DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

TRABAJO DE TITULACIÓN, PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN ELECTRÓNICA Y TELECOMUNICACIONES

TEMA: IMPLEMENTACIÓN DE UN CLASIFICADOR DE GÉNEROS MUSICALES ECUATORIANOS MEDIANTE DEEP LEARNING.

AUTOR: TAYUPANTA BARRENO, LEONARDO STEPHANO

DIRECTOR: ING. BERNAL OÑATE, CARLOS PAÚL

SANGOLQUÍ

2019

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

CERTIFICACIÓN

Certifico que el trabajo de titulación, ***"IMPLEMENTACIÓN DE UN CLASIFICADOR DE GÉNEROS MUSICALES ECUATORIANOS MEDIANTE DEEP LEARNING"*** fue realizado por el señor ***Tayupanta Barreno, Leonardo Stephano*** el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 19 de noviembre 2019

Ing. Bernal Oñate, Carlos Paúl
DIRECTOR

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

AUTORÍA DE RESPONSABILIDAD

Yo, ***Tayupanta Barreno, Leonardo Stephano***, declaro que el contenido, ideas y criterios del trabajo de titulación: ***"Implementación de un clasificador de géneros musicales ecuatorianos mediante deep learning."*** es de mi autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Consecuentemente el contenido de la investigación mencionada es veraz.

Sangolquí, 19 de noviembre 2019

Tayupanta Barreno, Leonardo Stephano
C.C.: 1718165051

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

AUTORIZACIÓN

Yo, ***Tayupanta Barreno, Leonardo Stephano*** *autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación:* ***"Implementación de un clasificador de géneros musicales ecuatorianos mediante deep learning."*** *en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.*

Sangolquí, 19 de noviembre 2019

Leonardo Tayupanta

Tayupanta Barreno, Leonardo Stephano

C.C.: 1718165051

DEDICATORIA

A Dios y a mis padres por su cariño y su apoyo incondicional, ha sido un esfuerzo en común ya que ellos con sus consejos me han sabido guiar hasta donde hoy por hoy me encuentro.

Agradezco a mis profesores y amigos que compartimos aulas en medio de risas y enseñanzas, en especial a mi director de tesis que supo guiarme en el desarrollo de este trabajo de titulación.

Leonardo Stephano Tayupanta Barreno.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

En el campo de la investigación musical el análisis de pequeños clips de música en categorías como artistas, géneros, años, instrumentos han sido bastante estudiados, en particular por sus aplicaciones comerciales en el mundo real como exploradores de música, servicios de *streaming*, tiendas de música en línea. En este trabajo de titulación se investiga el desempeño del aprendizaje profundo usando redes neuronales convolucionales en el etiquetado automático de géneros musicales propios de Ecuador como son Marimba, Bomba, Pasillos, Sanjuanito. Para este propósito se generó una base de datos propia para el entrenamiento y otra para evaluación de la red. Utilizando el software Matlab® se implementó una red neuronal, el entrenamiento se lo realizó con una ventana de espectrograma de diferentes tamaños de 1, 2, 3, 4 y 5 segundos respectivamente. Para el clasificador de géneros musicales se realizó un programa que genera un archivo .txt por cada ítem en el cual se guarda la ubicación de cada clip de audio y de esta manera se visualizan los resultados. Una vez entrenada la red se la evaluó en el programa con la segunda base de datos y se analizaron los datos obtenidos midiendo el desempeño de la red neuronal con los diferentes tamaños de ventana utilizados.

PALABRAS CLAVE:

- **ESPECTROGRAMA**
- **DEEP LEARNING**
- **MÚSICA ECUATORIANA**
- **ETIQUETADO AUTOMÁTICO**
- **REDES NEURONALES CONVOLUCIONALES (CNN)**

ABSTRACT

In the field of music research, the classification of short music clips into categories such as artists, genres, year of departure, instruments has been quite studied, for their commercial applications in real life such us music explorer, streaming services, online music stores. This work investigates the performance of deep learning using convolutional neural networks in the autotagging of Ecuador's own musical genres such as Marimba, Bomba, Pasillos, Sanjuanito. For this purpose, an own database for training and another for network evaluation were generated. Using the Matlab® software, a neural network was implemented, the training was carried out with a spectrogram window of different sizes of 1, 2, 3, 4 and 5 seconds respectively. For the music genre sorter, a program was created that generates a .txt file for each item in which the name of each audio clip is saved and in this way the results are displayed. Once the network was trained, it was evaluated in the program with the second database and the data obtained was analyzed by measuring the performance of the neural network with the different window sizes used.

KEYWORDS:

- **SPECTROGRAM**
- **DEEP LEARNING**
- **ECUADORIAN MUSIC**
- **AUTOTAGGING**
- **CONVOLUTIONAL NEURONAL NETWORKS (CNN)**

CAPÍTULO I

1. Introducción del proyecto de Investigación

1.1. Antecedentes y justificación del Proyecto

La industria de la música ha evolucionado dramáticamente , el primer formato en aparecer fue el fonógrafo cilíndrico inventado por Thomas Edison en 1877, luego apareció el vinilo el cual podía almacenar hasta 45 minutos de música, la cinta de casete permitió a los oyentes hacer sus propias grabaciones, el CD dejó obsoleto al vinilo, con el formato mp3 aparecieron sitios para compartir archivos, en 2011 las descargas digitales superaron a las ventas físicas, años más tarde aparecen los servicios de *streaming* como Apple Music, Spotify, Tidal, entre otros la posibilidad para los usuarios de escuchar cualquier canción sin tener que descargarla (Lee, 2018).

El contenido digital se intercambia, se compra, los servicios *streaming* de música cada día son más populares y las bases de datos más grandes, una forma de acceder a estas grandes colecciones de música es etiquetar los recursos musicales, estas etiquetas pueden ser añadidas manual o automáticamente, sin embargo, debido al gran esfuerzo requerido por el humano, la implementación automática es más efectiva. (Sordo, 2011)

El etiquetado automático de música es un problema reciente, en comparación con otras áreas de investigación como procesamiento de señal o procesamiento de voz, está incluido en el campo de investigación de recuperación de información musical (MIR *Music Information Retrieval*), el cual ha recibido mucha atención por los investigadores en los últimos años. (Guaus, 2009)

La comunidad MIR desarrolla un evento anual de intercambio, evaluación y recuperación de información musical (MIREX) que se lleva a cabo desde 2004, las tareas directamente relacionadas

con la clasificación de música se citan a continuación, clasificación de géneros, clasificación del modo, identificación de artistas, reconocimiento de instrumentos , anotación musical. (Fu, Lu, Ting, & Zhang, 2011)

Las primeras investigaciones sobre la clasificación automática de géneros musicales se remontan a 2001 (Tzanetakis, 2001), y la clasificación del estado de ánimo en 2003 (Liu, 2003) . En cuanto a la clasificación automática de etiquetas de música, los primeros hallazgos de investigación se remontan a finales de 2005 (Mandel & Ellis, 2005) (Turnbull, Barrington, & Lanckriet, 2006). Desde entonces, se han propuesto varios algoritmos para la tarea de etiquetado automático.

Una investigación en 2011 realizada en más de 149 papers sobre etiquetado y clasificación de música indica que la mayoría se basaron en el aprendizaje automático convencional, esto implica extraer características con *pipeline* y aprendizaje de clasificadores, de manera sucinta las características fueron en su mayoría diseñadas manualmente. Sin embargo, los recientes avances en el uso de redes neuronales profundas han cambiado el paradigma hacia el aprendizaje de las representaciones de manera integral, lo que ha abierto la era del aprendizaje profundo (Deep learning) (Hinton, et al., 2012) .

En recientes años, los métodos de *Deep Learning* se han hecho populares en el campo de investigación de MIR, por ejemplo, en la conferencia ISMIR del 2010 solo se publicaron dos artículos con *Deep learning*, en 2015 6 artículos, en 2016 16 artículos, esta tendencia es inclusive más fuerte en otros campos del aprendizaje de máquina cuyas comunidades son más grandes y competitivas. (Choi, 2018).

Existen varios avances que han contribuido al éxito del aprendizaje profundo moderno. La innovación más importante ocurrió en la técnica de optimización. La velocidad de entrenamiento de los DNN (*Deep neural network*) se mejoró significativamente al usar unidades lineales rectificadas (ReLU) en lugar de funciones sigmoideas (Glorot, Bordes, & Bengio, 2011). Esto llevó a innovaciones en el reconocimiento de imágenes (Krizhevsky, Sutskever, & Hinton, 2012) y el reconocimiento de voz (George E Dahl, 2013). Otro cambio importante es el avance en hardware. La computación paralela en unidades de procesamiento de gráficos (GPU) habilitó a Krizhevsky para ser pionero en una clasificación de imágenes visuales a gran escala en (Krizhevsky, Sutskever, & Hinton, 2012).

Muchas características del sonido de nivel superior se relacionan con las energías en diferentes bandas de frecuencia. Esto explica la utilidad de las representaciones de tiempo-frecuencia de audio, como los espectrogramas, que se utilizan con frecuencia en la literatura (Lee, Pham, Largman, & Ng, 2009) (Henaff, Jarrett, Kavukcuoglu, & LeCun, 2011) (Wulfing & Riedmiller, 2013) (Dieleman & Schrauwen, 2013). CNN han sido utilizadas para la clasificación de géneros musicales en (Lee, Pham, Largman, & Ng, 2009) (Li, Chan, & Chun, 2010) (Dieleman, Brakel, & Benjamin Schrauwen, 2011)

Sergey, Hamza, & Alexei (2017) aplicaron CNN a un gran conjunto de datos y evaluaron su desempeño empíricamente en la clasificación de audio, demostraron un desempeño preciso en dichas tareas de clasificación cuando se presentan ejemplos de música de 5 segundos obtenidos por simples transformaciones de la forma de onda de audio, como la transformada de matriz aleatoria y la aplicación de un banco de filtros espaciados logarítmicamente.

Choi, Fazekasñ, & Sandler (2016) presentaron un algoritmo de etiquetado automático de música basado en contenido que utiliza redes neuronales completamente convolucionales (FCN). Los experimentos muestran que el espectrograma de melodía es una representación efectiva de tiempo-frecuencia para el etiquetado automático y que modelos más complejos se benefician con más datos de entrenamiento.

Dong (2018) propone un nuevo método que combina el conocimiento del estudio de la percepción humana en la clasificación del género musical y la neurofisiología del sistema auditivo. Divide el espectrograma de la señal de música en segmentos consecutivos de 3 segundos, hace predicciones para cada segmento y finalmente combina las predicciones juntas. Los filtros aprendidos en la CNN se asemejan al campo espectrotemporal receptivo (STRF) en el sistema auditivo.

Todos los países latinoamericanos son representados o reconocidos a nivel internacional principalmente por su cultura, dentro de la misma, por sus distintos géneros musicales propios de la nación, en el artículo 102 de la ley orgánica de comunicación se menciona que la música emitida en las estaciones de radiodifusión sonora deberá representar al menos el 50% en todos sus horarios, por lo que este trabajo es un aporte para la música ecuatoriana tratando de fomentar la cultura y folklore, además de aportar géneros que no son muy conocidos en grandes plataformas de música entre las cuales lideran empresas como Apple, Google, Amazon Spotify , Pandora Radio.

Este proyecto está enfocado en desarrollar un sistema que pueda ser utilizado como herramienta para organizar música, clasificar canciones en cuatro géneros bastantes populares a nivel ecuatoriano Pasillo, Sanjuanito, Bomba, Marimba, se creará una base de datos reales compuesta de

fragmentos musicales para luego entrenar al sistema y poder probar la robustez al mismo con un grado de aceptación razonable.

