PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

RECUPERACIÓN DE INFORMACIÓN MUSICAL POR SIMILITUD USANDO REDES NEURONALES

Tesis para optar por el Título de Magíster en Ciencias de la Computación, que presenta:

Jael Nora Rojas Miguel

ASESOR: Ph.D. Kong Wong, Maynard

Miembros del comité examinador:

Pow Sang Portillo, Jose A.

Melgar Sasieta, Héctor A.

Lima, Perú
Noviembre del 2012

Resumen

En los últimos años, la distribución de música digital en la web ha permitido a los usuarios acceder a grandes cantidades de información musical, con ello surge la necesidad de obtener esa información de manera eficaz y eficiente. En la actualidad, los sistemas de recuperación han ayudado a los usuarios a encontrar información basada en texto, pero esos modelos tradicionales no son adecuados si deseamos encontrar canciones que se parezcan en contenido de audio, de allí la necesidad de modelar e implementar métodos de recuperación basado en audio musical.

En este estudio se describe un sistema que permite recuperar y clasificar canciones por similitud basado en contenido de audio musical. Se aplica un modelo de red neuronal a características de canciones. Primero se obtiene descriptores de canciones polifónicas en formato mp3 con características tales como: Análisis Espectral, Patrones de ritmo, Histograma de ritmo. Segundo, se realiza un análisis estadístico para seleccionar los descriptores válidos. Finalmente se ingresa a una red neuronal estos descriptores y se entrena.

El objetivo de este trabajo es implementar el sistema y determinar, a partir de los resultados experimentales, la eficiencia de acierto o no para clasificar y recuperar contenido de audio musical por similitud.

Tabla de contenidos

Listado de Figuras

Listado de Tablas

Introducción

Hoy en día, es posible acceder a grandes archivos de música digital en audio y texto en la web. La demanda de acceso a esa información de manera clasificada y rápida, es un tema de gran importancia; por ello, se han desarrollado aplicaciones que ayudan a identificar canciones y clasificación automática. Para encontrar y clasificar de millones de temas musicales de audio digital, no es suficiente considerar aspectos como género musical, autor, título de la canción, entre otros. Los usuarios desean encontrar rápidamente temas que sean conocidos y de su agrado.

Género musical y similitud son tipo de metadata para búsqueda de música. La clasificación por género musical ha recibido bastante atención de investigadores de música y audio especialmente en la Comunidad de Recuperación de Información Musical [1 2]. El género es intrínsecamente definido para clasificar, pero sufre de ambigüedad porque es visto bajo dos conceptos. Usado como concepto *intencional*, desde este punto de vista es una categoría lingüística (interpretación de un título) y usado como concepto *extensional,* desde este enfoque el género está estrechamente relacionado con el análisis.

En un estudio realizado por Jean-Julien Aucouturier y François Pachet, 2003 [1] describe 3 enfoques para extracción musical por género. Primero, la clasificación manual de títulos, pero no tan realística para grandes base de datos. Segundo, el género puede ser evaluado a partir de los atributos intrínsecos de la señal. Tercero, de un análisis, extraer similitudes entre títulos de artistas, pero no es fiable. De los enfoques mencionados, este estudio se orienta a la clasificación que evalúa los atributos de la señal de audio. En similitud musical los conceptos son parecidos, la percepción depende de varios fenómenos como timbre, ritmo, cultura, contexto social, entre otros. Aspectos relacionados a características de la señal se verán en esta tesis.

Este estudio se enfoca en la recuperación y clasificación de canciones por similitud basada en contenido de audio. Para este fin se ha propuesto un sistema simple que desarrolla técnicas que pueden aprovechar la información que se evalúa. Se usa para ello redes neuronales de Retropropagación, entrenadas con datos de características de audio. Previo al entrenamiento en esta red, se realiza un pre-entrenamiento donde se analizan los datos que se ingresan a la red filtrando información no relevante. La Figura 0.1 muestra el sistema propuesto.

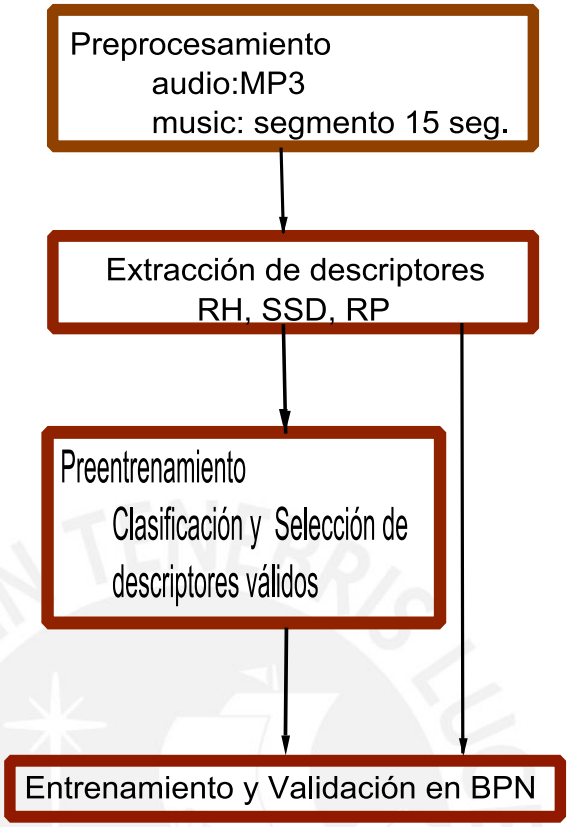

Figura 0.1: Estructura general del sistema propuesto.

Fuentes bibliográficas relacionadas

Los artículos más relevantes relacionados con el campo de Recuperación de Información musical se publican en C*omputer Musical Journal, Journal of new music Research, IEEE Transactions on Acoustics Speech and Signal Processing.*

Además se publican artículos relacionados con la Informática musical en Congresos como *International Conference of Music and Artificial Intelligence e International Symposium on Music Information Retrieval, Music Information Retrieval Evaluation Exchange, International Society for Music Information Retrieval Conference.*

Los centros de investigación más importantes son, el *IRCAM Institute Recherche Coordination Acoustique Musique, CCRMA (Center for Computer Research in Musics and Acoustics) de la Universidad de Stanford, y el Medialab del MIT (Massachussets Institute of Technology).*

Capítulo 1: Generalidades

En este capítulo se describe brevemente el fondo y contexto, seguido de los alcances, objetivos, la metodología que se desarrollará y una breve descripción del documento.

1.1 Fondo y Contexto

Se han implementado sistemas de recuperación musical basados en texto y audio [3]. Podemos encontrar muchos artículos que describen diversos métodos, técnicas, modelos, algoritmos y aplicaciones que ayudan a usuarios a navegar en grandes catálogos musicales tales como; descriptores de audio [3], música por similitud, búsqueda por tarareo [4], clasificación musical basada en contenido de audio entre otros.

Se usa el modelo de red neuronal de Retropropagación; esta metodología es un modelamiento que permite controlar sistemas usando data adquirida por entrenamiento. El modelo de red neural de Retropropagación es ampliamente usado en reconocimiento de patrones porque puede clasificar patrones complejos además de la capacidad de aprender y almacenar información.

1.2 Alcances y Objetivos

El objetivo del presente trabajo es diseñar un sistema para clasificar y recuperar canciones por similitud basado en audio e implementar el modelo de red neuronal usando el algoritmo de Retropropagación para recuperar una canción similar a partir del análisis de los descriptores de la señal de audio obtenidos para cada canción. Finalmente evaluar los resultados y determinar la eficiencia.

