Separación de fuentes en pistas de audio con múltiples instrumentos mediante redes convolucionales

Brisa Antuña B., Sofía Escudero y Nael Pighin
Universidad Nacional Del Litoral, brisantuna@gmail.com

***Resumen*—En este documento se detalla el procedimiento para la separación de fuentes de una pista de audio a partir de redes neuronales convolucionales. El mismo será presentado como un informe final de proyecto para la materia "Inteligencia Computacional" de la carrera Ingeniería Informática en la Facultad UNL-FICH. Si bien la problemática puede resolverse desde distintos ámbitos, abordarlo desde la inteligencia computacional puede obtener mejores resultados y, por lo tanto, ser de mejor ayuda.**

Este documento tiene como punto de partida el artículo "Monoaural Audio Source Separation Using Deep Convolutional Neural Networks" de Chandna et al[1], y a partir del mismo se aplican los métodos pertinentes para lograr la separación. Se realiza la STFT para preparar las entradas a la red, se modela, entrena y utiliza la red convolucional y luego se procesa la salida con la ISTFT para obtener nuevamente los audios. Finalmente se realiza el análisis de los resultados obtenidos y se comprueba la efectividad del método mediante métricas de calidad, obteniendo resultados decentes para la complejidad de la solución propuesta.

Para su evaluación, se tomaron de a pares las fuentes bajo, batería y vocales. Según los análisis objetivos y subjetivos, la fuente que mejor desempeño tuvo en este trabajo fueron las vocales.

***Palabras clave*— redes convolucionales, inteligencia computacional, separación de fuentes.**

I. Introducción

En los últimos años, la inteligencia artificial ha evolucionado a grandes pasos, y se han comenzado a utilizar cada vez más como herramientas para abordar diferentes problemas de forma más eficiente y efectiva.

En nuestro caso, se utilizará la inteligencia computacional, más específicamente las redes convolucionales (CNN), para la separación de fuentes de distintas obras musicales. Dicho objetivo fue explorado en un trabajo previo, donde se utilizó factorización en matrices no negativas (NMF), pero los resultados fueron poco favorables.

Dado esto, se buscaron nuevas formas de resolver el problema utilizando esta vez inteligencia computacional.

Las redes neuronales convolucionales nos permiten analizar las señales de audio por secciones más pequeñas, logrando obtener características específicas de las mismas, que luego se utilizan para separar las fuentes.

Obtener pistas de audios individuales puede ser de ayuda en múltiples ámbitos. Una implementación puede ser como paso previo para la separación del habla con ruido de fondo, aplicado principalmente en dispositivos como los audífonos amplificadores para personas con disminución auditiva. La separación también puede ser de ayuda en las lecciones para aprender a tocar distintos instrumentos, aportando la posibilidad de aislar el mismo para un análisis más detallado.

II. Material y métodos

Como se mencionó en la introducción, se utilizará una red neuronal convolucional para realizar la separación de fuentes en señales de audio.

Para poder utilizar este método, es pertinente realizar algunos tratamientos a la señal de audio previamente.

Teniendo la señal de audio en estéreo, se realiza la Transformada de Tiempo Corto de Fourier (STFT) para así obtener una representación en tiempo y frecuencia. La entrada de la red convolucional será la matriz correspondiente a la magnitud de la representación tiempo-frecuencia, pero también se guarda la fase de la misma, ya que será necesaria al momento de aplicar la Transformada de Tiempo Corto de Fourier Inversa (ISTFT) para la reconstrucción de las señales tras la separación de fuentes.

Es importante destacar que para el entrenamiento y la validación no será necesario guardar la fase de los audios que le pasemos a la red, ya que la red trabaja con la representación Tiempo-Frecuencia. Por otro lado, cuando se deseen obtener los audio reconstruidos, teniendo ya entrenada la red, si será necesario guardar la fase para dicha reconstrucción.