1.2. Objetivos de la Investigación

1.2.1. Objetivo General

- Implementación de un clasificador automático de géneros musicales ecuatorianos mediante el aprendizaje con *Deep Learning*.

1.2.2. Objetivos Específicos

- Realizar un estudio previo de los diferentes algoritmos para clasificación de música existentes.
- Elaborar una base de datos con los diferentes géneros musicales definidos.
- Obtener el espectrograma de los diferentes géneros musicales definidos
- Elaborar un clasificador automático a partir de redes neuronales convolucionales.
- Evaluar la calidad del aprendizaje del clasificador con canciones nuevas para el sistema.

CAPÍTULO II

2. MARCO TEÓRICO

2.1. Preprocesamiento

Las investigaciones relacionadas con el estudio del audio por lo general realizan un preprocesamiento en las formas de onda para enfatizar las características auditivas. Los enfoques más comunes para el preprocesamiento de la señal son la Transformada de Fourier de Término Reducido (*Short-Time Fourier Transform*, STFT), Mel-Espectrograma, los Coeficientes Cepstrales en las Frecuencias de Bark (BFCC (*bark frequency cepstral coefficients*)) (Yang, 2018).

2.1.1. Transformada de Fourier

El análisis de Fourier es un conjunto de técnicas matemáticas que se utilizan para descomponer señales en ondas sinusoidales. La transformada de Fourier básicamente convierte una señal del dominio temporal al dominio de frecuencia (Lihua, 2010). La mecánica fundamental de los métodos de preprocesamiento de audio se basa en la Transformada de Fourier, cuya definición es la siguiente:

$$\mathcal{F}(\omega) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \omega} dx \quad (1)$$

Donde $f(x)$ es la función original del tiempo, x representa el tiempo. $\mathcal{F}(\omega)$ es la función transformada de la frecuencia donde ω representa la frecuencia. La transformada de Fourier se inspiró en gran medida en el estudio de las series de Fourier, que descomponen una función complicada en sumas de ondas simples (Yang, 2018). Cuando se trata de análisis de sonidos, revela la información de frecuencia dentro de las señales de sonido. Para la extracción de características

de sonido / música, se utiliza una forma especial de la transformada de Fourier, la transformada discreta de Fourier de término reducido (STFT) (Lihua, 2010).

2.1.2. Transformada de Fourier de Término Reducido (STFT)

Proporciona una representación tiempo-frecuencia con frecuencias centrales espaciadas linealmente. El cálculo de la representación STFT suele ser más rápido que otras representaciones de tiempo-frecuencia gracias a la transformada rápida de Fourier (FFT) que reduce el costo o (N^2) a o (N log (N)) con respecto al número de puntos FFT y computación paralela (Choi, 2018)

Se utiliza para el estudio de música digital ya que son señales discretas, y el análisis de frecuencia solo tiene sentido cuando se trata de una ventana de tiempo corto, las señales de sonido como el habla y la música son generalmente muy cambiables con el tiempo. La siguiente fórmula muestra el cálculo de STFT

$$\{x[n]\} \equiv X(m,w) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \quad (2)$$

En la ecuación anterior, x [n] representa la señal de entrada y w [n] representa la función de ventana. La Figura 1 muestra el proceso genérico de extracción STFT.

La señal de audio original primero realiza una convolución con un tipo de función de ventana. En esta tesis, la función de ventana utilizada es la de Hamming para eliminar el efecto de los bordes (García Durán, 2011) (Gulzar, Singh, & Sharma, 2014). Las señales en ventana se transforman utilizando la ecuación 2. Por lo general, esta etapa se reemplaza con un algoritmo más rápido: Transformada rápida de Fourier. El resultado de la transformación son los valores STFT. Después del proceso de STFT, las señales de audio son transformados en espectrogramas (Lihua, 2010).

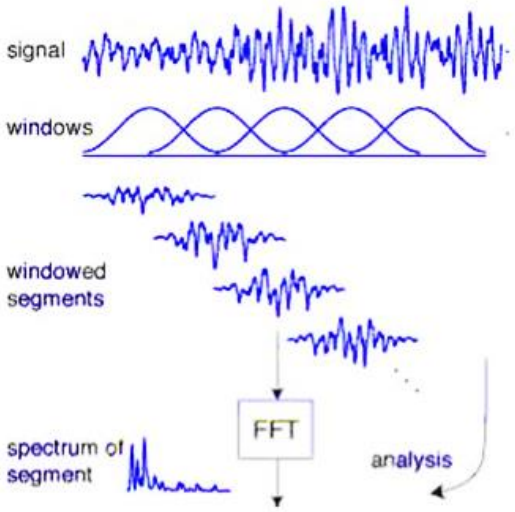

Figura 1. Proceso Transformada de Fourier de Término Reducido
Fuente: (Lihua, 2010)

2.1.3. Espectrograma

Una forma muy popular de usar la Transformada de Fourier es denominada Transformada de Fourier de Término Reducido (STFT). Se aplica la transformada de Fourier a una pequeña ventana de la señal y la combina con los resultados en una matriz bidimensional. Con una frecuencia de muestreo adecuada, un gran archivo de audio puede dividirse en varios fragmentos y cada uno puede transformarse por separado. La matriz combinada muestra la relación tiempo-frecuencia, y los valores de cada cuadrícula representan la magnitud de una determinada frecuencia en un momento determinado. El resultado de una STFT se llama espectrograma (Yang, 2018).

2.1.4. BFCC (Bark frequency cepstral coefficients)

BFCC es un método para extraer características de la señal, en la Figura 2 se muestra el diagrama de bloques del algoritmo BFCC. Este método es similar al MFCC (*Mel Frequency Cepstral Coefficients*), implementa una escala de filtros bark.

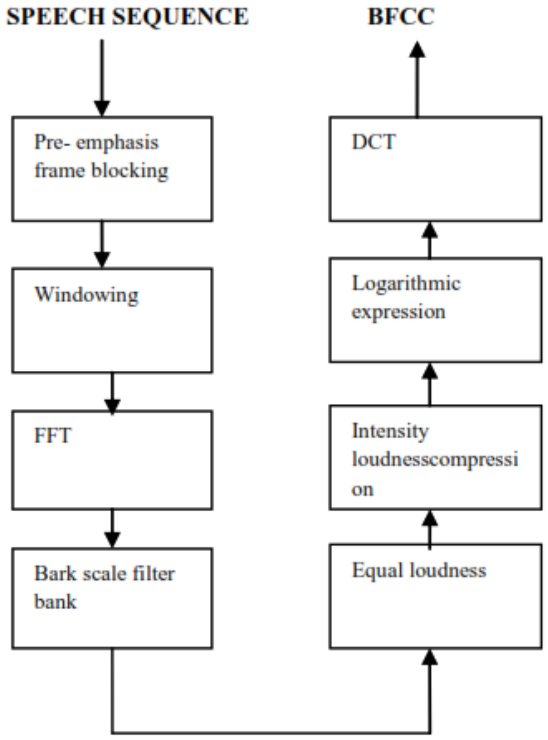

Figura 2. Diagrama de bloques del proceso BFCC.
Fuente: (Gulzar, Singh, & Sharma, 2014)

La escala Bark, diseñada por Eberhard Zwicker en 1961, divide el rango auditivo del oído en áreas que corresponden aproximadamente a las bandas críticas de la cóclea, lo que da como resultado una escala no lineal con la mayor resolución en torno a 2kHz, aproximadamente, donde las bandas críticas son más cercanas (Zwicker, 1961). Por lo tanto, la escala de Bark ofrece una escala mucho más psicoacústicamente pertinente que la división del espectro en distancias de Hertz igualmente espaciadas para el análisis de señales musicales.

Matemáticamente el filtro escala bark se representa por la siguiente fórmula:

$$f_{bark} = 6ln\left[\frac{f}{600} + \left[\frac{f}{600}\right]^2 + 1\right]^{0.5} \quad (3)$$

Donde f_{bark} corresponde a la frecuencia en Barks y f es la frecuencia lineal en Hertz. Las salidas del filtro se ponderan de acuerdo con la curva de igual sonoridad, que se aproxima a la sensibilidad de la audición humana. La señal se comprime luego bajo la ley de potencia de intensidad-volumen, donde la raíz de la señal es comprimida por la raíz cúbica para que coincida con la relación no lineal entre la intensidad del sonido y el volumen percibido. Finalmente, la señal se comprime primero a través de la función logarítmica y, se usa DCT(*Discrete Cosine Transform*) para decorrelar las características como en el caso de MFCC (Gulzar, Singh, & Sharma, 2014).

2.2. Etiquetado de Música

2.2.1. Etiquetas de Música

Son un conjunto de palabras claves descriptivas que transmiten información de alto nivel del clip de música, como emociones (tristeza, felicidad, ira), género (jazz, rock), instrumentos (guitarra, cuerdas, instrumental, voz). Estos tipos de etiquetas están profundamente relacionados con la percepción subjetiva de la música de los oyentes o las comunidades. (Choi, Fazekasñ, & Sandler, 2016). En el etiquetado de música se puede considerar una o múltiples etiquetas como las que se mencionaron anteriormente, géneros, reconocimiento de instrumentos, emociones, tiempo y posiblemente otros.

2.2.2. Etiquetado Automático

Es una tarea de clasificación que se encarga de predecir etiquetas de música usando la señal de audio. Requiere extraer características que sean buenos estimadores del tipo de etiqueta que es de interés, seguido de una clasificación de una etiqueta única o múltiple (Choi, Fazekasñ, & Sandler, 2016).

2.3. Identidades Musicales Ecuatorianas

En este trabajo de titulación me enfoco en las Identidades Musicales Ecuatorianas para dar un realce en el etiquetado de música y aportar al MIR. En Ecuador existen diferentes culturas con mucha diversidad por lo tanto posee una gran riqueza musical indígena, afroecuatoriana y mestiza, para el etiquetado de música se escogieron los siguientes géneros que se explican a continuación.

- Música Indígena: San Juanito
- Música Afroecuatoriana: La bomba, La Marimba
- Música Mestiza: Pasillo

2.3.1. Música Indígena

Representa la mayor diversidad étnica del Ecuador, la música andina y de la Amazonía se consideran música Indígena, en la composición presenta una pentafónica en modo menor (en la sierra) y una escala de sonidos trifónica en modo mayor (en la Amazonía). Entre los diferentes estilos que existen en la música indígena se encuentran: yaraví, sanjuanito, yumbo, dánzate, carnaval y la Amazonía (Vasco & Magdalena, 2009).

Para el etiquetado de música se escogió el Sanjuanito que es un género musical que se interpreta en la actualidad por grupos musicales indígenas, tiene una pequeña introducción o estribillo y está compuesto en un compás binario simple, 2/4, con ritmo alegre y rápido, allegretto, en la que la negra dura 114, se usa la escala pentafónica menor, es el ritmo de fiesta por excelencia. Según investigadores sus orígenes se remiten en San Juan Bautista de Chambo, en la provincia de Chimborazo y otros a la comunidad de San Juan de Ilumán, en Imbabura. Los instrumentos musicales que se usan son el tambor, la flauta o rondador los cuales tocan la melodía principal, el

arpa, la guitarra, el charango también se utilizan en los sanjuanes mestizos (Vasco & Magdalena, 2009).

