Clasificación de Instrumentos Musicales en Audios utilizando Coeficientes Cepstrales y Redes Neuronales Artificiales

Classification of Musical Instruments in Audios using Cepstral Coefficients and Artificial Neural Networks

Ing. Alan Salomón Vázquez Robledo[1] Dra. Rocío Alfonsina Lizárraga Morales[2] Dr. Misael López Ramírez[3]
[1] Departamento de Estudios Multidisciplinarios, Campus Irapuato-Salamanca (sede Yuriria), Universidad de Guanajuato.
[2] Departamento De Arte y Empresa División De Ingenierías Del Campus Irapuato-Salamanca, Universidad de Guanajuato.
[3] Departamento de Estudios Multidisciplinarios, Campus Irapuato-Salamanca (sede Yuriria), Universidad de Guanajuato.
as.vazquezrobledo@ugto.mx[1] , ra.lizarragamorales@ugto.mx[2], lopez.misael@ugto.mx[3]

Resumen

La música juega un papel crucial en el desarrollo de la identidad cultural de cada persona, ayuda a expresar emociones, narrando historias y creando conexiones entre individuos y comunidades. El desarrollo digital de la música en la actualidad nos presenta un nuevo campo con nueva información, características distintivas y con ello nuevas problemáticas a resolver.

Los Sistemas de Recuperación de Información Musical (MIR, por sus siglas en inglés, Music Information Retrieval) son un campo emergente basado en sistemas de software diseñados para extraer, analizar y recuperar información de archivos de audio musical. Estos sistemas combinan varias técnicas del procesamiento de señales, aprendizaje automático, inteligencia artificial y análisis de datos para trabajar con información musical. han surgido como solución innovadora para diferentes problemáticas sobre la organización de características e información de la música digital. Estos sistemas, que emplean técnicas de procesamiento de señales y aprendizaje automático, permiten analizar automáticamente la estructura de las señales de audio y extraer información relevante, como la identificación de instrumentos musicales. En este artículo, se propone un sistema de clasificación de instrumentos musicales basado en redes neuronales artificiales, utilizando archivos de audios de una base de datos con el nombre AAM. A través del análisis espectral de las señales de audio y la extracción de características como los Coeficientes Cepstrales de Frecuencia Mel (MFCC por sus siglas en inglés, Mel Frequency Cepstral Coefficients) y el Perceptrón Multicapa (MLP por sus siglas en inglés, Multilayer Perceptron), se logró desarrollar un modelo capaz de identificar con alta precisión (98.66%) tres instrumentos musicales: guitarra, bajo y batería.

Palabras clave: Identificación de instrumentos musicales, MFCCs, Machine Learning, Perceptrón Multicapa, clasificación de audio.

Introducción

El análisis de la recuperación de información musical es sumamente importante, para comprender la estructura y el funcionamiento de cada esencia de diferentes notas e instrumentos musicales en un solo archivo digital de audio. En este momento, los profesionales del área de la música tienen un gran desafío debido al crecimiento de datos disponibles en este campo y a causa de la complejidad de la organización de los mismos. Es necesario el desarrollo de una herramienta innovadora de análisis automático de datos. La tarea de los sistemas de Recuperación de Información Musical (MIR) se define como el área encargada del análisis de la estructura y extracción de la información necesaria de una señal de audio como lo plantea Alexandre M. Lucena (2020). Algunas de las tareas principales de los sistemas MIR se centran en funciones como identificación de artistas, clasificación de géneros, clasificación de estados de ánimo, notación musical y el reconocimiento de instrumentos musicales. Esto es crucial por varias razones, incluida la recuperación de información musical, la separación de fuentes de sonido y la transcripción automática de música.

