Separación de fuentes en pistas de audio con múltiples instrumentos

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 la factorización de matrices no negativas. El mismo será presentado como un informe final de proyecto para la materia "Procesamiento Digital de Señales" de la carrera Ingeniería Informática en la Facultad UNL-FICH.**

Este documento tiene como punto de partida el artículo "Audio Source Separation Using Non-Negative Matrix Factorization (NMF)" y a partir del mismo se aplican los métodos pertinentes para lograr la separación. Finalmente se realiza el análisis de los resultados obtenidos y se comprueba la efectividad del método mediante métricas de calidad.

***Palabras clave*— factorización no negativa de matrices, procesamiento digital de señales de audio, separación de fuentes.**

I. Introducción

El sonido se transmite por el aire mediante ondas sonoras. Estas ondas o señales pueden describirse mediante tres características principales que no se aprecian como fenómenos inmutables sino en base a la percepción del oyente, estas son: intensidad (amplitud de la onda), tono (frecuencia de la onda) y timbre (forma de la onda). En este trabajo se presta especial atención a los instrumentos musicales los cuales pueden producir ondas sonoras con distintas características como las ya mencionadas. En base a las mismas se buscará separar una canción compuesta por dos instrumentos en pistas que contengan únicamente la información de cada uno, basándonos en las diferentes frecuencias de los mismos.

El obtener pistas de audios individuales, puede ayudar, por ejemplo, como ayuda para aprender a tocar la melodía de algunas canciones, escuchando solamente el instrumento de interés.

Para lograr este objetivo, se realiza un análisis tiempo-frecuencia de la señal de la pista mezclada y una vez obtenida la matriz se implementa el método de factorización no-negativa de matrices. Este método separa la señal en dos nuevas matrices las cuales, al multiplicarlas entre sí, vuelven a formar la matriz original. Se parte de la hipótesis de que estas nuevas matrices representan el análisis tiempo-frecuencia de cada instrumento y que, realizando el procedimiento inverso, se puede reconstruir la señal del instrumento separado.

Para realizar el análisis pertinente de los resultados se cuenta desde un principio con las señales de los instrumentos aislados y se consigue la señal mezclada a partir de la superposición de las mismas. Finalmente, se realizan diferentes pruebas para corroborar la eficacia del método comparando las señales iniciales con las conseguidas mediante el método de NMF.

II. Material y métodos

Como se mencionó en la introducción, se utilizará el método de factorización no-negativa de matrices para realizar la separación de los instrumentos presentes en una misma señal de audio.

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

Al tener la señal digitalizada, lo primero que se debe hacer es pasar la misma de stereo a mono para que así la señal esté representada por vectores numéricos.

Luego, se realiza la Transformada de Tiempo Corto de Fourier (STFT) para así obtener una representación en tiempo y frecuencia. Para poder aplicar el método de NMF debemos trabajar con 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.

A. Transformada de Tiempo Corto de Fourier

La Transformada de Fourier de Tiempo Corto (STFT), permite realizar un análisis de la frecuencia de la señal a través del tiempo.

Dada una ventana simétrica $g(t) = g(-t)$, se la desplaza y modula con una frecuencia ξ obteniendo así un "átomo tiempo-frecuencia"

$$g_{u,\xi}(t) = e^{i\xi t} g(t - u) \qquad (1)$$

Suponiendo que dicha ventana ha sido normalizada, entonces se define a la STFT como el producto interno entre la señal $f(t)$ y el átomo tiempo-frecuencia. Es decir:

$$\int_{-\infty}^{\infty} f(t) \cdot g(t - u) \cdot e^{-i\xi t} dt \qquad (2)$$

De esta forma, al aplicar la STFT se obtiene una matriz que contiene información de la frecuencia de la señal a lo largo del tiempo.

B. Matrices no Negativas (NMF)

La factorización en matrices no negativas es una técnica de descomposición de matrices. La misma separa a la matriz original de la señal en dos matrices no negativas *W* y *H*.

La matriz *W* se conoce como la "*Matriz de bases*" o "*Matriz de características*" ya que, al combinar linealmente sus columnas se puede formar nuevamente la matriz original.

La matriz *H*, en cambio, es conocida como la "*Matriz de coeficientes*" y contiene, como su nombre lo indica, los coeficientes por los que hay que multiplicar las bases de la

matriz W para volver a formar la matriz original. Es decir, llamando 'V' a la matriz original, $V \approx W \cdot H$.