A. Redes neuronales convolucionales

Las redes neuronales convolucionales son un tipo de arquitectura de red neuronal que se aplica a matrices, en nuestro caso bidimensionales, siendo esto sumamente útil para el procesamiento de imágenes (matrices según dimensiones en píxeles) y, lo que nos interesa particularmente en el presente informe, el manejo de audio (con representaciones tiempo-frecuencia que se obtienen mediante la STFT previamente explicada).

Se componen de tres tipos principales de capas: capas convolucionales, capas de agrupación o pooling y capas totalmente conectadas.

Las capas convolucionales están caracterizadas por un kernel o filtro, que se desplaza por la matriz de audio o de la imagen para detectar características y patrones locales. El tamaño del kernel determina el tipo de características que se buscan agrupar,

Las capas de agrupación o pooling permiten reducir el número de parámetros de la entrada mediante un filtro, perdiéndose información pero ganando eficiencia y limitando riesgos de sobreajuste.

Las capas totalmente conectadas son las capas clásicas que realizan la clasificación basada en las características aprendidas por las capas anteriores.

Una red convolucional se forma con distintas organizaciones de estas capas dependiendo del problema a resolver. En general, las primeras capas se centran en características simples de la entrada y, a medida que los datos avanzan en las capas, la CNN aumenta en complejidad y comienza a reconocer en mayor magnitud.

B. Estructura de la red propuesta

Para el presente trabajo se propone la siguiente estructura de red convolucional, inspirado por el trabajo de Chandna et al[1]. La red se compone de dos partes, *codificación* y *decodificación*. Dentro de la etapa de codificación se encuentran 3 capas:

1. Capa de convolución vertical: Esta capa tiene 50 neuronas (N1) con una entrada de 513x25, donde se realiza la convolución con un kernel de tamaño (513,1), es decir, un kernel con una altura similar a la entrada y ancho 1. Además se agrega un padding de ceros por encima y por debajo de la matriz con el objetivo de obtener una salida de mayor verticalidad sin modificar el kernel. Esta capa tiene como función principal captar características del timbre de la entrada.
2. Capa de convolución horizontal: Esta capa posee 30 filtros o neuronas (N2) y tiene como entrada la salida de la capa de convolución vertical, que son matrices de 41x25. En esta capa se realiza la convolución con un kernel de 12x1. En este caso el kernel es horizontal, pero no ocupa todo el ancho de la entrada. Esta capa tiene como finalidad identificar características temporales de la entrada.
3. Capa totalmente conectada: La salida de la capa de convolución horizontal se pasa primero por una capa flatten, que “aplana” los datos, es decir, los convierte en un vector para así poder pasar a la capa totalmente conectada. Esta capa cuenta con 128 neuronas (NN).

Luego, en la etapa de decodificación, se realizan dos capas de cada tipo (tres en total) para obtener así una salida por fuente a separar. Se realiza para cada fuente el proceso inverso que se realizó para obtener las características, es decir, primero se pasa la salida de la capa totalmente conectada de la etapa de codificación a dos capas totalmente conectadas para comenzar con la separación. Luego se pasan ambas salidas por sus respectivas capas de deconvolución horizontal y, por último, por la de deconvolución vertical. De esta forma, se obtienen dos salidas del mismo tamaño de la entrada, cada una con las características del respectivo instrumento a separar.

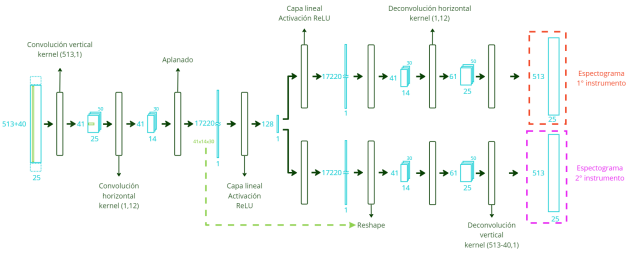

Fig. I: Diagrama de la red convolucional.

