Universidad de Lima

Facultad de Ingeniería y Arquitectura

Carrera de Ingeniería de Sistemas

ANÁLISIS COMPARATIVO DE LOS MÉTODOS REPET+ Y UNET PARA LA SEPARACIÓN DE LA VOZ CANTADA EN UNA PISTA MUSICAL

Tesis para optar el Título Profesional de Ingeniero de Sistemas

Jorge Luis Ramon Zuta

Código 20161200

Asesor

Oscar Efrain Ramos Ponce

Lima – Perú

Enero de 2023

Análisis comparativo de los métodos REPET+ y UNet para la separación de la voz cantada en una pista musical

Ramon Zuta, Jorge Luis
20161200@aloe.ulima.edu.pe
Universidad de Lima

Resumen: La separación de fuentes musicales es la tarea de aislar las frases musicales ejecutadas por diferentes instrumentos grabados individualmente y dispuestos juntos para formar una canción. A la actualidad se han desarrollados diversos métodos para abarcar la separación de fuentes musicales, los cuales se pueden clasificar en métodos supervisados y no supervisados; sin embargo, no se ha desarrollado una investigación en la cual se analice la efectividad de usar diferentes métodos en conjunto. Por este motivo, el presente trabajo busca medir los resultados de la utilización de dos métodos, REPET+ (no supervisado) y UNet (supervisado), de manera conjunta y aislada para separar las ondas musicales producidas por un cantante y las ondas provenientes de los instrumentos. Los resultados muestran un puntaje general (SDR) de los métodos para la separación vocal para la red UNet fue de 5.38 dB, REPET+ de -4.3 dB, -2.55 dB para REPET+ & UNet, y, -0.38 dB para UNet & REPET+, -6.16 dB para REPET+ & REPET+ y 5.17 dB para UNet & UNet, demostrando la superioridad de la red UNet para la separación de ondas vocales frente al método REPET+. Además, la utilización de los métodos en forma conjunta muestra una leve mejoría en ciertas métricas de evaluación; sin embargo, tomando en cuenta todas las métricas (SDR, SIR y SAR), se pone en evidencia que esto conlleva a una pérdida de información que recae en un bajo puntaje general de la solución.

Palabras Clave: Recuperación de Información Musical (MIR), Separación de fuentes de sonido, Transformada de Fourier

Abstract: Music source separation is the task of isolating the musical phrases played by different instruments recorded individually and arranged together to form a song. Nowadays, several methods have been developed to cover the separation of music sources, which can be classified into supervised and unsupervised learning, however, no research has been developed in which the effectiveness of using different methods together are analyzed , that's the reason the present work seeks to measure the results of the use of two methods, REPET + (unsupervised) and UNet (supervised), jointly and in isolation to separate the music waves produced by a singer and the waves from the instruments. The results show an overall score (SDR) of the methods for vocal separation for the UNet network was 5.38 dB, REPET+ -4.3 dB, -2.55 dB for REPET+ & UNet, -0.38 dB for UNet & REPET+, -6.16 dB for REPET+ & REPET+ and 5.17 dB for UNet & UNet, demonstrating the superiority of the UNet network for the separation of vocal waves compared to the REPET+ method. In addition, the use of the methods together shows a slight improvement in certain evaluation metrics, however, considering all the metrics (SDR, SIR and SAR), it is evident that this leads to a loss of information that results in a low overall score of the solution.

Keywords: Music Information Retrieval (MIR), Music source separation, Fourier Transform

1. INTRODUCCIÓN

Hoy en día, la música se ha vuelto parte de nuestras vidas debido a la disponibilidad que tenemos para su acceso, ya sea desde la radio, CDs, programas de TV, o incluso en servicios de streaming online. Algunas veces, como oyentes de la música, deseamos modificar el balance de ciertos sonidos pertenecientes a una pista musical, por ejemplo, modificar el volumen a ciertos instrumentos, suprimir sonidos no deseados, volver una pista de audio mono a estéreo, o incluso volverlo multicanal (5.1 surround sound). Querer realizar estas modificaciones en pistas musicales es dificultoso si tenemos, como usualmente lo es, pistas musicales sin las fuentes de sonido debidamente diferenciadas. Sin embargo, el oído humano tiene la capacidad de aislar el sonido de uno o varias fuentes acústicas, como en el caso de la música, donde este puede enfocarse en la presencia de la voz cantada acompañada por sonidos armónicos (Bregman, 1994). Esta tarea para el oído humano requiere de poco esfuerzo, pero, para un sistema computacional no es tarea fácil.

Esta característica del oído humano es altamente estudiada en el área de investigaciones MIR (*Music Information Retrieval*), donde buscan alcanzar dicha capacidad por medio de sistemas computacionales. Lograr esta característica da pie a otras investigaciones como la transcripción de música automática (Plumbley et al., 2002), identificación de instrumentos (Heittola et al., 2009), identificación de cantantes (Sharma et al., 2019), entre otros.

Al día de hoy, se han realizado muchas aproximaciones para la separación de fuentes de sonido, clasificándose en métodos supervisados y no supervisados; sin embargo, no existe investigación alguna que analice de manera profunda la efectividad de dichos métodos en conjunto. Es por ello que el presente trabajo tiene como objetivo comparar la aplicación y combinación del método no supervisado de separación vocal utilizado por Rafii & Pardo (2012), REPET+, y el modelo de aprendizaje profundo (supervisado), UNet, presentado por Jansson et al. (2017).

En la sección 1 este artículo se comenzará analizando investigaciones relacionadas que sirven de base para entender los métodos utilizados y sus aplicaciones. En la sección 2 se revisarán los antecedentes que brindarán una base teórica sobre las áreas del conocimiento involucradas en la investigación. En la sección 3 se presentarán los métodos utilizados para llevar a cabo la investigación, se describirán los resultados obtenidos y se mostrará las interpretaciones sobre estos. Por último, en la sección 4, se brindará un análisis crítico y objetivo de los resultados obtenidos en este trabajo.

2. ESTADO DEL ARTE

La separación de fuentes de sonido se ha enfocado en la manera en cómo el sistema auditivo humano es capaz de diferenciar las señales que emiten diferentes fuentes de sonido. En el libro *Auditory Scene Analysis* de Bregman (1994), se explica que el sistema auditivo realiza un proceso denominado análisis de la escena auditiva (ASA por sus siglas en inglés), en el cual separa los sonidos individuales, provenientes del mundo natural, en situaciones en las cuales estos sonidos suelen estar superpuestos en tiempo y en frecuencia. El proceso ASA, se compone de dos etapas principales, la etapa de agrupación secuencial y la etapa de agrupación simultánea. La primera agrupa toda la información auditiva obtenida respecto al tiempo en que fue detectada y su acústica, mientras que la agrupación simultánea selecciona, aquellos datos que se captan al mismo tiempo, y los agrupa de tal manera de identificarlos como sonidos de una misma fuente. Este proceso de descomposición y reconocimiento de los sonidos, por parte del sistema auditivo, ha inspirado a investigadores a construir sistemas de separación de fuentes de sonido, reconocimiento de voz, entre otros. Diversos métodos se han aplicado para poder separar la voz del acompañamiento musical, los cuales se nombrarán a continuación.

2.1 Métodos no supervisados

También son llamados métodos tradicionales. Estos métodos, por lo general, se basan en la premisa que las canciones presentan una estructura repetitiva que corresponde a la parte armónica (acompañamiento musical), y otra melódica, no repetitiva que varía a lo largo del tiempo la cual corresponde a la voz del cantante (Rafii & Pardo, 2012). Por ello, estos métodos tradicionales buscan identificar aquellos patrones repetitivos para poder filtrarlos del audio original bajo métodos estadísticos.

Uno de los métodos más conocidos es el *REpeating Pattern Extraction Technique* (REPET) implementado por Rafii & Pardo (2012). Este método busca identificar periodos globales de estructuras repetitivas para construir una máscara. Posteriormente, esta máscara es utilizada para segmentar los componentes repetitivos (acompañamiento musical) de la voz. Sin embargo, solo es posible aplicar este método para segmentos musicales cortos, por ejemplo 10 segundos. Para trabajar sobre una pieza musical completa, se ha realizado una versión extendida de este método (Rafii & Pardo, 2012), en la cual se aplican los algoritmos de REPET a segmentos cortos de una pieza musical, denominándolo REPET+.