Las obtención de las matrices W y H se realiza de manera iterativa mediante las siguientes ecuaciones:

$$H = H \cdot \frac{(W' \cdot |V|)}{W' \cdot W \cdot H} \qquad (3)$$

$$W = W \cdot \frac{|V| \cdot H'}{W \cdot H \cdot H'} \qquad (4)$$

Se destaca que, en la implementación del algoritmo, es pertinente sumar un error (por ejemplo, $1e^{-12}$) al denominador con el fin de evitar divisiones por cero.

Una de las formas para trabajar con NMF para la separación de elementos, y la que se utiliza en este trabajo, es la NMF supervisada. Consiste en aplicar primero el método a señales que tengan las mismas características que los elementos que se quiere separar.

Dado que W conformaría una "matriz de bases", se utilizan las matrices obtenidas de cada señal individual y se generan los coeficientes correspondientes de forma de obtener con esas bases la señal combinada.

Para esto, se unen las matrices W_i resultantes de aplicar NMF a las n señales individuales, formando una matriz $W = [W_1, W_2, \dots, W_n]$. Luego, se aplica el método a la señal con los n elementos mezclados pero manteniendo fija W y actualizando solamente H.

Así, se pueden obtener aproximaciones de las señales separadas mediante el uso de los coeficientes correspondientes en H.

C. *Métricas de calidad*

En un principio, se hará una calificación subjetiva de la calidad de las reconstrucciones a través de la escucha de los audios resultantes por parte de 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 pero más técnica, la SNR (Signal to Noise Ratio), que mide la relación entre la energía de la señal original y la energía del error residual, esto es, la diferencia entre la señal original y la aproximada por NMF.

D. *Bases de datos*

Para el desarrollo de este trabajo, se utilizaron tres bases de datos con diferentes.

La primera base de datos (B1) consiste de dos pistas de audio. La primera compuesta únicamente por piano y la segunda únicamente por flauta. En la tercer pista de audio se superponen las dos ya mencionadas En la pista combinada se pueden apreciar tres tramos bien diferenciados:

Tramo 1: [t=0.6 ; t=3.0] (Sólo flauta)
Tramo 2: [t=4.9 ; t=8.3] (Sólo piano)
Tramo 3: [t=9.0 ; t=11.3] (Ambos instrumentos)

La segunda base de datos (B2) consiste de una pista de audio compuesta por piano, flauta y bajo, de forma similar a B2 pero con distintas notas en el caso del piano y la flauta, junto con las respectivas pistas de audio de cada instrumento por separado.

La tercera y última base de datos (B3) es similar a B1 pero se le agrega a la mezcla un ruido gaussiano sutil, de media 0 y desvío estándar 0.03l, manteniendo los audios de los instrumentos separados sin ruido para poder aplicar NMF.

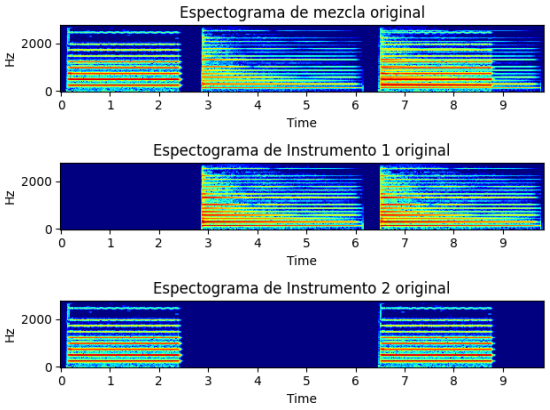

Fig. I: Análisis tiempo-frecuencia de B1.

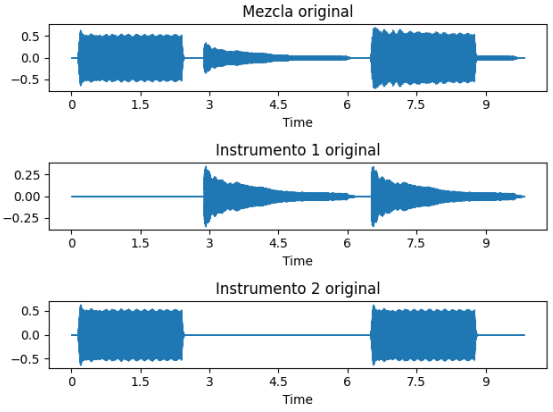

Fig. II: Señales de B1 en el dominio temporal.