1.3 Logros

En este trabajo, se propone un sistema de clasificación y recuperación de audio implementado en una red neuronal. Este sistema es similar a SOMeJB, basado en redes neuronales Kohonen [15 16]. En el sistema propuesto se desarrolló un método para obtener un vector de descriptores, que representa un género musical para clasificación, que es usado en la red neuronal de retropropagación.

Otro método basado en el algoritmo de retropropagación se implementó para reconocimiento de audio por similitud.

1.4 Metodología

La construcción del sistema se desarrollará considerando las siguientes etapas:

1.4.1 Primera parte: Obtención de descriptores de audio

Los descriptores, son las características de la señal de audio extraído del espectro normado por el estándar MPEG-7, destacan los descriptores de bajo y alto nivel, para este caso se considera de bajo nivel.

Se seleccionó el método usado por investigadores de la Universidad Tecnológica de Viena y MIR [5], para extraer información semántica de la música usando procesamiento de señal digital y psicoacústica. Estas técnicas de extracción de características, analizan la acústica de la señal tales como; ritmo, presencia de voz, timbre entre otros.

Tipos de descriptores usados:

- Patrones de ritmo (RP): Describe la amplitud de la modulación para un rango o banda de frecuencias del rango auditivo humano.
- Estadística del Espectro (SSD): Calcula sobre los valores del sonograma de cada banda la media, mediana, varianza, asimetría, kurtosis, mínimo y máximo valor.
- Histograma del Ritmo (RH): Es calculado tomando la mediana del histograma cada 6 segundos de segmento procesado. Llamado también Histograma de energía rítmica por modulación de frecuencia.

1.4.2 Segunda Parte: Entrenamiento y validación de la red

Red asociador de patrones por Retropropagación del error.

Consiste de dos fases:

Fase de Entrenamiento

1. El aprendizaje es de modo supervisado y los valores de los pesos son aleatorios.
2. Se selecciona 5 patrones para el entrenamiento y se coincide las salidas con la matriz identidad.
3. Se calcula el error con la salida, hasta minimizar.
4. Se guardan los datos (pesos).

Fase de Validación

1. Se ingresa el vector de descriptores de consulta.
2. Se reutiliza los pesos de la fase de entrenamiento
3. Se selecciona el mayor del vector de salida, el cual corresponde al vector similar buscado.

1.4.3 Tercera Parte: Métricas para determinar el desempeño de la evaluación en relación a otros modelos.

Las métricas usadas son: *Exactitud (Accuracy),* donde los datos obtenidos se representan en una Matriz de Confusión y *geometric mean (g-mean).* Ambos miden el rendimiento de recuperación de información musical para el sistema propuesto.

1.5 Descripción del documento

En este documento se ha incluido los siguientes capítulos:

Capitulo 2, Estado del Arte se describe los algoritmos más usados para recuperación y clasificación de información musical por similitud. Descripción breve de los modelos de red usados en estos trabajos.

Capitulo 3, Redes Neuronales Retropropagación se describe el modelo de red usado y el algoritmo.

Capitulo 4, Resultado Experimentales, se describe el hardware usado para las pruebas, descripción de la implementación y los resultados computacionales de las pruebas.

Capítulo 5, Conclusión se describe las conclusiones y trabajo a futuro para esta tesis.

Capítulo 2: Estado del Arte

En este capítulo se resume los algoritmos usados en recuperación musical basados en contenido de audio y se describe brevemente conceptos de Redes neuronales y el usado en este trabajo.

Definición de terminología usada

La siguiente terminología se mencionará con frecuencia.

***Descriptores/patrones*:** Es una representación de una o más características que proporciona información de la señal.

MPEG-7: Es una representación estándar de la información audiovisual que permite la descripción de contenidos. Estándar de la Organización Internacional para la Estandarización ISO/IEC.

MIR (Music Information Retrieval): Ciencia interdisciplinaria que estudia la recuperación de información de una pieza musical.

Pitch Musical: Es la frecuencia fundamental de una señal. Mide el número de ocurrencias por unidad de tiempo.

Sonograma: Conocido como espectrograma, usado para identificar sonidos.

Descriptores de audio: Vienen a ser las características de la señal.

FFT (Fast Fourier Transform): Algoritmo para calcular la Transformada de Fourier.

Capas Ocultas: Neuronas que reciben entradas de capas anteriores y salidas que pasan a capas posteriores.

Scala Bark: Escala Psicoacústica (24 bandas críticas del oído).

Mediana: Dada una muestra, esta se considera el valor de la posición central.

Kurtosis: Medida de la forma de la distribución de probabilidad. Estudia la proporción de la varianza, se explica por la combinación de datos extremos respecto a la media en contraposición con datos alejados de la misma.

MFCC (Mel Frequency Cepstral Coefcients): *Coeficientes* cepstrales en las frecuencias de Mel; son coeficientes usado en tareas como reconocimiento automático del habla y procesamiento de música.

Edit Distance: Permite medir la similitud cuantitativa. Es el mínimo número de modificaciones (inserción, supresión, sustitución) realizado en una cadena origen con el fin de ser similar a la cadena objetivo.

String Contour: Es la cadena de la forma de la melodía.

Textura: Origen de sonido creado.

Monofónica: Un solo sonido.

Polifónica: Muchos sonidos.

Beat: Constante de tiempo marcado. Dividido en unidades pequeñas llamadas pulsos.

2.1 Métodos de Recuperación de Información Musical

En el dominio de audio, existen diferentes métodos de búsqueda; los sistemas MIR consideran dos grupos principales: Recuperación de datos simbólicos y Recuperación de datos de audio.

2.1.1 Recuperación de datos simbólicos

En este grupo, la melodía es analizada según la secuencia de las notas. Para música monofónica, la secuencia de las notas son descritas una cada vez y los métodos usados son: representación de secuencia de intervalos, secuencia de *pitch* y algoritmos de búsqueda de cadenas como el algoritmo para calcular *edit distance*, encontrar la más larga subsecuencia común o encontrar ocurrencias de una cadena.

Para música polifónica, la secuencia de las notas es descrita a la vez, la música es vista como una secuencia de eventos con propiedades como: tiempo, *pitch* y duración. Los métodos por coincidencia probabilística, tratan de determinar las propiedades probabilísticas de cada melodía y compararlas con las propiedades de búsqueda.

Estas metodologías no son usadas en este trabajo debido a que la melodía que se analizará no está basada en texto sino en audio.

2.1.2 Recuperación de datos de audio

La literatura actual menciona 3 metodologías.

- Extracción de características relevantes.

  La melodía analizada es comparada con otras a partir de la extracción de descripciones abstractas de una señal de audio. Algunas características comunes extraídas de una ventana de audio son: *pitch*, volumen, *chroma*, tono, MFCC y derivados [6]. En la siguiente sección describimos trabajos que usan las características relevantes mencionadas.

- Firmas de audio.

  Usado para identificar grabaciones también para estimar la calidad de una grabación. Esta metodología no es usada en este trabajo, se aplica mayormente para identificar grabaciones con ruido causado por parlantes, teléfonos celulares, micrófonos baratos.

- Mapa auto-organizativo (SOM).

  Llamado también Mapa de Kohonen. Es un algoritmo de red neuronal usado para clusterizar y clasificar piezas musicales ordenados en dos dimensiones. El vector modelo que contiene cualquier característica es ajustado al vector correspondiente de tal manera que las distancias sean minimizadas. Trabajos similares se describen en la sección 2.3.1.

Para el modelo propuesto se usa una metodología similar a SOM, se emplea el algoritmo de la red neuronal de Retropropagación para aprendizaje supervisado y el vector modelo contiene descriptores de ritmo y energía.

2.2 Algoritmos usados en recuperación y clasificación de audio.

Las metodologías existentes y algoritmos usados para sistemas de recuperación de información musical (MIR), se basan en estudios estadísticos, máquinas de aprendizaje, métricas de similitud musical, forma musical entre otros.