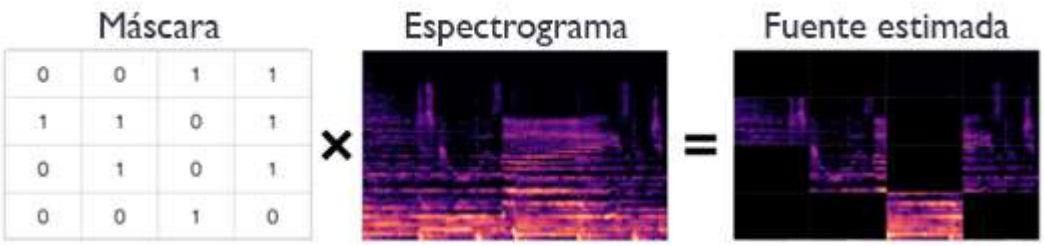
Otros métodos de separación tradicionales intentan modelar la señal musical de segmentos vocales extrayendo las frecuencias predominantes como es el caso de los métodos presentados por Hsu & Jang (2009) y Li & Wang (2007); o realizan un filtrado sobre el espectrograma de una mezcla musical con técnicas como la factorización matricial no negativa (NMF) como en los trabajos presentados por Virtanen (2007) y Vembu & Baumann (2005) y también métodos bayesianos (Ozerov et al., 2007).

2.2 Métodos supervisados

Los avances del aprendizaje profundo han permitido desarrollar modelos que permiten obtener mejores resultados frente a los métodos tradicionales. Por lo general, estos métodos, trabajan con representaciones musicales en el dominio de tiempo-frecuencia, también llamados espectrogramas. La finalidad de estos métodos es encontrar una máscara que represente las frecuencias generadas por la voz, para que, al multiplicarse con el espectrograma original de la canción a querer procesar, se obtenga un espectrograma estimado de la voz como se muestra en la Figura 2.1.

Figura 2.1

Máscara binaria aplicada a espectrograma de un mix para poder separar la fuente deseada.



Nota: Adaptado de Manilow et al. (2020).

Una primera aproximación de la aplicación de métodos supervisados para la separación vocal fue realizada por Grais et al. (2014), donde se utilizó una arquitectura DNN para clasificar los espectros del espectrograma de una pista musical para cada fuente. Simpson et al. (2015) presentaron una red neuronal prealimentada que buscaba estimar una máscara binaria ideal que represente las regiones en donde la voz es más predominante que el acompañamiento musical. Un enfoque similar fue propuesto por Huang et al. (2014), en donde presentaron una arquitectura de red neuronal recurrente para predecir una máscara que es multiplicada por la señal original de la canción para obtener la voz separada del componente instrumental.

Modelos más recientes han sido trabajados en el dominio de forma onda, sin embargo, estos presentan peor rendimiento que los desarrollados en el dominio de tiempo-frecuencia. Un ejemplo de esto es el modelo desarrollado por Jansson et al. (2017) en el que utilizaron la arquitectura UNet, inicialmente desarrollada para segmentación de imágenes, para la separación de fuentes musicales en el dominio de tiempo-frecuencia, y, posteriormente, en el año 2018, Stoller et al. adaptaron dicha arquitectura para que trabajara en el dominio de forma onda.

El problema con los métodos que hacen uso de un aprendizaje supervisado es la falta de grandes conjuntos de datos para entrenamiento debido a que casi la totalidad de la música está sujeta a la protección de derechos de autor, pero, un pequeño número de artistas optan por licencias como *Creative Commons*, que permite compartir sus grabaciones, permitiendo generar datasets de acceso libre con los componentes musicales debidamente diferenciados. Autores han recopilado dichas grabaciones de audio para la creación de datasets orientados a la separación de fuentes de sonido, como MUSDB18 realizado por (Rafii et al., 2017), DSD100 desarrollado por Liutkus et al. (2017), iKala de Chan et al. (2015), por mencionar algunos. Cabe resaltar que estos datasets no son de gran amplitud (MUSDB18 – 150 canciones – 236± 95(s); DSD100 – 100 canciones – 251 ± 60(s); iKala – 206 canciones – 30(s)), es por ello por lo que algunos autores optan por construir su propio dataset, como Jansson et al. (2017), cuyo dataset contiene aproximadamente 20,000 pistas musicales, obteniendo un aproximado de dos meses de reproducción continua, argumentando que es el dataset más grande utilizado para la separación de fuentes musicales.

El presente trabajo tiene como objetivo comparar la aplicación y combinación del método estadístico de separación vocal utilizado por Rafii & Pardo (2012), REPET+, y el modelo de aprendizaje profundo, basado en la segmentación de imágenes, UNet, presentado por Jansson et al. (2017). Este último será entrenado con el repositorio de datos DSD100 (Liutkus et al., 2017).

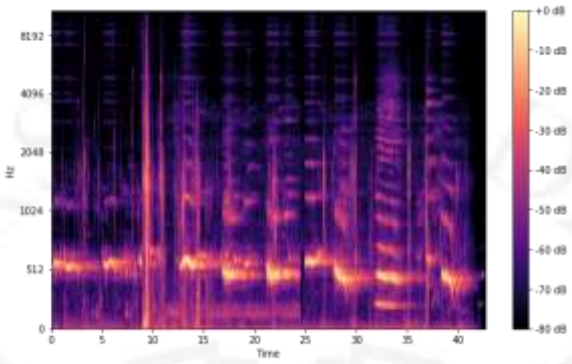
3. ANTECEDENTES

3.1 Espectrogramas

Un espectrograma es una representación visual de una fuente de audio en el dominio de tiempo-frecuencia. Básicamente es una matriz bidimensional que contiene las frecuencias contenidas en una señal de audio a lo largo del tiempo (Smith, 2011). Se puede leer como un mapa de calor, donde los colores más claros representan frecuencias con una mayor amplitud, y los colores más oscuros, las frecuencias de menor amplitud, como se ilustra en la Figura 3.1.

Figura 3.1

Espectrograma de una fuente de audio.



3.2 Transformada de Fourier de Tiempo Reducido (STFT)

Los sonidos que nosotros escuchamos son la composición de varias ondas sonoras juntas, sin embargo, nuestro oído tiene la capacidad de diferenciar los sonidos de manera automática. En el año 1822, un matemático francés llamado Joseph Fourier, descubrió que hasta el movimiento ondulatorio periódico más complejo se puede descomponer en ondas senoidales más sencillas que se suman (Bracewell, 1989).

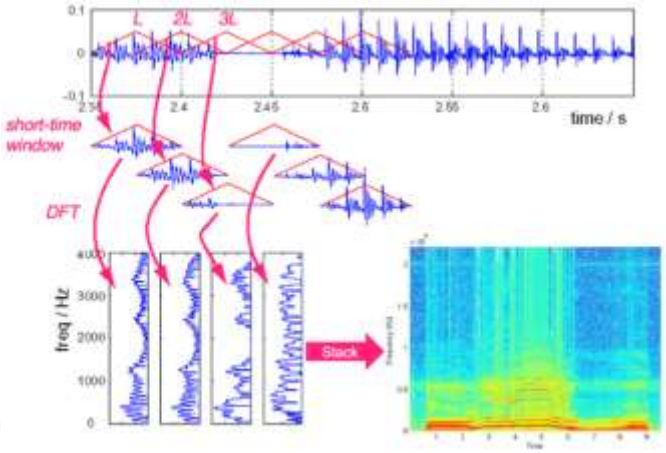
Esta descomposición de ondas, denominada “transformada de Fourier” (FT, por sus siglas en inglés), convierte señales que se encuentran en el dominio del tiempo al dominio de frecuencia; pero, para su aplicación, es necesario que la señal que se va a descomponer sea periódica e infinita en el tiempo. Para señales que son finitas en el tiempo, existe la DFT (transformada discreta de Fourier) que convierte una señal discreta a una continua. Sin embargo, utilizar la DFT requiere que la señal en cuestión sea una muestra de una señal periódica.