D. Entrenamiento

Para poder procesar las canciones en el entrenamiento, primero se extraen 30 segundos de cada una a modo de obtener entradas del mismo tamaño y simplificadas para agilizar el procesamiento experimental. Luego se realiza la STFT y se separan los espectrogramas de a 25 frames, formando así la entrada para la red con las dimensiones necesarias.

La optimización de los parámetros de la red se realiza minimizando la pérdida de la función de error

$$loss = L_{sq} - alpha * L_{diff} \quad (3),$$

donde L_{sq} se calcula como:

$$L_{sq} = \sum_{i=1}^{N} \left|\left| \overline{y_n} - y_n \right|\right|^2 \quad (3)$$

y L_{diff} como el error cuadrático entre las predicciones de las dos fuentes, penalizando a la función de pérdida si ambos son muy diferentes. Esto es dado que se busca separar los instrumentos de una misma canción, por lo que los sonidos de los mismos deben ser coherentes entre sí.

E. Obtención de la salida

En última instancia, para comprobar el funcionamiento de la red y generar los archivos de audio de las fuentes aisladas se preprocesa la canción a separar. Se separa la misma en grupos de 25 frames con 2 de superposición. Para las columnas que se superponen se realiza un promedio y luego se concatenan los fragmentos de espectrograma. Una vez conseguido el espectrograma completo se realiza la antitransformada de tiempo corto para obtener la señal final.

F. Métricas de calidad

En un principio, se hará una calificación subjetiva (en una escala del 1 al 5) de la calidad de las reconstrucciones a través de la escucha de los audios resultantes por parte de diez distintos individuos, a modo de tener una métrica que considere la percepción humana de las frecuencias.

Se complementará el análisis del desempeño con una métrica objetiva: la SDR (Source-Distortion Ratio) que utiliza Chandna[3] en su interpretación del problema, y es una de las métricas más usadas para la separación de fuentes.

G. Bases de datos

Para el desarrollo de este trabajo se utilizó la base de datos MUSDB18[4] desarrollada para evaluar métodos de separación de fuentes en grabaciones musicales. Contiene 150 pistas musicales completas de distintos géneros (aproximadamente 10 horas de duración), diferenciándose en cinco canales: mix, bajo, batería, acompañamientos (una selección de instrumentos muy variada) y finalmente voz. Contiene una carpeta de "train" con 100 pistas y otra de test con 50 pistas.

III. Resultados

Se entrenó la red descrita con un total de 10 épocas, logrando una pérdida de aproximadamente 2535. Se realizó el análisis de los resultados con un audio de la base de datos en particular, "Arise - Run Run Run", tomando de a pares de fuentes para observar los resultados al intentar separar distintas combinaciones de las mismas.

Como resultados, se presentan los distintos espectrogramas entre las fuentes originales y las separadas, y las métricas tanto objetivas como subjetivas registradas para cada caso.

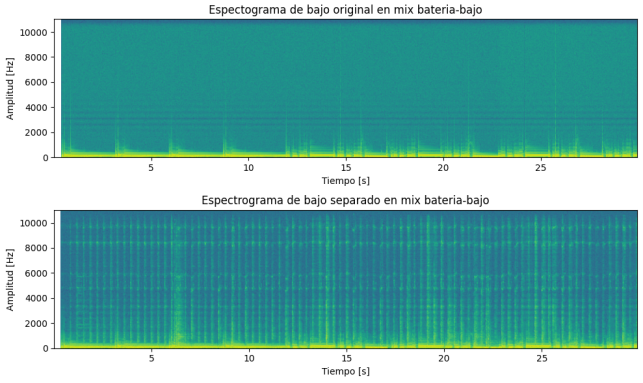

Fig. II: Espectrogramas del bajo original y el bajo separado de la combinación bajo-batería.