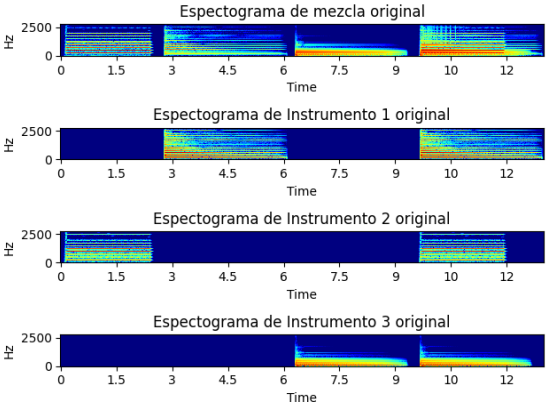

Fig. III: Análisis tiempo-frecuencia de B2.

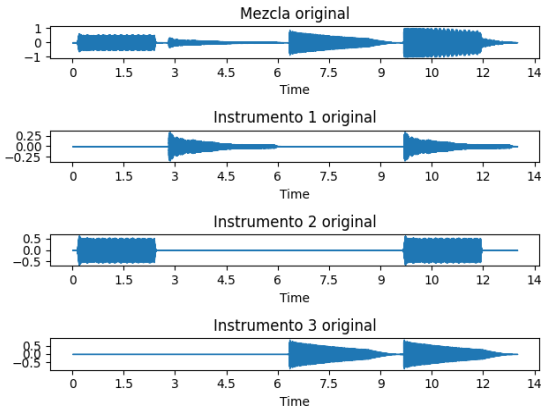

Fig. IV: Señales de B2 en el dominio temporal.

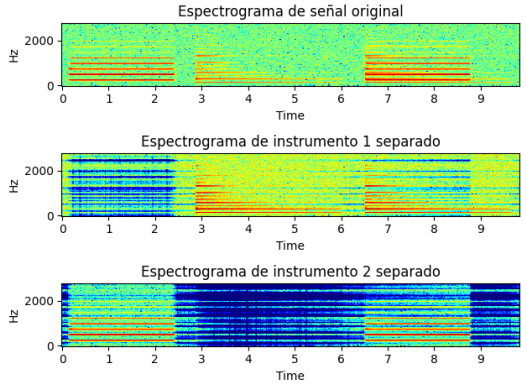

Fig. V: Análisis tiempo-frecuencia de B3.

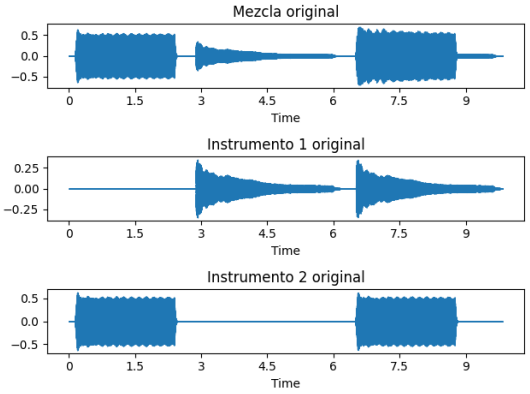

Fig. VI: Señales de B3 en el dominio temporal.

III. Resultados

Se realizaron tres pruebas distintas con las bases de datos disponibles.

En primer lugar, se aplicó el método a las pistas de la base de datos 1 y 2, de donde se obtuvieron los siguientes resultados:

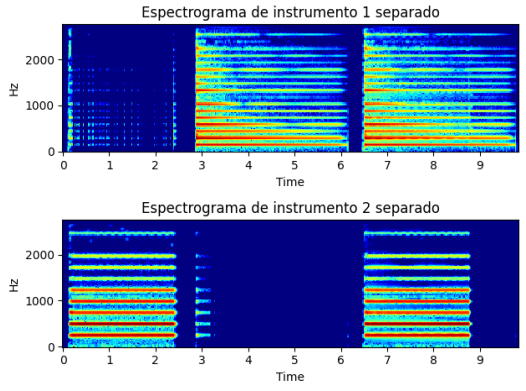

Fig. VII: Análisis tiempo-frecuencia obtenido a partir de B1.

