ISSN 1870-4069

Transferencia de aprendizaje de una arquitectura de aprendizaje profundo para la separación de frecuencias musicales

Esteban Uriel Ildefonso-Orozco[1], Alberto Jorge Rosales-Silva[1], Armando Adrián Miranda-González[2], Jena Marie Vianney Kinani[3], Dante Mújica Vargas[4], Ponciano Jorge Escamilla Ambrosio[5]

[1] Instituto Politécnico Nacional, Escuela Superior de Ingeniería Mecánica y Eléctrica Zacatenco, México

[2] Instituto Politécnico Nacional, Escuela Superior de Ingeniería Mecánica y Eléctrica Culhuacán, México

[3] Instituto Politécnico Nacional, Unidad Profesional Interdisciplinaria de Ingeniería Campus Hidalgo, México

[4] Tecnológico Nacional de México, Departamento de Ciencias de la Computación, México

[5] Instituto Politécnico Nacional, Centro de Investigación en Computación, México

{eildefonsoo1500, amirandag1100}@alumno.ipn.mx, {arosaless, jkinani}@ipn.mx, dantemv@cenidet.edu.mx, pescamilla@cic.ipn.mx

Resumen. Este artículo propone un enfoque de aprendizaje por transferencia para abordar la separación de frecuencias musicales en pistas de audio no disponibles públicamente. La arquitectura combina una red convolucional (U-NET) y una red recurrente (LSTM) para segmentar y organizar la información a lo largo de una línea de tiempo utilizando coeficientes cepstrales de frecuencia Mel (MFCC). Esta arquitectura permite la extracción de pistas de audio individuales correspondientes a diferentes instrumentos (bajo, batería, voz y melodía), comúnmente conocidas como STEMS. Además de abordar la escasez de recursos en la separación de fuentes y satisfacer la creciente demanda de habilidades de producción musical, también facilita el aprendizaje y la práctica musical, fomentando la creatividad y la exploración de nuevas ideas musicales. Se destaca que los beneficios de esta arquitectura se enfocan exclusivamente a fines educativos y de obtención de pistas de canciones inaccesibles. Esta arquitectura de aprendizaje profundo propuesta representa una alternativa automatizada para la obtención de STEMS.

Palabras clave: LSTM, MFCC, neural networks, U-Net.

Transfer Learning of a Deep Learning Architecture for Musical Frequency Separation

Abstract. This article proposes a transfer learning approach to address the separation of musical frequencies in publicly unavailable audio tracks. The architecture combines a convolutional network (U-NET) and a recurrent network (LSTM) to segment and organize information along a timeline using Mel frequency cepstral coefficients (MFCC). This architecture enables the extraction of individual audio tracks corresponding to different instruments (bass, drums, vocals, and melody), commonly known as STEMS. In addition to addressing the scarcity of resources in source separation and meeting the growing demand for musical production skills, it also facilitates music learning and practice, fostering creativity and the exploration of new musical ideas. It is emphasized that the benefits of this architecture are exclusively focused on educational purposes and obtaining tracks from inaccessible songs. This proposed deep learning architecture represents an automated alternative for obtaining STEMS.

Keywords: LSTM, MFCC, neural networks, U-Net.

1. Introducción

Este artículo presenta un enfoque innovador para la separación de frecuencias musicales en pistas de audio utilizando aprendizaje profundo. A diferencia de otros métodos que requieren acceso a pistas de audio de un estudio de grabación, este enfoque propone el uso de redes neuronales para separar automáticamente las pistas de bajo, batería, voz y melodía a partir de audios disponibles en línea.

Esto hace que sea más accesible para su uso en diferentes aplicaciones. En este trabajo se exploran diferentes criterios de evaluación para validar la efectividad del enfoque propuesto. Las pistas de audio digital se dividen en STEMS, que son componentes musicales independientes que incluyen el bajo, la batería, la voz y la melodía.

Cada uno de estos componentes está definido por su contenido de frecuencias específicas y se pueden manipular individualmente para crear mezclas personalizadas [13]. En este trabajo se propone una arquitectura que combina una red neuronal convolucional (U-NET) y una red neuronal recurrente (LSTM) para abordar el problema conocido como el 'Efecto de fiesta de cóctel' [2].

Este fenómeno se refiere a la habilidad de enfocar la atención auditiva en un estímulo particular mientras se filtra un conjunto más amplio de estímulos, de manera coloquial sería como una persona puede concentrarse en una sola conversación en medio de una sala ruidosa durante una fiesta.

La arquitectura propuesta busca solucionar este problema mediante la identificación y separación de cada instrumento en una mezcla de audio, lo que resulta en pistas de audio separadas para cada uno de ellos. El entrenamiento se lleva a cabo utilizando la base de datos denominada sigsep musdb18, que consta de un total de 150 canciones completas de diferentes estilos.