Una de las problemáticas considerables que podemos encontrar dentro de la identificación de instrumentos musicales se encuentra relacionada con la labor intensiva y poco eficiente de los productores musicales al tener que identificar manualmente los instrumentos musicales en las grabaciones de audio. Esta tarea en especial es costosa y consume mucho tiempo, tiende a ser muy propensa a errores, especialmente en

volúmenes grandes de conjuntos de datos. Debido al gran aumento en la cantidad de información que se maneja, se ha convertido en una tarea con un grado alto de dificultad en la actualidad.

Una solución a la problemática antes planteada podemos encontrarla haciendo uso de la extracción de Coeficientes Cepstrales de Frecuencia Mel (MFCC), ya que se ha convertido en una herramienta muy útil para el procesamiento de señales de audio debido a su capacidad de capturar las características más relevantes del espectro auditivo. Este articulo propone una aproximación que combina dicha técnica con un modelo de Machine Learning basado en un algoritmo de Perceptrón Multicapa con salida Softmax para clasificar automáticamente tres instrumentos musicales: guitarra, bajo y batería.

Estado del arte

En el área de la clasificación de instrumentos musicales podemos considerar las técnicas del campo de aprendizaje de máquina (ML por sus siglas en inglés, Machine Learning), una de estas técnicas es la Máquina de Soporte Vectorial (SVM, por sus siglas en inglés, Support Vector Machine) que es un modelo de aprendizaje supervisado que se utiliza principalmente para tareas de clasificación y regresión. Algunos trabajos que han implementado dichas técnicas proporcionan sus resultados, por su parte S. Prabavathy (2020) nos propone la clasificación automática de instrumentos musicales como son el trombón, tuba, trompeta y piano utilizando SVM y la técnica de k vecinos más cercanos (KNN, por sus siglas en inglés, K-Nearest Neighbor), como parte de sus resultados nos muestra una precisión con SVM del 99.37 % utilizando dichas técnicas. Sanraga Kingkor (2021) presenta un modelo de Red Neuronal Artificial (RNA por sus siglas en inglés, Artificial Neural Network) la cual fue entrenada para realizar clasificación en veinte clases diferentes de instrumentos musicales de la Orquesta Filarmónica de Londres en conjunto de la extracción de Coeficientes Cepstrales de Frecuencia Mel (MFCC), en su trabajo se logró una precisión del 97% en el conjunto de datos completo que contiene las 20 clases de diferentes instrumentos musicales. Además, podemos mencionar algunas otras técnicas incluyendo una Red neuronal de impulsos (SNN por sus siglas en inglés, Spiking Neural Network), que nos permite modelar más estrechamente la forma en que los humanos distinguen los instrumentos musicales al identificar diferencias en las características temporales de los instrumentos a medida que se reciben neuronalmente. La principal ventaja de utilizar Machine Learning en Identificación de Instrumentos Musicales en Señales de Audio es la capacidad para identificar patrones complicados que pueden ser difíciles de detectar utilizando otras técnicas, una desventaja es la alta dependencia a los datos de entrenamiento, los cuales deben ser extensos y con características viables para la clasificación.

