estilos, matices, progresiones armónicas y melódicas características de cada género para poder nutrir al clasificador con información diversa y evitar sobre ajustar el modelo a un solo estilo de música.

Si analizamos los resultados de las clasificaciones de los participantes en las Figuras 26 y 27 podemos notar que la mayoría de canciones fueron clasificadas como canciones de alegría, seguido de tristeza, enojo y miedo. Esta cantidad de clasificaciones se mantiene en casi todos los géneros a excepción del rock, en donde existe igual número de canciones clasificadas como de enojo y alegría. El género ambiente también fue una excepción, en donde la mayoría de canciones se clasificaron como canciones de tristeza. Estos resultados nos pueden dar una idea de que las emociones expresadas por la música no necesariamente van estar relacionadas a los géneros, y que a pesar de que una pieza tenga ciertas características sonoras que las clasifica dentro de un género, existen otros elementos sonoros más sutiles existentes dentro de los mismos géneros que hacen que las personas clasifiquen una canción en alguna emoción.

La Figura 28, nos muestra los tiempos de escucha de los participantes en las clasificaciones de las canciones y muestran un promedio de 40.7 segundos. Este resultado nos indica que las personas no necesitan escucha una pieza musical completa para poder emitir un juicio sobre la emoción que expresa. Además, esto puede hacer pensar que este tiempo es el tiempo necesario para extraer toda la información relevante de una canción para poder crear un modelo de clasificación de emociones. Sin embargo, es importante menciones que, aunque la mayor cantidad de música mantiene o desarrolla una misma idea musical durante toda la pieza, existen piezas que cambian dramáticamente a medida que se desarrollan y podrían resultar en cambios en los juicios de las personas sobre la emoción que expresan. Estas piezas podrían dar problemas si utilizamos un clasificador que utiliza solo el tiempo promedio de escucha. A pesar de esto, los resultados presentados apuntan a que utilizar solo el tiempo promedio de escucha mejora el rendimiento del clasificador y puede eliminar datos que causen ruido en la data de entrenamiento ya que no fueron procesadas por los participantes y por tanto no fueron tomadas en cuenta para emitir su juicio sobre la emoción de la canción.

Para poder ilustrar de mejor manera las ideas de las características extraídas se eligieron cuatro canciones de cada emoción y se presentan las gráficas de cada característica en las Figuras 29 a 68. Aunque no todas las canciones tendrán precisamente las mismas diferencias que las cuatro elegidas, estas fueron elegidas por ejemplifican de mejor manera como cada característica aporta información que ayuda a diferenciar las canciones entre las distintas emociones. Si a estas características añadimos estadísticas como la desviación estándar, tenemos aún más datos para poder diferencias entre cada emoción al momento de analizar el contenido de audio de una canción. Aunque en algunos casos las diferencias no son tan perceptibles en una gráfica, se interpretaran las posibles diferencias entre cada emoción para cada una de las características y como estas aportan en aspectos que resultan ser diferenciadores para la percepción humana del sonido y su relación con las emociones expresadas en una canción.

La tasa de cruzamiento de ceros para las emociones de alegría, tristeza, enojo y miedo (Figuras 29, 30, 31 y 32), presentan varios aspectos que da algunos factores distintivos entre las emociones. Si comparamos las emociones de alegría y tristeza podemos observar claramente que la emoción de tristeza posee valores mucho más altos que los valores de la emoción de alegría, llegando hasta 0.35 para tristeza y valores cercanos a 0.20 para alegría. Estas diferencias nos indican que la emoción de tristeza posee frecuencias ligeramente

más altas que la canción de alegría. Las emociones de enojo y miedo poseen valores similares a los valores de tristeza. Sin embargo, es bastante notable que la tasa de cruzamiento de ceros en la emoción de miedo posee cambios abruptos bastante notables que pueden ser importantes para diferenciar esta emoción del resto. Esta característica entonces muestra valores bajos para la emoción de alegría, valores más altos para la tristeza y el enojo y una alta dispersión para la emoción de enojo. Es importante notar que, aunque los valores medios mostrados en las gráficas pueden ser similares en la emoción de enojo y tristeza, si añadimos la dispersión al análisis de las gráficas, la diferenciación se puede volver más clara. Todas estas diferencias aportan elementos para poder empezar a distinguir las cuatro clases en el modelo a de clasificación a construir.

Pasando al análisis de la energía cuadrática media, en las Figuras 33, 34, 35 y 36 podemos notar claras diferencias entre las emociones de tristeza/miedo con las emociones de alegría/enojo. Si observamos las canciones de alegría y enojo, podemos observar valores entre 0.15 y 0.175. Estos valores son considerablemente más altos si los comparamos con los valores de las emociones de tristeza/miedo, los cuales no superan a 0.1 en magnitud. Hay que notar que estos valores bajo son evidentes solamente en el inicio de la canción de miedo, luego de algún tiempo, la energía cuadrática media se incrementa a valores bastante similares a los de alegría y enojo. Estos cambios abruptos en la energía cuadrática media pueden ser un elemento clave en las emociones de miedo, ya que los cambios abruptos en la energía de una pieza pueden estar asociado a elementos de sorpresa, los cuales son utilizados comúnmente por compositores para generar emociones de miedo y mantener suspenso en los oyentes. Estas diferencias en la energía cuadrática media pueden llevar a inferir que la magnitud energía cuadrática media puede ser un buen factor para diferenciar la tristeza del resto de emociones y que la desviación estándar podría ser un elemento importante para diferenciar la emoción de miedo las otras tres emociones.