En esta sección mencionamos algunas características de los sistemas MIR.

2.2.1 Sistemas MIR

- ***CubyHum***

  Consiste en detectar el *pitch* en una canción y compararlos con representación simbólica de melodías conocidas. Melodías similares a los *pitches* de la canción son recuperadas. Nueve intervalos (distancia entre dos *pitches*) de clases son usados, intervalos largos sobre los 6 semitonos no son considerados. El método de filtración usado es el algoritmo LET (*Linear expected time*) [7], está basado en la edición de la distancia [8].

  Presentado en la *conferencia International Society for Music Information Retrieval ISMIR Proceedings.*

- ***Cuidado music browser***

  Basado en medidas por similitud y en características tales como ritmo, energía, timbre y voz. Análisis de la distancia basado en el modelo *Gaussiano*. [9].

  Presentado en *Proceedings of the 3rd International Workshop on Content-Based Multimedia Indexing, 2003.*

- ***Musipedia***

  El buscador recupera 100 entradas acorde a *edit distance*, de la forma de las cadenas, que son una secuencia de *pitches* para una melodía o ritmo dado. Para indexar usa el método descrito por [10 11].

  Presentado en, *ACM Transactions on Computer-Human Interaction, 2001.*

- ***Notify Whistle***

Consultas de música monofónica se comparan con un conjunto de notas polifónicas. Un rastreador de ritmo se adapta aún si hay variaciones o diferencias en tiempo [11].

Presentado en, *112th Convention of the Audio Engineering Society, 2002.*

- ***Query by Humming***

Se basa en la observación de la forma de la melodía definida como la secuencia de diferencias relativas de *pitch* entre notas sucesivas. Usa un alfabeto de 3 relaciones posibles entre cada *pitch* (U, D y S), estos representan donde se encuentra la nota musical: sobre, abajo o en igual que la nota anterior. El Baeza-Yates/Perleberg [12] algoritmo de coincidencia de patrones es usado para encontrar de una cadena patrón $P = p1p2p3....pm$ en una cadena de texto $T = t1t2t3....tn$ de modo que haya a lo mucho k desajustes (caracteres que no son iguales), para cada instancia de P en T [4].

La figura 2.1 muestra la arquitectura del sistema.

Presentado en *Proceedings ACM Multimedia, 1995.*

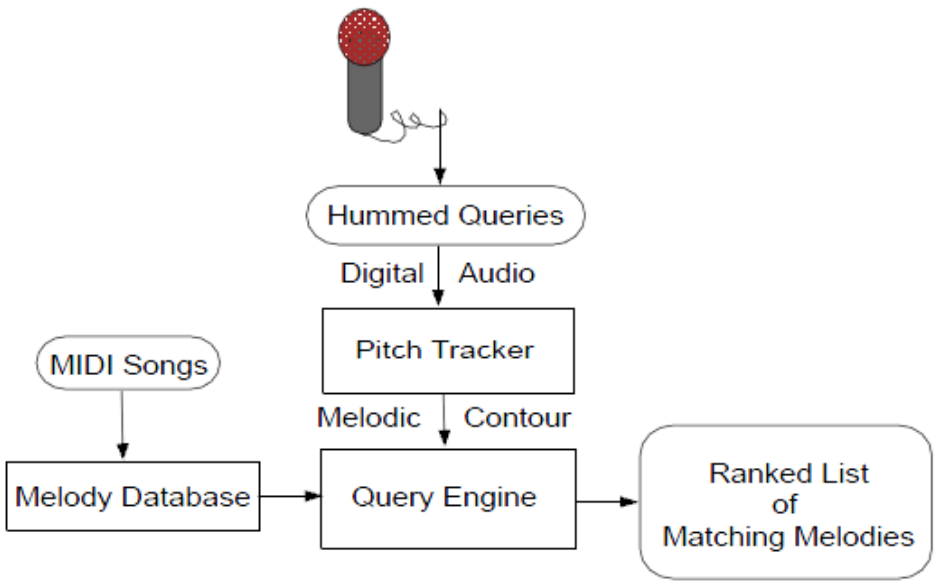

Figura 2.1: Arquitectura del sistema *Query by Humming (propuesto por* [11]).

- ***SoundCompas***

  Los usuarios establecen un metrónomo para indicar el tiempo conveniente y luego tararean su melodía de modo que los *beats* (ritmo, pulsación) coincidan con los clics del metrónomo. Tres vectores (tono parcial, tono transición, tono distribución) son almacenados por superposición de ventanas de la canción. *Sound Compass* realiza el cálculo de la *distancia Euclideana* [11].

  Presentado en *Proceedings ACM Multimedia, 2000*

2.2.2 Cuadro comparativo de Sistemas MIR

La tabla 2.1 muestra el algoritmo y las características relevantes tales como: *pitch*, energía, timbre, *string contour*, tono y ritmo, usados en algunos sistemas MIR y en el propuesto (MusicS). Se observa que las características mayormente usadas son energía, ritmo y *pitch*, estos son considerados los atributos básicos de la música y que mejor describen a una señal de audio

Tabla 2.1 Descripción de las características y algoritmos de sistemas MIR mencionados y el sistema propuesto.

Características	*CubyHum* [7 8]	*Cuidado* [9]	*Musipedia* [10 11]	*Notify* [11]	*Query by Humming* [12 4]	*Sound compas* [11]	***MusicS***
Pitch	x		x		x	x	X
Ritmo		x	x	x		x	X
Energía		x		x			X
Timbre		x		x			
String contour			x				
Tono						x	
Algoritmo	*LET (Linear excpected time)*	Análisis de distancia modelo Gaussiano	*Edit distance*	Comparación música monofónica con polifónica	Baeza-Yates, algoritmo de coincidencia de patrones	Distancia Euclideana	***BPN (Neural Backpropagation)***

2.3 Arquitecturas de Redes Neuronales usados en recuperación y clasificación de audio.

En esta sección describimos las redes neuronales, componentes, usos y aplicaciones.

Las redes neuronales vienen siendo estudiadas desde los años 1980; en las últimas décadas se ha venido desarrollando muchas aplicaciones para resolver problemas como clasificación de patrones, clusterización/categorización, aproximación a una función, predicción, optimización, control, recuperación por contenido, optimización y filtro de ruido.

A continuación se menciona aspectos teóricos definidos por la literatura actual.

Neurona

Vienen a ser unidades de procesamiento interconectados, los que reciben como entradas señales discretas o continuas, es ponderada y envía el resultado a las neuronas conectadas. Matemáticamente se representa:

$$y = \theta(\sum_{j=1}^{n} w_j x_j - u)$$

,

donde $\theta(.)$ es una función de transferencia a 0 y w_j es el peso asociado con la *j-ésima* entrada, se considera el umbral *u* como otro peso sujeto a la neurona con entrada constante igual a *1*.

Las funciones de activación son: *lineal, sigmoidal o gaussiana*.

El proceso es simple y único para la neurona, consiste en recibir entradas de células vecinas y calcular la salida la cual se envía a todas las células restantes [13].

Arquitecturas de red

Son agrupados en dos categorías:

- *Feed-forward networks,* en los grafos no ocurren bucles, por lo general son redes estáticas produce una sola salida.

- *Recurrent network,* en los cuales los bucles ocurren, son sistemas dinámicos.

Aprendizaje

Es descrito como un procedimiento en el cual las reglas de aprendizaje son usadas para ajustar los pesos y bias. Se actualiza la red en este proceso para desarrollar la tarea específica. Para mejorar el desempeño de la red se debe adaptar los pesos en el tiempo. Las redes aprenden usando reglas que han sido dadas.