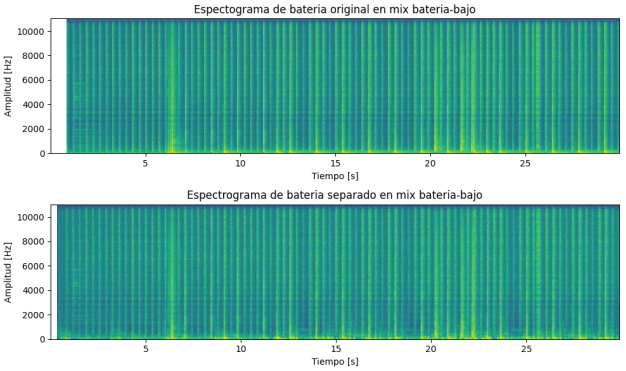

Fig. III: Espectrogramas de la batería original y la batería separada de la combinación bajo-batería.

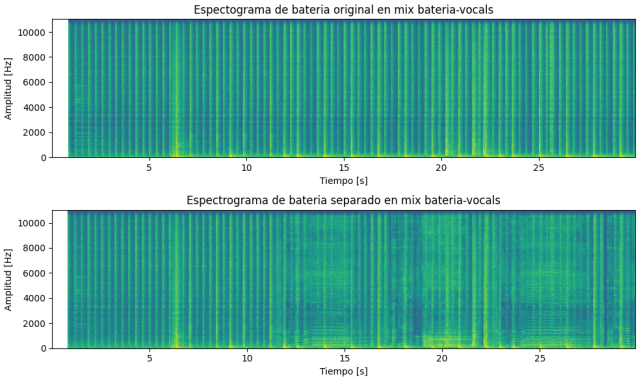

Fig. IV: Espectrogramas de la batería original y la batería separada de la combinación bajo-voz.

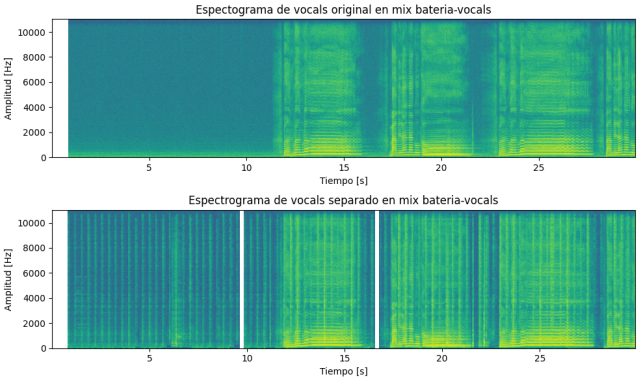

Fig. V: Espectrogramas de la voz original y la voz separada de la combinación bajo-voz.

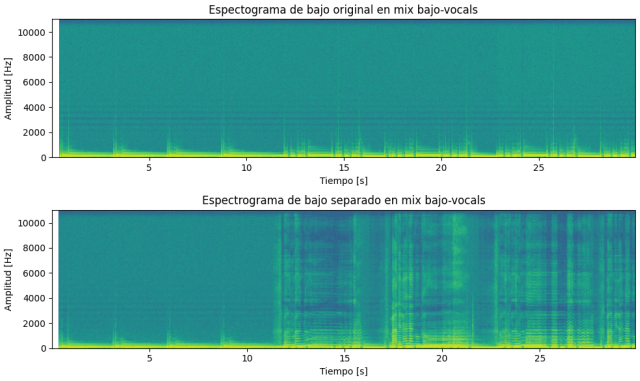

Fig. VI: Espectrogramas del bajo original y el bajo separado de la combinación bajo-voz.

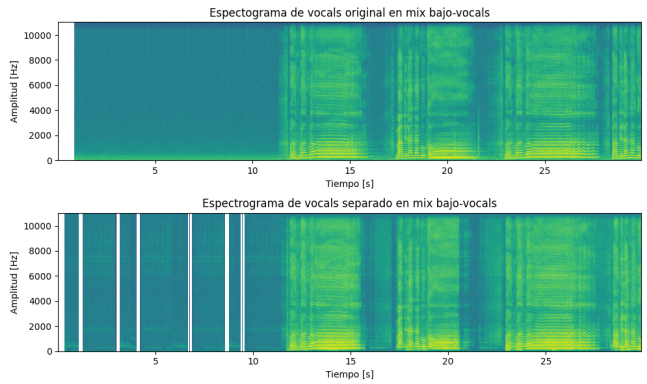

Fig. VII: Espectrogramas de la voz original y la voz separada de la combinación bajo-voz.