Pasando al análisis de la entropía de la energía, se podrá analizar de mejor manera la relación de los cambios en la energía y cada una de las emociones. Con base en las Figuras 37, 38, 39 y 40 se puede notar un claro factor diferenciador entre la gráfica de la emoción de miedo y la del resto de emociones. Si se observa con detenimiento la gráfica de la emoción del miedo, se puede encontrar que la cantidad de picos en los valores bajos y casi nula en la mayor parte de la canción. En cambio, tanto para la alegría, la tristeza y el enojo, se pueden encontrar picos en valores bajos de la entropía que están en los rangos de 1 a 1.5. Este comportamiento nos muestra que la alta entropía podría ser un factor diferenciador para la emoción del miedo de las emociones de alegría, tristeza, enojo y miedo. A pesar de esto, la entropía podría no ser muy útil para diferenciar entre alegría, tristeza y enojo ya que los valores de las magnitudes de estas son bastante parecidos, aunque ligeramente mayores para la emoción del enojo. Por tanto, la entropía es una característica que puede ser clave para la diferenciación de la emoción del miedo, pero que por sí sola, no podría realizar un buen trabajo para discernir entre las cuatro emociones. Esta característica en conjunto con las previamente analizadas puede empezar a dar una mejor idea de cómo distinguir entre cada emoción Otra manera de ver el análisis de la entropía de la energía es a través de la dispersión de los datos. Las emociones de alegría, tristeza y enojo tendrán una dispersión más alta que la emoción del miedo, ya que las magnitudes de la entropía del

miedo se mantienen en niveles altos durante la mayor parte de la pieza, el resto de emociones presentan picos en valores bajos en repetidas ocasiones, lo cual provoca una mayor dispersión y por tanto valores de desviación estándar un tanto más altos que los que se podrán presentar en la emoción del miedo.

Todas las características analizadas hasta el momento se relacionan principalmente con los niveles de energía del audio. Este elemento va relacionado a características de alto nivel como la intensidad del sonido, niveles de ruido y matices en la música. Sin embargo, hay muchísima más información que se puede extraer de analizar los comportamientos de las frecuencias de onda en las piezas. La frecuencia puede ser la base para construir ideas de alto nivel como la tonalidad, armonía, consonancia y disonancia. Las siguientes características que se analizaran se enfocan más en estos aspectos y en conjunto con las previamente analizadas, dan mayor riqueza de información para poder construir un modelo robusto que logre clasificar de manera exitosa las canciones en las 4 distintas emociones.

Entre las características de frecuencias de bajo nivel, encontramos el centroide espectral en las Figuras 41, 42, 43 y 44. Esta característica nos da indicios sobre el timbre de una canción y como se puede observar en las figuras, puede ser una característica que diferencia claramente las 4 emociones. Si observamos la gráfica de alegría, notamos que las magnitudes de esta característica van desde 0.10 hasta 0.25, siendo una de las emociones con más dispersión en sus magnitudes y con magnitudes más altas. La tristeza por otro lado, presenta magnitudes considerablemente más bajas entre 0.02 y 0.05, asimismo la dispersión de esta emoción también es muchísimo más baja que la de la alegría. La emoción de enojo se mantiene en valores medios que van entre 0.15 hasta 0.20, pero también encontramos algunos picos que llegan hasta 0.35. Finalmente, la emoción del miedo nos muestra una alta dispersión, con magnitudes que van desde 0.1 hasta 0.30. Como se puede ver esta característica puede ser clave para la correcta diferenciación de los 4 géneros, ya que posee comportamientos característicos que pueden diferenciarse fácilmente al observar la gráfica. Es importante mencionar que a pesar de que en este caso el centroide espectral muestra diferencias claras para cada una de las 4 emociones en las elecciones de canciones para el análisis, esto podría no ser cierto en otras elecciones. Sin embargo, es este tipo de análisis el que puede ser crucial para la correcta elección de características y para desarrollar una intuición sobre el comportamiento de los datos y la relación entre lo que se percibe a nivel sonoro y lo que se muestra a nivel numérico en el análisis de la señal de audio.

La siguiente característica es el ancho de banda, presentadas en las Figuras 45, 46, 47 y 48. Esta característica puede llegar a darnos una intuición sobre la variación en las frecuencias de una pieza y por tanto puede presentar con valores altos en piezas que tengan ideas musicales cambiantes y dinámicas o armonías dinámicas, como es el caso de la música jazz. Si observamos la emoción de alegría, los rangos de valores de magnitud son bastantes similares a los de la emoción de miedo y tristeza. El enojo se diferencia claramente del resto de las emociones al tener valores que están entre 0.15 y 0.20 en la mayoría de marcos de audio de la pieza. La alegría, el enojo y la tristeza en cambio, poseen magnitudes que oscilan entre 0.05 y 0.25. La emoción de miedo posee un rango de valores mucho más pequeño que el resto de valores por lo que

podemos deducir que el rango de variación de frecuencias se mantiene alta en todo momento, mientras que el resto de emociones tienen momento donde las variaciones se reducen.