Mencionado lo anterior, una pista musical, no cumple con ninguno de los requisitos para aplicar la FT o la DFT, por este motivo, se opta por aplicar la transformada de Fourier de Tiempo Reducido (STFT) que divide la señal de audio en pequeños segmentos de mismo tamaño, denominados “ventanas”, a las cuales se les aplica la DFT permitiendo obtener el espectro de cada segmento, formando así, el espectrograma.

En otras palabras, un espectrograma es el resultado de aplicar la transformada de Fourier a pequeñas ventanas de una representación sonora en el dominio de forma de onda, estas ventanas van recorriendo a lo largo del tiempo para poder generar una representación en el dominio tiempo-frecuencia, tal y como se muestra en la Figura 3.2.

Figura 3.2

Aplicación de la transformada de Fourier.



Nota: Aplicación de la Transformada de Fourier para transformar una señal de audio que se encuentra en el dominio de forma de onda al dominio de tiempo-frecuencia (espectrograma). En la figura, se le es aplicada la transformada de Fourier (DFT) a pequeñas ventanas de una fuente de audio (L, 2L, 3L) que se encuentra en el dominio de forma de onda, para luego unirlas y generar una representación en el dominio de tiempo-frecuencia (espectrograma). Tomado de Manilow et al. (2020).

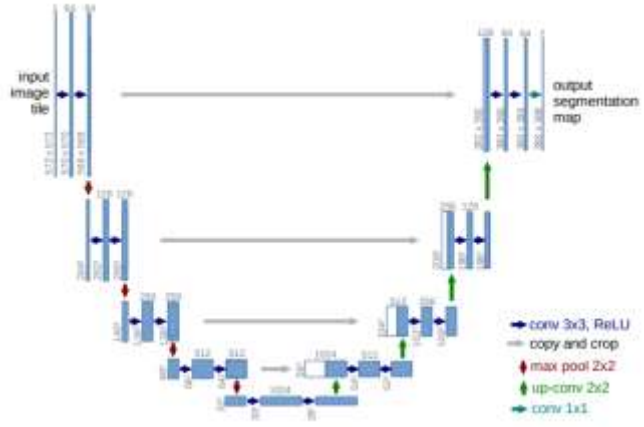
La aplicación de STFT resulta en números complejos, lo que significa que cada entrada tiene un componente de magnitud y fase. Estos dos componentes son necesarios para convertir un espectrograma, que se encuentra en el dominio de tiempo-frecuencia, al dominio de forma de onda para que así podamos oírla. Este proceso de inversión se denomina transformada de Fourier inversa de corta duración o iSTFT.

3.3 Red UNet

UNet es una arquitectura de redes neuronales convolucionales desarrollada y aplicada en la segmentación de imágenes biomédicas (Ronneberger et al., 2015). Por lo general, una red neuronal convolucional se enfoca en la clasificación de imágenes, donde la salida del procesamiento corresponde a una etiqueta; sin embargo, la arquitectura UNet, permite localizar y distinguir información asignando a cada píxel, una etiqueta.

Figura 3.3

Arquitectura UNet



Nota: Arquitectura UNet (ejemplo con una resolución mínima de 32x32 píxeles). Cada caja azul corresponde a un mapa de características. El número de canales es denotado sobre cada caja. Los tamaños x-y están indicados en la zona inferior izquierda de cada caja. Las cajas blancas representan los mapas de

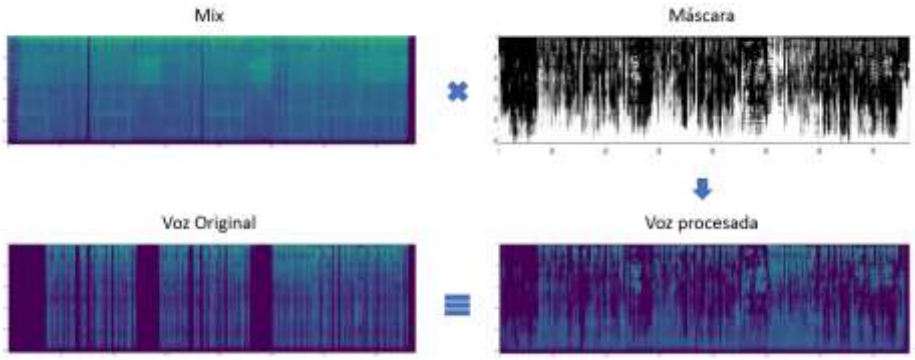
características copiados. Las flechas denotan las operaciones ejecutadas. Tomado de Ronneberger et al. (2015).

La razón por lo que esta red es llamada UNet, es porque presenta una forma de "U", como se observa en la figura 3.3. Esta arquitectura se puede dividir en 2 segmentos, el primero (izquierda) se denomina ruta de contracción (codificador) que se compone principalmente de procesos convolucionales seguidos principalmente de operaciones *ReLU* y *max pooling* Durante esta fase, la información espacial es reducida, mientras que la información de características es aumentada. El segundo segmento (derecho), denominado ruta de expansión (decodificador) combina la información espacial y de características a través de una secuencia de deconvoluciones y concatenaciones con las características obtenidas en la parte codificadora (Ronneberger et al., 2015). Estas concatenaciones de información, permite a la red localizar mapas de características con mayor precisión.

La red utilizada en este trabajo pretende generar una máscara como salida, la cual represente las señales generadas por la voz (la inversa de dicha máscara, representan las señales generadas por los instrumentos), esta máscara es brindada por la última capa decodificadora y contiene las mismas dimensiones que la señal de entrada. Por ello, la máscara obtenida se multiplica con el espectrograma del mix como se muestra en la Figura 3.4, obteniendo de esta manera la voz procesada, es decir, una fuente de audio con la voz separada de los instrumentos. Sin embargo, un espectrograma no puede ser reproducido como audio. Para ello, se le adiciona la fase, que fue previamente reservada y se le aplica la inversa de la transformada de Fourier (iSTFT), para pasar el espectrograma que se encuentra en el dominio de tiempo-frecuencia al dominio forma de onda.

Figura 3.4

Aplicación de máscara brindada por la red UNet a una fuente de audio.



Nota: Mix: Espectrograma de la canción original; Máscara: representación de las ondas vocales producidas por la red profunda; Voz procesada: Espectrograma del resultado del sistema de separación de voz; Voz original: Espectrograma de las ondas vocales de la canción original. Elaboración propia.

3.4 REPET+

El método REPET+, presentado por Rafii & Pardo (2012), es un método que separa las ondas vocales e instrumentales de una fuente de audio en base a la identificación de patrones repetitivos que son considerados como parte de las ondas instrumentales. Este método se divide en 4 etapas:

En la primera etapa, se brinda una señal de audio x como entrada, a la cual se le aplica la STFT para transformar la señal, que se encuentra en un dominio de forma de onda, a un dominio de tiempo-frecuencia (espectrograma). Como los valores de un espectrograma son valores complejos, se realiza una separación de estos en: magnitud del espectrograma (V) que contiene los valores absolutos de x, y fase (P) que contiene los valores imaginarios. Una vez, que el espectrograma V ha sido obtenido, se define la matriz de similitud S, normalizando cada *frame* de V con la normalización Euclideana, donde, la matriz de similitud es una representación bidimensional donde cada punto (a, b) mide la similitud entre dos elementos a y b de una secuencia dada de acuerdo con la similitud coseno entre los puntos dados (Rafii & Pardo, 2012); esto permitirá identificar elementos que son considerados repetitivos para *frames* j de V.