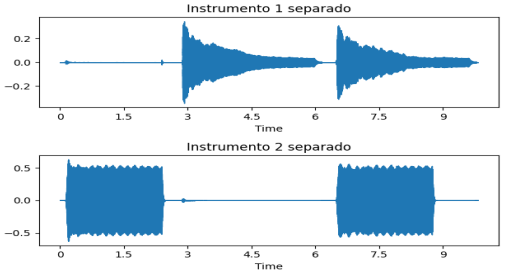

Fig. VIII: Señales en el dominio temporal obtenida a partir de B1.

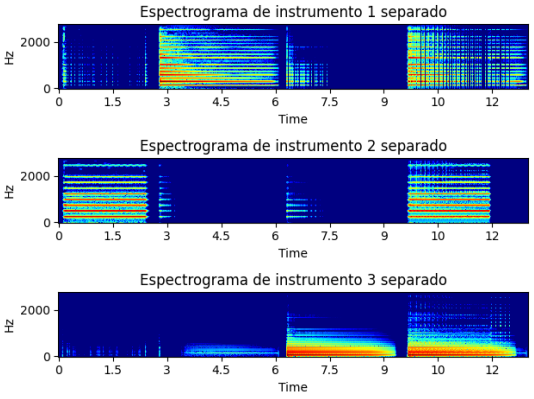

Fig. IX: Análisis tiempo-frecuencia obtenido a partir de B2.

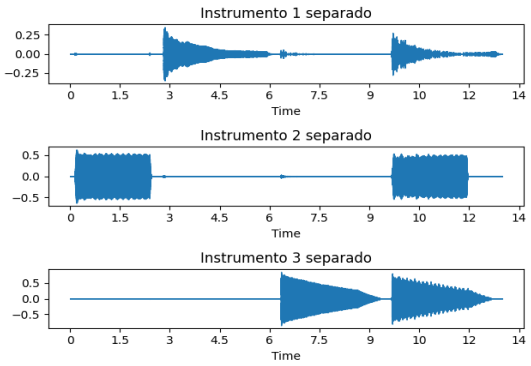

Fig. X: Señales en el dominio temporal obtenida a partir de B2.

A partir de las separaciones, se realizó una comparación con las pistas separadas originales y, aplicando la métrica SNR, se obtuvieron los siguientes resultados:

Base de datos B1	SNR
Instrumento 1 (Piano)	13.54
Instrumento 2 (Flauta)	25.02

Base de datos B2	SNR
Instrumento 1 (Piano)	5.55
Instrumento 2 (Flauta)	17.10
Instrumento 3 (Bajo)	15.56

Por último, se trabajó con la base de datos B3 y se realizó el mismo análisis que para B1.

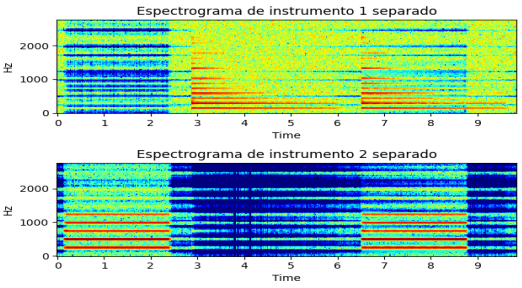

Fig. XI: Análisis tiempo-frecuencia obtenido a partir de B3.

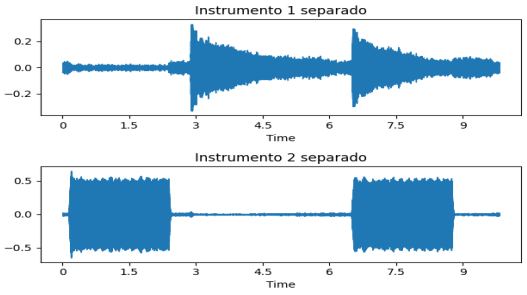

Fig. XII: Señales en el dominio temporal obtenida a partir de B1.

Base de datos B3	SNR
Instrumento 1 (Piano)	4.63
Instrumento 2 (Flauta)	18.72

IV. Discusión

Para la base de datos B1 comparando los resultados obtenidos (Fig.VIII) con los originales (Fig. II) se puede observar el resultado final de la descomposición de la onda en el dominio temporal.

Observando el tramo 1 es claro que la señal de la flauta filtrada representa casi la totalidad de la onda original mientras que el mismo tramo de la señal de piano tiene una amplitud casi nula. De igual manera se puede observar esto para el tramo 2 pero en este caso quien contiene casi la totalidad de la onda es el piano, como era de esperarse.