Establecer reglas bien definidas para la solución del problema de aprendizaje se llama algoritmo de aprendizaje [14].

Existen 3 principales categorías de aprendizaje:

- Supervisado: La red se asigna a una respuesta correcta llamada salida a cada patrón de entrada. Mencionamos algunas redes: *Perceptron* simple o multicapas, Recurrente, Multicapas hacia adelante, Competitiva, Red ART, etc.

- No supervisado: No se requiere respuesta correcta asociada con cada patrón de entrada. Mencionamos algunas: Multicapas hacia adelante, Competitiva, Red *Hopfield*, Competitivo, *Kohonen´s* SOM, Red ART, etc.

- Hibrido: Se combina con el aprendizaje supervisado y no supervisado. Red RBF (*Radial Basis Function*) [13].

Error

El principio básico es usar la señal de error *(d-y)* para modificar los pesos de conexión y gradualmente reducir el error [13].

Actúa como mecanismo de control, el propósito es aplicar una serie de ajustes correctivos a la sinapsis de los pesos de la neurona [14].

Entradas

Valores discretos o continuos que son ingresados para el aprendizaje de la red.

Pesos

Intensidad de la sinapsis al conectar una neurona con otra. Los valores iniciales son pequeños aleatorios.

Bias

Es una entrada definida por 0 ó 1 que puede alterar el valor de la salida.

Tasa de aprendizaje

Determina el tamaño del paso que usa el algoritmo para moverse entre los pesos. Los valores pequeños provocan convergencia lenta y los valores altos provocan divergencia.

Red de Retropropagación

Creada por E. Rumelhart, Geofrfrey E. Hinton y Ronald J. Williams en el año 1986, esta red neuronal multinivel aprende la asociación que existe entre los patrones de entrada y las clases correspondientes, está basado en la generalización de la regla delta o regla del mínimo cuadrado medio.

Algoritmo

1. Inicializar los pesos con pequeños valores aleatorios.
2. Aleatoriamente escoger el ingreso de patrones $X^{(\upsilon)}$.
3. Propagar hacia adelante a través de la red.
4. Calcular θ_i^L en la capa de salida $(o_i = y_i^L)$.

   $$\theta_i^L = g'(h_i^L)[d_i^u - y_i^L]$$

   Donde h_i^L representa la entrada a la i-ésima unidad en la l-ésima capa y g' es la derivada de la función de activación g.

5. Calcular los deltas de las capas anteriores propagando los errores hacia atrás.

   $$\theta_i^l = g'(h_i^l)\sum_j [w_{ij}^{l+1} - y_j^{l+1}]$$

   Para l = (L-1),…., 1.

6. Actualizar los pesos usando

   $$\Delta w_{ji}^l = n\theta_j^l y_j^{l-1}$$

7. Regresar al paso 2 y repetir para el siguiente patrón hasta que el error en la capa de salida esté por debajo del umbral pre-especificado o se alcance un número máximo de iteraciones.

2.3.1 Algoritmos de Redes Neuronales usado para recuperación y clasificación de audio por similitud.

- **SOMeJB The SOM-enhanced JukeBox**

Mapa auto-organizativo es usado para *clusterizar* y visualizar colecciones musicales. Usando el patrón de ritmo, se aplica el algoritmo SOM para organizar piezas musicales en un mapa de dos dimensiones en las que las piezas similares son agrupadas.

SOM *Kohonen map*, es un modelo de red neuronal con aprendizaje no supervisado que permite reducir la dimensionalidad y encontrar la similitud entre los datos de entrada. Durante el entrenamiento, los vectores son ajustados a los datos tal que la distancia entre los datos y los vectores de salida son minimizados.

El proceso de entrenamiento de SOM inicia con el ingreso de patrones y vector de pesos, cada iteración *t* inicia con la selección aleatoria de una entrada de patrones x (t), este es presentada al SOM y cada unidad determina esta activación. La unidad con la más baja activación es la ganadora de la iteración del entrenamiento, se representa: $m_c(t) = \min_i \left\| x(t) - m_i(t) \right\|$. Finalmente, el vector de pesos del ganador, así como los vectores de pesos de unidades seleccionadas en la vecindad del ganador son adaptados.

Durante el entrenamiento, el GHSOM (*Growing Hierarchical Self-organizing maps*) adapta su arquitectura de acuerdo a los requerimientos de los datos de entrada, permitiendo la representación jerárquica de los datos. [15][16].

- **Audio-Based Music Classification with a pretrained Convolutional Network**

La red de convolución es usado para realizar el reconocimiento de artistas y género. La red resume las características de audio a través de escalas de

tiempo. Se usa un pre-entrenamiento no supervisado, el entrenamiento se realiza en una convolucional *deep belief network* (DBN), luego se usa los parámetros aprendidos para inicializar la red *multilayer perceptron* (MLP).

Se ingresa a la red las características de timbre y *chroma* por cada pista, luego separa en capas de convolución, en cada capa aprende las características de la pista. Finalmente las salidas máximas de las capas se concatenan. Para clasificar una pista se realizan separadamente y se promedia el resultado de la distribución de cada clase; la clase probable es la seleccionada [17].

2.3.2 Cuadro comparativo de algoritmos de redes neuronales de sistemas MIR

La tabla 2.2 da una vista general de características de sistemas usando Redes neuronales y el sistema propuesto en esta tesis (II). Los trabajos que se menciona, (I) y (II) fueron expuestos en la *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, realizado en Florida (USA); esta conferencia se inició el 2000 y se viene realizando cada año. Es uno de los principales foros mundiales para la investigación en búsqueda, creación, procesamiento y uso de datos musicales.

El primer trabajo, *Audio_based music classification with a pretrained convolutional network* (I) hace uso de algoritmos de Red de convolución y Perceptrón multicapa de tipo no supervisado, 20 géneros musicales fueron seleccionados. Entrenaron la red usando como dato características de timbre y *chroma*, el ratio de aprendizaje para la red es de 0.5-0.00005 con 30 iteraciones, para el modelo propuesto el ratio de aprendizaje es 0.00001 con 999 iteraciones, esto significa una desventaja; es ideal que el número de iteraciones sea pequeño para reducir el tiempo de procesamiento. El porcentaje de acierto en reconocimiento de género es 29.52% menor al obtenido por el modelo propuesto, sin dejar de considerar que la cantidad de pistas o audios evaluados son distintos en ambos casos.

Literatura: Sander Dieleman, Philémon Brakel and Benjamin Schrauwen.

El segundo trabajo, *Recuperación de Información musical por similitud – MusicS* (II); sistema propuesto en esta tesis, se describe en el capítulo 4.

El tercer trabajo *The SOM-enhanced JukeBox: Organization and Visualization of Music Collections based on Perceptual Models* y *Genre-oriented Organization of Music Collections using the SOMeJB System: An Analysis of Rhythm Patterns and Other Features Proceedings of the DELOS* (III), usa el algoritmo de red *SOM-GHSOM de Kohonen Map* de tipo no supervisado y las

características relevantes de audio: ritmo, melodía, *pitch* usados asimismo en el sistema propuesto. El proceso de entrenamiento en ambos casos es similar se ingresan vectores con las características de audio que son multiplicados con el vector de pesos aleatorios. La desventaja de la red Kohonen es que puede presentar una arquitectura compleja si se usa GHSOM, variante de SOM, se tendría una arquitectura organizada acorde con la distribución de los datos; el sistema propuesto no plantea una arquitectura compleja, la distribución de los datos es organizada en el pre-entrenamiento. Los resultados de recuperación de datos para 86 pistas señalan: música clásica 54.7% de acierto y para 20 pistas: música electrónica 100% de acierto.