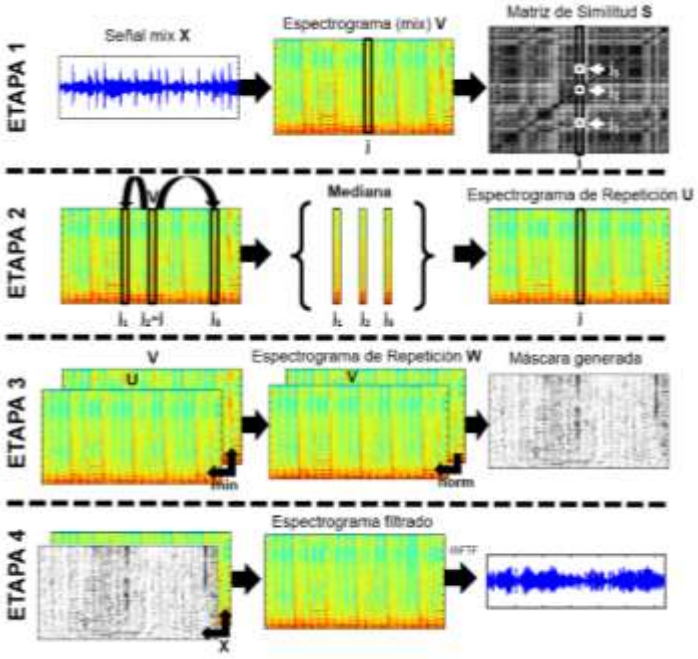
En la segunda etapa, se busca generar el espectrograma de repetición U, para esto, para cada *frame* j en V, se busca aquellos otros *frames* que son similares en V con ayuda de S, y se deriva el *frame* j en el espectrograma de repetición U, tomando la mediana de los *frames* que son considerados similares para el *frame* j de V dado, generando así, el espectrograma de repetición U. Hay que recordar que se asume que los elementos no repetitivos son parte de la voz y los repetitivos del acompañamiento.

En la tercera etapa, se busca generar la máscara que permitirá separa la voz de los instrumentos, para lo cual, se debe refinar U, tomando los valores mínimos de las magnitudes entre U y V para cada *frame*, resultando en un nuevo espectrograma de repetición W, esto es, porque los valores de salida del filtro no deben ser mayores que el input dado, ya que asumimos que las señales son aditivas (voz e instrumentos). Luego, la máscara es generada normalizando W por V para cada *frame.*

Finalmente, en la etapa 4, dicha máscara se multiplica con la señal original obteniendo así el espectrograma filtrado, se le aplica la iSTFT (haciendo uso de P) y de esta manera, el audio ya puede volver a reproducirse. Cabe resaltar, que los valores inversos de la máscara corresponden a secuencias no repetitivas, es decir, la voz. La metodología mencionada de REPET+ se encuentra ilustrada en la Figura 3.5.

Figura 3.5

Metodología REPET+.



Nota: Proceso de obtención del espectrograma de repetición W a partir de un espectrograma de un mix V, por medio de una matriz de similitud S. Adaptado de Raffi & Pardo (2012).

3.5 Métricas de evaluación

Los sistemas de separación de fuentes de sonido pueden ser evaluados de manera objetiva o subjetiva. Las evaluaciones objetivas miden la calidad de la separación mediante la realización de cálculos matemáticos que comparan la salida de un sistema de separación de fuentes de sonido y un resultado esperado; por otra parte, las medidas subjetivas implican que evaluadores humanos puntúen la salida del sistema de separación de fuentes de sonido de acuerdo con su propia percepción.

Ambos tipos de métricas tienen ventajas y desventajas, por ejemplo, las medidas objetivas luchan porque hay aspectos de la percepción humana que para medios computacionales es difícil de capturar. Sin embargo, en comparación con las medidas subjetivas, son mucho más rápidas y económicas

de obtener. Por otra parte, las medidas subjetivas consumen más tiempo que las objetivas, y están sujetas a la variabilidad de los evaluadores, pero pueden ser más confiables que las objetivas.

3.5.1 Métricas de evaluación objetivas

Los sistemas de separación de fuentes de sonido son evaluados bajo las métricas objetivas propuestas por Vincent et al. (2006). Este tipo de evaluaciones miden la calidad del sistema realizando cálculos que comparan las señales de salida (resultado) con las señales verdaderas de la fuente separada.

SAR (Source-to-Artifact Ratio), SIR (Source-to-Interference Ratio) y SDR (Source-to-Distortion Ratio) son las métricas más utilizadas para medir el rendimiento de un sistema de separación de fuentes de sonido. Una fuente estimada $\hat{S}_i$, se asume que está compuesta por 4 componentes separados:

$$\hat{S}_i = S_{target} + e_{interf} + e_{noise} + e_{artif} \quad (1)$$

donde s_{target} es la señal verdadera de la fuente separada, es decir, la que se quiere obtener (Ejm. Voz), e_{interf} es el error en términos de interferencia, e_{noise} es el error en términos de ruido, y e_{artif} es el error en términos de artefactos agregados.

Dado estos cuatro términos, se pueden definir las métricas de evaluación en términos de decibeles (dB), donde mayores valores significan un mejor resultado. Tomar en cuenta que es necesario poseer la señal objetiva.

La métrica SAR indica la cantidad de artefactos no deseados que tiene una fuente estimada con relación a una fuente verdadera y se define como:

$$SAR := 10\log_{10} \frac{||s_{target} + e_{interf} + e_{noise}||^2}{||e_{artif}||^2} \quad (2)$$

Por otro lado, la métrica SIR, indica la cantidad de sonido proveniente de otras fuentes que se pueden escuchar en una fuente estimada, definido por:

$$SIR := 10\log_{10} \frac{||s_{target}||^2}{||e_{interf}||^2} \quad (3)$$

Por último, la métrica SDR, se considera como la medida general de un sistema de separación de fuentes de sonido, indica qué tan bueno suena la fuente estimada.

$$SDR := 10\log_{10} \frac{||s_{target}||^2}{||e_{interf} + e_{artif} + e_{noise}||^2} \quad (4)$$

A la fecha, el mejor SDR reportado ha sido obtenido por el sistema de separación de voz de Takahashi & Mitsufuji (2020), con un valor de 7.80 dB.

3.5.2 Métricas de evaluación subjetivas

Las métricas de evaluación subjetivas se basan prácticamente en que un grupo de personas calificadas evalúen la calidad del resultado de un sistema de separación de fuentes de sonido. Por ejemplo, para este tipo de sistemas, se recomienda utilizar las pruebas de MUSHRA (*Multiple Stimuli with Hidden Reference and Anchor*), donde ingenieros de audio califican el resultado de un sistema de separación de fuentes de sonido (Series, B, 2014). En este tipo de pruebas, se le presenta al oyente la pista de referencia (indicándola como tal) y varias otra pistas de audio, siendo una de estas el resultado del sistema y las restantes modificaciones de la pista original, para que puntúen la calidad del sonido respecto a la referencia con una escala del 1 al 100. Si bien pueden ser más confiables que las técnicas objetivas, rara vez se usan debido a que son lentas y costosas de realizar.

En la literatura, es raro encontrar investigaciones con evaluaciones subjetivas, es por ello, que en el presente trabajo se medirán los métodos presentados con métricas de evaluación objetivas, debido a que se compararán los resultados obtenidos con otros resultados de la literatura.

4. METODOLOGÍA Y EXPERIMENTACIÓN

Se desea probar la efectividad de la utilización de los métodos UNet y REPET+ evaluándolos de manera independiente y de manera conjunta. Como se mencionó en secciones anteriores, se hará uso del dataset

DSD100 (Liutkus et al., 2017). Este dataset contiene 100 pistas de música de diferentes géneros musicales como rap, pop/rock, country, heavy metal, electrónica, entre otros, en formato estéreo con muestreo de 44100Hz. Cada pista musical contiene, la canción completa (mix), el componente vocal y los instrumentales (bajo, percusión y otros) de forma aislada en diferentes archivos de sonido. Con el fin de aumentar la cantidad de muestras para entrenar el modelo, se generaron canciones aleatorias, tomando los componentes vocales e instrumentales de diferentes canciones y uniéndolas, creando un nuevo mix, obteniendo un total de 140 canciones. De esta totalidad de canciones se segmentó el 80% en canciones para entrenamiento y el resto como canciones para pruebas. La tabla 4.1 muestra la distribución porcentual de los géneros musicales del dataset utilizado.