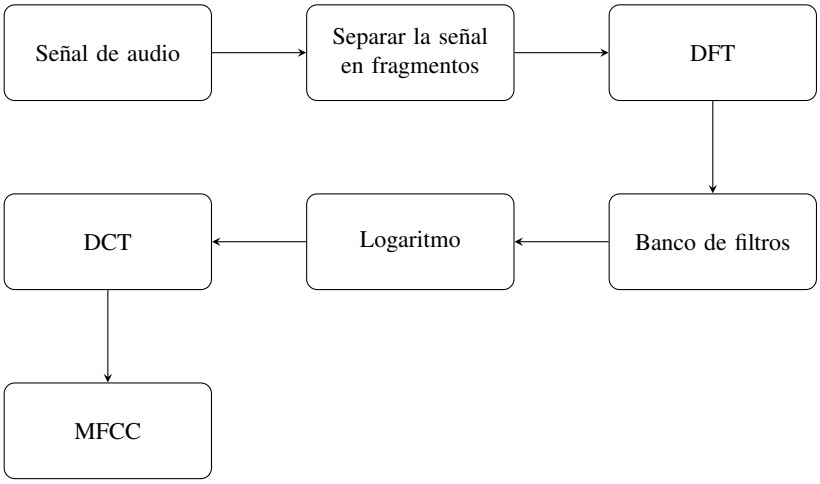

Fig. 1. Diagrama de bloques para obtener los MFCC.

Esta base de datos incluye tanto las mezclas estéreo como las fuentes originales, que son las pistas individuales conocidas como STEMS. Las grabaciones tienen una frecuencia de muestreo de 44.1 KHz y están divididas en conjuntos de entrenamiento y prueba.

Esta base de datos ha sido diseñada y evaluada como referencia para algoritmos de separación de fuentes, y es proporcionada por la Campaña de Evaluación de Separación de Señales de 2018 (SiSEC 2018) [14]. Se propone una arquitectura que combina una estructura U-Net y LSTM para separar pistas de audio.

La red U-Net es comúnmente utilizada para la segmentación semántica de imágenes, pero aquí se utiliza para separar pistas de audio. La red LSTM se utiliza para solucionar la línea del tiempo de las pistas de audio, lo que es crucial para la correcta separación de las pistas.

1.1. Red U-NET

Una Red U-Net es un tipo especial de red neuronal convolucional utilizada para la segmentación semántica de imágenes. Durante su proceso, aplica convoluciones para reducir dimensionalmente la información y, posteriormente, la recupera con el uso de deconvoluciones [12]. Además de las convoluciones y deconvoluciones, la red U-Net también contiene funciones de activación que se encargan de devolver una salida a partir de un valor de entrada.

Estas funciones pueden ser ReLU (Unidad Lineal Rectificada), sigmoide, tangente hiperbólica, entre otras [15]. La arquitectura de la Red U-Net permite que la red aprenda tanto características locales como globales de la imagen, lo que la hace especialmente útil en tareas de segmentación semántica.

La arquitectura propuesta utiliza la segmentación semántica de los MFCC para separar cada pista de los instrumentos.Aplicando la red U-Net logrando separar cada componente musical en pistas separadas.

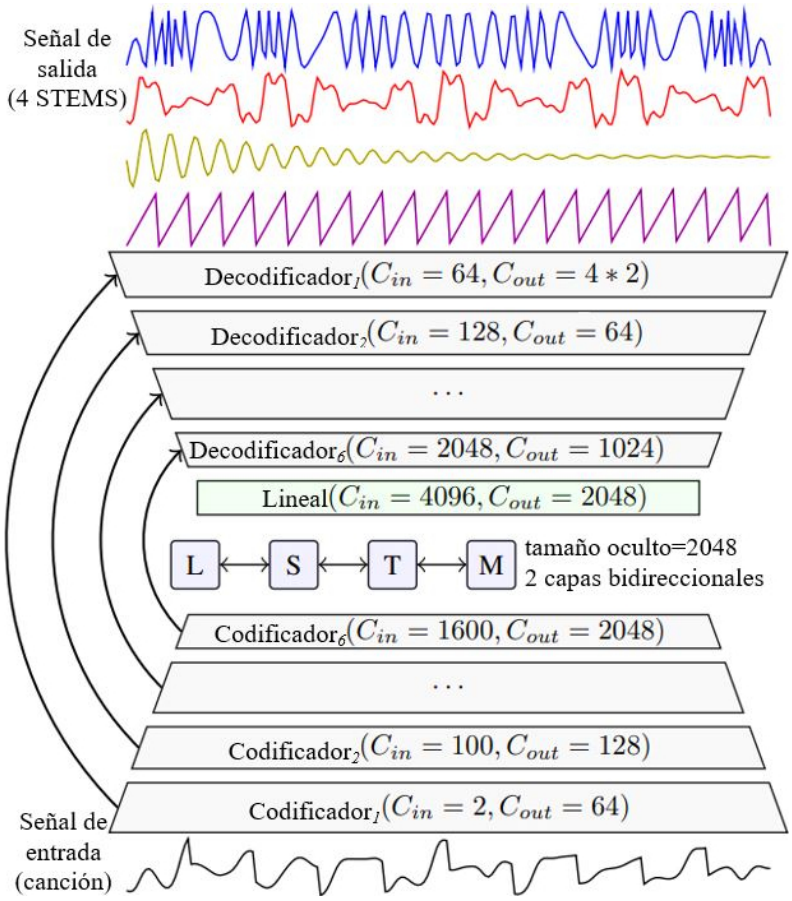

Fig. 2. Arquitectura de separación de frecuencias musicales.