Recientemente, algunas de las técnicas con más auge son las basadas en el campo de Aprendizaje Profundo (DL por sus siglas en inglés, Deep Learning). Dentro de las técnicas más relevante de Aprendizaje Profundo es la aplicación de Redes Neuronales Convolucionales (CNN por sus siglas en inglés, Convolutional Neural Networks), estas se refieren a un tipo especializado de Red Neuronal Artificial (RNA por sus siglas en inglés, Artificial Neural Network) diseñada específicamente para procesar datos como imágenes o señales de audio, los espectrogramas se envían a las Redes Neuronales Convolucionales (CNN por sus siglas en inglés, Convolutional Neural Networks) para aprender patrones de cómo se visualizan los diferentes instrumentos musicales. Maciej Blaszke (2022) presenta la construcción de un algoritmo para la automatización e identificación de instrumentos presentes en un extracto de audio utilizando conjuntos de Redes Neuronales Convolucionales (CNN por sus siglas en inglés, Convolutional Neural Networks) individuales por instrumento, los cuales son bajo, batería, guitarra y piano. En dicho trabajo se logró una precisión de aproximadamente del 100%, pero cuando se observan los resultados de reconocimiento de instrumentos musicales en conjunto, los valores métricos son más bajos. Una arquitectura similar es la que presenta Chinmay Relkar (2019) que costa de una Red Neuronal Convolucional (CNN) de 4 capas inspirada en la arquitectura de AlexNet, el cual lleva el nombre de Red de grupos de geometría visual (VGGNet por sus siglas en ingles Visual Geometry Group Network). Chinmay Relkar (2019) también menciona la técnica de Red Neuronal Convolucional basada en regiones (RCNN, por sus siglas en inglés), esta técnica fue una de las primeras arquitecturas en abordar el problema de detección de objetos en imágenes utilizando Redes Neuronales Convolucionales. Una técnica variante es la Red Neuronal Recurrente Convolucional (CRNN, por sus siglas en inglés). El uso del Aprendizaje Profundo en la Identificación de Instrumentos Musicales, tiene la ventaja de desempeñar a un nivel más alto la precisión cuando se trata de la clasificación de instrumentos musicales, esto sucede cuando el modelo está entrenado en un conjunto adecuado de datos y situaciones. Con respecto a las desventajas, se requieren grandes cantidades de datos etiquetados en el entrenamiento efectivo de modelos de Aprendizaje Profundo para la clasificación de instrumentos musicales, lo que consume mucho tiempo en recopilar y etiquetar.

Metodología

La metodología que se implemento es de carácter cuantitativo, ya que se usaron datos numéricos y métricas para evaluar los resultados obtenidos. En la figura 1 se muestra la metodología propuesta, en donde podemos observar las dos fases principales las cuales son entrenamiento y evaluación, ambas fases constan de 3 partes importantes: Archivos de audio, Extracción de características MFCC y MLP. En las siguientes secciones se describirá de una forma más completa cada una de las siguientes partes.

Figura 1. Metodología propuesta.
Fuente: Autoría Propia.

Archivos de audio

El análisis de las señales de audio comienza con el uso de la base de datos conocida como Conjunto de datos de pistas múltiples de audio artificial (AAM por sus siglas en ingles Artificial Audio Multitracks Dataset) es introducida por Ostermann et al. (2023). Este conjunto de datos contiene 3000 pistas de audio de música artificial, basadas en muestras de instrumentos reales y generadas mediante composición algorítmica con respecto a la teoría musical. Proporciona mezclas completas de canciones, así como pistas de un solo instrumento. Para este trabajo se utilizaron 100 audios para cada instrumento, guitarra, batería y bajo.

Extracción de características Coeficientes Cepstrales de Frecuencia Mel (MFCC)

La extracción de características consta de extraer al menos 13 de los Coeficientes Cepstrales de Frecuencia Mel (MFCC) de cada uno de los audios digitales de la base de datos antes señalada, los cuales son necesarios para el análisis del espectro de audio, ya que modelan la forma en que los humanos perciben el sonido, proporcionando una representación compacta y robusta de la señal de cada audio que analizaremos. Los Coeficientes Cepstrales de Frecuencia Mel (MFCC) capturan las características espectrales más relevantes de cada audio y son muy útiles para la clasificación de señales de audio, ya que representan tanto la envolvente del espectro como sus cambios a lo largo del tiempo.

Los Coeficientes Cepstrales de Frecuencia Mel (MFCC) describen la forma espectral cada uno de los audios digitales que analizaremos de la base de datos. Como primera parte, las bandas de frecuencia se posicionan logarítmicamente. Lo que podemos llamar escala Mel. Se utiliza un método que tiene capacidad de compactación de energía denominado con el nombre de Transformada de Coseno Discreta (DCT por sus siglas en inglés, Discrete Cosine Transform), que solo considera los números reales. De forma predeterminada, para así tomar los primeros 13 MFCC componentes de cada audio analizado (Majeed et al., 2015).