Tabla 4.1

Distribución porcentual de los géneros musicales presentes en el dataset DSD100.

Género	Porcentaje (%)
Pop/Rock	71.0%
Rap	6.0%
Country	2.0%
Heavy Metal	11.0%
Electronic	6.0%
Jazz	2.0%
Reggae	2.0%

4.1 Tecnologías utilizadas

El desarrollo de la experimentación se llevó a cabo en la plataforma de Google, *Google Colaboratory* haciendo uso del lenguaje Python 3.7.

Las librerías utilizadas fueron las siguientes:

- *Librosa*: librería que contiene funciones para analizar pistas de música y audio.
- *Numpy*: librería que contiene funciones para crear y operar con vectores y matrices multidimensionales.
- *Chainer*: *framework* para diseñar y entrenar redes neuronales.
- *Mir_*eval: librería desarrollado y presentado por Raffel et al. (2014), la cual contiene funciones para la evaluación de técnicas MIR (Music Information Retrieval).
- *Matplotlib*: librería que permite la generación de gráficos a partir de información contenida en listas.

4.2 Preparación de la data de entrenamiento

Primero, cada fuente de audio se remuestrea a una tasa de 22050Hz, con la finalidad de mejorar la velocidad de procesamiento de la red. Luego, a cada una de las muestras se le aplica la transformada de Fourier de Tiempo Reducido (STFT) con una ventada de 1024 *frames* y un salto de 512. De esta manera, se pasa la muestra que se encuentra en el dominio de forma de onda, al dominio de tiempo-frecuencia. La aplicación de esta transformada brinda un espectrograma con valores complejos, por lo que se descompondrá en su magnitud (real) y fase (complejo), siendo la magnitud con la que se entrenará la red. La fase del espectrograma se reservará con el fin de reconstruir la señal a su dominio de forma de onda.

4.3 Entrenamiento de la red

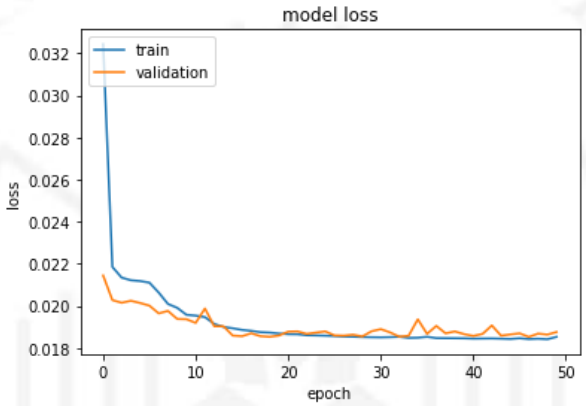
Una vez que la data está preparada, se entrenará la red UNet con la data designada para el entrenamiento; y cuya configuración está dada por una parte codificadora que está compuesta de 6 bloques, donde cada bloque contiene una capa de convolución 2D, con k*ernel* de tamaño 5, *stride* de 2, normalización de lotes

y función ReLU como función de activación. Por otro lado, la parte decodificadora, también compuesta de 6 bloques, contiene por cada uno de ellos una capa de deconvolución 2D, *kernel* de tamaño 5, *stride* de 2, normalización de lotes y función ReLU como función de activación; 50% de *dropout* para los 3 primeros bloques y, el último bloque presenta una función sigmoidal como salida. La red se entrena con el optimizador Adam a través de 50 épocas. Los parámetros de configuración fueron tomados de la investigación Jansson et al. (2017).

El entrenamiento se realizó en plataforma de Google Colaboratory el cuál brinda una GPU NVIDIA Tesla T4 de dos núcleos de 1.59GHz cada uno y 16GB de memoria y el tiempo de ejecución fue de un aproximado de 2 horas con 10 minutos. En la figura 4.1 se muestra las curvas de pérdida generadas durante el entrenamiento del modelo, para los conjuntos de datos de entrenamiento y validación (20% de validación), lo cual indica que el modelo, con los parámetros y data proporcionados, está aprendiendo de una manera correcta, sin llegar al overfitting (sobre-entrenamiento), debido a que ambas curvas, convergen entre sí y permanecen estables durante las épocas de entrenamiento.

Figura 4.1

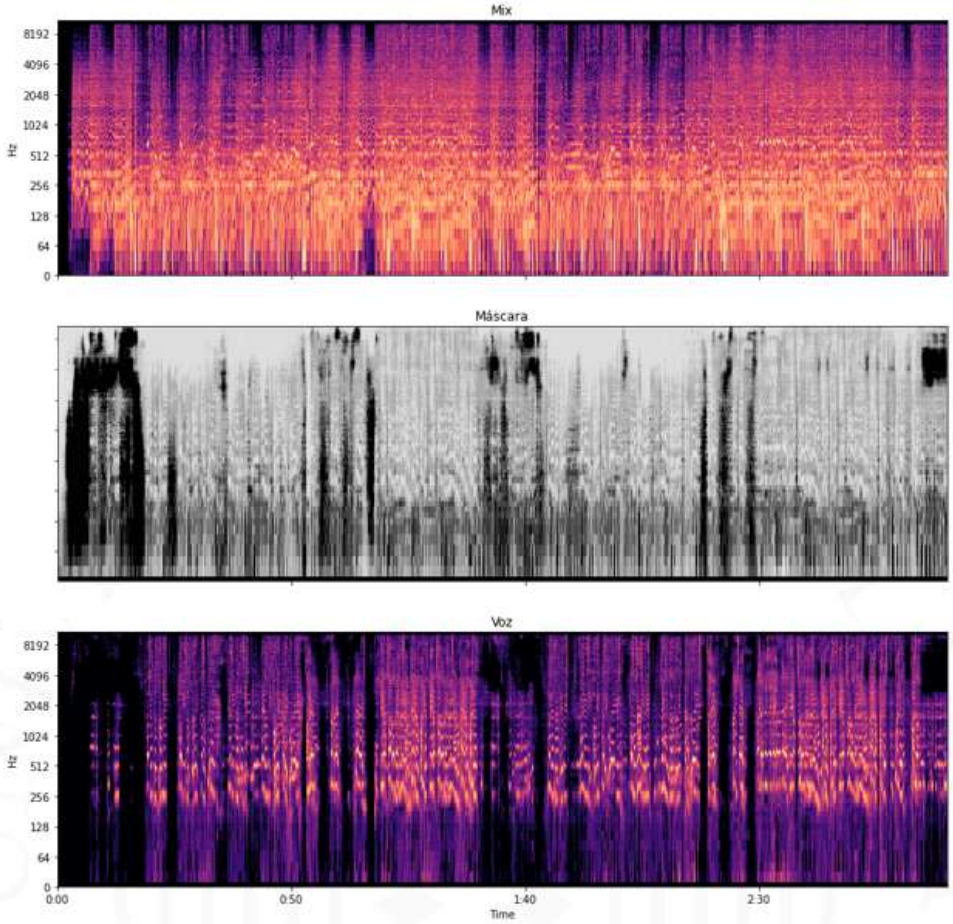
Curvas de pérdida de los conjuntos de entrenamiento y validación.



Asimismo, en la figura 4.2 se muestra el espectrograma de una pista de audio, la máscara generada por el modelo UNet, y finalmente, el espectrograma de la voz cantada.

Figura 4.2

Espectrogramas de una pista de audio, máscara generada por UNet y voz cantada.