El análisis de la entropía espectral nos da una idea de la variación en la energía en un sistema de audio. En este caso, mientras más grande sea el número, encontramos una menor entropía. Si observamos las figuras 49, 50, 51 y 52, notaremos que la entropía espectral de la alegría tiene un rango de valores entre 0.1 y 1.5, diferenciándose de todas las demás emociones. La entropía espectral de la tristeza oscila en valores que van desde 0.1 hasta 2.0 mostrando un rango de valores bastante más amplio que la emoción de alegría. La entropía espectral de enojo se mantiene más alta que el resto de las emociones en todo momento, teniendo un valor mínimo de 0.5 y máximo de hasta 2.5. La emoción de miedo posee magnitudes bastante similares a la emoción de tristeza, aunque los valores máximos llegan hasta 2.5. Según estas gráficas, podemos decir que las entropías espectrales pueden diferenciar bastante bien la emoción de enojo del resto de emociones, ya que tendrá un valor promedio superior al resto. La desviación estándar de esta característica puede ser de utilidad para diferenciar entre todas las emociones ya que los rangos de valores de cada emoción varían considerablemente y pueden ser fáciles de diferenciar.

Los valores observados en la entropía espectral pueden interpretarse de mejor manera si son relacionados con los audios de las piezas. La canción de enojo posee sonidos mucho más estridentes, por lo que es razonable observar una entropía mayor ya que las variaciones en la energía de la pieza son mucho más drásticas. La canción de tristeza es mucho más suave y con matices mucho más sutiles, por esta razón podemos observar una entropía que tiene valores mínimos más bajos que el resto de emociones. La emoción de alegría tiene a ser música mucho más predecible y común para las personas, por lo que la entropía es menor en este caso. Finalmente podemos observar una entropía más alta en la emoción de miedo por los mismos efectos de suspenso y sorpresa que este tipo de música intentan crear, aunque sin llegar a los mismos niveles que la emoción del miedo.

Pasando al análisis del flujo espectral en las Figuras 53, 54, 55 y 56, debemos recordar que esta característica busca relacionar las diferencias en todos los posibles valores de frecuencias entre dos marcos consecutivos. La emoción que más destaca al analizar estas gráficas es la emoción de tristeza, ya que posee valores considerablemente más altos que el resto de emociones. Esto nos indica que esta emoción presenta cambios fuertes en gran parte del espectro de frecuencias entre cada marco de audio. Las tres emociones restantes no presentan mayor diferencia entre ellas, por lo que el flujo espectral podría ser útil para diferenciar únicamente la tristeza del resto de emociones en base a las canciones analizadas. Hay que notar que para las cuatro emociones se observan dos grandes picos tanto al inicio como al final de la pieza, esto se debe a que por la naturaleza de esta característica las diferencias en el espectro frecuencia cuando empieza la canción y cuando terminan son iguales a las frecuencias en los primero y últimos marcos de audio, ya que antes y después de estos marcos no existe presencia alguna de frecuencias.

El rodamiento espectral, en las figuras 57, 58, 59 y 60, muestra valores bastante estables entre 0.01 y 0.2 para la emoción de alegría. Los rangos de valores de la emoción de enojo presentan magnitudes similares

a la emoción de alegría, pero con picos repetidos que pueden llegar hasta 0.5. Estos picos pueden representar la variabilidad del timbre entre la música asociada a estas dos emociones, ya que tener picos en el rodamiento espectral se relaciona directamente espectros de frecuencia más amplios y por tanto sonidos con timbres más estridentes. El rodamiento espectral entre la emoción de tristeza y miedo es mucho más más similar por lo que en este caso se puede decir que los timbres de estas dos emociones pueden tener algo de parecido entre ellos.

Pasando a los Coeficientes Espectrales de Mel, en las figuras 61, 62, 63 y 64, se puede observar que, aunque las gráficas no pueden dar mucha información más que algunos patrones en la aparición de ciertas frecuencias, con base en los resultados de la precisión de los modelos de los cuadros 1 al 3, los coeficientes de Mel son una de las características que más aportan a elevar las precisiones de los modelos de clasificación. Por esta razón se considera una de las características más valiosa en la clasificación de música por emociones. Algunas de las razones que pueden llevar a esta idea es la gran relevancia que se le da a la forma en que se escucha el ser humano. El proceso de obtener el poder espectral de cada marco de audio se relaciona con la manera en que se comporta la cóclea del ser humano. Esta vibra en diferentes lugares dependiendo de la frecuencia a la que es expuesta de manera similar a como se muestran las magnitudes de cada frecuencia al realizar en el períodograma. Aunque este resultado posee una gran cantidad de información, tiene demasiada información que no es percibida por la cóclea del ser humano. En particular, la cóclea tiene dificultades para diferenciar entre dos frecuencias muy cercanas. Esta es la razón por la que se definen pequeños contenedores que corresponden a cada una de las doce características de los coeficientes de Mel, donde cada coeficiente va a representar cuanta energía existe en una región de frecuencias determinada, de manera similar a la cóclea. Esta división se define con la ayuda del banco de filtros de Mel en la Figura 8 la cual representa claramente como la percepción de diferencia entre frecuencias es más pronunciada en frecuencias bajas y menos notable en las frecuencias altas.