En el siguiente apartado se muestran los pasos necesarios para poder obtener los MFCC de cada uno de nuestros archivos de audio.

1. Framing (segmentando): La señal de cada archivo de audio que tomamos se divide en bloques cortos llamados frames (segmentos o cuadros). La longitud típica del frame (segmento o cuadro) es de 20-30 ms (milisegundos) y el desplazamiento es de 10 ms (milisegundos).

2. Windowing (visualización de información en una ventana o recuadro): Cada frame (cuadro) se multiplica por una ventana que utiliza la función matemática para suavizar los bordes de un segmento llamada Hamming, para reducir discontinuidades: donde N es la longitud del frame (cuadro), la formula se presenta a continuación.

$$h(n) = x(n)w(n) \quad (1)$$

$$w\,(n) = 0.54 - 0.46\,cos\left(\frac{2\pi}{N-1}\right) \quad (2)$$

3. Estimación espectral: para este paso se aplica la Transformada Discreta de Fourier (DFT por sus siglas en inglés Discrete Fourier Transform) descrita por Cooley y Tukey (1965), a cada frame (cuadro) para obtener los coeficientes espectrales, la formula se presenta a continuación:

$$x(k) = \sum_{n-0}^{N-9} y(n)e^{-\frac{j2\pi}{N}} \qquad 0 \leq n, k \geq N-1 \qquad (3)$$

4. Aplicación de la escala Mel: Se transforma la escala de frecuencia de cada audio a la escala Mel, esta escala se centra en como los humanos perciben las frecuencias del sonido, que es más cercana a la percepción humana. Esto se hace utilizando un banco de filtros Mel, donde cada filtro se aplica a la magnitud del espectro, la formula se presenta a continuación:

$$f_M = 2525 \times log\left(1 + \frac{J}{7}\right) \qquad (4)$$

5. Cálculo de los Coeficientes Cepstrales: Se toma el logaritmo de la energía de la magnitud de la respuesta del filtro Mel que se presenta a continuación:

$$E_j^x = \Sigma_{k-1}^n |x(k)|^2 * \psi_t(k) \qquad (5)$$

6. Transformada Discreta del Coseno (DCT por sus siglas en inglés, Discrete Cosine Transform): Finalmente, se aplica la Transformada Discreta del Coseno (DCT por sus siglas en inglés, Discrete Cosine Transform) a los logaritmos de la energía para obtener los Coeficientes Cepstrales de cada audio, la formula se presenta a continuación:

$$C_t^x = \sum_{t=1}^{M} lo\, g(E_t^x) C\left[l \cdot \frac{(2\pi - 1)\pi}{2M}\right] \qquad (6)$$

Los Coeficientes Cepstrales de Frecuencia Mel fueron obtenidos utilizando estos pasos mediante la función "mfccs" disponible de Matlab.

Clasificador Perceptrón Multicapa (MLP)

Los Coeficientes Cepstrales de Frecuencia Mel (MFCC) de cada uno de los 100 audios de guitarra, bajo y batería se utilizan como entrada en un Perceptrón Multicapa (MLP por sus siglas en inglés, Multilayer Perceptron), que costa de un algoritmo de inteligencia artificial de clasificación perteneciente a la categoría de Red Neuronal Artificial (RNA por sus siglas en inglés, Artificial Neural Network), este algoritmo está diseñado para clasificar 100 audios de 3 instrumentos musicales: guitarra, bajo y batería. El 80% de los datos se usa para entrenar el modelo y el 20% restante para evaluarlo, con 100 audios por instrumento. La arquitectura de la red está formada de una capa oculta con 32 neuronas, respectivamente, estas neuronas se encuentran activadas por la función ReLU, lo que permite al modelo aprender representaciones intermedias y patrones complejos de los datos. La capa de salida utiliza una función Softmax para convertir las salidas en probabilidades, con tantas neuronas como clases a predecir (por ejemplo, guitarra, bajo y batería). El modelo se entrena durante 1000 épocas utilizando lotes de 32 muestras, lo que permite un ajuste continuo de los pesos después de cada lote, optimizando la eficiencia del proceso de entrenamiento.