2.3.2. Música Afroecuatoriana

Principalmente asentada en las provincias de Esmeraldas e Imbabura, es uno de los componentes más importantes de las identidades musicales ecuatorianas. La música se caracteriza por la variación e improvisación de la melodía, el predominio rítmico, la supremacía de la percusión sobre las melodías y armonías. Los instrumentos que se utilizan son el cununo, la marimba, el guasá, la charrasca, el bombo, las maracas entre otros. La música Afroecuatoriana se puede dividir en cantos sagrados: los arrullos, alabaos y chigualos; y los cantos profanos: la marimba y el conjunto de marimba, la bomba y la banda mocha (Vasco & Magdalena, 2009).

Para el etiquetado de música se escogieron los géneros musicales la bomba y la marimba.

- La Bomba surge como un género musical de la combinación de la rítmica, movimiento, timbre y líricas de la música africana, los estribillos y formas musicales europeas, las armonías pentafónicas, propias de la música indígena. Es un género musical y una danza propia de la zona que proviene del Chota-Mira. La fórmula rítmica es similar al albazo, es escrita en compás binario compuesto, la velocidad con la que se ejecuta es rápida al ritmo de la bomba-instrumento musical. Los instrumentos que se utilizan son la bomba, dos guitarras o una guitarra y un requinto, un guasá y un arpa, pueden variar (Vasco & Magdalena, 2009).
- Dentro de la música afroecuatoriana la Marimba es el instrumento más importante y es todo un referente cultural, aunque el origen de la marimba es incierto, la marimba

esmeraldeña es propia del Ecuador y es construida con maderas de la zona. El teclado produce un sonido pentafónico, tiene de dieciocho a treinta teclas y son dos personas las que tocan este instrumento, el tiplero se encarga de las partes agudas y el bordonero se encarga de las partes graves. Otros instrumentos que se ocupan son el cununo que es una especie de tambor, el guasá es un instrumento idiófono de sacudimiento, el bombo es un tambor de dos parches (Vasco & Magdalena, 2009).

2.3.3. Música Mestiza

Aparece el momento en que se genera la escala mestiza de 7 sonidos ya que los indígenas usaban solo la escala pentafónica, la diferencia entre la música indígena y mestiza es netamente musical. La música mestiza tiene sus melodías armonizadas con el quinto grado mayor que lo diferencia completamente de la música indígena. La guitarra fue otro factor encargado del mestizaje musical. Las primeras composiciones mestizas aparecen a finales del siglo XIX (Vasco & Magdalena, 2009). Los ritmos son los siguientes: el pasacalle, el pasillo, el alza que te han visto, el aire típico y sus derivaciones, el albazo, la tonada, el cachullapi y el saltashpa.

Para el etiquetado de música se escogió el pasillo, género musical que marcó la dinámica de la música popular del siglo XX, es una innovación del valls europeo que se origina en el siglo XIX, se caracteriza por el acompañamiento de guitarras y requinto. Su estructura corresponde a una forma A-B-A con una introducción de cuatro, ocho o doce compases. Los instrumentos que se utilizan son la guitarra y el requinto, en la actualidad también se utilizan los instrumentos orquestales como clarinetes, violonchelos, violines, contrabajos, traversas, flautas etc (Vasco & Magdalena, 2009).

2.4. Bases de Datos

Por el tema de derechos de autor relacionados con el contenido de música, los conjuntos de datos públicos que se presentan a continuación evitan el problema mediante el uso de clips de audio recortados, bajados la calidad o sin derechos de autor.

2.4.1. GTZAN

Es una de las bases de datos más utilizadas para la clasificación de géneros musicales (Tzanetakis, 2001). Contiene 1000 piezas de audio de 30 segundos cada una, cuenta con 10 géneros musicales los cuales son blues, clásica, country, disco, hip-hop, jazz, metal, pop, reggae, rock. Cuenta con 100 canciones por cada género musical. Los archivos de audio tienen las siguientes características 22050 Hz, Mono, 16-bit.

2.5. Deep Learning

Con el resurgimiento de las redes neuronales en el 2000, Deep Learning se ha convertido en un área de investigación extremadamente activa (Buduma & Locascio, 2017), ha ganado gran popularidad debido a los avances en la capacidad computacional, mejora en la eficiencia en los modelos de entrenamiento y la gran cantidad de datos disponibles (Wiley, 2016). Deep Learning busca automatizar la inteligencia bit a bit, es un subconjunto de Machine Learning, el cual se encarga del desarrollo y uso de algoritmos que aprenden de datos sin procesar para hacer predicciones con la capacidad de aprender sin ser programado explícitamente (Trask, 2019), elimina el engorroso y limitante proceso de selección de características (Buduma & Locascio, 2017). Deep Learning es una poderosa herramienta con una arquitectura multi-capa que se utiliza para reconocimiento de patrones, detección de señales, clasificación y predicción (Wiley, 2016).

2.6. Redes Neuronales Convolucionales (CNN convolutional neural networks).

La idea para el desarrollo de las CNNs nació de los sistemas de visión biológica, donde la información de las regiones locales es capturada por células sensoriales que se encargan de procesar la información de nivel superior (LeCun & Bengio, 1995). Las CNNs fueron diseñadas como una forma para aprender características robustas que corresponden a ciertos objetos con una translación o distorsión local (Choi, Fazekasñ, & Sandler, 2016). Las CNNs son redes neuronales que consisten en tres tipos de capas: capas convolucionales, capas de agrupación (*Polling Layer*) y capas completamente conectadas.

2.6.1. Capa Convolucional

La capa convolucional calcula una convolución de su entrada bidimensional con un *kernel* de tamaño fijo, seguida de una no linealidad de elementos. La entrada puede consistir en múltiples canales del mismo tamaño, en cuyo caso convoluciona a cada uno con un *kernel* separado y suma los resultados. Del mismo modo, la salida puede consistir en múltiples canales computados con distintos conjuntos de *kernel*. Normalmente, los *kernels* son pequeños en comparación con la entrada, lo que permite a las CNN procesar grandes entradas con pocos parámetros aprendibles (Ullrich, Schülter, & Grill, 2014).

Las capas convolucionales se consideran los componentes básicos de la arquitectura CNN. Como se ilustra en la Figura 3, las capas convolucionales transforman los datos de entrada mediante un parche de neuronas que se conectan localmente de la capa anterior. La capa calcula el producto punto entre la región de las neuronas en la capa de entrada y los pesos a los que están conectados localmente en la capa de salida.

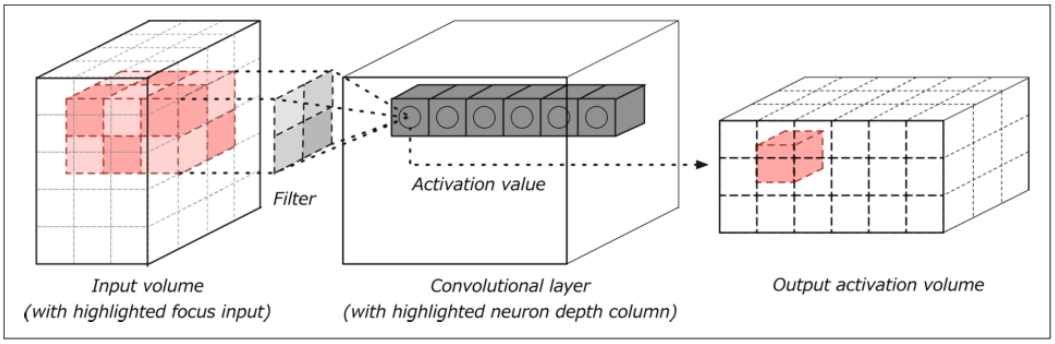

Figura 3. Capa Convolucional con volúmenes de entrada y salida
Fuente: (Buduma & Locascio, 2017)

La salida resultante generalmente tiene las mismas dimensiones espaciales (o dimensiones espaciales más pequeñas) pero a veces aumenta el número de elementos en la tercera dimensión de la salida (dimensión de profundidad).

2.6.2. Capa de agrupación (*Pooling Layer*)

Las capas de agrupación se insertan comúnmente entre capas convolucionales sucesivas para reducir el tamaño espacial (alto y ancho) de la representación de datos progresivamente en la red y ayudan a controlar el sobreajuste, la idea esencial es dividir cada mapa de características en mosaicos de igual tamaño, la capa de agrupación funciona independientemente en cada segmento de la entrada (Patterson & Gibson, 2017), toma submuestras de la capa convolucional para alimentar la siguiente capa, específicamente, se crea una celda para cada mosaico, se calcula el valor máximo en el mosaico y se propaga este valor máximo en la celda correspondiente del mapa de entidades, este proceso se ilustra en la Figura 4. Los esquemas de reparto de peso y agrupación permiten que la CNN genere propiedades de conservación como la invariancia de traducción (Vieira, 2018) (Buduma & Locascio, 2017)

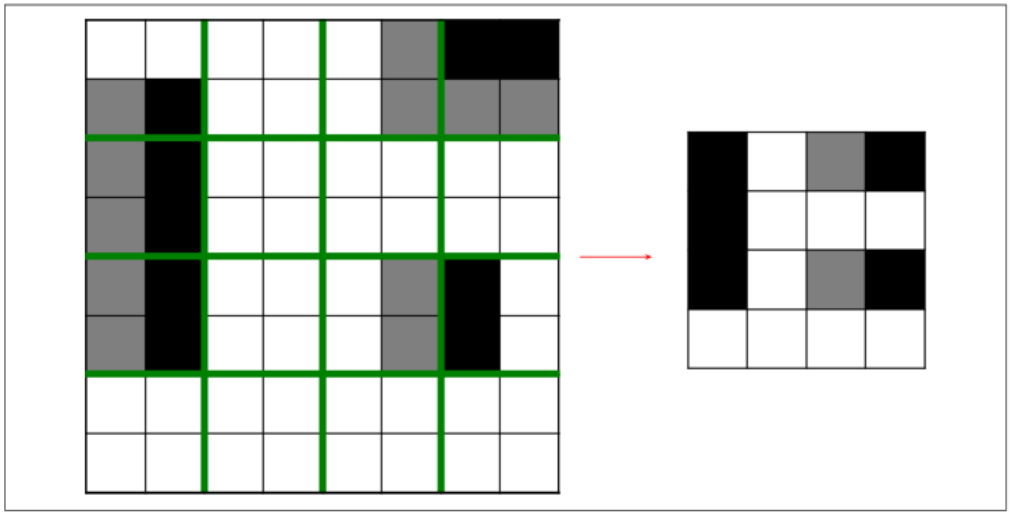

Figura 4. Max Pooling reduce significativamente los parámetros
Fuente: (Buduma & Locascio, 2017)

2.6.3. Capa Totalmente Conectada

Finalmente, una capa totalmente conectada descarta cualquier disposición espacial de su entrada al cambiarla de forma en un vector, calcula un producto de puntos con una matriz de ponderación y aplica una no linealidad de elementos al resultado. Por lo tanto, a diferencia de los otros tipos de capas, no está restringido a operaciones locales y puede servir como etapa final integrando toda la información para tomar una decisión (Ullrich, Schülter, & Grill, 2014).

2.7. Datos de Entrenamiento Validación y de Prueba

La evaluación de un modelo siempre se reduce a dividir los datos disponibles en tres conjuntos: entrenamiento, validación y prueba. El modelo se entrena con los datos de entrenamiento, este incluye el conjunto de ejemplos de entrada en los que el modelo se ajustará, o se capacitará, ajustando los parámetros (es decir, los pesos en el contexto de las Redes Neuronales) (Francois, 2017).