Literatura: *Journal of New Music Research (JNMR), 32(2):193-210, Swets and Zeitlinger, 2003. Rauber, E. Pampalk, D. Merkl.*

Tabla 2.2 Descripción de características y algoritmos de Redes Neuronales de sistemas MIR mencionados y el sistema propuesto.

Trabajos	**I**		**II**	**III**
Algoritmo	Red de convolución	Perceptrón multicapa	**Red Backpropagation**	SOM Kohonen map
Tipo	No supervisado		**Supervisado**	No supervisado
Formato			**MP3**	MP3
Géneros	20		**5**	
Ritmo			**X**	X
Melodía			**X**	X
Pitch	X		**X**	X
Timbre	X			
Ratios	0.05-0.00005		**0,00001**	
Iteraciones	30		**999**	
Exactitud	Reconocimiento Género: 29.52% Reconocimiento Artista: 35.74%		**Reconocimiento Género: 60%**	Reconocimiento Género: 86 pistas clásica: 54.7% 20 pistas electrónicas: 100%.

2.4 Observaciones

Existen trabajos similares que no se mencionan en este estudio porque no se basan en las técnicas del MIR, los descritos han sido propuestos e implementados como sistemas comerciales. Algunos de estos productos vienen siendo usados por usuarios en la web; otros como SOMEJB, *Query by Hummming* con su producto comercial MIDOMI, son aplicaciones usados en móviles.

Se menciona algunas observaciones sobre la base de trabajos relatados.

Usar solo ciertas características relevantes de audio es un factor que va a definir el tipo de búsqueda que se realizará, se podría considerar el utilizar en todos los casos las características fundamentales de la música como ritmo, armonía y melodía.

Búsqueda por género y similitud caen aparentemente en el mismo concepto, se podría considerar excluyentes si se trata de búsquedas independientes una de la otra. Sería necesario buscar algoritmos que representen y fusionen estos dos conceptos.

Capítulo 3: Desarrollo

En este capítulo se describe la metodología usada para la obtención de descriptores de canciones y el algoritmo de la Red Neuronal de Retropopagación para el modelo propuesto.

3.1 Primera etapa: Obtención de descriptores de Audio

Los descriptores de audio son las características de la señal normado por el estándar MPEG-7, destacan los descriptores de alto nivel que incluyen reconocimiento de sonido general y herramientas para descripción de timbres instrumentales, descripción de melodías, descripción del contenido hablado; los descriptores de bajo nivel que describen características del contenido espectral, paramétricas y temporales de la señal; para este caso se considera descriptores de bajo nivel. Los descriptores son extraídos de la representación espectral de una señal de audio particionados en segmentos de 6 segundos. Se describe las características del algoritmo para:

Patrones de ritmo (RP):

Llamado también fluctuación de patrones.

Compuesto por dos estados:

- Sonoridad en 24 bandas de frecuencia, es computarizado por *FFT*, agrupando las bandas de frecuencia resultantes a scala b*ark* (escala psicoacústica) y transformaciones sucesivas en decibelios, *phon* (unidad de nivel de volumen para tonos) y escalas *sone*. Se obtiene el sonograma de sensación de sonoridad
- En este estado se aplica un *FFT* al sonograma, resultando un espectro, modulación por frecuencia de modulación para cada banda crítica [18].

Estadística del Espectro (SSD):

Se calcula sobre los valores del sonograma de cada banda crítica. Los valores estadísticos calculados son: la media, mediana, varianza, asimetría, kurtosis, mínimo y máximo valor [18].

Histograma del Ritmo (RH):

Histograma de energía rítmica por modulación de frecuencia. Es calculado tomando la mediana del histograma por cada 6 segundos de segmento procesado [18].

La base de datos musical que se usará para el entrenamiento y testeo contiene descriptores de 50 canciones polifónicas, alrededor de 15 segundos de duración por cada una. La colección corresponde a canciones de 5 géneros musicales: bossa-nova, clásica, pop, rock y jazz.

Los descriptores obtenidos forman un vector de 1668 valores. Para reducir la cantidad de datos por vector que se ingresará a la red, se realiza un análisis previo se detalla a continuación.

3.1.1 Determinación del vector de descriptores:

a. Para el clasificador:

Se describe el procedimiento para determinar los descriptores válidos de cada canción.

Se selecciona 10 temas por género musical.

Dado el vector $x = [x_1, x_2 \ldots\ldots x_{1168}]$ con *n* elementos, donde $x_i \in \mathbb{R}$ y n = 1668.

Se tiene la matriz:

$$X_{genero\,(1..50)} = \begin{pmatrix} [X_{rock\,(1..10)}[x_1..x_{1168}]] \\ [X_{bossanova\,(1..10)}[x_1..x_{1168}]] \\ [X_{clasica\,(1..10)}[x_1..x_{1168}]] \\ [X_{jazz\,(1..10)}[x_1..x_{1168}]] \\ [X_{pop\,(1..10)}[x_1..x_{1168}]] \end{pmatrix},$$

donde, x_{rock} , $x_{bossanova}$, $x_{clasica}$, x_{jazz} , x_{pop} son vectores de descriptores.

Se obtiene la media ($\bar{X}$) y varianza (σ_X^2) de X para cada género:

- Rock: $\bar{X}_{rock}$, $\sigma^2_{X\,rock}$
- Jazz: $\bar{X}_{jazz}$, $\sigma^2_{X\,jazz}$
- Pop: $\bar{X}_{pop}$, $\sigma^2_{X\,pop}$
- Clásica: $\bar{X}_{clasica}$, $\sigma^2_{X\,clasica}$
- Bossanova: $\bar{X}_{bossanova}$, $\sigma^2_{X\,bossanova}$

Luego, se descarta las varianzas mayores y menores de las medias de la muestra, para ello se obtiene:

$$D = \bar{X}_{\sigma_X^2} \pm \sigma_{\sigma_X^2}$$

Los números mayores y menores a D se eliminan, estos valores deben coincidir para los 5 géneros musicales en filas y columnas; así, se elimina (asigna el valor cero) las columnas que presentan valores comunes entre ellos y se obtiene el vector representativo con descriptores no comunes por cada género.

La figura 3.1 muestra un ejemplo de algunos descriptores del género pop. Los descriptores no comunes en punto azul y los descriptores comunes en blanco.

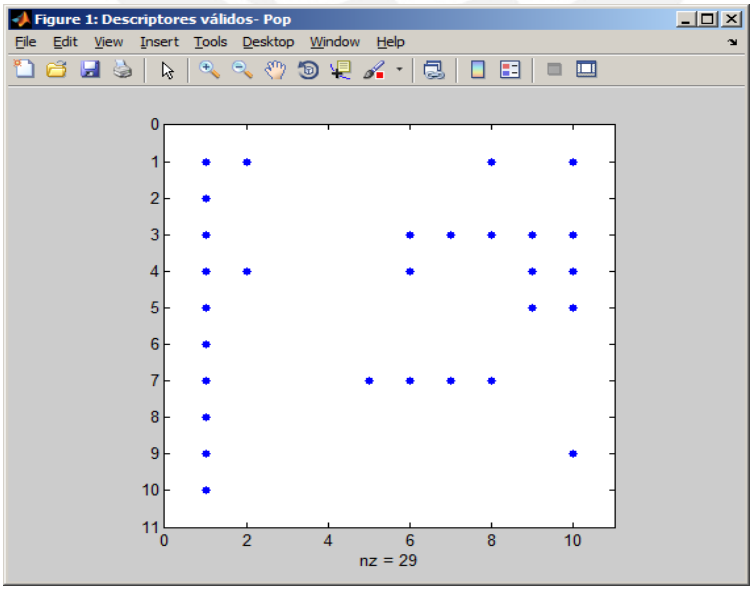

Figura 3.1: Descriptores comunes y no comunes para género *pop*.

b. Para el recuperador musical:

Se extrae en un vector los descriptores Patrones de Ritmo (RP), Estadística de Espectro (SSD) e Histograma de Ritmo (RH) de cada canción.