1.2. Funciones de activación

La función de activación ReLU (Unidad Lineal Rectificada) se define como:

$$f(x) = \text{máx}(0, x) = \begin{Bmatrix} 0 \text{ para } X < 0 \\ 0 \text{ para } X \geq 0 \end{Bmatrix}, \tag{1}$$

donde $f(x)$ representa la salida de la función de activación, donde y x es el valor de entrada. La función ReLU es comúnmente utilizada en redes neuronales debido a su simplicidad y eficiencia computacional. Además, su comportamiento no lineal permite a la red aprender relaciones no lineales entre los datos de entrada y salida.

En resumen, la función ReLU se encarga de anular los valores negativos de la entrada y mantener los valores positivos para ser utilizados como entrada en la siguiente capa de la red. La función GLU (Gated Linear Unit) es una función de activación que se utiliza en redes neuronales. Esta función reduce a la mitad el número de canales de la capa anterior de la red [6]. La ec. 2 que define la función GLU es la siguiente:

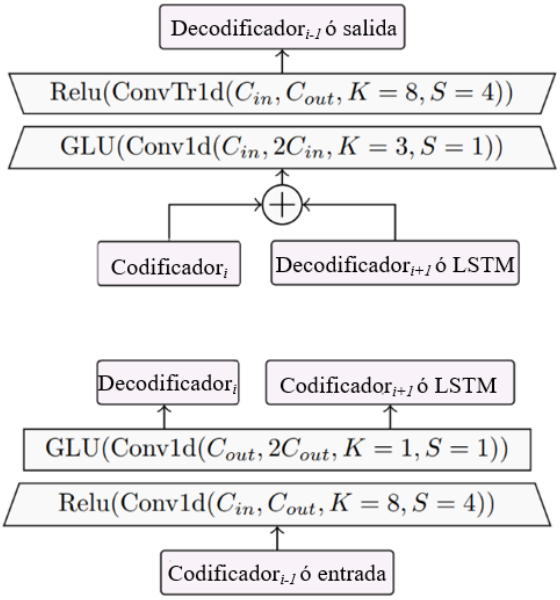

Fig. 3. Bloque i: Codificador y decodificador.

$$h(X) = (X \cdot W + b) \bigotimes \sigma\left(X \cdot V + c\right), \tag{2}$$

donde $h(X)$ representa la salida de la función de activación GLU, σ es la función de activación sigmoide, $\bigotimes$ representa la operación de multiplicación puntual entre matrices, W y V son matrices de pesos que se aprenden durante el entrenamiento de la red, b y c son parámetros de bias que también se aprenden durante el entrenamiento de la red, y X es la entrada a la capa de la red que utiliza la función GLU como función de activación [3].

Se utilizan las funciones de activación ReLU y GLU en la arquitectura propuesta debido a sus propiedades de no linealidad y eficiencia computacional. La función ReLU es simple y eficiente, lo que la hace adecuada para su uso en redes neuronales profundas. Además, ayuda a evitar el problema del desvanecimiento del gradiente, que puede afectar negativamente el rendimiento del modelo. Por otro lado, la función de activación GLU se utiliza para resaltar las características más importantes de los datos de entrada y ayudar a la red a aprender patrones más complejos en los datos.

1.3. Red LSTM

Una red LSTM es una arquitectura de red neuronal recurrente que se enfoca en mantener y recordar información a largo plazo. La principal ventaja de una red LSTM sobre otras arquitecturas recurrentes es su capacidad para evitar el problema del desvanecimiento del gradiente, que puede ocurrir cuando se propagan los errores a través de muchas capas de la red.

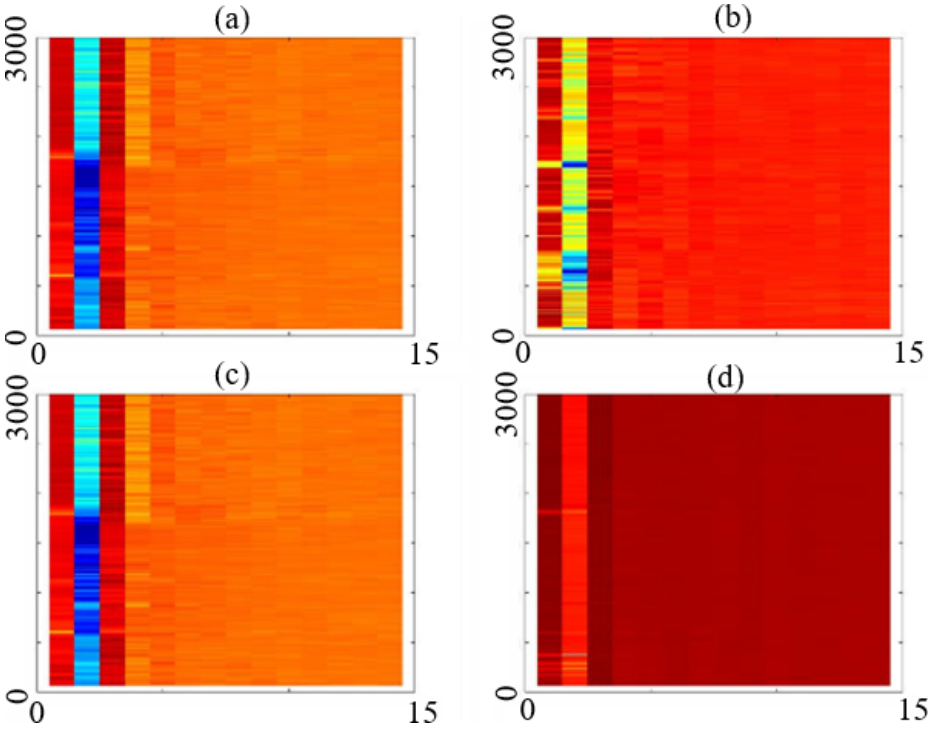

Fig. 4. Comparativa de STEMS: (a) Voz en estudio vs (b) Voz con la arquitectura. (c) Bajo en estudio vs (d) Bajo con la arquitectura.

