ANÁLISIS Y CARACTERIZACIÓN DE INSTRUMENTOS MUSICALES USANDO LA TRANSFORMADA DE FOURIER.

A PREPRINT

Ignacio Manuel Lebrero Rial
Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Unversidad de Buenos Aires
ignaciolebrero@gmail.com
751/13

Nestor Dario Ocles Garcia
Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Unversidad de Buenos Aires
dario.ocles@gmail.com
633/15

Matias Millassón
Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Unversidad de Buenos Aires
matiasmillasson@gmail.com
LU 131/13

Joaquín Romera
Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Unversidad de Buenos Aires
joakromera@gmail.com
183/16

November 28, 2019

ABSTRACT

En este trabajo presentamos un análisis sobre audios de distintos instrumentos musicales e intentamos caracterizarlos definiendo una noción de **Identidad Instrumental** para más adelante intentar clasificarlos automáticamente según la misma. Para empezar recopilamos grabaciones de un conjunto de instrumentos con el objetivo tratar de entender sus propiedades individuales y tener un set para hacer los experimentos. De este set surgieron algunas clasificaciones básicas basadas en el material de construcción del instrumento (metales, maderas) y su forma de ejecución (percusión, con arco, viento), a su vez la fuente del sonido (cuerda, viento) y su medio de excitación (arco, boquilla) sugirieron otras categorías. Estas, en parte, determinarían las cualidades espectrales del sonido de cada instrumento. En segundo lugar decidimos centrar nuestro análisis en una serie de descriptores según la representación del sonido en su dominio espectral, temporal, y según su contenido armónico. Utilizamos estos descriptores para entrenar un modelo que permita predecir a qué familia pertenece cada muestra. Según el grado de éxito de este modelo podremos especular si estos descriptores son relevantes para describir un instrumento musical o la familia a la cual pertenece, o si son relevantes para una determinada familia y no para otra. Finalmente intentamos lograr la separación de instrumentos en grabaciones que incluyan dos ejecutando notas diferentes al mismo tiempo, utilizando la decomposición de componentes y activaciones de su espectograma.

***K*eywords** Transformada de Fourier · Procesaminto de señales · Instrumentos Musicales

1 Introduction

1.1 Problema General

El análisis de instrumentos musicales es un problema clásico en el campo de la Recuperación de Información Musical (MIR, Music Information Retrieval) que a su vez representa un desafío de gran complejidad e interés en el ámbito. La clasificación de muestras de instrumentos según su timbre es un campo de investigación todavía abierto debido a esta complejidad. La música es una actividad artística esencial de la sociedad humana, diversos estudios ya trataron la clasificación de los instrumentos en jerarquías basadas en los materiales de fabricación de los instrumentos, así como en su fuente de sonido y su método de excitación, estos derivaron en las familias clásicas de instrumentos que se estudian desde hace casi dos siglos: instrumentos de cuerda (percutida, frotada con arco), de viento (maderas, metales), percusión, etc.

Gracias a los avances tecnológicos este análisis hoy se puede hacer sobre grabaciones digitales, estudiando la representación del sonido en distintos dominios facilitados por la transformada de Fourier (del tiempo, o de la frecuencia). Este acercamiento está basado directamente en el contenido que se puede extraer de las características del mismo audio del instrumento. Si la distinción hecha sobre materiales de construcción es una de alto nivel, podemos considerar este análisis como uno de bajo nivel. El problema que estaríamos intentando resolver entonces, es encontrar una correlación entre descriptores de bajo y alto nivel que permita identificar una muestra de un instrumento o por lo menos acercarlo a su familia según su descriptor de alto nivel.

1.2 ¿Qué hacemos en este trabajo?

En la seccion 2 presentamos un breve análisis sobre audios de instrumentos aislados descargados del MIS dataset desarrollado en la Universidad de IOWA[1] para tratar de caracterizarlos, basándonos en una técnica ya conocida (Benetos, Kotti & Kotropoulos, 2006)[2] armamos un modelo de ML para clasificar audios de acuerdo al instrumento que se toca, es útil aclarar que en esta parte solo usamos audios aislados tanto para entrenar como para predecir. Podemos decir que la caracterización de los instrumentos en esta parte queda dada por los features usados y el modelo entrenado.