Se obtiene:

$$X_{genero\,(1..50)} = \begin{pmatrix} \left[X_{rock}(1..10) \left[X1..X1168 \right] \right] \\ \left[X_{bossanova}(1..10) \left[X1..X1168 \right] \right] \\ \left[X_{clasica}(1..10) \left[X1..X1168 \right] \right] \\ \left[X_{jazz}(1..10) \left[X1..X1168 \right] \right] \\ \left[X_{pop}(1..10) \left[X1..X1168 \right] \right] \end{pmatrix},$$

donde, X_{rock} , $X_{bossanova}$, $X_{clasica}$, X_{jazz} , X_{pop} son vectores de descriptores que se ingresan a la red neuronal para el entrenamiento.

3.2 Segunda Etapa: Entrenamiento

Se describe el algoritmo de entrenamiento.

Las redes de Retropropagación tienen un método de entrenamiento supervisado. Las entradas de la red se emparejan con un patrón de salida deseado.

Para cada iteración los pesos son ajustados de manera que disminuya el error entre la salida deseada y la respuesta de la red.

De dos fases:

Fase de Entrenamiento

1. Se ingresa a la red patrones, para este estudio se usa los descriptores obtenidos en la sección anterior.
2. Los pesos son aleatorios.
3. A cada entrada le corresponde una salida de la matriz de identidad.
4. Calcula el error con la salida hasta minimizar.
5. Se guardan los datos: pesos.

Los parámetros considerados para el entrenamiento son:

Ratio de aprendizaje, bias, número de neuronas, número de capas ocultas y número de iteraciones.

Fase de Validación

1. Se ingresa el vector descriptor consulta.
2. Se reutiliza los pesos de la fase de entrenamiento en la capa oculta.
3. Se selecciona el mayor del vector de salida que corresponde al vector similar buscado.
4. Se valida y devuelve el vector consultado o similar.

3.3 Métricas de evaluación

Se usa la medida llamada de Exactitud (*accuracy)*, viene a ser el grado de concordancia entre el valor medido y el verdadero valor medido [19][20].

Responde a la ecuación:

$$Exactitud = \frac{TP + TN}{FP + FN + TP + TN} \tag{3.1}$$

Donde:

- Exactitud: Es la proporción de predicciones positivos correctos.
- TP (Verdadero positivo): Es el número de aciertos correctos.
- FP (Falso positivo): Es el número de desaciertos que fueron clasificados como positivos.
- TN (Verdadero negativo): Es el número de desaciertos que fueron clasificados correctamente.
- FN (Falso negativo): Es el número de aciertos que fueron incorrectamente clasificados.

Los datos se muestran en la Matriz de confusión como se visualiza en la tabla 3.1

Tabla 3.1 Matriz de confusión

Verdaderos positivos TP	Falsos negativos FN
Falsos Positivos FP	Verdaderos negativos TN

Debido a que el desempeño de esta medida no es la adecuada cuando el número de casos negativos es mucho más grande que el número de casos positivos (Kubat et al, 1998), se podría considerar además, *geometric mean (g-mean)*; es el promedio que indica el valor típico del grupo de datos. (Kubat et al., 1998); definido por la ecuación:

$$g-mean = \sqrt{TP * \left(\frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \right)} \qquad (3.2)$$

Capítulo 4: Descripción de los Resultados

En este capítulo se describe el hardware y software usado para realizar las pruebas; la implementación y los resultados obtenidos de la evaluación, usando dos redes neuronales de Retropropagación.

4.1 Condiciones de Prueba

Las pruebas fueron ejecutadas en una PC HP con AMD Turion II Dual Core 2.2Ghz, 4 GB de memoria. El sistema operativo es Windows 7 (64 bits).

El software que se utilizó para la implementación es Matlab 7.12.0 (R2011a) [21].

4.2 Implementación

4.2.1 Clasificador de audio

- De los descriptores:

  Se inicia con la evaluación de los descriptores de 10 canciones por los 5 géneros musicales; se discrimina los descriptores no válidos y se obtiene los 5 vectores representativos para cada género, que serán finalmente, los valores de entrada a la red neuronal.

- De la red:

  La red neuronal está formada por tres capas, incluyendo una capa oculta.

  Durante la fase de entrenamiento, los 5 vectores de entrada son multiplicados por los pesos asignados con valores random, la función de activación usada es Gauss.

Para la fase de validación, los pesos estables obtenidos en el entrenamiento son usados. El valor de entrada es el vector de descriptores de la canción que se quiere consultar y, la salida es el valor que indica el estilo musical al que pertenece la canción.

Los parámetros usados en el entrenamiento son:

Bias: 1
Ratio de aprendizaje: 0.00001
Número de neuronas: 9000
Iteraciones: 999
Valores de entrada: 5 vectores de 1668 valores.

4.2.2 Reconocedor de audio

- De los descriptores:

Se inicia con la selección de 50 canciones que pertenecen a los 5 géneros anteriormente mencionados; se obtiene los descriptores por cada una, luego son ingresados a la red vectores de 1668 valores.

Se considera todos los descriptores como válidos pues, no son descriptores representativos como en el caso del clasificador.

- De la red:

La red neuronal está formada por tres capas con una capa oculta. Los 50 vectores obtenidos son las entradas de la red neuronal y, le corresponden 50 salidas; los pesos son asignados con números aleatorios y la función de activación usada es Gauss.

De la misma manera, los pesos estables obtenidos en el entrenamiento son usados en la fase de validación; aquí el valor de entrada es la canción que se quiere consultar y, el valor de salida es la canción más similar.

Los valores de los parámetros usados en el entrenamiento, son los mismos usados en la red de clasificación.

4.3 Resultados computacionales

4.3.1 Clasificador de audio

- **Fase de Entrenamiento**

Las figuras 4.1 y 4.2, muestran la estabilidad de la red en la fase de entrenamiento para 5 entradas y salidas respectivamente; el error tiende a cero en ambos casos, esta medida refleja el modo en que la red está logrando respuestas correctas a medida que la red aprende.

El error total es 8 e-014, siendo la tolerancia de aceptación del error menor a 0.5, es aceptable.

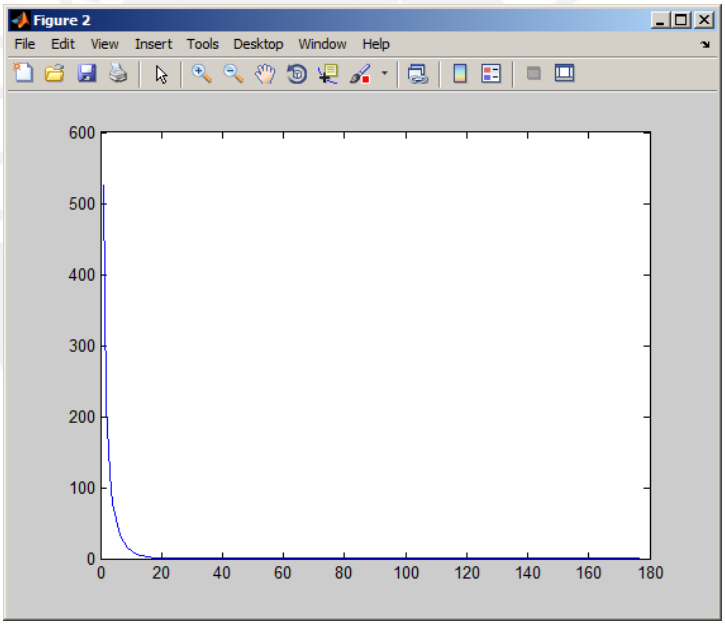

Figura 4.1: Convergencia del error

Las salidas y entradas deseadas son coincidentes en la fase de entrenamiento lo cual indica que los pesos son estables para ser evaluados en la fase de validación