Los LSTM utilizan una estructura de celdas de memoria con compuertas que regulan el flujo de información en la celda, lo que les permite recordar y olvidar información según sea necesario [11]. En el contexto de la arquitectura propuesta, la red LSTM se utiliza para organizar la salida de la red en una línea de tiempo, lo que permite que los STEMS de la canción tengan la misma duración de tiempo que la entrada y mantengan la coherencia temporal.

1.4. Coeficientes cepstrales en las frecuencias de Mel

A la entrada de la red U-NET se le proporciona un archivo de audio el cual es procesado para obtener los coeficientes Cesptrales en las frecuencias de mel (MFCC) para reducir la dimensionalidad y extraer características relevantes de los espectros de frecuencia.

Los MFCC son similares al espectro de frecuencia, pero en lugar de representar la amplitud en cada banda de frecuencia, se utilizan los coeficientes cepstrales de Mel para representar la energía en diferentes bandas de frecuencia. Esto permite una mejor identificación de los diferentes STEMS y su posterior segmentación en la U-NET [5]. La Figura 1. muestra el diagrama de flujo del proceso para obtener los MFCC:

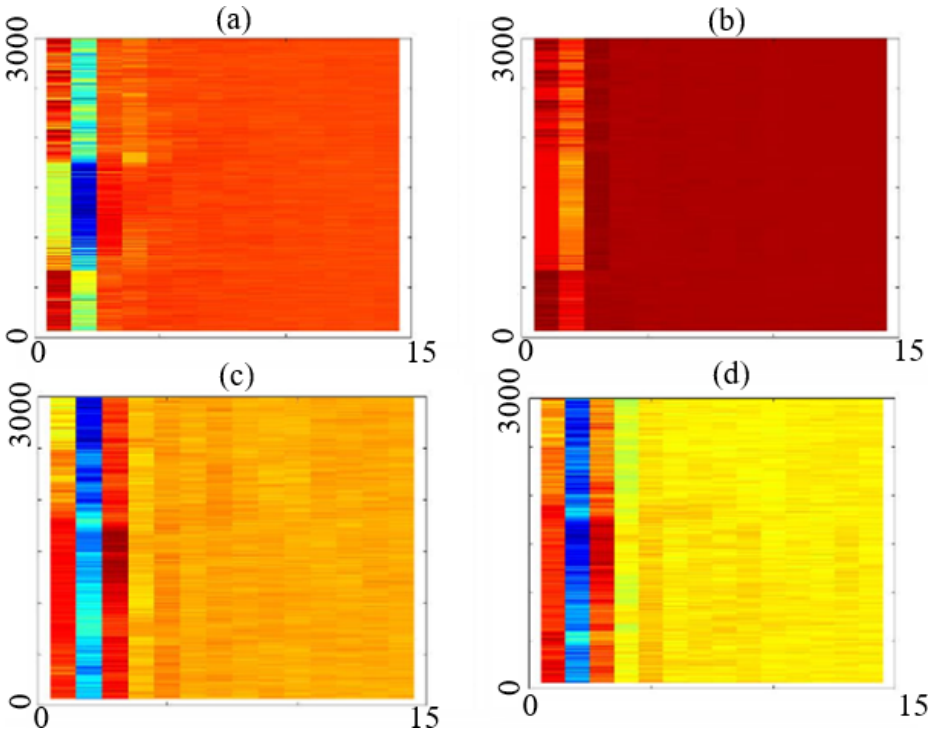

Fig. 5. Comparativa de STEMS: (a) Batería en estudio vs (b) Batería con la arquitectura. (c) Otros en estudio vs (d) Otros con la arquitectura.

1. Señal de audio: Esta entrada contiene la canción a la cual se le quieran extraer los MFCC.
2. Separar la señal en fragmentos cortos y de tiempos iguales: las frecuencias en una señal cambian con el tiempo, es por eso que se analiza en pequeños tramos de tiempo (comúnmente se realiza un corte de entre 20 y 40 ms).
3. Transformada discreta de Fourier (DFT): Se calcula la transformada, obteniendo así la representación en el dominio de la frecuencia (periodograma).
4. Banco de filtros: Lo que hace este filtro es tomar grupos de periodogramas y sumarlos, para tener una idea de cuanta energía existe en varias regiones de frecuencia.
5. Logaritmo: Los resultados son sometidos al logaritmo que representan una aproximación al comportamiento del odio humano al escuchar música.
6. Transformada de coseno discreta (DCT): Se aplica para obtener una correlación de las energías de potencia en dB de los MFCC.
7. Al finalizar el proceso se obtienen los MFCC.

Tabla 1. Análisis de SDR en dB.

	SDR Vocales	SDR Bajo	SDR batería	SDR Otros
STEMS obtenidos de estudio musical	8.91	5.94	11.05	9.80
STEMS obtenidos con red neuronal	7.99	4.65	8.99	9.11

2. Metodología / desarrollo

La arquitectura propuesta está diseñada para tomar una mezcla estéreo como entrada (denominada $C = 2$) y generar una salida estéreo para cada fuente, se puede considerar que la U-Net es la arquitectura principal, seguido de un codificador convolucional, un LSTM bidireccional y un decodificador convolucional.