Luego de esto, el cálculo del logaritmo también es motivado por la forma en que el ser humano escucha. Específicamente se puede notar en las Figura 2 como el comportamiento de la percepción del sonido no es lineal, sino tiene un comportamiento más cercano al de un logaritmo. Finalmente, la aplicación de la Transformada Discreta de Fourier es principalmente porque el banco de filtros de Mel se traslapan y esto hace que las energías calculadas entre los filtros este muy correlacionadas. Esta transformación ha ayudado a reducir la correlación y es por esta misma razón que en la mayoría de aplicaciones de esta característica también se seleccionan solo 12 de los 26 coeficientes.

La última característica es la de Vectores Cromáticos, en las figuras 65, 66, 67 y 68. Esta característica es de gran importancia para empezar a relacionar características de alto nivel como la tonalidad, la armonía e intervalos con las emociones asociadas.

Pasando a los resultados de las precisiones de cada modelo entrenado en los Cuadros 1,2 y 3, se puede observar que las precisiones de los modelos llegaron a un máximo de 81% y un 73% de clasificaciones correctas con los datos de prueba. Si se observa el Cuadro 1, el modelo utilizando K Vecinos más Cercanos,

se puede observar que las precisiones más altas se obtienen si se elimina la desviación estándar. Esta mejora puede estar relacionada a que el algoritmo de K Vecinos más cercanos toma en cuenta el concepto de distancia. Esto puede ser problemático si tenemos datos con mucho ruido ya que pueden existir características que tengan ruido cercano a los datos de prueba y por tanto dar clasificaciones erróneas. Al remover la desviación estándar podemos estar eliminando este ruido y aumentar la precisión para este algoritmo en específico. Además, hay que notar que la precisión también se incrementa al tomar en cuenta solo los marcos de audio dentro del tiempo de escucha promedio. Esto tiene sentido debido a que las personas no emitieron opinión sobre el resto de la canción, y puede ser posible que las partes que no escucharon hubiesen hecho que su juicio sobre la emoción de la canción sea distinto. Por tanto, agregar la pieza completa al modelo de entrenamiento puede resultar en información que sea tomada como ruido en lugar de ayudar a una mejor clasificación de emociones. Este modelo tuvo un desempeño máximo de 79% con un 70% de las piezas de prueba clasificadas correctamente.

El modelo de máquina de vectores de soporte, presentado en el Cuadro 2, presentó los resultados menos satisfactorios de los tres modelos construidos. Con un desempeño máximo de 57% y 68% de clasificaciones correctas, este modelo también muestra mejorías cuando se agrega el promedio de tiempo de escucha y se elimina la desviación estándar de las características extraídas. Las razones por las que este modelo puede estar teniendo los desempeños más bajos pueden ser debido a que la máquina de vectores de soporte fue diseñada inicialmente para clasificaciones binarias y sets de datos lineales. Se tuvieron que agregar elementos como los kernels y reducción de dimensiones para lidiar con problemas de más de dos clases. En algunos casos los kernels pueden ser inadecuados para los datos de entrenamiento y puede ser necesario estandarizar los datos a cierta escala, lo cual puede ser complicado cuando las características son distintas entre ellas. En los resultados se puede observar que el kernel que mejor dio resultados es el kernel polinomial, esto nos indica que el comportamiento de los datos no es lineal. Por otro lado, este modelo es uno de los más lentos para entrenar por lo que no es un algoritmo que pueda escalar para librerías de millones de canciones como encontramos hoy en día en las plataformas de música más importantes. Esto sugiere que este es uno de los modelos menos recomendados para la clasificación de música por emociones.

El modelo de bosques aleatorios, en el Cuadro 3, muestra los mejores desempeños encontrados con un máximo de 80% de desempeño con un 73% de las piezas de prueba clasificadas correctamente. Las ventajas que tiene el algoritmo de bosques aleatorios sobre el resto de modelos de clasificación es que se comporta bastante bien con valores en distintas escalas y también puede lidiar con datos con patrones no lineales e inclusos datos que no son numéricos. A esto podemos agregar la alta velocidad con la que se construye el modelo a comparación de las máquinas de vectores de soporte. También es importante notar que, en este caso, la desviación estándar ayudo a elevar el desempeño del algoritmo. Un argumento que puede explicar este hecho es que los bosques aleatorios toman características aleatorias, y cada árbol generado pueden estar anti correlacionados entre sí. Sin embargo, al generar varios árboles con varias características combinadas y promediarlos, resulta ser que el modelo se acerca bastante al comportamiento de los datos de entrenamiento.