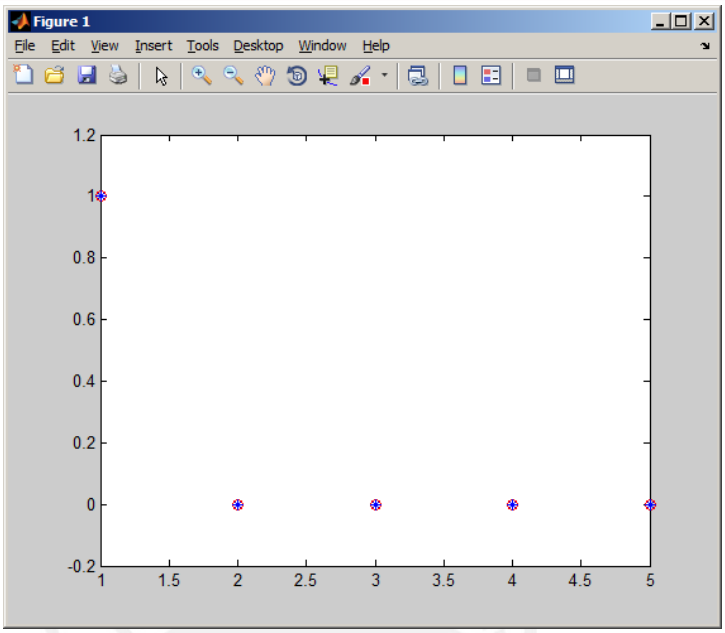

Figura 4.2: Salida obtenida y deseada

- **Fase de Validación**

La tabla 4.1, muestra los resultados obtenidos en la fase de validación, se presentan los valores en la Matriz de Confusión.

Se evalúa cada tema y la red retorna como resultado el género musical de la canción consultada.

Las filas representan los géneros actuales y las columnas los géneros predichos. La diagonal representa el número de aciertos de 10 canciones por cada género. Se tiene como máximo 6 aciertos para Jazz que representa el 60% de precisión como máximo y 0 aciertos para Pop que representa el 0% de precisión como mínimo, con un tiempo de respuesta en promedio de 0.6609 microsegundos.

Tabla 4.1: Matriz de Confusión para clasificador

Previsto

Actual	Jazz	Bossanova	Clásica	Pop	Rock
Jazz	6	1	0	2	1
Bossanova	2	2	3	2	1
Clásica	2	2	3	2	1
Pop	3	2	2	0	3
Rock	3	3	2	1	1

A la matriz de confusión le corresponden las tablas 4.2, 4.3, 4.4, 4.5, 4.6 de confusión de verdadero y falso positivos y, falso y verdadero negativos.

Tabla 4.2 Matriz de confusión para Jazz

TP	FN
6	4
FP	TN
10	30

Tabla 4.3 Matriz de confusión para Bossa-nova

TP	FN
2	8
FP	TN
8	32

Tabla 4.4 Matriz de confusión para Clásica

TP 3	FN 7
FP 7	TN 33

Tabla 4.5 Matriz de confusión para Pop

TP 0	FN 10
FP 7	TN 33

Tabla 4.6 Matriz de confusión para Rock

TP 1	FN 9
FP 6	TN 34

Los resultados se interpretan como sigue: Para la Tabla 4.2 se tiene 6 canciones correctamente clasificadas como jazz, 4 canciones jazz marcados como otros géneros, 10 canciones que fueron reconocidas incorrectamente como jazz y 30 que fueron clasificados correctamente como no jazz. En la tabla 4.3 se tiene 2 canciones correctamente clasificadas como bossa-nova, 8 canciones bossa-nova clasificados incorrectamente como otros géneros, 8 canciones que fueron reconocidas incorrectamente como bossa-nova y 32 que fueron clasificados correctamente como no bossa-nova. Para la tabla 4.4 se tiene 3 canciones correctamente clasificadas como clásica, 7 canciones clásica clasificados incorrectamente como otros géneros, 7 canciones que fueron reconocidas incorrectamente como clásica y 33 que fueron clasificados correctamente como no clásica. La tabla 4.5 tiene 0 canciones correctamente clasificadas como pop, 10 canciones pop clasificados incorrectamente como otros géneros, 7 canciones que fueron reconocidas incorrectamente como pop y 33 que fueron clasificados correctamente como no pop. Para la tabla 4.6 se tiene 1 canción correctamente clasificada como rock, 9 canciones rock clasificados incorrectamente como otros géneros, 6

canciones que fueron reconocidas incorrectamente como rock y 34 que fueron clasificados correctamente como no rock.

Finalmente la tabla 4.7 muestra el rendimiento del sistema calculado de la ecuación de *Exactitud y G-mean*, mencionado en el capítulo 3.

El éxito de detección ha sido favorable para los géneros: jazz, bossa-nova y clásica y desfavorable para rock y pop.

Tabla 4.7 Exactitud y G-mean

	Jazz	Bossa-nova	Clásica	Pop	Rock
Exactitud	0.64	0.68	0.74	0.66	0.69
g-mean	1.96	1.17	1.49	0	0.83

4.3.2 Reconocedor de audio

- **Fase de Entrenamiento**

Las figuras 4.3 y 4.4 muestran la estabilidad de la red durante el entrenamiento para 50 entradas y salidas respectivamente.

El error es 0.05, la tolerancia de aceptación del error es 0.5 por lo que es aceptable el resultado obtenido.