El codificador se encarga de reducir la dimensionalidad de la entrada mediante la aplicación de múltiples capas convolucionales, mientras que el decodificador se encarga de reconstruir la salida a su tamaño original mediante la aplicación de múltiples capas deconvolucionales.

El LSTM bidireccional se utiliza para capturar las relaciones temporales a lo largo del tiempo en la entrada. Además, la arquitectura utiliza conexiones U-Net para vincular el codificador y el decodificador, lo que permite que la información de la entrada se transmita directamente a la salida. Esto permite que la red pueda realizar una segmentación de los STEMS de la canción en la salida, la arquitectura general se representa en la Figura 2.

El codificador consta de bloques convolucionales ($L = 6$) apilados y numerados del 1 al L. Cada bloque (i) está compuesto por una convolución con tamaño de núcleo $K = 8$, zancadas $S = 4$, canales de entrada $Ci - 1$, canales de salida Ci y una función de activación ReLU.

Luego, se realiza una convolución con tamaño de núcleo $K = 1$, canales de salida $2Ci$ y se aplica la función de activación GLU. Como las GLU reducen a la mitad el número de canales, la salida final del bloque i tiene canales de salida Ci. En la Fig. 3 se describe este bloque, Debido a que nuestros canales de entrada son una mezcla estéreo, es decir, izquierda y derecha, corresponde $Ci = 2$.

El decodificador se compone de L bloques, numerados de forma inversa de L a 1. Cada bloque i comienza con una convolución de tamaño de núcleo 3, zancada 1, canales de entrada/salida Ci y activación ReLU. Luego, se aplica una convolución transpuesta con un ancho de núcleo de 8 y zancada de 4, canales de salida $Ci-1$ y activación ReLU.

En la capa final, se sintetizan las fuentes (S) después de todos los bloques del decodificador. La capa final es lineal con $S \cdot Ci$ canales de salida, uno para cada fuente (4 canales estéreo en nuestro caso), sin ninguna función de activación. Cada uno de estos canales genera directamente una representación de la forma de onda correspondiente a una pista de audio [4].

3. Resultados y discusión

3.1. Comparativa de SDR y MFCC

La relación de fuente a distorsión (SDR) es uno de los métodos más comúnmente utilizados para evaluar la calidad de la salida de un sistema de separación de fuentes de música.

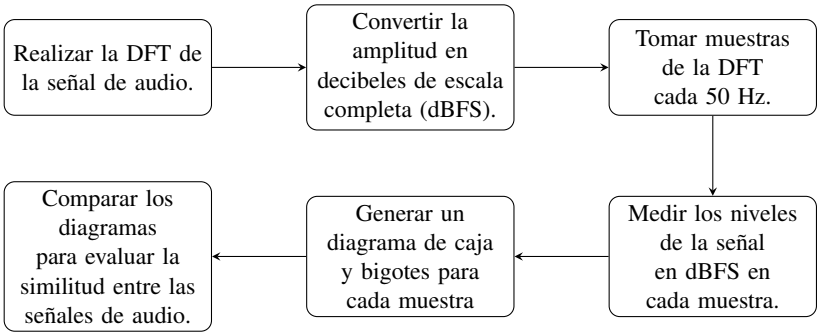

Fig. 6. Diagrama de bloques para obtener la comparativa en dBFS.

El SDR se utiliza para medir la relación entre las señales de las fuentes originales y las señales estimadas por el sistema de separación de fuentes [16]. Una mayor SDR indica una mejor separación de las fuentes y, por lo tanto, una mejor calidad de la pista de audio resultante, con ayuda de la ec. 3, donde:

$$\mathrm{SDR} = 10\log_{10}\left(\frac{\|S_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}\|^2}\right), \tag{3}$$

donde S_{target} se refiere a la fuente verdadera, e_{interf} al error de interferencia, e_{noise} al error de ruido y e_{artif} al error de artefactos. Basándonos en los criterios de la sexta campaña comunitaria de evaluación de separación de señales de música (SiSEC-Mus) [14], se presentarán los resultados del SDR para evaluar la calidad de la salida del sistema de separación de fuentes de música. Además, se analizará la forma de onda del resultado para obtener una mejor comprensión de la calidad de la separación [10]. Para obtener los resultados presentados, se siguieron los siguientes pasos:

1. Se obtuvieron 30 segundos de STEMS de una canción producidos en un estudio de grabación de música, y se obtuvieron STEMS de la misma canción producidos con la red neuronal.
2. Se realizaron comparaciones entre los STEMS obtenidos con la arquitectura propuesta en el paso 1 y los obtenidos en el estudio de grabación mediante la extracción de MFCC para analizar las características de frecuencia de los coeficientes. Se detectó una diferencia en potencia en dB entre ambos, y se identificó que los coeficientes de color rojo indican una saturación en el sonido, lo que puede ser causado por la presencia de ruido no deseado debido a la filtración de instrumentos o partes de audio no deseadas. Cabe destacar que la presencia de ruido indica que la arquitectura no proporciona una separación de pistas de audio completamente fiel al original.
3. Se obtuvieron los resultados de SDR de los STEMS obtenidos en el paso 1, en donde Un valor más alto de SDR indica que la señal es más clara y está menos distorsionada, lo que se traduce en una mayor calidad de audio. En otras palabras, cuanto mayor es el valor de SDR, menor es la cantidad de ruido y distorsión en la señal de audio, lo que indica una mayor calidad de los STEMS.