Para que el modelo sea entrenado, necesita ser evaluado periódicamente, es decir hacer que el modelo aprenda de sus errores, y para eso es exactamente el conjunto de validación. Al calcular la pérdida que el modelo produce en el conjunto de validación en cualquier punto dado, podemos saber qué tan precisa es. Esta es la esencia del entrenamiento. Posteriormente, el modelo ajustará sus parámetros en función de los resultados de la evaluación frecuente en el conjunto de validación (Francois, 2017).

Una vez que el modelo está listo lo prueba una última vez con los datos de prueba, corresponde a la evaluación final, después que el modelo completa la fase de entrenamiento. Este paso es crítico para probar la generalización del modelo llegar a una conclusión sobre qué tan bien funciona el modelo. Al usar este conjunto, podemos obtener la precisión de trabajo de nuestro modelo. No se debe exponer el modelo al conjunto de prueba hasta que termine la fase de entrenamiento. De esta manera, podemos considerar que la medida de precisión final es confiable (Francois, 2017).

CAPÍTULO III

3. METODOLOGÍA DEL PROYECTO DE INVESTIGACIÓN

3.1. Descripción general del proyecto de Investigación

El proyecto de investigación presenta un algoritmo de etiquetado de música automático basado en contenido que utiliza redes neuronales convolucionales, que etiqueta 4 géneros musicales ecuatorianos (pasillos, bomba, sanjuanito, marimba) con la herramienta Matlab®, la descripción general se puede observar en la Figura 5. Consta de 4 bloques, entrada, preprocesamiento, modelo de aprendizaje profundo y salida.

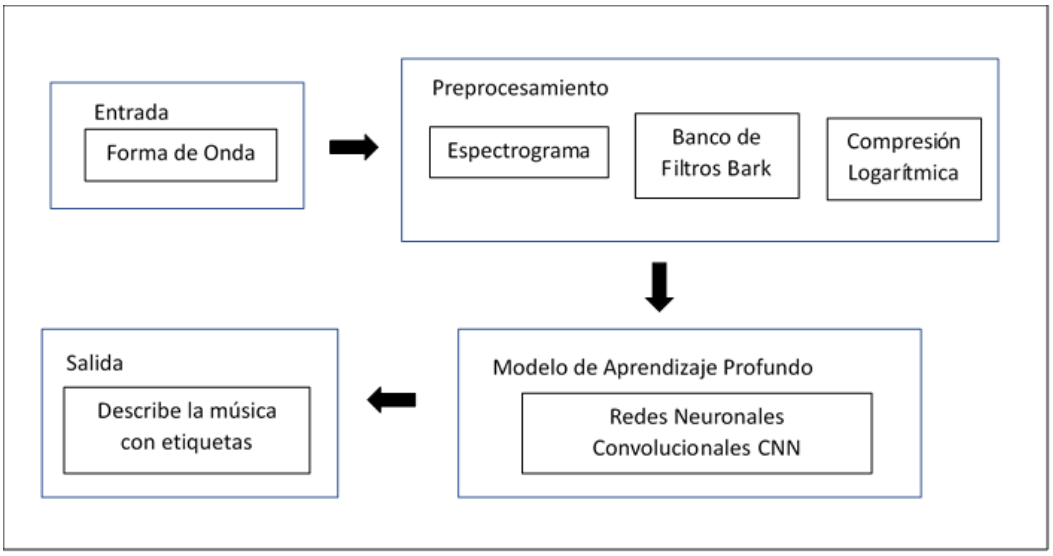

Figura 5. Diagrama de Bloques del proyecto de Investigación

En el bloque de entrada ingresan los clips de audio sin procesar de la base de datos para el entrenamiento, con estos datos se genera un espectrograma utilizando la transformada rápida de Fourier (STFT), se aplica un banco de filtros Bark y una compresión logarítmica, la red neuronal

convolucional es entrenada con estos datos y finalmente a la salida tenemos el etiquetado automático de música.

3.2. Entrada

La red neuronal convolucional es una representación de los métodos de aprendizaje que permite a la maquina ingresar datos sin procesar y automáticamente descubrir las representaciones necesarias para la clasificación o detección (LeCun, Bengio, & G. Hinton, 2015). Sin embargo, un preprocesamiento apropiado en los datos de entrada aún sigue siendo importante para mejorar el rendimiento del sistema. A continuación, se describe la base de datos personalizada, el ANEXO A incluye todos los artistas y canciones que se encontró pertenecientes a cada género musical lo cual es una limitación puesto que pueden existir más, todos los archivos fueron extraídos de la plataforma digital YouTube:

3.2.1. Base de datos para entrenamiento

Los clips de audio tienen las siguientes características, 16bits/16KHz/Mono.

- Pasillos: Consta de 64 canciones, 7 artistas con una duración de 3 horas 46 minutos.
- San Juanito: Consta de 38 canciones, 17 artistas con una duración de 2 horas 25 minutos.
- Marimba: Consta de 30 canciones, 3 artistas con una duración de 2 hora 30 minutos.
- Bomba: Consta de 16 canciones, 9 artistas con una duración de 1 hora.

Falsos Positivos: Contiene 8 géneros musicales (blues, clásica, country, hiphop, jazz, pop, reggae, rock), por género musical se seleccionó 30 clips. Para encontrar el mejor desempeño de la red neuronal, esta fue entrenada con clips de audio de 1, 2, 3, 4, 5 segundos respectivamente, en la Tabla 1 se muestra un resumen de los clips de audio. El aprendizaje con una ventana de 1, 2, 3

segundos tiene clips de 10 segundos y para una ventana de 4, 5 segundos, tiene clips de 16 segundos. Para el entrenamiento de la red se seleccionó el 30% de falsos positivos aleatoriamente.

Tabla 1
Base de Datos

Género	**# Seg**	**# Clips de Audio**
Bomba	1	3217
	2	1607
	3	1072
	4	802
	5	642
Marimba	1	6029
	2	3017
	3	2011
	4	1507
	5	1206
Pasillos	1	12365
	2	6182
	3	4120
	4	3090
	5	2472
Sanjuanito	1	4908
	2	4054
	3	2702
	4	2025
	5	1621

En total la base de datos consta de 147 canciones, con 39 artistas y una duración de 9 horas 20 minutos, 64649 clips de audio de los cuales 7340, 13770, 28229, 15310 pertenecen a bomba, marimba, pasillos, sanjuanito respectivamente.

La base de datos para el entrenamiento de la red neuronal se divide en tres grupos, un conjunto de entrenamiento, un conjunto de validación y un conjunto de prueba, de acuerdo al tamaño de la base de datos estos porcentajes pueden variar por ejemplo 60%, 20%, 20% respectivamente y si la cantidad de datos es muy grande 80%, 10%, 10% respectivamente (Pons, Slizovskaia, Gong, Gómez, & Serra, 2017). En este trabajo se utilizaron los valores de entrenamiento, validación y prueba con un porcentaje de 76%, 12%, 12% respectivamente como se muestra en la Tabla 2.

Tabla 2
Entrenamiento Validación y Prueba

# Seg	# Clips de Audio	Entrenamiento (76%)	Validación (12%)	Prueba (12%)
1	26996	20517	3240	3240
2	15136	11503	1816	1816
3	10329	7850	1239	1239
4	7666	5826	920	920
5	6118	4650	734	734

3.2.2. Base de Datos para evaluar el Clasificador

Los clips de audio tienen las siguientes características 16 bits/16KHz/Mono, 30 segundos

- Pasillos: Consta de 100 clips de audio extraídos de entre de 75 canciones, 22 artistas.
- San Juanito: Consta de 100 clips de audio, 29 artistas, 46 canciones.
- Marimba: Consta de 100 clips de audio, 44 canciones, 3 artistas.
- Bomba: Consta de 100 clips de audio, 28 canciones 8 artistas.
- Base de datos GTZAN: Consta de 100 clips de audio por género musical (blues, clásica, country, disco, hip-hop, jazz, metal, pop, reggae, rock).

En total la base de datos consta de 1400 clips de audio como se muestra en la Tabla 3.

Tabla 3
Entrenamiento Validación y Prueba

# Tag	# Clips de Audio
Pasillos	100
Sanjuanito	100
Marimba	100
Bomba	100
GTZAN	1000
Total	1400

3.2.3. Preprocesamiento del Audio

Se utiliza el espectrograma a través de un banco de logaritmos en una escala Bark como una representación de frecuencia tiempo para las características de contenido sin procesar de la música, comúnmente se usa en modelos de aprendizaje profundo que tratan con señales de audio (Oord, A., & Schrauwen, 2013) (Liu & Yang, 2016). Para un eficiente entrenamiento de la red neuronal convolucional las formas de onda deben tener un correcto preprocesamiento.

La señal de audio es segmentada en secuencias de 1, 2, 3, 4 y 5 segundos y se genera un espectrograma utilizando la transformada rápida de Fourier (STFT) con una ventana Hamming de 10 ms, un tamaño de ventana de 400 muestras y una longitud de superposición de 240 muestras. Después que el espectrograma es generado, la escala Bark es usada para agregar frecuencias de banda críticas del espectrograma en 40 bandas desde los 27.5 Hz hasta los 16 kHz. Para obtener datos con una distribución más uniforme se toma el logaritmo de los espectrogramas utilizando un pequeño desplazamiento de 1e-6 denominado epsil.

En las Figuras 6, 7, 8, 9 se muestran ejemplos de espectrograma de los 4 géneros musicales en los que se puede apreciar sus diferentes tamaños.

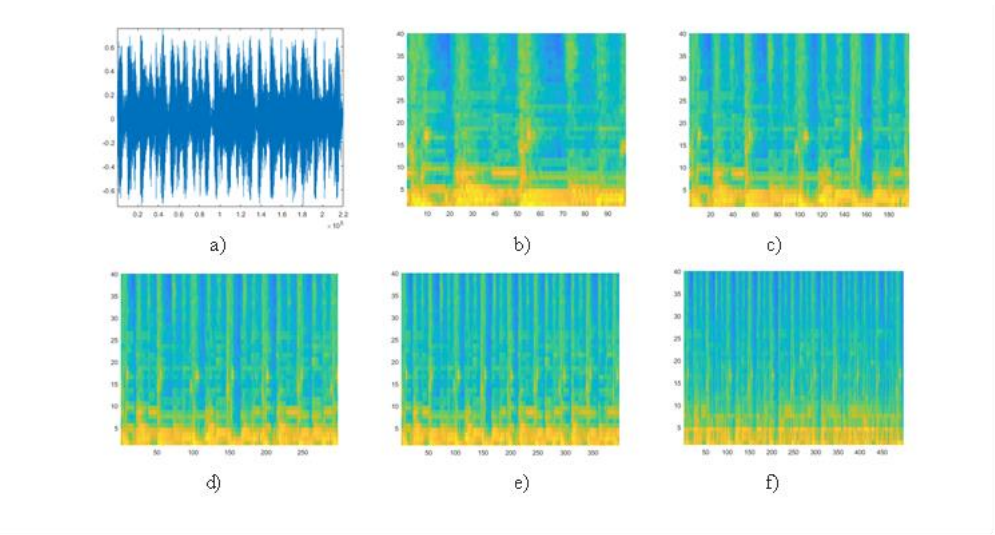

Figura 6. Espectrograma Bomba a) Señal de Audio b) 1 Segundo c) 2 Segundos d) 3 Segundos e) 4 Segundos e) 5 Segundos