Posteriormente, el modelo se evalúa con los datos de prueba para comprobar su capacidad de generalización, asegurando que las predicciones sean precisas en nuevos datos y no se limite a memorizar los datos de entrenamiento. El modelo será evaluado utilizando métricas estándar como la exactitud, precisión, recall y F1-score. La exactitud mide el porcentaje total de predicciones correctas, proporcionando una visión general del rendimiento del modelo. Sin embargo, dado que puede haber un desbalance en las clases (por ejemplo, más audios de un instrumento que de otro), la precisión y el recall son fundamentales: la precisión indica qué proporción de las predicciones positivas son correctas, mientras que el recall mide cuántos de los casos positivos reales fueron correctamente identificados. El F1-score, que es la media armónica entre precisión y recall, se utilizará para obtener un equilibrio entre ambos, especialmente en situaciones donde los datos de las clases no están equilibrados. Cada una de estas métricas serán evaluadas mediante el uso de la validación cruzada con 5 conjuntos de pruebas para mostrar el promedio de ellas, las fórmulas de estas métricas y de la validación cruzada se presentan a continuación:

$$Exactitud = \frac{Número\ de\ predicciones\ correctas}{Número\ total\ de\ predicciones}, \qquad (7)$$

$$Precisión = \frac{Verdaderos\ positivos}{Verdaderos\ positivos\ +\ Falsos\ positivos}, \qquad (8)$$

$$Recall = \frac{Verdaderos\ positivos}{Verdaderos\ positivos\ +\ Falsos\ negativos}, \qquad (9)$$

$$F1 = 2 \times \frac{Precision\ x\ Recall}{Precision\ +\ Recall}, \qquad (10)$$

$$Validación\ cruzada = \frac{1}{K}\sum_{i=1}^{K} Metrica_i, \qquad (11)$$

Resultados

A continuación, se muestran los resultados obtenidos al realizar las pruebas con el concentrado de 100 audios de cada instrumento musical: Guitarra, Bajo y Batería, los cuales fueron procesados para extraer sus 13 Coeficientes Cepstrales de Frecuencia Mel (MFCC por sus siglas en inglés, Mel Frequency Cepstral Coefficients) principales para realizar la posterior clasificación por medio del algoritmo de Perceptrón Multicapa (MLP por sus siglas en inglés, Multilayer Perceptron).

La exactitud del modelo clasificó correctamente el 98.66% de los datos de prueba de las tres clases. Por lo cual, de los 60 datos totales de los tres instrumentos, el modelo acertó en 59 de ellos. La precisión nos muestra que, de todas las predicciones positivas hechas por el modelo el 98.66% fueron correctas. El recall identificó correctamente el 98.81% de todos los datos que en realidad pertenecían a las clases correctas. El F1-score es del 98.66%. Esta métrica es una combinación de la precisión y el recall, proporcionando un equilibrio entre ambas métricas. El promedio macro y promedio ponderado nos muestra un resultado similar en las tres diferentes métricas del 98% que utilizamos, finalmente se implementó el uso de la validación cruzada en 5 pruebas promediando los resultados las métricas para dicho número de pruebas, podemos visualizar dichos resultados en la siguiente tabla que tiene el nombre de Tabla 1. Métricas de evaluación de resultados.

***Tabla 1**. Métricas de evaluación de resultados.*

Clase	Precisión	Recall	F1-Score	Exactitud (Porcentaje)
Guitarra	100.0000	100.0000	100.0000	100.0000%
Bajo	100.0000	96.4389	98.1168	98.1852%
Batería	96.0000	100.0000	97.8656	97.9552%%
Promedio	98.6667	98.8130	98.6608	98.6667%

Fuente: Autoría Propia.