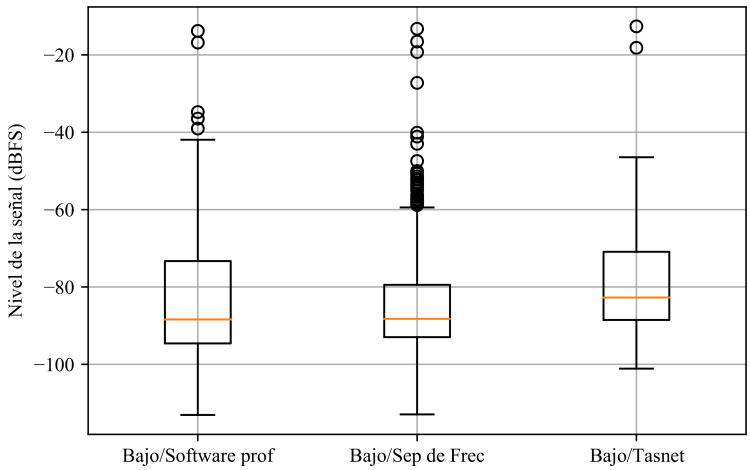

Fig. 7. Diagrama de caja del STEM de bajo.

Los resultados de SDR muestran que la arquitectura de separación de frecuencias de música proporciona resultados similares a los STEMS originales obtenidos en el estudio de grabación. Dichos resultados se pueden observar en la tabla 1.

3.2. Comparativa de nivel de señal

Una forma de analizar el nivel de una señal de audio es con la transformada discreta de Fourier (DFT) ec. 4, que es una técnica utilizada para analizar señales en el dominio de la frecuencia. La DFT nos proporciona un espectro que muestra la distribución de las diferentes componentes de frecuencia que componen una señal de audio con respecto a una magnitud de coeficientes adimensionales dados por la DFT [7]:

$$\text{DFT} = x(k) = \sum_{n=0}^{N-1} x(n) e^{-\mathrm{j}\,2\pi kn/2}, \tag{4}$$

donde, $x(k)$ es la k-ésima muestra en el dominio de la frecuencia, $x(n)$ es la n-ésima muestra en el dominio del tiempo, N es el número de muestras en la señal, k es el índice de la frecuencia en el dominio de la frecuencia y j es la unidad imaginaria.

Para obtener una representación más inteligible de dichos coeficientes adimensionales, se pueden transformar a un nivel en dBFS (decibeles en relación a la escala completa) con el uso de la ec. 5.

El dBFS es una medida utilizada comúnmente para conocer los niveles de audio digital, y permite expresar la amplitud de la señal en una escala logarítmica relativa al máximo nivel posible en el formato digital utilizado.

De esta manera, los valores en dBFS proporcionan una mejor comprensión del nivel de señal presente en la DFT, especialmente cuando se trabaja con señales de audio digitales y se requiere una medición más precisa y detallada de los niveles de señal:

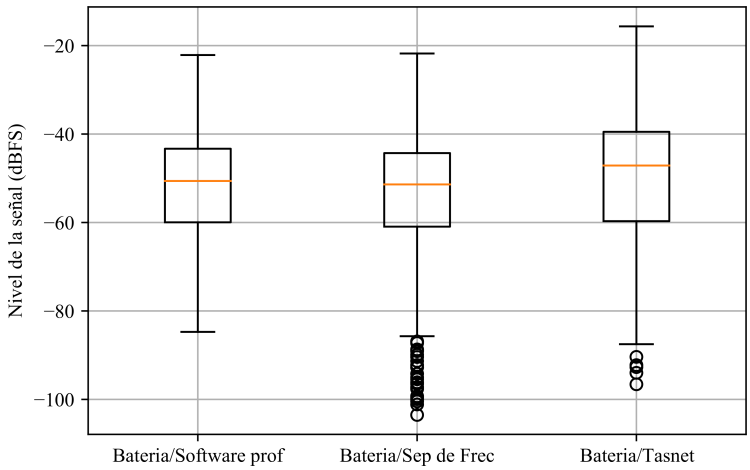

Fig. 8. Diagrama de caja del STEM de Batería.

$$\mathrm{dBFS}[k] = 20 \cdot \log_{10}(PR[k]), \tag{5}$$

donde, $\mathrm{dBFS}[k]$ representa el nivel de amplitud de la señal en una escala de decibeles a escala completa y $PR[k]$ es la potencia relativa de la señal en la muestra de la DFT. Posteriormente, se realizaron mediciones de los niveles en dBFS para cada intervalo de 50 Hz en nuestra DFT.

El análisis se llevó a cabo con la hipótesis de que la DFT obtenida a través de la arquitectura de aprendizaje profundo debería ser similar al STEM obtenido mediante un programa profesional. Con el fin de evaluar la calidad de la DFT obtenida por el modelo de aprendizaje profundo, se midieron los niveles de la señal en dBFS en intervalos regulares de frecuencia (Fig. 6).