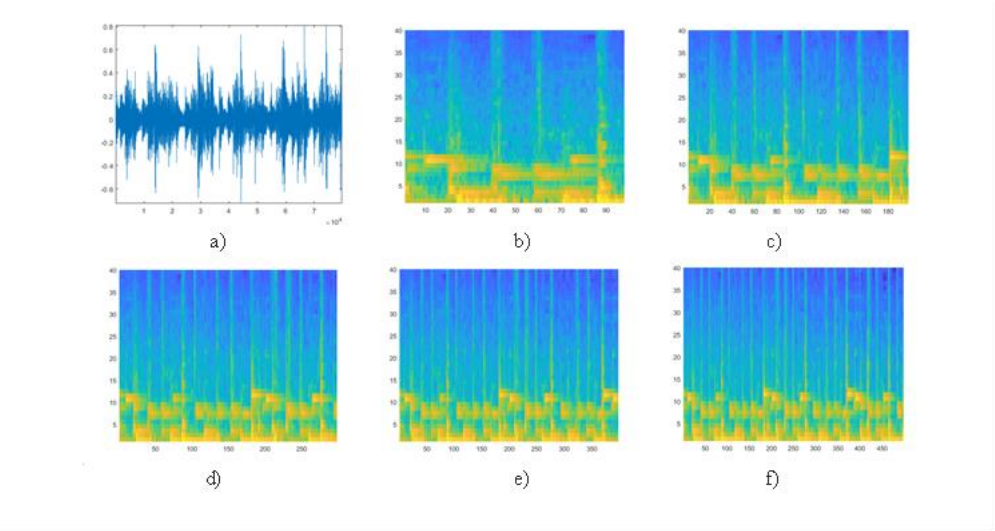

Figura 7. Espectrograma Marimba a) Señal de Audio b) 1 Segundo c) 2 Segundos d) 3 Segundos e) 4 Segundos e) 5 Segundos

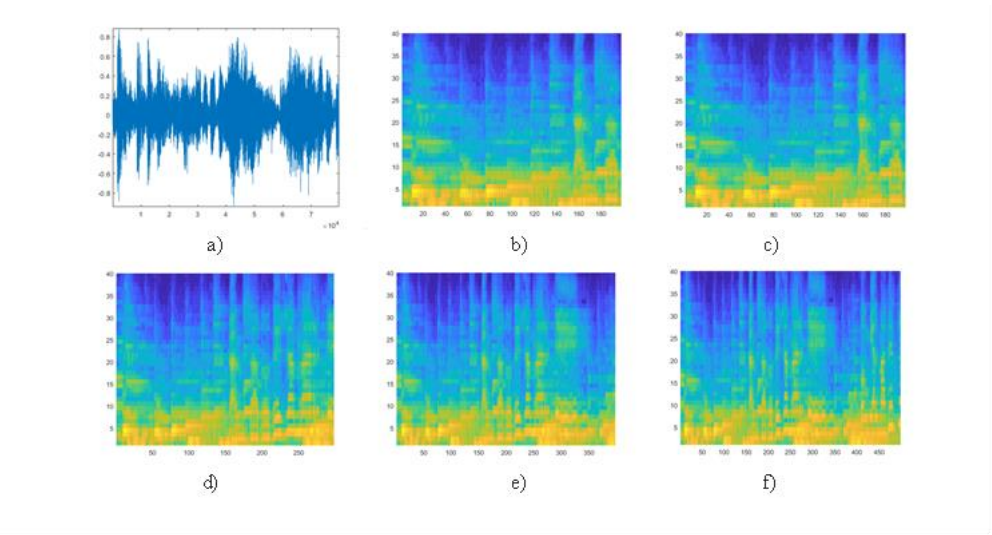

Figura 8. Espectrograma Pasillos a) Señal de Audio b) 1 Segundo c) 2 Segundos d) 3 Segundos e) 4 Segundos e) 5 Segundos

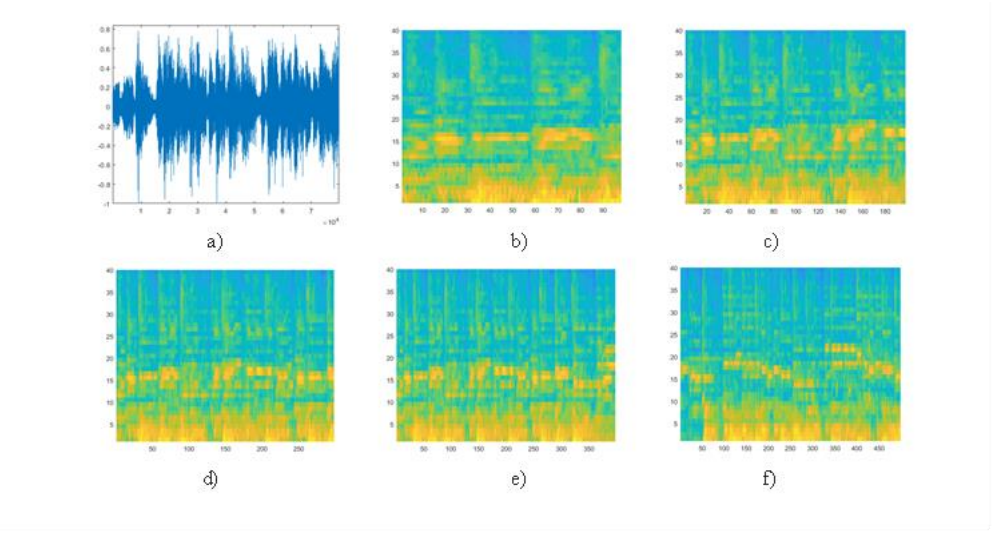

Figura 9. Espectrograma Sanjuanito a) Señal de Audio b) 1 Segundo c) 2 Segundos d) 3 Segundos e) 4 Segundos e) 5 Segundos

3.3. Arquitectura Propuesta

La entrada de la a red neuronal convolucional es un espectrograma-logarítmico, con 5 capas convolucionales, la primera contiene 12 filtros de tamaño 3x3 y una capa *max poling* con un *kernel* de 3x3, la segunda 24 filtros de tamaño 3x3 y una capa *max poling* con un *kernel* de 3x3, la tercera 48 filtros de tamaño 3x3 y una capa *max poling* con un *kernel* de 3x3, la cuarta 48 filtros de tamaño 3x3, la quinta 48 filtros de tamaño 3x3 y una capa *max poling* con un *kernel* de 1x13, después de las capas de convolución la representación se optimiza mediante la función *softmax* y la entropía cruzada ponderada en una capa completamente conectada. La arquitectura se muestra a continuación en la Figura 10 y de forma detallada en el ANEXO B.

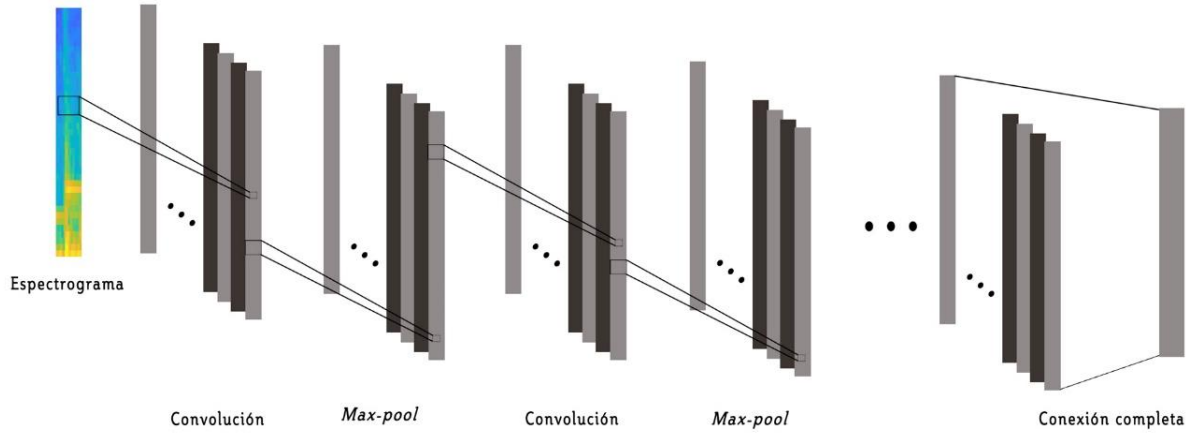

Figura 10. Arquitectura Red Neuronal Convolucional

3.4. Entrenamiento y Evaluación de la red

La red neuronal fue entrenada con la base de datos explicada en la sección 3.2.1 con extractos de 98, 198, 298, 398, 498 *frames* (1, 2, 3, 4, 5 segundos respectivamente) de espectrograma. Se utilizó la optimización adaptativa de ADAM (Kingma & Ba, 2015) durante los experimentos con un *mini-batch* de tamaño 128, durante 25 épocas, después de 20 épocas se reduce la tasa de aprendizaje por un factor de 10. Luego se evalúa la red entrenada calculando la precisión final de la red con los datos de entrenamiento y validación.

El entrenamiento de la red neuronal durante los diferentes experimentos se muestra en la Figura 11, un resumen de las características en la Tabla 4 y la matriz de confusión de los datos de validación en la Figura 12

Tabla 4
Características entrenamiento de la red neuronal

	Error de entrenamiento	Error de Validación	Tamaño de la Red	Tiempo de predicción de una sola imagen en la CPU
1 Segundo	0.56648%	2.4043%	248.6895 kB	5.7058 ms
2 segundos	0.13059%	1.8391%	331.8145 kB	9.4883 ms
3 segundos	**0.045479%**	**1.3049%**	**420.5645 kB**	**13.7367 ms**
4 segundo	0.28041%	1.6841%	503.6895 kB	15.5727 ms
5 segundos	0.070089%	1.3784%	592.4395 kB	17.7564 ms