Matriz de Confusión (Confusion Matrix):

Esta matriz muestra cuántos datos fueron clasificados de manera correcta o incorrecta en cada una de nuestras clases: Guitarra, Bajo y Batería. Estos datos pertenecen a una sola prueba, parte de un concentrado de 5 para realizar la validación cruzada.

Para la clase Guitarra: 20 datos fueron correctamente clasificados, ninguno fue mal clasificado.
Para la clase Bajo: 19 fueron correctamente clasificados, y 1 fue mal clasificado como Guitarra.
Para la clase Batería: Todos los 20 fueron correctamente clasificados.

Matriz de Confusión			
Datos Reales Guitarra	20	0	0
Datos Reales Bajo	0	19	1
Datos Reales Batería	0	1	19
	Datos de Predicción Guitarra	Datos de Predicción Bajo	Datos de Predicción Batería

Figura 2. Matriz de confusión.
Fuente: Autoría Propia.

Conclusiones

El modelo muestra un rendimiento de clasificación con una precisión global evaluada con validación cruzada del 98.66%, de 60 audios de instrumentos musicales se acertó en 59 de ellos, mostrándonos un solo error con el uso de los Coeficientes Cepstrales de Frecuencia Mel (MFCC por sus siglas en inglés, Mel Frequency Cepstral Coefficients) y el Perceptrón Multicapa (MLP por sus siglas en inglés, Multilayer Perceptron), en cuanto en las métricas de precisión, recall y F1-Score se muestra que se encuentran muy cercanas al 100%, para los tres tipos de instrumentos musicales como son la guitarra, el bajo y la batería, en cuanto a la matriz de confusión indica un solo error en la clasificación de bajos, y las métricas de precisión, recall y F1-score están equilibradas alrededor del 98%. Algunos aspectos a mejorar son la clasificación del bajo eléctrico, ya que podría llevar a un resultado mejor. Además, se puede experimentar con más datos o diferentes instrumentos o ajustar la arquitectura para ver si es posible mejorar estos resultados.

Referencias

Blaszke, M., & Kostek, B. (2022). Musical Instrument Identification Using Deep Learning Approach. *Sensors (Basel, Switzerland)*, 22(8), 3033. https://doi.org/10.3390/s22083033

Cooley, J. W., & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation, 19*(90), 297. https://doi.org/10.1090/S0025-5718-1965-0178586-1

Lucena, A., Pires, C., Nose-Filho, K., & Suyama, R. (2020, October 26). Musical Instruments Recognition Using Machine Learning. *Brazilian Technology Symposium*, Brasil.

Mahanta, S. K., Khilji, A. F. U. R., & Pakray, P. (2021). Deep Neural Network for Musical Instrument Recognition Using MFCCs. *Computación y Sistemas, 25*(2), 351–360. https://doi.org/10.13053/cys-25-2-3946

Majeed, S. A., Husain, H., Abdul Samad, S., & Idbeaa, T. F. (2015). MEL frequency cepstral coefficients (MFCC) feature extraction enhancement in the application of speech recognition: A comparison study. *Journal of Theoretical and Applied Information Technology, 79*(1), 2005–2015. https://doi.org/ISSN: 1992-8645

Ostermann, F., Vatolkin, I., & Ebeling, M. (2023). AAM: a dataset of Artificial Audio Multitracks for diverse music information retrieval tasks. *J Audio Speech Music*. https://doi.org/10.1186/s13636-023-00278-7

Prabavathy, S. V. R. (2020, May). Musical Instruments Classification using Pre-Trained Model. *International Research Journal of Engineering and Technology (IRJET)*, *7*(5), 585–589. https://www.irjet.net/

Relkar, V. T. Chinmay (2019, september). Musical Instrument Identification Using. International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET) Volumen 2, issue 9, pages: 1826-1829. https://www.ijmrset.com/