Esta ventaja permite que los bosques aleatorios puedan comportarse bien a pesar de que existan algunos datos que tengan comportamientos atípicos. Este modelo presentó la mayor cantidad de ventajas y por tanto es uno de los más recomendados para realizar clasificación de música en base a emociones utilizando únicamente el contenido de audio de la pieza como fuente de datos de entrenamiento.

Para concluir el análisis, vale la pena mencionar que, en todos los casos, tomar en cuenta únicamente el promedio de escucha de las canciones ayudó a elevar los desempeños de los modelos. Además, hay que notar que en los tres casos se crearon modelos utilizando únicamente los coeficientes de Mel y los vectores cromáticos. En estos casos, los desempeños de los modelos se redujeron drásticamente, por lo obtenemos un buen indicador de que estas dos características en específico son de gran importancia para la clasificación de música por emociones. El hecho de que utilizan ideas que se acercan bastante a las ideas musicales y a la percepción humana del sonido, permite que los modelos se creen de una manera cercana a la que los humanos relacionan la música con las emociones y elevan significativamente los resultados de las clasificaciones. Sin embargo, no se descartan las características de bajo nivel ya que elevan aún más los desempeños y son lo suficientemente genéricas como para poder aplicarse a distintos géneros musicales, mientras que algunas características de nivel medio pueden no aplicar o dar resultados negativos con ciertos estilos o géneros de música.

VIII. Conclusiones

Se logró construir un conjunto de datos de entrenamiento de 860 canciones de distintos géneros clasificadas únicamente por guatemaltecos para crear los criterios de diferenciación de los modelos de clasificación automática de música en base a emociones. Además, se construyeron distintos modelos de aprendizaje de máquina utilizando los algoritmos de K Vecinos Más Cercanos, Máquinas de Vectores de Soporte y Bosques Aleatorios, utilizando un vector de 34 características de audio junto con la desviación estándar de cada característica. El mejor desempeño lo tuvo el modelo de Bosques Aleatorios utilizando los marcos de audio del tiempo de escucha promedio e incluyendo la desviación estándar de las características de cada marco de audio. Este modelo resultó con un desempeño del 80% y un 73% de las canciones de prueba clasificadas correctamente.

Además, se lograron aplicar algoritmos de extracción de características de audio de bajo nivel y nivel medio. Entre estos la tasa de cruzamiento de ceros, coeficientes espectrales de Mel, vectores cromáticos, flujo espectral, centroide espectral, ancho de banda y rodamiento espectral para la creación de los modelos presentados.

Finalmente se concluye que las características de bajo nivel en conjunto con las de nivel medio combinadas ayudan a elevar el desempeño del algoritmo, pero que las características principales y más efectivas para la clasificación de música por emociones son los Coeficientes Espectrales de Mel y los Vectores Cromáticos, ya que toman en cuenta muchos elementos de la percepción humana de la frecuencia y la teoría musical del sistema occidental.

IX. Recomendaciones

En cuanto a la elección de características del audio, se recomienda experimentar y explorar otras características de bajo nivel y nivel medio con las utilizadas en este trabajo para intentar elevar los desempeños de clasificación. Por otro lado, se puede realizar un análisis más profundo por género y aplicar características de nivel medio sabiendo de antemano que se está trabajando bajo un género musical específico, ya las características de nivel medio funcionan mejor si sabemos de alguna manera el tipo de música que se analizará. Característica como las progresiones armónicas, bailabilidad, e inclusive las letras de las canciones pueden aportar en la correcta diferenciación de música por emociones. También se recomienda explorar los otros acercamientos a la clasificación de música, utilizando elementos contextuales del usuario y propiedades específicas del usuario quien escucha la música. Esto puede ayudar a crear clasificadores mucho mejor adaptados a las necesidades personales de cada usuario y al lugar y tiempo en el que se encuentre el mismo al momento de estar explorando u buscando la música.

En cuanto al uso de modelos, se recomienda explorar otros algoritmos como las redes neurales, estimaciones de Kernels o Naive Bayes. Además, puede ser muy útil explorar las técnicas de Aprendizaje Profundo con herramientas como Tensorflow para la construcción de redes neuronales recurrentes, ya que se ha visto en los últimos meses grandes avances en la resolución de problemas utilizando este tipo de técnicas cuando la cantidad de datos y variables es muy amplia. Finalmente se recomienda unir esfuerzos con ramas de psicología y antropología para poder incluir características relacionadas con la cultura del individuo o las expresiones faciales del mismo, de manera que se puedan relacionar factores biométricos en los datos de entrenamiento para explorar clasificadores que incluyan este tipo de elementos

X. BIBLIOGRAFÍA

Ali, N. 2010. Generalized Discrete Fourier Transform with nonlinear Phase, IEEE Transactions on Signal Processing, 58(9) p.1-10

Argstatteer, Heike. 2015. Perception of basic emotions in music: Culture specific or multicultural? Psycology of Music. 18 páginas.