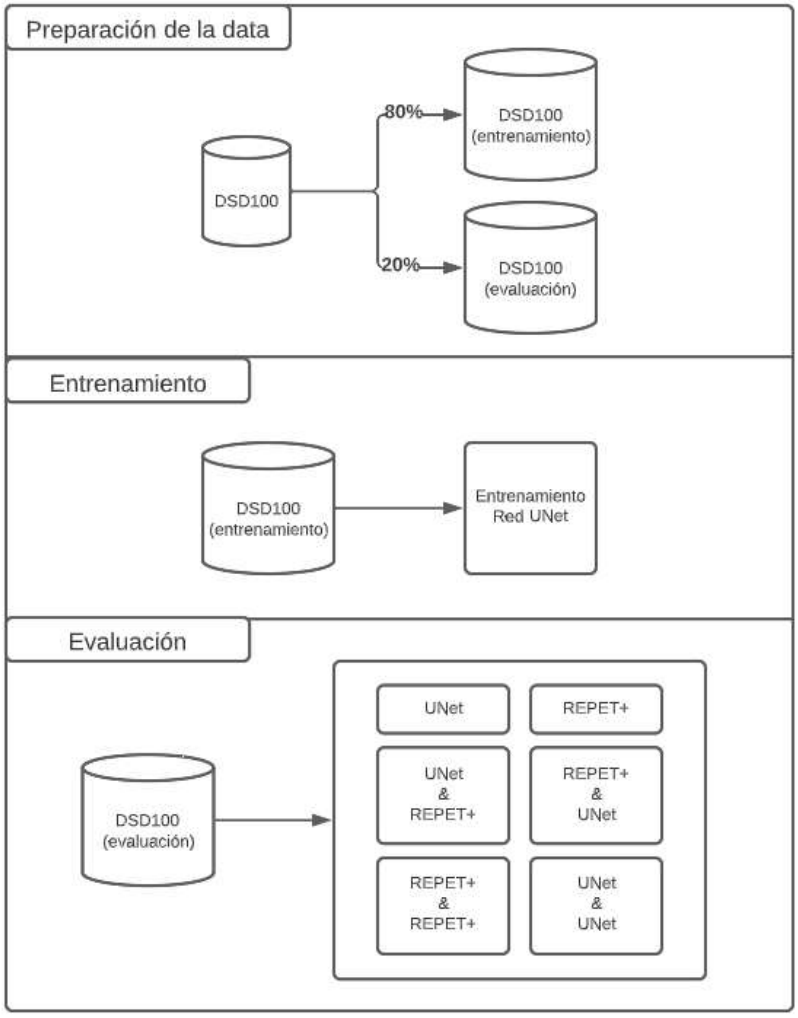
El método REPET+ no necesita entrenamiento ya que este se basa en métodos estadísticos para poder segmentar los sonidos vocales de los instrumentales por medio de la identificación de patrones repetitivos.

4.4 Evaluación

Debido a que el objetivo del trabajo de investigación es medir si la utilización de un método no supervisado (REPET+) junto a uno supervisado (UNet) presenta mejoras en cuanto a sus resultados obtenidos de forma individual, se evaluarán primeramente REPET+ y UNet de manera independiente. Posteriormente, los resultados obtenidos de REPET+ serán procesados, nuevamente, a través del modelo de UNet, a este proceso se le denominará REPET+ & UNet; los resultados de UNet serán procesados con REPET+, este denominado, UNet & REPET+; de la misma manera, los resultados de REPET+ serán procesados nuevamente por el mismo método, el cual se denomina como REPET+ & REPET+; y, por último, los resultados de UNet serán procesados nuevamente por UNet, denominado UNet & UNet. Igualmente, los resultados que se obtengan serán comparados con otros métodos encontrados en la literatura. Un resumen de la metodología puede observarse en la figura 4.3.

Figura 4.3

Metodología planteada.



Nota: Metodología planteada. Se divide en 3 pasos: preparación de la data, en la cual se segmenta en data para entrenamiento y evaluación; entrenamiento, donde se entrena la red UNet, REPET+ no necesita entrenamiento ya que se basa en métodos estadísticos; evaluación, en la cual se evalúan los métodos UNet y REPET+ de forma independiente y conjunta.

5. RESULTADOS

Para poder obtener los puntajes de evaluación para los métodos estudiados, se ha utilizado la librería de Python *mir_eval* desarrollado y presentado por Raffel et al. (2014), el cual es un estándar para la evaluación de técnicas MIR (Music Information Retrieval).

Las métricas SDR, SIR y SAR fueron calculadas para cada uno de los casos propuestos (UNet, REPET+, UNet & REPET+, REPET+ & UNet, REPET+ & REPET+ y UNet & UNet) y, cada una de las fuentes de sonido (voz e instrumentos) pertenecientes al conjunto de datos designados para pruebas. A continuación, se mostrarán los resultados obtenidos para los métodos estudiados (a mayor valor, mejores resultados), dichos valores están medidos en decibeles (dB), donde, a mayor corresponde un mejor desempeño.

La tabla 5.1 muestra la métrica SAR la cual indica el nivel de artefactos no deseados que tiene una fuente estimada con relación a una fuente verdadera (a mayor valor, menos artefactos no deseados). Se puede observar que la red UNet presenta un mejor valor frente a los otros métodos.

Tabla 5.1

Resultados de la métrica SAR para la separación Vocal.

Modelo	Media	Desv. Est.	Min	Max	Mediana
REPET+	-0.1	2.27	-5.32	3.17	0.51
UNet	**7.3**	5.84	-15.47	14.49	8.63
REPET+ & UNet	-0.1	2.27	-5.32	3.17	0.51
UNet & REPET+	0.71	4.59	-16.18	6.17	1.62
REPET+ & REPET+	-3,21	3.02	-9.9	1.23	-2.36
UNet & UNet	6.77	6.38	-17.76	14.41	8.06

Por otra parte, la tabla 5.2, indica la métrica SIR, la cual mide el nivel de sonidos proveniente de otras fuentes, en este caso, provenientes de los instrumentos. A diferencia de las tablas anteriores, un mayor valor lo obtiene el método UNet utilizado de forma conjunta con REPET+, en ese orden, con un valor de 17.51 dB.

Tabla 5.2

Resultados de la métrica SIR para la separación Vocal

Modelo	Media	Desv. Est.	Min	Max	Mediana
REPET+	3.51	9.47	-31.91	14.17	5.77
UNet	14.21	8.43	-18.86	22.08	16.22
REPET+ & UNet	9.1	10.13	-30.34	18.85	11.15
UNet & REPET+	**17.51**	8.96	-18.15	24.28	19.57
REPET+ & REPET+	5.19	8.85	-26.41	15.91	6.74
UNet & UNet	15.18	8.1	-16.13	25.51	16.45

En la tabla 5.3 se muestra los resultados de las métricas SDR para la separación de la voz. Cabe recordar que la métrica SDR es una medida general de qué tan bueno es el resultado brindado por un modelo dado frente a una fuente de audio objetiva. Como se puede observar, el método UNet presenta un mayor valor (5.38 dB) frente a los otros métodos (el mayor valor SDR obtenido hasta la fecha ha sido obtenido por el sistema de separación de voz de (Takahashi & Mitsufuji, 2020), con un valor de 7.80 dB).

Tabla 5.3

Resultados de la métrica SDR para la separación Vocal

Modelo	Media	Desv. Est.	Min	Max	Mediana
REPET+	-4.3	7.98	-36.45	2.16	-1.74
UNet	**5.38**	8.43	-18.86	22.08	16.22
REPET+ & UNet	-2.55	8.24	-36.78	2.76	-0.35
UNet & REPET+	-0.38	8.22	-34.49	6.06	1.53
REPET+ & REPET+	-6.16	7.6	-33.98	0.98	-3.64
UNet & UNet	5.17	9.44	-34.07	13.13	6.97

La tabla 5.4 presenta los resultados SAR para la separación de sonidos instrumentales, donde la red UNet utilizado de forma individual, presenta por un amplio margen, un mejor resultado (17.7 dB).

Tabla 5.4

Resultados de la métrica SAR para la separación Instrumental.