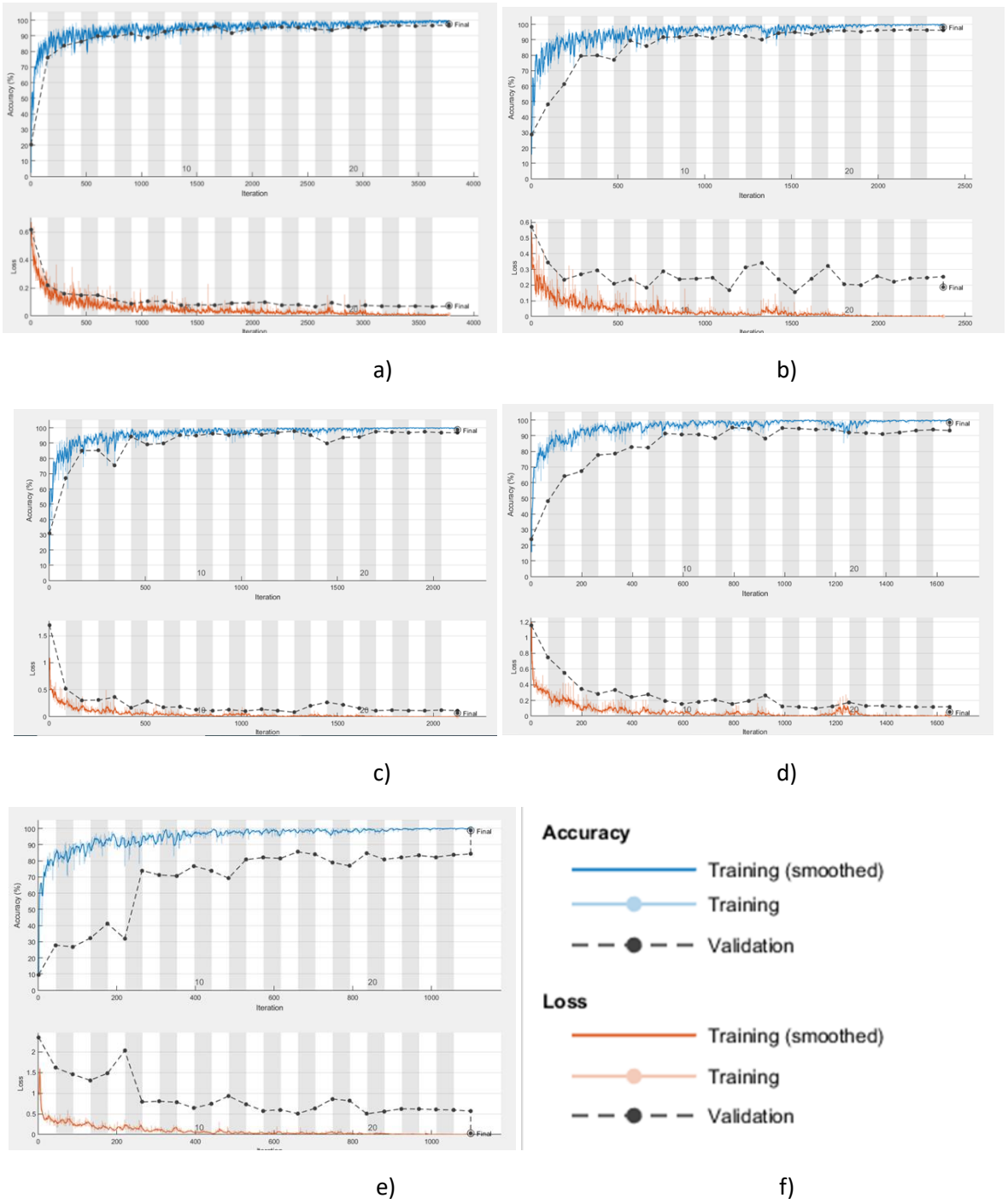

Figura 11. Entrenamiento red neuronal a) Experimento 1 b) Experimento 2 c) Experimento 3 d) Experimento 4 e) Experimento 5 f) Significado de las líneas

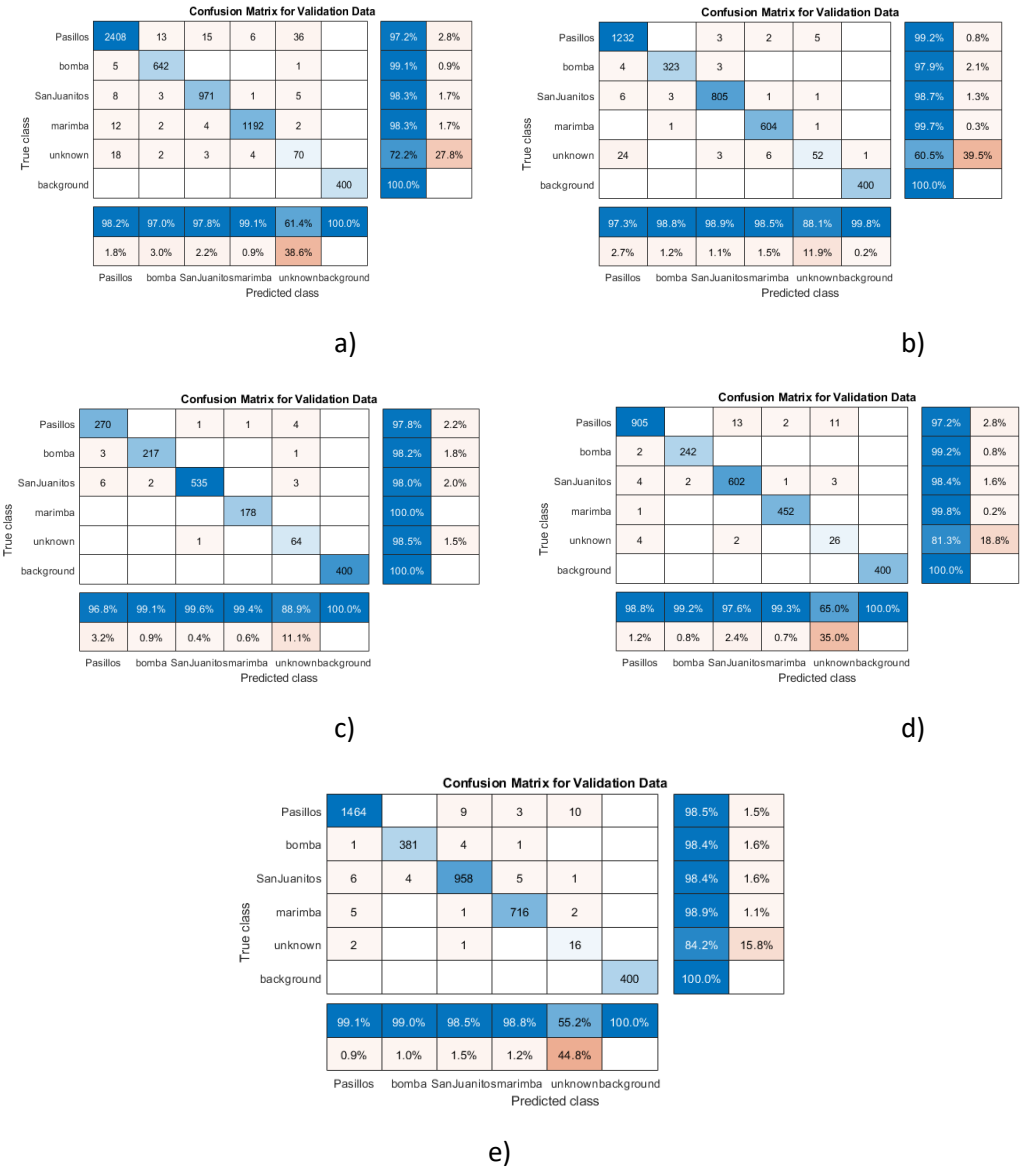

Figura 12. Matriz de Confusión a) Experimento 1 b) Experimento 2 c) Experimento 3 d) Experimento 4 e) Experimento 5

3.5. Elaboración del Clasificador

Una vez que la red neuronal convolucional está lista se elabora el clasificador de géneros musicales mediante el uso del software computacional Matlab®, el cual funciona como se muestra en la Figura 13. Primero el audio entra al clasificador, de acuerdo con el tamaño de la ventana que se está utilizando se segmenta en 1, 2, 3, 4 o 5 segundos, se crea el espectrogama de esta ventana, se analiza y genera una etiqueta que puede ser bomba, marimba, pasillos, sanjuanito o desconocido. La ventana avanza un segundo y el proceso se repite hasta acabar. Finalizado, se suman todas las etiquetas, se muestra el resultado y se genera un archivo .txt por etiqueta en el que se almacena la ubicación del archivo ingresado.

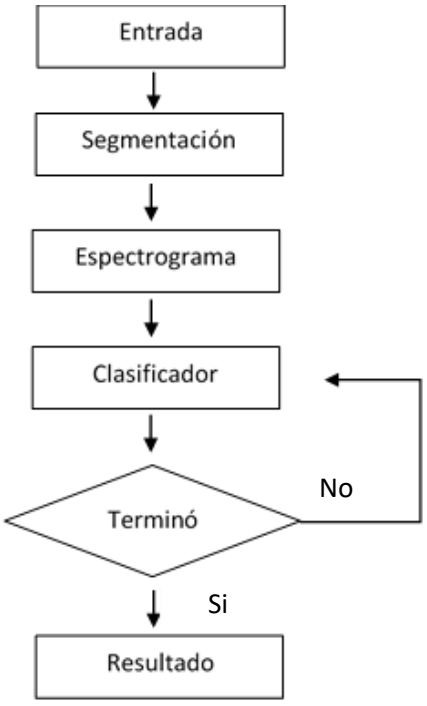

Figura 13. Diagrama de Flujo del Clasificador

CAPÍTULO IV

4. PRUEBAS Y RESULTADOS

4.1. Análisis de resultados

La siguiente sección muestra el rendimiento del sistema de clasificación desarrollado evaluando la base de datos propuesta en la Sección 3.2.2.

Los parámetros que se van a medir son el número total de clips de audio (N_T), número de clips de audio clasificados correctamente (N_E), verdaderos positivos (N_{VP}), falsos positivos (N_{FP}), verdaderos negativos (N_{VN}), falsos negativos (N_{FN}) para analizar así la Exactitud, Precisión, Sensibilidad y Especificidad del clasificador de géneros musicales. A continuación, se muestran las ecuaciones que serán utilizadas para lograr este propósito:

Exactitud (A)

$$A(\%) = \frac{N_E}{N_T} \times 100 \quad (4)$$

Precisión (P)

$$P(\%) = \frac{N_{VP}}{N_{VP} + N_{FP}} \times 100 \quad (5)$$

Sensibilidad (R)

$$R(\%) = \frac{N_{VP}}{N_{VP} + N_{FN}} \times 100 \quad (6)$$

Especificidad (S)

$$S(\%) = \frac{N_{VN}}{N_{VN} + N_{FP}} \times 100 \quad (7)$$

4.2. Análisis del desempeño del clasificador de géneros musicales

4.3. Análisis del experimento 1

El experimento 1 consiste en medir el desempeño del clasificador con un *frame* de 1 segundo. Los resultados de la red neuronal muestran un error de entrenamiento de 0.56% y un error de validación de 2.40 %, el archivo generado a partir del entrenamiento de la misma tiene un tamaño de 248.68 kB, el tiempo de predicción de una sola imagen en la CPU es de 14.5058 ms, por lo tanto el tiempo total en la predicción aplicada en la base de datos en la sección 3.2.2 fue de 9 minutos 48 segundos, los resultados del desempeño del clasificador se muestran en la Tabla 5 y Figura 14 respectivamente.

Tabla 5
Desempeño del Clasificador Experimento 1

	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)
Pasillos	40,33	98,00	98,00	99,85
SanJuanito	81,31	87,00	87,00	99,00
Bomba	85,85	91,00	91,00	99,31
Marimba	85,58	89,00	89,00	99,15
Desconocidos	97,14	81,60	81,60	85,85

De la Tablas 5 se obtienen los siguientes resultados, la predicción con el porcentaje de exactitud más bajo es en pasillos con 40.33%, mientras que bomba, marimba, y sanjuanito no superan el

86%, el porcentaje más alto de la predicción del clasificador es etiquetando géneros musicales desconocidos con un 97.14 % de exactitud.