Finalmente, en la sección 3 experimentamos con la *factorizacion no negativa de matrices* para descomponer un audio y tratar de analizar si es posible usar esto para aislar instrumentos.

Las herramientas utilizadas en el trabajo fueron todas open source como Python, librosa [1], Jupyter Notebook, sklearn de Scipy.

Ademas, dentro de un repositorio es pueden encontrar los sources y datasets necesarios, en particular aqui para la seccion 2 y aqui para la seccion 3.

En el primero se pueden encontrar:

- *Src:* Contiene los archivos en python para correr el codigo.
- *Audios:* Contiene los audios para generar el dataset de entrenamiento y de prueba.
- *Notebooks:* Contiene los jupyter notebooks con los mismos experimentos que en este trabajo.

En el segundo simplemente la carpeta *Notebooks* con la notebook y los audios necesarios para correr el experimento, notar que los archivos resultantes tambien se guardan en ese directorio.

[1] Librería open source sobre manipulación de sonido en Python https://librosa.github.io/librosa/

2 Caracterizacion de un instrumento

2.1 Extracción de Features

Hoy se encuentran muchos descriptores de bajo nivel basados en el contenido espectral usados en este campo de investigación. Para intentar determinar el timbre de los instrumentos usamos cuatro: Zero Crossing Rate, Spectral Rolloff, Spectral Centroid, Spectral Flatness.

- *Zero-Crossing Rate:* es la velocidad en la que el signo de una señal cambia entre positivo y negativo. Este feature suele dar buenos resultados para clasificar instrumentos percusivos.
- *Spectral Rolloff:* es una medida de la forma de la señal. Se define para cada muestra como la frecuencia central para un bin del espectrograma de modo que al menos el 85% de la energía del espectro en esta muestra está contenido en este bin y los de abajo. Esto se puede utilizar para, por ejemplo, aproximar la frecuencia máxima (o mínima) ya que mide la frecuencia donde se concentra un porcentaje de la distribución de magnitud.
- *Audio Spectrum Centroid:* es una medida utilizada en el procesamiento digital de señales para caracterizar un espectro. Indica dónde está el centro de masa del mismo. Perceptualmente tiene una fuerte correlación con la impresión del brillo de un sonido.
- *Audio Spectrum Flatness:* o el coeficiente de tonalidad, es otra medida del procesamiento de señales que sirve para caracterizar el espectro de un audio. Se mide típicamente en decibeles y provee una forma para cuantificar que tan tonal es un sonido, contrapuesto a que tan parecido al ruido es.

2.2 Modelo

Parar el modelo usamos la *factorización no negativa de matrices* (En adelante NMF). Esto es, dada una matriz no negativa V de nxm (m vectores de n dimenciones), es posible encontrar matrices no negativas W y H tal que

$$V \approx WH$$

donde W, de la forma nxr, contiene una base de vectores y H, de la forma rxm, contiene los pesos necesarios para aproximar la correspondiente columnas de V. En particular, r se elige de manera arbitraria. Usualmente es bueno tomar un r tal que $(n + m)r < nm$ de manera que la matriz resultante sea una versión compresa de la original.

2.3 Procesamiento de datos y experimentos

2.3.1 Dataset

El set de sonidos utilizados incluyó platillos, violines, flautas, trombones, guitarra y trompetas, con apróximadamente 70 muestras de cada uno, todas del mismo largo y discretizadas usando 44.1 kHz de sampling rate. De estas tomamos un 10% para test y el resto para entrenamiento. Por un lado podemos distinguir entre estos instrumentos aquellos que son percutidos -platillos y al ser cuerda percutida, la guitarra- de los que no, y los que son no-tonales -los platillos- de los tonales -el resto-.

A su vez las familias clásicas basadas en los descriptores de alto nivel los distinguen en cuerdas: guitarra, violín; vientos: trompeta, trombón, flauta; percusión: platillos. En la familia de los vientos también se distingue entre los metales (trombón y trompeta) y las maderas (flauta).

Estas muestras son distintas notas únicas de cada instrumento, o golpes si se trata de un platillo, con varias articulaciones dependiendo de las posibilidades de cada instrumento. Por lo que cada archivo de audio corresponde a un único instrumento, ejecutando una sola nota.