Tomando como nota que en el caso de una frecuencia de muestreo de 44100 Hz, el teorema de Nyquist-Shannon establece que la frecuencia máxima que se puede representar es de 22050 Hz (la mitad de la frecuencia de muestreo). Esto implica que, para realizar mediciones de niveles en dBFS cada 100 Hz en una señal digital con frecuencia de muestreo de 44100 Hz, el rango de frecuencias que se puede analizar es de 0 a 22050 Hz [1].

Además, también se realizó la comparativa con Conv-TasNet [9], que es un algoritmo de separación de fuentes de audio que utiliza capas convolucionales unidimensionales (1D) en su red neuronal. TasNet, que significa Red de separación de audio en el dominio del tiempo, es la arquitectura general utilizada en Conv-TasNet.

Esta arquitectura consiste en capas convolucionales 1D seguidas de capas de activación y agrupamiento. El objetivo de Conv-TasNet es separar señales de audio en fuentes separadas, como la separación de voces en una grabación. Las capas convolucionales 1D aplican filtros a lo largo del eje temporal de la señal de entrada, capturando características locales y patrones en datos secuenciales. La salida de una capa convolucional 1D puede tener múltiples canales, lo que significa que se aplican varios filtros para capturar diferentes características.

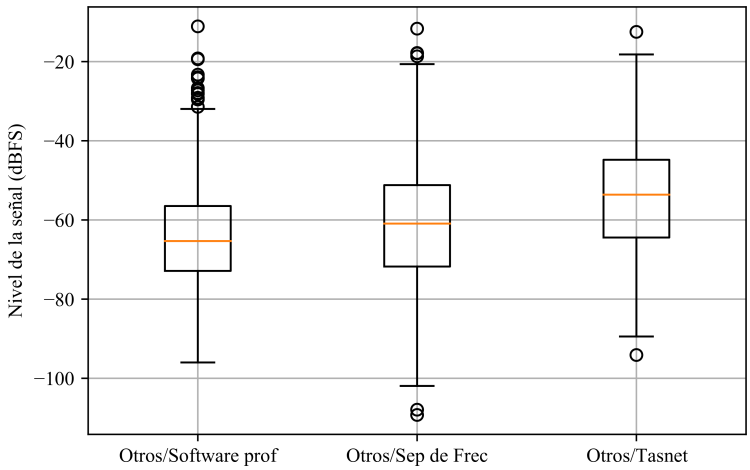

Fig. 9. Diagrama de caja del STEM de Otros.

Después de la convolución, se puede aplicar una función de activación, como ReLU, para introducir no linealidad en la red. Estas capas convolucionales 1D son especialmente adecuadas para el procesamiento de señales de audio debido a la estructura temporal de las formas de onda de audio, lo que las convierte en una herramienta útil en la separación de fuentes de audio, incluyendo la separación de frecuencias musicales. Datos por resaltar:

– La mediana en los audios obtenidos con la arquitectura de frecuencias musicales es similar a la mediana de la fuente original esto sugiere que esta arquitectura ha logrado preservar la ubicación central de la información en la señal original. La similitud en las medianas indica que la arquitectura de frecuencias musicales ha sido efectiva para capturar y mantener la tendencia central de la señal de audio, mientras que por otra parte en Conv-Tasnet se observa un desfase de esta mediana.
– En el diagrama del bajo, se puede observar que el valor máximo y mínimo es más amplio en la arquitectura de separación de frecuencias en comparación con Conv-TasNet por lo que nos habla de que tiene un mayor rango de nivel de señal con respecto a la potencia. Además, se pueden identificar varios valores atípicos en la arquitectura de separación de frecuencias, los cuales casi completan el parecido al rango del bajo de la fuente original. Esto podría sugerir que la arquitectura de separación de frecuencias captura una mayor variabilidad o componentes inusuales en comparación con la fuente original y Conv-TasNet.
– Es importante tener en cuenta que la presencia de valores atípicos puede tener diferentes causas, como artefactos de separación de fuentes o la presencia de componentes de bajo similares en otras fuentes. Por lo tanto, aunque los valores atípicos en la arquitectura de separación de frecuencias se asemejen al rango del bajo original, es necesario analizar con mayor detalle la naturaleza de estos valores atípicos y su relación con la fuente original.

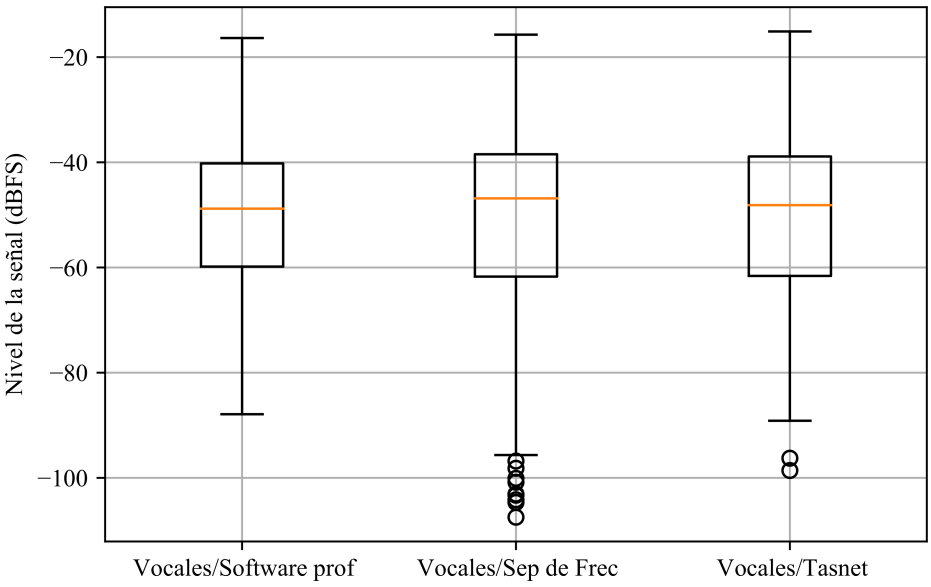

Fig. 10. Diagrama de caja del STEM de Vocales.