Para el análisis del tramo 3 es conveniente observar cada señal por separado, es decir, comparar los tramos 1 y 3 de la flauta y los tramos 2 y 3 del piano, ya que las notas producidas en ambos instrumentos fueron las mismas. Teniendo en cuenta esto se ve una similitud en cuanto a la forma de onda en ambos casos.

El análisis en el dominio temporal se condice perfectamente con el análisis tiempo-frecuencia de las tres señales. Se puede comprobar a simple vista que la señal original puede componerse como la sumatoria de las frecuencias producidas por ambos instrumentos que, como es lógico tratándose de instrumentos musicales, producen sonidos que se pueden descomponer en un número reducido de frecuencias según la nota.

En este aspecto es destacable como la influencia de la flauta pasa a ser prácticamente 0 a partir de t=6, por lo que se puede afirmar que la descomposición de instrumentos en el tramo [t=6. t=8] es casi perfecta.

Cabe destacar la efectividad del método NMF. En la experimentación se comprueba que a medida que se aumentan las iteraciones del método progresivamente se reduce la diferencia entre la Matriz original V y su aproximación $W \cdot H$ siendo E la sumatoria de la diferencia cuadrática de cada elemento de la matriz.

$$E = \sum_i \sum_j (V_{ij} - B_{ij})^2 \qquad (5)$$

Considerando

$$B_{ij} = W_{ik}.H_{kj} \qquad (6)$$

A su vez se puede apreciar que el error disminuye notablemente a medida que disminuye el tamaño de las ventanas. A partir de esto se determina una de 256 muestras y con esta se analiza el overlapping concluyendo de igual manera que cuanto menor sea el mismo más preciso es el método.

Aproximadamente a partir de las 60-70 iteraciones se puede ver que la mejora entre pasos disminuye notablemente sin dejar de ser constante. Concretamente para 1000 iteraciones el número que se obtuvo en la experimentación fue $E \simeq 0.0225$

Respecto al análisis para la base de datos B2, a simple vista parece tener la misma forma que cuando estaban individualmente. Sin embargo, al mirar un poco más en detalle, puede notarse un serruchado que en la señal original no existía, sobre todo en el caso del piano. Esto puede notarse también cuando se escucha la pista de audio, la cual está entrecortada, y mirando los valores del SNR: muy bajo para el piano, con 5.55 dB, y valores más aceptables para la flauta (17.10 dB) y el bajo (15.56 dB).

Finalmente, en la separación de señales con ruido, sucede lo contrario al caso anterior. Nuevamente, la separación es deficiente en comparación a la separación sin ruido, pero esta vez la señal posee más información y el método dejó pasar ruido. En cuanto a la métrica SNR, se registró un valor de 4.63 dB para el piano, el más bajo entre las tres bases de datos y todos los instrumentos analizados, y 18.72 dB para la flauta.

V. Conclusiones

Luego de aplicar la factorización en matrices no negativas a pistas de audio compuestas por notas de diferentes instrumentos, se pudieron notar algunas ventajas y desventajas del mismo.

Se pudo observar cómo, al aplicar el método a pistas de solo dos instrumentos, el mismo funciona de manera óptima, quedando muy poco resto del otro instrumento en las pistas separadas.

Sin embargo, si se requieren separar pistas con más de dos instrumentos, o si las mismas tienen ruido, el método pierde calidad de reconstrucción.

Adicionalmente, se notó que el algoritmo funciona mejor para ciertos instrumentos que para otros. En general, no tuvo un buen desempeño al intentar separar el piano de la mezcla. En las tres bases de datos se registró una SNR mucho menor a las demás y, del lado auditivo, más interferencias del ruido o de los otros instrumentos presentes. Por otro lado, mantuvo un buen desempeño para la flauta en los tres casos presentados, aunque con leves caídas de calidad en las bases de datos B2 y B3 en comparación a B1.

VI. Referencias

[1] Lee, D. D., y Seung, H. S., *"Learning the parts of objects by non-negative matrix factorization"*, *Revista Nature*, 1999.

[2] Anupama, G., "NMF — A visual explainer and Python Implementation", *Towards Data Science*, 2021.

[3] Benslimane, Z, "Audio Source Separation Using Non-Negative Matrix Factorization (NMF)", en *Proc. of the 7th European Conference on Speech Communication and Technology*, vol. 1, pp. 267-270, 1999.

[4] Févotte, C., Vincent E. y Ozerov A., "Single-channel audio source separation with NMF: divergences, constraints and algorithms", en *Audio Source Separation. Signals and Communication Technology,* cap. 1. 2018.