Bigand, E et al. 2005. Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. Cognition & Emotion. Psycology Press [Francia]. 19(8): 1113–1139.

Caetano, M., Burred, J., Rodet, X. 2010. Automatic segmentation of the temporal evolution of isolated acoustic musical instrument sounds using spectro-temporal cues. Proceedings of the 13th International Conference on Digital Audio Effects (DAFx). 10 páginas.

Carmen, G., & Daniel, F. (2012). Aportaciones Recientes Al Estudio De La Motivacion Y De Las Emociones, Edition: 2012, Chapter: Regulación De Emociones: Una Vision Pragmatica E Integradora Desde El Modelo Circumplejo La Regulación De Las Emociones: Concepto Y Fundamentos. Felix Ediciones. P.261-268

Dalla Bella, S., Peretz, I., Rousseau, L., y Gosselin, N. 2001. A developmental study of the affective value of tempo and mode in music. Cognition. 80(3): B1-10

Gouyon, Fabien, Pachet, Francosi, Delerue, Olivier. 2000. On the use of zero-crossing rate for an application of classification of percussive sounds. Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx). Verona. 6 páginas

Ha, Jin, *et al. 2012. What Does Music Mood Mean for Real Users?. University of Washington. 6 páginas.*

Halpern, Andrea, et al. 2016. Perceived and Induced Emotion Responses to Popular Music: Categorical and Dimensional Models. Centre for Digital Music. 48 páginas.

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. 2009. The elements of statistical learning: data mining, inference, and prediction. New York: Springer.

Imbrasaité Vaiva. Multi-modal dimensional emotion tracking in Music. [En línea]. En: https://www.cl.cam.ac.uk/~vi206/files/1st%20year%20report.pdf. *Consultado el 15-11-2016*

In T. Dalgleish, M. Power (Eds.). 1999 Handbook of Cognition and Emotion. Sussex, Reino Unido. John Wiley & Sons, Ltd. 13 páginas.

Juslin, P.N.2013. What does music express? Basic emotions and beyond. Front. Psychol. 4(596)

Knees, Peter. 2016. << 2. Basic Methods on Audio Signal Processing>>. Music Similarity and Retrieval. Austria. Springer. págs 33-50

Knees, Peter. 2016. << 3. Basic Methods on Audio Signal Processing>>. Music Similarity and Retrieval. Austria. Springer. págs 51-80

Laukka, P et al. 2013. Universal and culture-specific factors in the recognition and performance of musical affect expressions. Emotion. Suecia. 13(3): 434–449.

Lerch, Alexander. 2012. << 3. Instantaneous Features>> An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics. Berlin. John Wiley & Sons. Inc. págs 31-71

Majeed, Sayf; Husain, H.2015. Mel Frequency Cepstral Coefficients (Mfcc) Feature Extraction Enhancement In The Application Of Speech Recognition: A Comparison Study. University Kebangsaan Malaysia. Journal of Theoretical and Applied Information Technology. 19 páginas

Ng, Andrew. CSS229- Lecture Notes Part V: Support Vector Machines. [En línea]. En: http://cs229.stanford.edu/notes/cs229-notes3.pdf. Consultado el 29-10-2016

Pikrakis, Agelos<<4. Audio Features>>. 2014. Introduction to Audio Analysis: A Matlab Approach. Estados Unidos. ElSevier. págs 74-86

Schubert, E., Wolfe, J., Tarnopolsky, A. 2002. Spectral centroid and timbre in complex, multiple instrumental textures. In: Proceedings of the 8th International Conference on Music Perception and Cognition (ICMPC). Evanston, IL: Northwestern University. 654-657

Serra, X. Audio Signal Processing For Music Applications. [En Línea]. En: https://www.coursera.org/learn/audio-signal-processing/. Consultado el 15-10-2016

Vuoskoski, Jonna. 2012. <<4. Music And Emotion>>. Emotions Represented and Induced by Music The Role of Individual Differences. Finlandia. Publishing Unit, University Library of Jyvaskyla. págs 27-33.

Wiggins, G.A. 2010. Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music. Proceedings of the 11th IEEE International Symposium on Multimedia (ISM), San Diego

XI. ANEXOS

El código desarrollado para este trabajo de graduación se puede encontrar en el siguiente repositorio público: https://github.com/PJEstrada/song-emotion-collector