Modelo	Media	Desv. Est.	Min	Max	Mediana
REPET+	7.43	2.41	1.96	10.93	7.77
UNet	**17.7**	7.94	10.52	49.81	17.24
REPET+ & UNet	-5.08	2.6	-11.41	-0.55	-5.15
UNet & REPET+	-0.27	3.8	-13.41	5.37	0.46
REPET+ & REPET+	5.82	4.34	-10.75	10.09	6.35
UNet & UNet	16.95	5.92	10.28	39.19	16.96

La tabla 5.5 muestra los resultados SIR obtenidos para la separación de sonidos instrumentales, obteniendo, nuevamente, un mejor valor para la red UNet (25.82 dB). La utilización de los métodos UNet y REPET+ usados de forma conjunta y en ese orden obtienen un valor muy por debajo de los otros métodos (-11.57 dB).

Tabla 5.5

Resultados de la métrica SIR para la separación Instrumental

Modelo	Media	Desv. Est.	Min	Max	Mediana
REPET+	15.38	9.41	0.49	47.38	15.89
UNet	**25.82**	13.57	12.06	81.93	22.35
REPET+ & UNet	21.94	7.31	5.91	34.62	21.85
UNet & REPET+	-11.57	8.18	-20.62	19.11	-14.6
REPET+ & REPET+	17.93	8.31	7.2	47.17	17.95
UNet & UNet	25.27	10.52	12.11	67.15	22.49

Finalmente, la tabla 5.6 muestra los resultados de la métrica SDR, siendo la red UNet la que obtiene un mayor valor frente a los otros métodos (16.67 dB).

Tabla 5.6

Resultados de la métrica SDR para la separación Instrumental

Modelo	Media	Desv. Est.	Min	Max	Mediana
REPET+	5.91	3.36	-3.22	10.92	6.03
UNet	**16.67**	8.18	9.35	49.81	16.33
REPET+ & UNet	-5.21	2.64	-11.61	-0.56	-5.24
UNet & REPET+	-16.17	3.96	-22.72	-9.45	-17.22
REPET+ & REPET+	5.14	4.42	-11.12	10.08	5.56
UNet & UNet	15.99	6.12	9.33	39.19	16.23

Posteriormente, en la tabla 5.7 se muestran los resultados obtenidos por otros métodos de la literatura (solamente la métrica SDR, el cual es la medida estándar con la cual se comparan los diferentes métodos), por ejemplo, el modelo D3Net (Takahashi & Mirsufuji, 2020), el cual tiene el puntaje SDR más alto hasta la fecha para la separación de la voz cantada, Spleeter (Hennequin et al., 2020), el cual hace uso de la combinación de varios modelos pre entrenados, y Wavenet (Lluís et al., 2018) cuyo modelo fue entrenado en el dominio de forma de onda.

Tabla 5.7

Resultados SDR comparados con otros métodos de la literatura.

Modelo	Voz	Instrumentos
REPET+	-4.3	5.91
UNet	5.38	**16.67**
REPET+ & UNet	-2.55	-5.21
UNet & REPET+	-0.38	-16.17
REPET+ & REPET+	-6.16	5.14
UNet & UNet	5.17	15.99
D3Net	**7.80**	14.26
Spleeter	6.68	12.54
Wavenet	3.67	10.64

Nota: Promedio de los resultados de la métrica SDR para los métodos utilizados en el presente trabajo (REPET+, UNet, REPET+ & UNet, UNet + REPET+, REPET+ & REPET+, UNet & UNet) y los métodos D3Net (Takahashi & Mirsufuji, 2020), Spleeter (Hennequin et al., 2020) y Wavenet (Lluís et al., 2018)

6. DISCUSIÓN DE RESULTADOS

Como se pudo observar en las tablas anteriores, la red UNet, utilizada de manera individual, presenta, por lo general, en todas las métricas mejores resultados frente a los otros métodos (REPET+, REPET+ & UNet, UNet & REPET+, REPET+ & REPET+, UNet & UNet). A continuación, analizaremos cada una de las métricas y su rendimiento para cada uno de los métodos utilizados.

Durante el proceso de separación de fuentes de sonido, los resultados pueden presentar sonidos que no existían en la fuente de sonido original debido a las transformaciones que se realizan durante dicho proceso. Estos tipos de sonidos son denominados “artefactos”. La métrica SAR, indica el nivel de artefactos no deseados en una fuente estimada. De todos los métodos utilizados (REPET+, UNet, REPET+ & UNet, UNet & REPET+, REPET+ & REPET+, UNet & UNet), UNet fue el que obtuvo un mejor rendimiento frente a los otros, con un puntaje promedio de 7.3 dB (Tabla 5.1) para la separación vocal y 17.7 dB (Tabla 5.4) para la separación instrumental, presentando un menor nivel de artefactos. Esto se debe a su naturaleza de aprendizaje profundo, cuyo modelo permite la generación de una máscara en base a un análisis profundo de los datos brindados, a diferencia del método REPET+ la separación de sonidos en base a la identificación de patrones repetitivos a través métodos estadísticos, lo cual ocasiona que se genere una máscara de menor calidad que conlleva a un peor rendimiento. Respecto a los métodos usados en conjunto para la separación vocal, REPET+ & UNet, UNet & REPET+ y REPET+ & REPET+, presentan valores bajos debido a que cada fuente de audio se procesa dos veces, lo cual conlleva que, en el segundo proceso de separación, se trabaje sobre espectrogramas con data que ya contienen sonidos no esperados. Sin embargo, esto no sucede en la aplicación de UNet & UNet obteniendo un puntaje promedio de 6.77 dB para la separación vocal y 16.95 dB para la separación instrumental, relativamente cerca al puntaje obtenido por UNet (7.3 dB y 16.95 dB, respectivamente), esto sucede porque en el segundo procesamiento de la pista de audio, la señal ya se encuentra filtrada, por lo que solo sufre alteraciones que conllevan a la generación de artefactos, este mismo comportamiento se replica en REPET+ & REPET+ para la separación instrumental donde obtiene un puntaje de 5.82 dB.

En las tablas 5.2 y 5.5 se muestran los resultados de la métrica SIR para todos los métodos utilizados. Esta métrica indica el nivel de sonidos pertenecientes a otras fuentes en la fuente de audio estimada y/o el nivel de sonidos perdidos para una fuente estimada. Para la separación vocal, el puntaje más alto obtenido fue de 17.51 dB para el método UNet & REPET+, siendo una mejora para cuando se utilizó el método UNet de forma individual. De la misma manera, el resultado del método REPET+ (3.51 dB) mejoró cuando se utilizó de forma conjunta con el método UNet (REPET+ & UNet). Esto se debe a que se realiza un doble filtrado de sonidos a la fuente de audio que se brinda como entrada, primero, removiendo patrones repetitivos y luego, realizando un filtrado con el modelo de aprendizaje profundo. Sin embargo, esta mejora de rendimiento no se refleja para la separación instrumental (Tabla 5.5). En este caso, el mejor resultado lo obtuvo la red UNet con un valor de 25.82 dB utilizado de forma individual. REPET+ obtuvo una mejora cuando se le adicionó a su resultado el procesamiento con la red UNet (de 15.38 dB – REPET+ a 21.94 dB – REPET+ & UNet), esto se da porque, UNet, toma el resultado brindado por REPET+, y sobre este realiza un filtrado de sonidos lo cual conlleva a una mejora de resultados, sin embargo, UNet & REPET+ presenta un decrecimiento en su rendimiento (de 25.82 dB – UNet a -11.57 dB – UNet & REPET+), debido a que REPET+ no realiza un filtrado de sonidos sobre el resultado brindado por la red UNet, sino que busca patrones repetitivos para posteriormente eliminarlos, y como dicho resultado de la red UNet presenta sonidos instrumentales, los cuales, por lo general son repetitivos, son eliminados conllevando a una pérdida de información. Respecto al método REPET+ & REPET+, este presenta mejoras en ambas separaciones (vocal e instrumental) ya que se realiza un doble filtrado a la señal, recabando en una señal con patrones repetitivos mejor aislados. Por otra parte, el resultado del método UNet & UNet es muy similar comparado con UNet, por lo que demuestra que el modelo, en el segundo procesamiento no realiza muchas transformaciones a la señal debido a que esta ya se encuentra filtrada y el modelo ya la detecta como correcta.