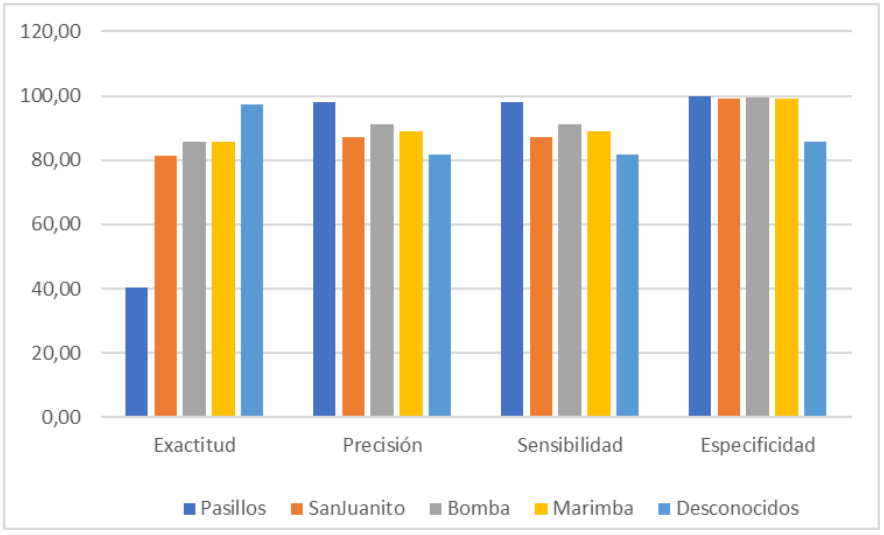

Figura 14. Desempeño del Clasificador Experimento 1

4.4. Análisis del experimento 2

El experimento 2 consiste en medir el desempeño del clasificador con un *frame* de 2 segundos. Los resultados de la red neuronal muestran un mejor entrenamiento en relación al experimento 1, los datos que se obtuvieron son los siguientes: error de entrenamiento de 0.13059%, error de validación 1.8391% , el tamaño del archivo generado a partir del entrenamiento de la misma aumentó a 331.8145 kB, el tiempo de predicción de una sola imagen en la CPU es de 21 ms , por lo tanto el tiempo total en la predicción aplicada en la base de datos en la sección 3.2.2 fue de 13 minutos 43 segundos, los resultados del desempeño del clasificador se muestran en la Tabla 6 y Figura 15 respectivamente.

Tabla 6
Desempeño del Clasificador Experimento 2

	Exactitud	Precisión	Sensibilidad	Especificidad
Pasillos	33,681	97,000	97,000	99,769
SanJuanito	54,348	75,000	75,000	98,077
Bomba	88,571	93,000	93,000	99,462
Marimba	80,583	83,000	83,000	98,692
Desconocidos	96,475	73,900	73,900	79,923

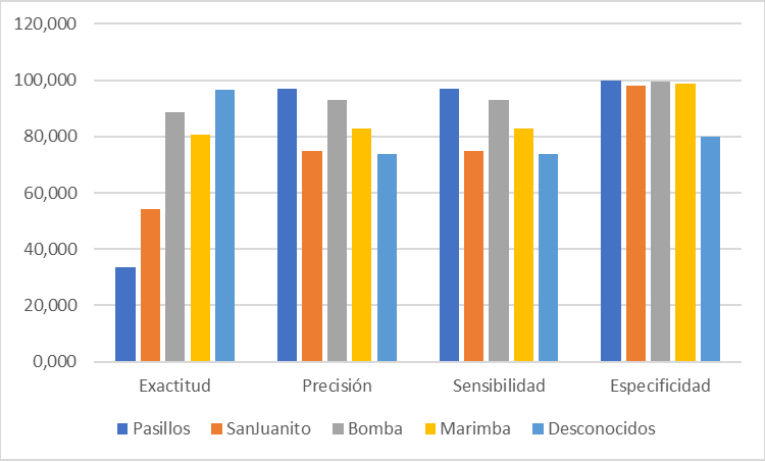

Figura 15. Desempeño del Clasificador Experimento 2

De las Tabla 6 se obtienen los siguientes resultados, la predicción con el porcentaje de exactitud más bajo se presenta en el etiquetado de pasillos y sanjuanito con 33.6% y 54.34% respectivamente, mientras que marimba y bomba son similares al experimento 1 con 80% y 88% respectivamente, el porcentaje más alto de la predicción del clasificador es etiquetando géneros musicales desconocidos con un 96.47 % de exactitud.

4.5. Análisis del experimento 3

El experimento 3 consiste en medir el desempeño del clasificador con un *frame* de 3 segundos. Los resultados de la red neuronal muestran un mejor entrenamiento en relación al experimento 1 y 2, los datos que se obtuvieron son los siguientes: error de entrenamiento de 0.045479%, error de validación 1.3049% , el tamaño del archivo generado a partir del entrenamiento de la misma aumentó a 420.5645 kB, el tiempo de predicción de una sola imagen en la CPU es de 31.5 ms , por lo tanto el tiempo total en la predicción aplicada en la base de datos en la Sección 3.2.2 fue de 19 minutos 50 segundos, los resultados del desempeño del clasificador se muestran en la Tabla 7 y Figura 16 respectivamente.

Tabla 7
Desempeño del Clasificador Experimento 3

	Exactitud	Precisión	Sensibilidad	Especificidad
Pasillos	69,06	96,00	96,00	99,69
SanJuanito	73,96	71,00	71,00	97,77
Bomba	96,74	89,00	89,00	99,15
Marimba	86,60	84,00	84,00	98,77
Desconocidos	95,59	93,30	93,30	94,85

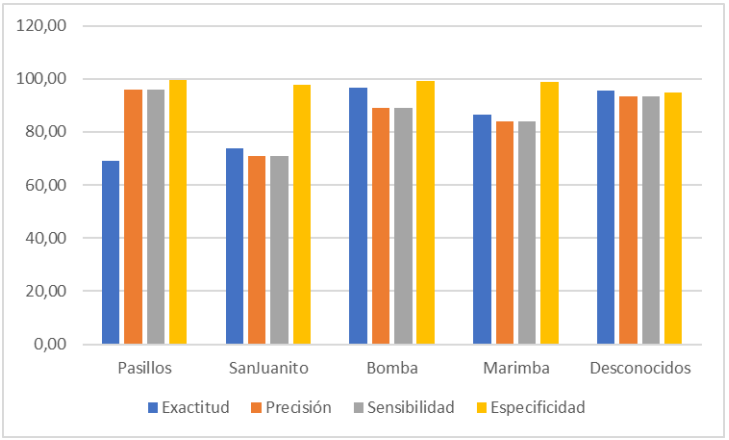

Figura 16. Desempeño del Clasificador Experimento 3

De la Tabla 7 se obtienen los siguientes resultados, la predicción en el porcentaje de exactitud más bajo es en el etiquetado de pasillos 69%,mientras que bomba, sanjuanito y marimba poseen un porcentaje superior al 70% con 96%, 74%, 87% respectivamente, la predicción del clasificador etiquetando géneros musicales desconocidos es de 96.47 % de exactitud.

4.6. Análisis del experimento 4

El experimento 4 consiste en medir el desempeño del clasificador con un *frame* de 4 segundos. Los resultados de la red neuronal son los siguientes: error de entrenamiento de 0.28041%, error de validación 1.6841%, el archivo generado a partir del entrenamiento de la misma tiene un tamaño de 503.6895 kB, el tiempo de predicción de una sola imagen en la CPU es de 40 ms, por lo tanto, el tiempo total en la predicción aplicada en la base de datos en la Sección 3.2.2 fue de 24 minutos 16 segundos, los resultados del desempeño del clasificador se muestran en la Tabla 8 y Figura 17 respectivamente.

Tabla 8
Desempeño del Clasificador Experimento 4

	Exactitud	Precisión	Sensibilidad	Especificidad
Pasillos	54,76	92,00	92,00	99,38
SanJuanito	72,04	67,00	67,00	97,46
Bomba	97,92	94,00	94,00	99,54
Marimba	73,95	88,00	88,00	99,08
Desconocidos	94,59	87,40	87,40	90,31

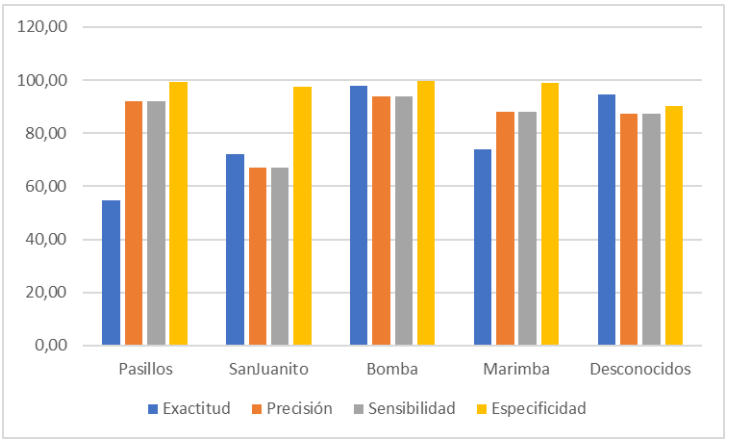

Figura 17. Desempeño del Clasificador Experimento 4

De las Tabla 8 se obtienen los siguientes resultados, la predicción con el porcentaje de exactitud más bajo es en el etiquetado de pasillos 54%, la predicción más alta es bomba con 97.92%, mientras que sanjuanito y marimba 72%, 74% respectivamente, la predicción del clasificador etiquetando géneros musicales desconocidos es de 95 % de exactitud.

4.7. Análisis del experimento 5

El experimento 5 consiste en medir el desempeño del clasificador con un *frame* de 5 segundos. Los resultados de la red neuronal son los siguientes: error de entrenamiento de 0.100089%, error de validación 1.3784%, el archivo generado a partir del entrenamiento de la misma tiene un tamaño de 592.4395 kB, el tiempo de predicción de una sola imagen en la CPU es de 47 ms, por lo tanto, el tiempo total en la predicción aplicada en la base de datos en la Sección 3.2.2 fue de 28 minutos, los resultados del desempeño del clasificador se muestran en la Tabla 9 y Figura 18 respectivamente.

Tabla 9
Desempeño del Clasificador Experimento 5

	Exactitud	Precisión	Sensibilidad	Especificidad
Pasillos	41,85	95,00	95,00	99,62
SanJuanito	47,59	69,00	69,00	97,62
Bomba	93,75	90,00	90,00	99,23
Marimba	75,93	82,00	82,00	98,62
Desconocidos	94,66	78,00	78,00	83,08

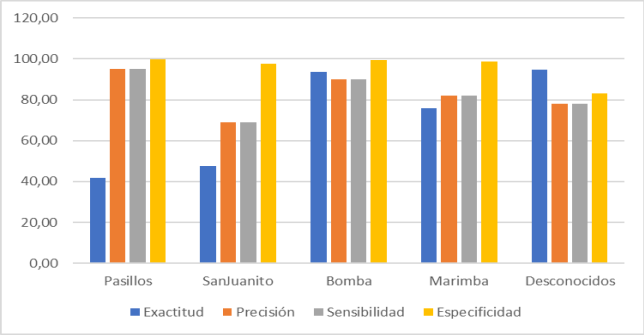

Figura 18. Desempeño del Clasificador Experimento 5

De la Tabla 9 se obtienen los siguientes resultados, la predicción con el porcentaje de exactitud más bajo es en el etiquetado de pasillos y sanjaunito con 42% y 48% respectivamente, la predicción más alta es bomba con 93.7% y marimba 76%, la predicción del clasificador etiquetando géneros musicales desconocidos es de 95 % de exactitud.

4.8. Análisis total de los experimentos

Los resultados para la tarea de predicción de etiquetas obtenida con cada uno de los diferentes experimentos realizados se muestran resumidos en la Tabla 10 y Figura 19.