Figura 70: Formulario de inicio de participante

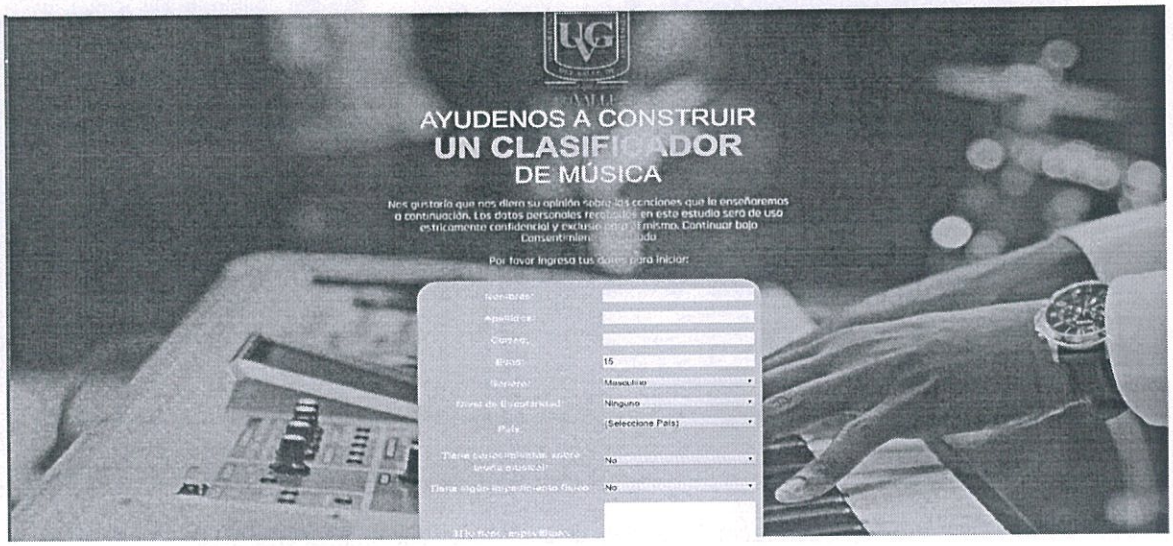


Figura 71: Formulario de inicio de participante (continuación)