Finalmente, las tablas 5.3 y 5.6 muestran los resultados SDR para la separación vocal e instrumental, respectivamente, los cuales indican la medida general de qué tan bueno es el resultado de la fuente estimada para cada uno de los métodos presentados. UNet ha presentado un resultado muy por encima los otros métodos (5.38 dB – voz y 16.67 dB - instrumentos), debido a que la máscara procesada es multiplicada sobre la fuente de audio original, lo cual permite una menor pérdida de información y un nivel menor de distorsión. Los otros métodos presentados reflejan un bajo rendimiento para la separación de fuentes de sonido. Si comparamos los resultados obtenidos con otros métodos de la literatura (tabla 5.7), observamos que los mejores resultados de los métodos utilizados en este trabajo es de la red UNet utilizada de manera aislada, otros métodos como D3Net (Takahashi & Mirsufuji, 2020), y Spleeter (Hennequin et al., 2020), obtuvieron valores más altos 7.80 dB y 6.68 dB, respectivamente, una de las razones es que utilizaron un largo banco de datos para el entrenamiento, 1500 canciones con un aproximado de 93 horas de reproducción continua para D3Net y 24097 canciones con un aproximado de 79 horas de reproducción continua para spleeter. WaveNet (Lluís et al., 2018) obtuvo un SDR de 3.67 dB; a diferencia de los métodos anteriormente mencionados, este trabaja en el dominio de forma de onda, hoy en día, los métodos que trabajan en este dominio son pocos y se encuentran en investigación, por ahora presentan peores resultados de los que trabajan en el dominio de tiempo-frecuencia.

7. CONCLUSIONES

En el presente trabajo se ha realizado un análisis comparativo de la utilización de los métodos REPET+, el cual es un método tradicional que utiliza técnicas estadísticas para la identificación y separación de segmentos repetitivos de una canción, y UNet, el cual es un método supervisado basado en el aprendizaje profundo que utiliza la arquitectura UNet, la cual fue desarrollada inicialmente para la segmentación de imágenes, utilizándolos aisladamente y en conjunto para medir en qué medida mejoran o empeoran sus valores de métricas de evaluación.

Durante los experimentos realizados, se ha demostrado que los métodos REPET+ y UNet, han presentado un mejor resultado cuando estos son utilizados de forma individual. Si bien, algunas de las métricas de los resultados mejoraron cuando fueron utilizados de forma conjunta, no significaba que presentaran un mejor desempeño, sino que al mejorar algunas, otras empeoraban. En general, estos métodos presentaron un mejor performance siendo utilizados de forma individual, sobre todo la red UNet, por lo tanto, se concluye que los métodos supervisados, basados en el aprendizaje profundo, son superiores frente a los métodos no supervisados que se basan en algoritmos estadísticos. Sin embargo, un inconveniente de los métodos supervisados es la poca disponibilidad de data a disposición del público. Si bien existen algunos bancos de datos disponibles, estos no se comparan con la cantidad de data utilizada por otros trabajos de investigación, como se mencionó anteriormente, en los métodos D3Net y Spleeter que utilizaron un dataset de 93 y 79 horas de música respectivamente.

8. REFERENCIAS

Bracewell, R. N. (1989). The fourier transform. Scientific American, 260(6), 86-95.

Bregman, A. S. (1994). Auditory scene analysis: The perceptual organization of sound. MIT press.

Chan, T. S., Yeh, T. C., Fan, Z. C., Chen, H. W., Su, L., Yang, Y. H., & Jang, R. (2015, April). Vocal activity informed singing voice separation with the iKala dataset. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 718-722). IEEE.

Grais, E. M., Sen, M. U., & Erdogan, H. (2014, May). Deep neural networks for single channel source separation. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3734-3738). IEEE.

Heittola, T., Klapuri, A., & Virtanen, T. (2009, October). Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In ISMIR (pp. 327-332).

Hennequin, R., Khlif, A., Voituret, F., & Moussallam, M. (2020). Spleeter: a fast and efficient music source separation tool with pre-trained models. Journal of Open Source Software, 5(50), 2154.

Hsu, C. L., & Jang, J. S. R. (2009). On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. IEEE transactions on audio, speech, and language processing, 18(2), 310-319.

Huang, P. S., Kim, M., Hasegawa-Johnson, M., & Smaragdis, P. (2014, October). Singing-Voice Separation from Monaural Recordings using Deep Recurrent Neural Networks. In ISMIR (pp. 477-482).

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., & Weyde, T. (2017). Singing voice separation with deep u-net convolutional networks.

Li, Y., & Wang, D. (2007). Separation of singing voice from music accompaniment for monaural recordings. IEEE Transactions on Audio, Speech, and Language Processing, 15(4), 1475-1487.

Liutkus, A., Stöter, F. R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., ... & Fontecave, J. (2017, February). The 2016 signal separation evaluation campaign. In International conference on latent variable analysis and signal separation (pp. 323-332). Springer, Cham.

Lluís, F., Pons, J., & Serra, X. (2018). End-to-end music source separation: is it possible in the waveform domain? arXiv preprint arXiv:1810.12187.

Manilow, E., Seetharman, P., & Salamon, J. (2020). Open Source Tools & Data for Music Source Separation. Opgehaal van https://source-separation.github.io/tutorial

Ozerov, A., Philippe, P., Bimbot, F., & Gribonval, R. (2007). Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs. IEEE Transactions on Audio, Speech, and Language Processing, 15(5), 1564-1578.

Plumbley, M. D., Abdallah, S. A., Bello, J. P., Davies, M. E., Monti, G., & Sandler, M. B. (2002). Automatic music transcription and audio source separation. Cybernetics &Systems, 33(6), 603-627.

Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "mir_eval: A Transparent Implementation of Common MIR Metrics", Proceedings of the 15th International Conference on Music Information Retrieval, 2014.

Rafii, Z., & Pardo, B. (2012, October). Music/Voice Separation Using the Similarity Matrix. In ISMIR (pp. 583-588).

Rafii, Z., Liutkus, A., Stöter, F. R., Mimilakis, S. I., & Bittner, R. (2017). MUSDB18-a corpus for music separation.

Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

Series, B. (2014). Method for the subjective assessment of intermediate quality level of audio systems. International Telecommunication Union Radiocommunication Assembly.

Sharma, B., Das, R. K., & Li, H. (2019). On the Importance of Audio-Source Separation for Singer Identification in Polyphonic Music. In INTERSPEECH (pp. 2020-2024).

Simpson, A. J., Roma, G., & Plumbley, M. D. (2015, August). Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In International Conference on Latent Variable Analysis and Signal Separation (pp. 429-436). Springer, Cham.

Smith, J. O. (2011). Spectral audio signal processing. W3K.

Stoller, D., Ewert, S., & Dixon, S. (2018). Wave-u-net: A multi-scale neural network for end-to-end audio source separation. arXiv preprint arXiv:1806.03185.

Takahashi, N., & Mitsufuji, Y. (2020). D3net: Densely connected multidilated densenet for music source separation. arXiv preprint arXiv:2010.01733.

Vembu, S., & Baumann, S. (2005, September). Separation of Vocals from Polyphonic Audio Recordings. In ISMIR (pp. 337-344).

Vincent, E., Gribonval, R., & Févotte, C. (2006). Performance measurement in blind audio source separation. IEEE transactions on audio, speech, and language processing, 14(4), 1462-1469.

Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE transactions on audio, speech, and language processing, 15(3), 1066-1074.

Tesis

INFORME DE ORIGINALIDAD

16%	15%	10%	11%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	Submitted to Universidad de Lima Trabajo del estudiante	1%
2	www.springerprofessional.de Fuente de Internet	1%
3	Berrak Ozturk Simsek, Aydin Akan. "Audio Melody Extraction from Monophonic Turkish Maqam Music", 2020 28th Signal Processing and Communications Applications Conference (SIU), 2020 Publicación	1%
4	zaguan.unizar.es Fuente de Internet	1%
5	link.springer.com Fuente de Internet	1%
6	dorienherremans.com Fuente de Internet	1%
7	export.arxiv.org Fuente de Internet	1%

arxiv.org