	Bajo	Batería	Voz
Bajo/Batería	-0.6587	-7.246	
Bajo/Voz	-0.9694		-0.45331
Batería/Voz		-4.9993	0.575

Tabla 1: Registro de la relación Señal-Distorsión para las fuentes separadas según la combinación de la que se obtuvo.

	Bajo	Batería	Voz
Bajo/Batería	2.9	4.1	
Bajo/Vocals	1.9		4.9
Batería/Vocals		2.9	3

Tabla 2: Registro del puntaje promedio por criterio humano para las fuentes separadas según la combinación de la que se obtuvo.

IV. Discusión

Observando los espectrogramas de las distintas combinaciones, se puede concluir que el método no es perfecto, ya que se agregan y/o atenúan frecuencias en relación a los espectrogramas originales. Con una simple observación de los gráficos, el peor caso de separación parece ser el bajo para la combinación bajo-voz (figura VI), y el mejor es la voz para esta misma combinación a partir de un cierto frame (figura VII).

En particular, en ambas separaciones de la voz se han registrado pérdidas o segmentos que no se han podido separar correctamente en el algoritmo (frames blancos en los espectrogramas de las figuras V y VII), lo cual establece una importante diferencia con los espectrogramas de las otras fuentes al ser separadas.

Si, por otro lado, analizamos los resultados de las métricas objetivas (tabla 1), vemos que la separación de la voz ha tenido mejores resultados en relación a las demás fuentes. Le sigue el bajo y, por último, la batería.

En cuanto a las métricas subjetivas (tabla 2), el promedio de las calificaciones realizadas ha determinado que el instrumento con mejores separaciones es la voz con una calificación de 5 y 3 para sus separaciones. Esto coincide con lo que se observó en los análisis previos.

Por otro lado, el bajo ha mostrado peores resultados que la batería en la métrica subjetiva, a diferencia de la métrica objetiva. Sin embargo, sí coincide que la separación peor calificada es aquella con una mayor diferencia entre sus espectrogramas, el bajo en la combinación bajo-voz.

V. Conclusiones

En este informe se presentó el uso de redes neuronales convolucionales para poder resolver el problema de separación de fuentes en señales de audio, llegando a resultados decentes para una estructura simplificada como la propuesta. En particular, este método se mostró conveniente para el problema abordado ya que permitió trabajar con las matrices de la representación tiempo-frecuencia de manera directa y realizando simplificaciones a las mismas a modo de poder realizar los cálculos matriciales más rápidamente.

Si bien hacer combinaciones de a pares de fuentes para realizar las separaciones supone un análisis más intrínseco del problema, para futuras extensiones de este trabajo se podría trabajar con mezclas de todas las fuentes disponibles en la base de datos elegida y una red ampliada que permita separar las 4 fuentes (o más). También se propone analizar el desempeño del algoritmo al agregar ruido aditivo.

VI. Referencias

[1] Chandna, P., Miron, M., Janer, J., Gómez, E. “Monoaural Audio Source Separation Using Deep Convolutional Neural Networks”, in *Tichavský, P., Babaie-Zadeh, M., Michel, O., Thirion-Moreau, N. (eds) Latent Variable Analysis and Signal Separation.* LVA/ICA 2017. Lecture Notes in Computer Science(), vol 10169. Springer, Cham.

[2] Mirion, M. “Source Separation Methods for Orchestral Music: Timbre-Informed and Score-Informed Strategies”, UPF 2017. Doctoral thesis.

[3] Chandna, P. “Audio Source Separation Using Deep Neural Networks”, UPF 2016. Doctoral thesis.

[4] URL, “https://sigsep.github.io/datasets/musdb.html,” 2017.