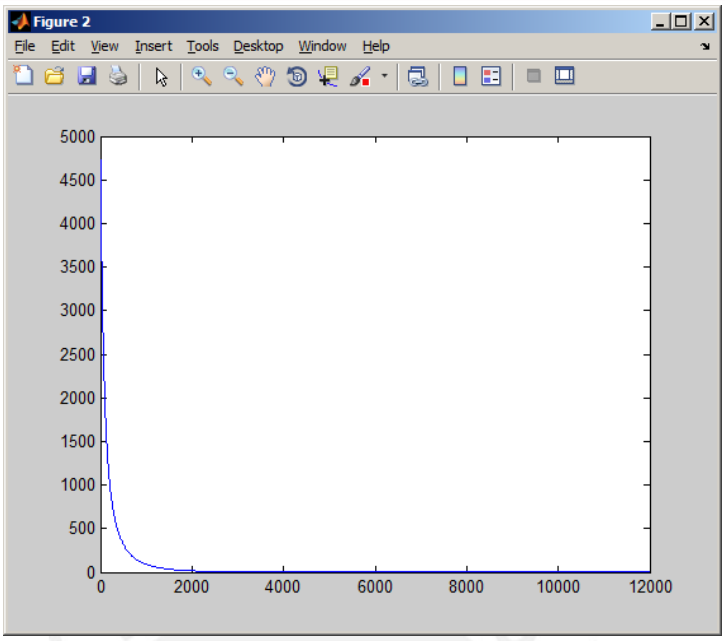

Figura 4.3: Comportamiento del Error Total

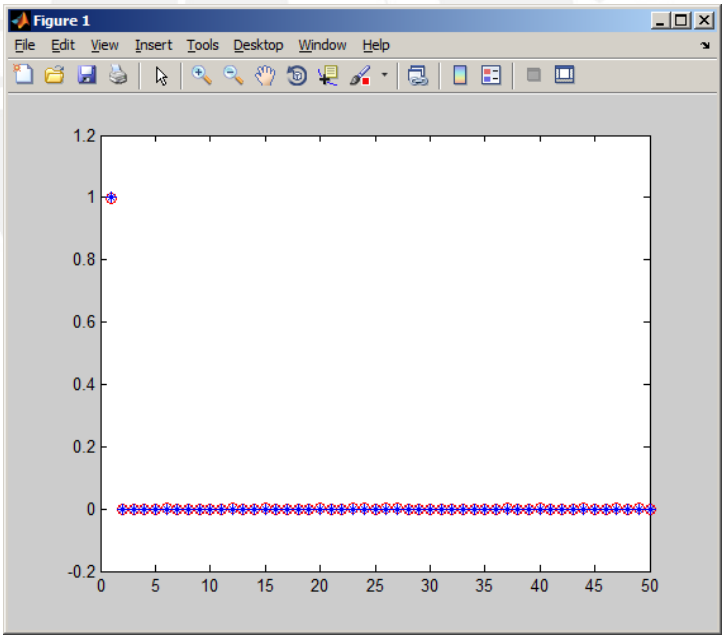

Figura 4.4: Salidas obtenidas y deseadas

- **Fase de Validación**

Para evaluar el rendimiento en este caso se indica la tasa de acierto. La tabla 4.8 muestra el porcentaje de acierto para 50 canciones consultadas las mismas que han sido ingresadas en el entrenamiento; la red devuelve la misma canción o similar, obteniéndose como máximo 100% de exactitud de recuperación y como mínimo 96% con un tiempo de 0.6407 microsegundos en promedio.

Tabla 4.8 Tasa de acierto para reconocedor musical.

Género musical	Jazz	Bossa-nova	Clásica	Rock	Pop
	100%	100%	96%	100%	100%
Muestra	50	50	50	50	50
# encontrados	50	50	50	50	50
Recuperados	50	50	48	50	50

Capítulo 5: Conclusión

En este capítulo se describe algunas conclusiones del trabajo y una breve justificación, ventajas y desventajas observadas. Además mencionamos trabajos a futuro que propone la investigación realizada.

En esta tesis se ha propuesto un sistema que reconoce y clasifica audio por similitud. Se usaron dos redes neuronales de Retropropagación uno para el reconocedor y otro para el clasificador, con datos de entrada que consideran características relevantes de audio tales como: pitch, ritmo, melodía; obtenido de la extracción de piezas musicales. La red de Retropropagación permite obtener un resultado similar o el valor ingresado. Se realizó un pre-entrenamiento que filtra información irrelevante que se ingresa a la red el cual mejora la medida de exactitud del resultado, la velocidad de convergencia y disminuye el tiempo de respuesta por consulta.

La principal desventaja de la red de Retropropagación es que, en algunos casos no converge a un estado estable y cae en mínimos locales; se ha logrado superar esta desventaja al reducir los datos de ingreso, usar ratios de aprendizaje que van de 0.00001 a 0.01, incrementar el número de neuronas; por otro lado nos encontramos con un grado de dificultad al tratar de encontrar los parámetros correctos para un óptimo entrenamiento cuando variamos los valores de entrada y el número de neuronas.

De los resultados obtenidos para el clasificador por similitud la métrica Exactitud indica que el porcentaje de acierto fluctúa entre 40% - 60% en la recuperación y clasificación, lo cual es aceptable con tendencia a mejora. Para el recuperador musical por similitud se tiene en promedio 100% de exactitud solo en el caso de que la consulta haya sido la misma usada en el entrenamiento. La desventaja para este caso radica en que se debe ingresar la colección de canciones a la red y por cada nueva colección realizar un nuevo entrenamiento esto, demandaría mayor tiempo. Para lograr obtener una red estable; se sugiere en este caso trabajar con redes paralelas.

5.1 Trabajo a futuro

Para el trabajo a futuro sería proponer mejorar algunos aspectos de la arquitectura de la red de tal manera que la información ingresada no tenga que ser procesada cada vez que se actualice la base de datos. Se propone usar redes neuronales de Retropropagación en paralelo para disminuir los cálculos, obtener menor redundancia y disminuir el tiempo de procesamiento. Observar el desempeño de esa red y comparar con los resultados obtenidos en la primera evaluación.

Implementar el problema en una red neuronal Asociativa Hopfield, que tiene como ventaja indicar a la red inicialmente el comportamiento adecuado ya que estas redes no son entrenadas sino construidas.

Bibliografía

[1] J.J. Aucouturier and F. Pachet, "Representing musical genre: A state of the art," *Journal of New Music Research*, 32(1), 2003.

[2] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, 10(5), 2002.

[3] Perfecto Herrera, Xavier Serra, "Audio Descriptors and Descriptor Schemes in the Context of MPEG-7,". *Proceeding of the ICM99.*

[4] Asif Ghias, Jonathan Logan, David Chamberlin, Brian C. Smith, "Query by Humming Musical Information Retrieval in an Audio Database," 2005.

[5] T. Lydy, R. Mayer, A. Rauber, P.J. Ponce de León, A. Pertusa, J. M. Iñesta, "A Cartesian Ensemble of Feature Subspace Classifiers for Music Categorization," *ISMIR* 2010.

[6] Rainer Typke, Frans Wiering, Remco C. Veltkamp, "A survey of Music Information Retrieval," 2005.

[7] Chang, W. y Lawler, E. (1994), "Sublinear approximate string matching and biological applications," *Algorithmica*,12, 4/5, 327-344, 1994.

[8] Steffen Pauws, "CubyHum: A fully Operational Query by Humming System," *ISMIR* 2002.

[9] François Pachet, Jean-Julien Aucouturier, Amaury La Burthe, Aymeric Zils y Anthony Beurive. "The Cuidado Music Browser an end-to-end Electronic Music Distribution System," International Workshop on Content-Based Multimedia. Indexing 2003.

[10] Rainer Typke, Remco C. Veltkamp, Frans Wiering. "Searching Notated Polyphonic digital library toolkit," *ISMIR Proceeding*, p42-43, 2004.

[11] Rainer Typke, "Music Retrieval based on Melodic Similarity," 2007.

[12] Ricardo Baeza-Yates and G.H. Gonnet, "Fast string matching with Mistmatches," Information *and Computation*, 1992.

[13] Anil K. Jain, Jianchang Mao, "Artificial Neural Network: a Tutorial," 1996 *IEEE*.

[14] Simon Haykin, "Neural Network a Comprehensive Foundation," Second Edition. p72, 2005.

[15] A. Rauber, E. Pampalk y D. Merkl. "The SOM-enhanced JukeBox: Organization and Visualization of Music Collections based on Perceptual Models," *Journal of New Music Research* (JNMR), 32(2):193-210, 2003.

[16] T. Lidy y A. Rauber. "Genre-oriented Organization of Music Collections using the SOMeJB System: An Analysis of Rhythm Patterns and other Features," *Proceedings of the DELOS Workshop on Multimedia Contents in Digital Libraries*, 2003.

[17] Sander Dieleman, Philémon Brakel and Benjamin Schrauwen, "Audio-Based Music Classification with a pretrained Convolutional Network*," 12th International Society for Music Information Retrieval Conference (ISMIR),* 2011.

[18] T. Lydy, R. Mayer, A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification," *Proceedings ISMIR,London, UK,* 2005.

[19] International vocabulary of metrology – Basic and general concepts and associated terms (VIM) *Edition*. p21, 2012.

[20] Annsi Klapuri & Colby Leider, *Proceeding of the 12th International Society for Music Information Retrieval Conference ISMIR 2011.*

[21] MatLab, "nntool" http://www.mathworks.com/help/nntool.html, 2011.