Tabla 10
Desempeño del Clasificador

	Exactitud	Precisión	Sensibilidad	Especificad
1 Segundo	84,36 %	84,36 %	84,36 %	83,15 %
2 segundos	77,64 %	77,64 %	77,64 %	75,92 %
3 segundos	**90,93** %	**90,93 %**	**90,93 %**	**90,23 %**
4 segundo	86,79 %	86,79 %	86,79 %	85,77 %
5 segundos	79,71 %	79,71 %	79,71 %	78,15 %

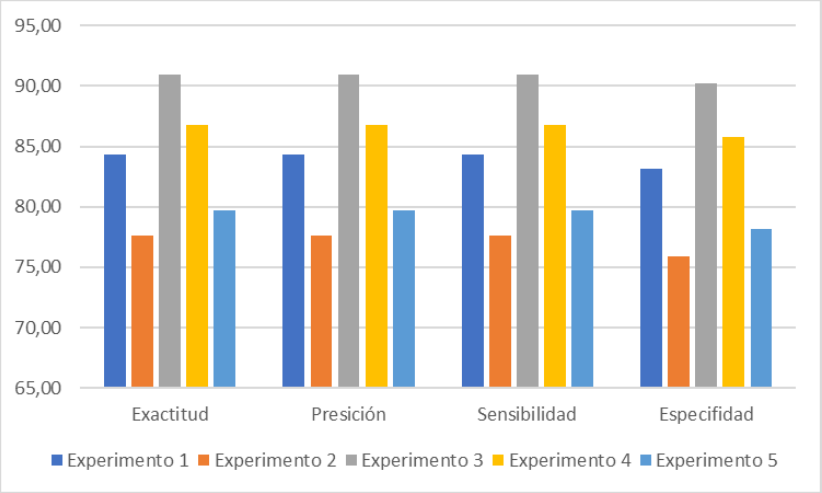

Figura 19. Desempeño del Clasificador

De acuerdo a los resultados que se muestran en la tabla 10 entre el experimento 1 y 2 no se observó una mejoría, el experimento 2 obtuvo valores menores a los obtenidos en el experimento 1, el experimento 3 logró un mejor desempeño que el experimento 1 y 2, el experimento 4 expuso un mejor desempeño que el experimento 1 y 2 pero no mejor que el experimento 3 y el experimento 5 tuvo un desempeño parecido al experimento 2.

En los experimentos se fue aumentando el tamaño de la ventana y se determinó que el clasificador con un *frame* de 3 segundos tiene los mejores resultados con un 90% exactitud, precisión, sensibilidad y especificidad, al aumentar o disminuir el tamaño de ventana el rendimiento comienza a disminuir y la carga computacional comienza a subir.

CAPÍTULO V

5. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones y Recomendaciones

- Se determinó que el clasificador de géneros musicales tiene el mejor desempeño con una ventana de 3 segundos de acuerdo con la Tabla 10, si se aumenta o disminuye la ventana no se consigue mejoras sustanciales.
- De acuerdo con la Tabla 3, el clasificador también puede trabajar con una ventana de 4 segundos, pero el desempeño disminuye un 4% además aumenta la carga computacional, con el resto de las ventanas el desempeño es mucho menor y los resultados ya no son aceptables.
- De los géneros musicales que se analizaron se determinó que el de mayor exactitud fue Bomba de acuerdo con la Tabla 7 por su compás binario compuesto y su fórmula rítmica.
- De los géneros musicales que se analizaron se determinó que el de menor exactitud fue Pasillos de acuerdo con la Tabla 7 debido al tempo propio que tiene este género que se caracteriza por el acompañamiento de guitarras y requinto.
- El tiempo de procesamiento de un espectrograma con una ventana de 3 segundos toma aproximadamente 31.5 ms, para el procesamiento del clip de audio de 30 segundos, el tiempo total estimado es de 0.8505 segundos dichos valores son proporcionados por Matlab®.
- Es importante que la comunidad que se encarga del desarrollo de la investigación en el campo de la recolección de información de música añada a sus bases de datos públicas estos géneros musicales ya que fue un limitante, en este estudio todos los archivos fueron bajados

de Youtube y al dar un formato único se pierde información debido a los conversores que se utilizan.

- Para incrementar el rendimiento de la red se recomienda aumentar la profundidad de esta agregando bloques convolucionales idénticos, el número de filtros o ambas, esto requiere un procesamiento alto y por ende de un hardware más robusto.
- Se recomienda que todos los clips de audio estén en el mismo formato antes de empezar a trabajar con la red neuronal y la base de datos por ítem para el entrenamiento no debe ser menor a una hora para tener un desempeño aceptable.

5.2. Trabajos Futuros

- Como trabajos futuros se propone realizar el clasificador de géneros musicales utilizando la escala mel y comparar el desempeño con la escala Bark.
- Aumentar el tamaño de la base de datos con otros géneros musicales ecuatorianos como el yumbo, dánzate, yaraví, arrullos, alabaos, chigualos, albazo, tonada, pasacalle, música montubia, y el amorfino.
- Realizar un clasificador de géneros musicales que funcione online, identifique el audio mediante un micrófono y muestre el resultado en el dispositivo en phyton.
- En la literatura se encontraron trabajos que mencionan la reducción de la tasa de muestreo para reducir el costo computacional, al trabajar a tasas de muestreo más bajas no afecta en el desempeño del clasificador, esto se podría verificar en trabajos futuros.

BIBLIOGRAFÍA

Buduma, N., & Locascio, N. (2017). Fundamentals of deep learning: Designing next-generation machine intelligence algorithms. O'Reilly Media, Inc.

Choi, K. (2018). Deep Neural Networks for Music Tagging. (Doctor of Philosophy). University of London United Kingdom.

Choi, K., Fazekasñ, G., & Sandler, M. (2016). Automatic tagging using deep convolutional neural networks. ISMIR.

Dieleman, S., & Schrauwen, B. (2013). Multiscale approaches to music audio feature learning. Proceedings of the 14th International Conference on Music Information Retrieval (ISMIR).

Dieleman, S., Brakel, P., & Benjamin Schrauwen. (2011). Audio-based music classification with a pretrained convolutional network. Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR).

Dong, M. (2018). Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification. arXiv e-prints.

Francois, C. (2017). Deep learning with Python. Manning.

Fu, Z., Lu, G., Ting, K. M., & Zhang, D. (2011). A Survey of Audio-Based Music Classification and Annotation. IEEE, 13(2), 303–319.

García Durán, A. (2011). Diseño de un clasificador de géneros musicales basado en modelos dinámicos. Master's thesis.

George E Dahl, T. N. (2013). Improving deep neural networks for lvcsr using rectiffied linear units and dropout. Acoustics,Speech and Signal Processing (ICASSP), 8609-8613.

Glorot, X., Bordes, A., & Bengio, a. Y. (2011). Deep sparse rectifier neural networks. Aistats, 15, 275.

Guaus, E. (2009). Audio content processing for automatic music genre classification : descriptors, databases, and classifiers. (Tesis Doctoral). Universitat Pompeu Fabra. Departament de Tecnologies de la Informació i les Comunicacions. Obtenido de https://www.tdx.cat/handle/10803/7559;jsessionid=3256054E54A8B3C084748FB0818A76CF#page=1

Gulzar, T., Singh, A., & Sharma, S. (2014). Comparative analysis of LPCC, MFCC and BFCC for the recognition of Hindi words using artificial neural networks. International Journal of Computer Applications, 22-27.

Gulzar, T., Singh, A., & Sharma, S. (2014). Comparative Analysis of LPCC, MFCC and BFCC for the Recognition of Hindi Words using Artificial Neural Networks . International Journal of Computer Applications .

Gulzar, T., Singh, A., & Sharma, S. (2014). omparative analysis of LPCC, MFCC and BFCC for the recognition of Hindi words using artificial neural networks. International Journal of Computer Applications, 22-27.

Henaff, M., Jarrett, K., Kavukcuoglu, K., & LeCun, Y. (2011). Unsupervised learning of sparse features for scalableaudio classification. Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR).

Hinton, G., Deng, L., D. Yu, G. D., Mohamed, A., Jaitly, N., Senior, A., . . . Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Mag, 29(6), 82–97.

Kingma, D., & Ba, J. (2015). A method for stochastic optimization. Proceedings of the 6th International Conference on Learning Representations (ICLR).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 1097-1105.

LeCun, Bengio, Y., & G. Hinton. (2015). Deep learning. Nature, 436-444.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks.

Lee, H., Pham, P., Largman, Y., & Ng, A. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. Advances in NeuralInformation Processing Systems .

Lee, S. (13 de 5 de 2018). RedBull. Obtenido de https://www.redbull.com/ar-es/musica-formatos-desde-fonografo-thomas-edison-a-spotify

Li, T., Chan, A. B., & Chun, A. (2010). Automatic musical pattern feature extraction using convolutional neural network. Proc. Int. Conf. Data Mining and Applications.

Lihua, L. (2010). Audio musical genre classification using convolutional neural networks and pitch and tempo transformations. Doctoral dissertation. City University of Hong Kong.

Liu, D. L. (2003). Automatic mood detection from acoustic. Proceedings of the International Symposium on Music, pp. 81–87.

Liu, J. Y., & Yang, Y. H. (2016). Event localization in music auto-tagging. In Proceedings of the 24th ACM international conference on Multimedia , 1048-1057.

Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. ISMIR. , 1-11.

Mandel, M., & Ellis, D. (2005). Song-level features and support vector machines. Proceedings of the International Conference on.

Moore, B. C. (2012). An introduction to the psychology of hearing. Brill.

O'Shaughnessy, D. (2000). Speech communications: human and machine. Institute of Electrical and Electronics Engineers.

Oord, V. d., A., D. S., & Schrauwen, B. (2013). Deep content-based music recommendation. In Advances in neural information processing systems, 2643-2651.

O'Shaughnessy, D. (1987). Speech communication: human and machine. Universities Press .

Patterson, J., & Gibson, A. (2017). Deep learning: A practitioner's approach. O'Reilly Media, Inc.

Pons, J., Slizovskaia, O., Gong, R., Gómez, E., & Serra, X. (2017). Timbre analysis of music audio signals with convolutional neural networks. European Signal Processing Conference (EUSIPCO), 2744-2748.

Saha, M. S. (2012). Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. . Speech Communication, 543-565.

Sergey, S., Hamza, G., & Alexei, K. (8 de 12 de 2017). Representations of Sound in Deep Learning of Audio. Obtenido de https://arxiv.org/abs/1712.02898

Sordo, M. (2011). Semantic Annotation of Music Collections: A Computational Approach. (PhD Thesis). Universitat Pompeu Fabra, Barcelona (Spain).

Trask, A. W. (2019). Grokking deep learning. Manning.

Turnbull, D., Barrington, L., & Lanckriet, G. (2006). Modelling music and words using a multi-class naive bayes approach. Proceedings of the International Conference on Music Information Retrieval.

Tzanetakis, G. E. (2001). Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10, 293 - 302.

Ullrich, K., Schülter, J., & Grill, T. (2014). Boundary Detection in Music Structure Analysis using Convolutional Neural Networks . International Conference on Music Information Retrieval (ISMIR) .

Vasco, M., & Magdalena, M. (2009). Identidades musicales ecuatorianas : Diseño, mercadeo y difusión en Quito de una serie de productos radicales sobre musica nacional. Obtenido de http://dspace.ups.edu.ec/handle/123456789/2577

Vieira, A. &. (2018). ntroduction to Deep Learning Business Applications for Developers: From Conversational Bots in Customer Service to Medical Image Processing. Apress.

Wiley, J. F. (2016). R deep learning essentials. Packt Publishing Ltd.

Wulfing, J. W., & Riedmiller, M. (2013). Unsupervised learning of local features for music classification. Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR).

Yang, J. (2018). Music Genre Classification With Neural Networks: An Examination Of Several Impactful Variables.

Zwicker, E. (1961). Subdivision of the audiblefrequency range into critical bands. The Journal of the Acoustical Society of America, 248.