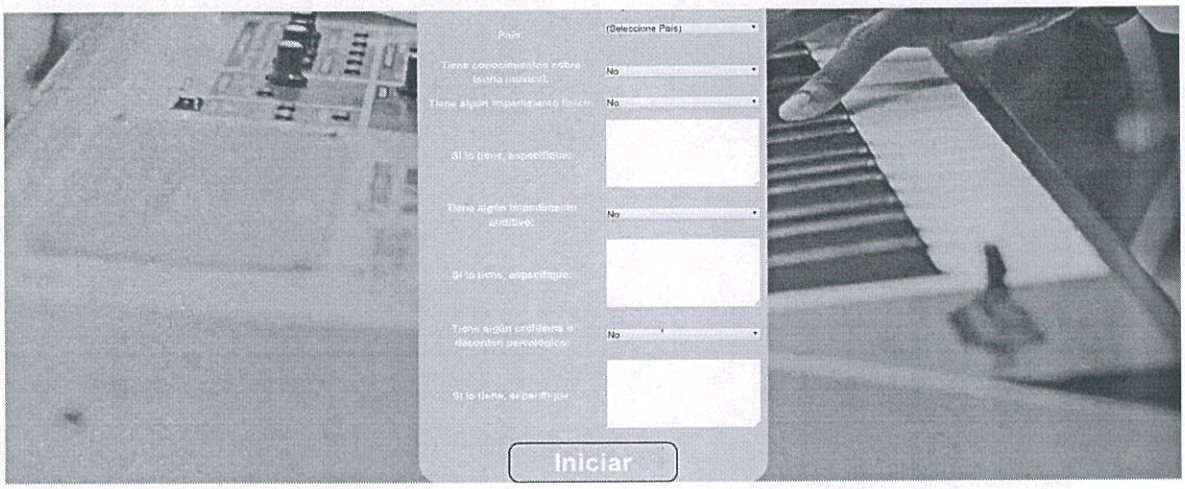


Figura 72: Clasificación de una canción

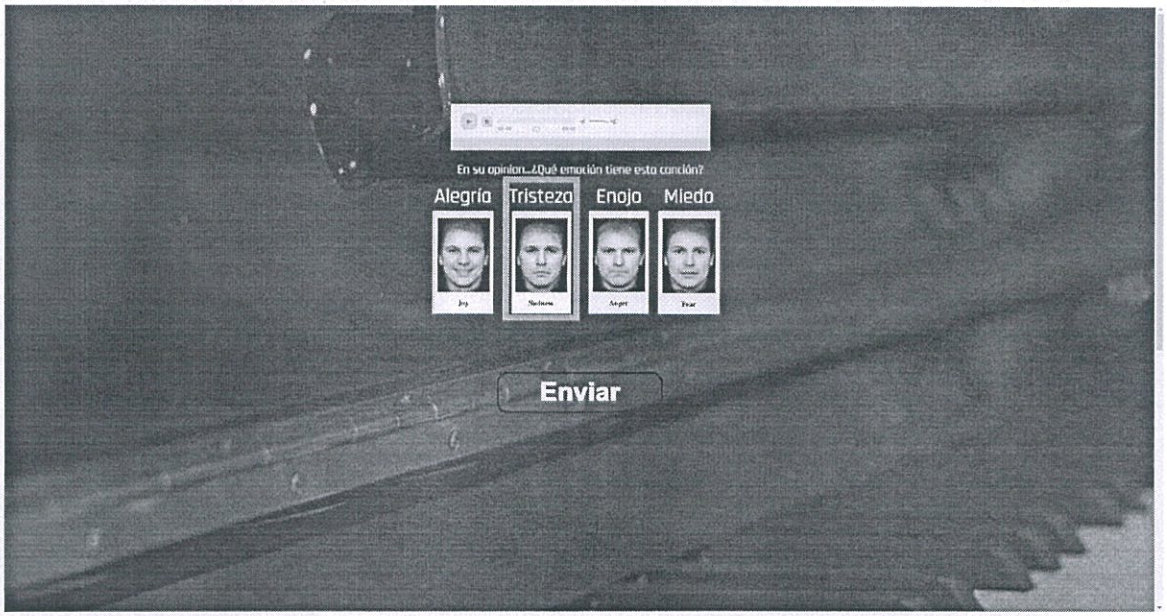


Figura 73: Subida de una nueva canción para clasificar

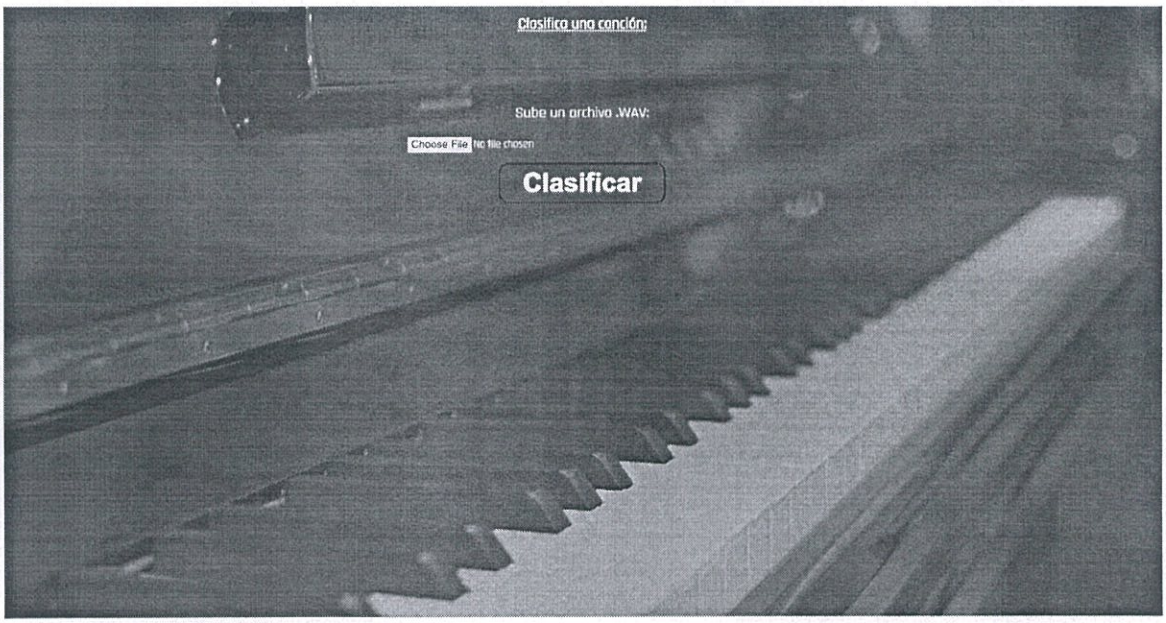


Figura 74: Resultado del clasificador

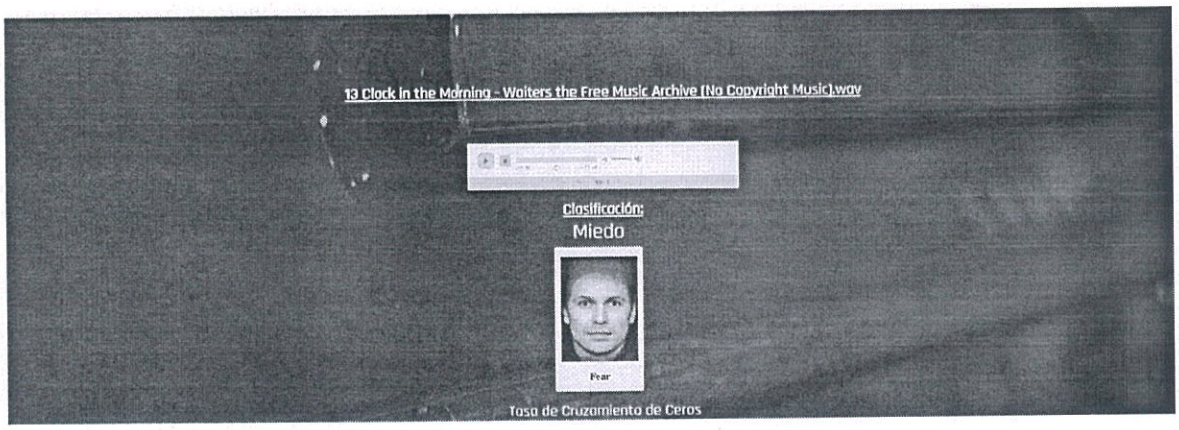


Figura 75: Listado de canciones