– El fenómeno de valores atípicos que se menciona anteriormente se observa con frecuencia en las arquitecturas de aprendizaje profundo utilizadas para el procesamiento de frecuencias musicales.
– El tamaño de la caja en el diagrama de las arquitecturas Conv-Tasnet y de separación de frecuencias musicales es similar al de la fuente original. Esta similitud en el tamaño de la caja sugiere que no existe una diferencia significativa en la dispersión de los datos entre ambas fuentes.

4. Conclusiones

Se ha logrado una transferencia de aprendizaje a una arquitectura que utiliza dos modelos de red neuronal para mejorar la calidad de la separación de las pistas de audio de los instrumentos principales que componen una canción.

Se ha analizado individualmente las dos arquitecturas que componen la red neuronal, comprendido su funcionamiento en el procesamiento de audio, y se ha entendido la información relevante para implementarla en la arquitectura de aprendizaje profundo.

Además, se ha ejecutado y aplicado la arquitectura propuesta para la separación de frecuencias de música la separación de frecuencias musicales. Estos logros permitirán avanzar en la mejora de la calidad de la separación de pistas de audio y en la aplicación de esta tecnología en la separación de fuentes de música.

Agradecimientos. Agradecimientos al Instituto Politécnico Nacional por el apoyo en la realización de este trabajo de investigación.

Referencias

1. Alibeigi, M., Hashemi, S., Hamzeh, A.: DBFS: An effective density based feature selection scheme for small sample size and high dimensional imbalanced data sets. Data and Knowledge Engineering, vol. 81-82, pp. 67–103 (2012) doi: 10.1016/j.datak.2012.08.001
2. Cherry, E. C.: Some experiments on the recognition of speech, with one and with two ears. The Journal of the Acoustical Society of America, vol. 25, no. 5, pp. 975–979 (1953) doi: 10.1121/1.1907229
3. Dauphin, Y. N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. In: Proceedings of the 34th International Conference on Machine Learning, vo. 70, pp. 933-941 (2017) doi: 10.48550/ARXIV.1612.08083
4. Défossez, A., Usunier, N., Bottou, L., Bach, F.: Demucs: Deep extractor for music sources with extra unlabeled data remixed (2019) doi: 10.48550/ARXIV.1909.01174
5. Huizen, R. R., Kurniati, F. T.: Feature extraction with mel scale separation method on noise audio recordings (2021) doi: 10.48550/ARXIV.2112.14930
6. Lederer, J.: Activation functions in artificial neural networks: A systematic overview (2021) doi: 10.48550/ARXIV.2101.09957
7. Lenssen, N., Needell, D.: An introduction to fourier analysis with applications to music. Journal of Humanistic Mathematics, vol. 4, no. 1, pp. 72–91 (2014) doi: 10.5642/jhummath.201401.05
8. Lostanlen, V., Cella, C. E.: Deep convolutional networks on the pitch spiral for musical instrument recognition (2016) doi: 10.48550/ARXIV.1605.06644
9. Luo, Y., Mesgarani, N.: Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 8, pp. 1256–1266 (2019) doi: 10.1109/taslp.2019.2915167
10. Nakajima, H., Takahashi, Y., Kondo, K., Hisaminato, Y.: Monaural source enhancement maximizing source-to-distortion ratio via automatic differentiation (2018) doi: 10.48550/ARXIV.1806.05791
11. Staudemeyer, R. C., Morris, E. R.: Understanding LSTM – a tutorial into long short-term memory recurrent neural networks (2019) doi: 10.48550/ARXIV.1909.09586
12. Stoller, D., Ewert, S., Dixon, S.: Wave-U-Net: A multi-scale neural network for end-to-end audio source separation (2018) doi: 10.48550/ARXIV.1806.03185
13. Stöter, F. R., Uhlich, S., Liutkus, A., Mitsufuji, Y.: Open-Unmix - A reference implementation for music source separation. Journal of Open Source Software, vol. 4, no. 41, pp. 1667 (2019) doi: 10.21105/joss.01667
14. Stöter, F. R., Liutkus, A., Ito, N.: The 2018 signal separation evaluation campaign. In: Latent Variable Analysis and Signal Separation, Lecture Notes in Computer Science, vol 10891, pp. 293-–305 (2018) doi: 10.1007/978-3-319-93764-9_28
15. Takahashi, N., Goswami, N., Mitsufuji, Y.: MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation. In: 16th International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 106-110 (2018) doi: 10.1109/IWAENC.2018.8521383
16. Vincent, E., Gribonval, R., Fevotte, C.: Performance measurement in blind audio source separation. In: IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 4, pp. 1462–1469 (2006) doi: 10.1109/tsa.2005.858005