2.3.2 Método de Clasificación según features

Extrayendo los features previamente dichos, generamos un vector de features v_j por cada audio y con eso armamos la matriz V donde cada v_j representa una columna de la misma. Además, al saber de qué instrumento es cada audio guardamos un vector l con los labels de cada columna de V.

El entrenamiento del modelo consiste en aplicar NMF a V para calcular W y H. Luego calculamos la pseudo inversa W^{-1}.

Finalmente, para estimar un nuevo audio, generamos su vector de features v_{test} y hacemos:

$$h_{test} = W^{-1} v_{test}$$

Solo nos queda matchear de alguna manera con los vectores h que ya calculamos y estimar el label. Para esto usamos KNN comparando h_test contra los vectores columna de H y consideramos los más cercanos a los que minimizan la *distancia euclidea* entre ellos, esto es, los que minimizan la norma de la resta. Originalmente probamos maximizando *Cosine Similarity*, pero esta medida no dió buenos resultados.

2.3.3 Evaluación

Los parametros con los que experimentamos fueron por un lado las *n componentes* sobre la factorizacion NMF, este sería el *r* de la factorizacion (que tanto comprimimos los datos o que factores tenemos en cuenta). Por otro lado probamos con distintos valores de *k* para los k vecinos mas cercanos. Finalmente maximizamos la performance en los valores **6** y **4** para **n_components** y **k** correspondientemente. A continuacion mostramos los resultados obtenidos corriendo el modelo con estos parametros y comparando el uso de distintos features isolados para tratar de entender como estos impactan en la predicción:

Instruments / hit rate using	Todos	Zero Crossing Rate	Spectral Rolloff	Spectral Centroid	Spectral Flatness
Violin	42.86	28.57	42.86	14.29	28.57
Flauta	75.00	0.00	0.00	0.00	0.00
Trombon	0.00	33.33	100.00	100.00	0.00
Guitarra	0.00	0.00	71.43	0.00	100.00
Clash Symbals	100.00	100.00	50.00	50.00	50.00
Trompeta	57.14	85.71	0.00	0.00	14.29

Figure 1: Resultados: Hit rate

Zero Crossing Rate

Era esperable que este descriptor de buenos resultados con instrumentos percusivos y en efecto los platillos fueron identificados correctamente en un 100%.

Audio Spectral Roll-off

El descriptor de spectral roll-off arrojó buenos resultados para identificar las muestras de trombones y guitarras en menor medida.

Audio Spectrum Centroid

Consideramos que el 'brillo' de un sonido es una de las formas más sencillas de caracterizar y diferenciar sonidos, la formalización de esta característica corresponde a una indicación de la cantidad de energía en las frecuencias más altas de un sonido y es lo que podemos medir mediante el centroide espectral.

Llevado este análisis a los instrumentos estudiados encontramos un buen grado de precisión para el trombón. Sin embargo, el resultado no tiene una buena correlación con los descriptores de alto nivel dado que el trombón pertenece a la misma familia que la trompeta (vientos y metales) y es de un registro más grave, por lo que si obtuvimos buenos resultados con el trombón, hubieramos esperado obtener aún mejores resultados con la trompeta, ya que al ser de la misma familia y de un registro más agudo, debería haber mejorado su predicción con respecto al brillo.

Audio Spectral Flatness

Cuantificar qué tan tonal es un sonido no debería permitir distinguir instrumentos cuando son todos tonales, en nuestro set el único que se diferencia en esta dimensión son los platillos, que al ser inarmónicos deberían presentar diferenciarse facilmente de los demás.

Sin embargo, este desriptor no fue suficiente para lograr identificar esta distinción, el grado de precisión que arrojaron las muestras de platillos sugieren que este descriptor no alcanza para evidenciar esta hipótesis. En cuanto a la precisión que obtuvimos al predecir con los samples de guitarras, podemos suponer que la cualidad tonal de los mismos resultó más homogénea para este instrumento que para el resto.

Todos los descriptores combinados

Combinando todos los descriptores obtuvimos buenos resultados para identificar platillos en nuestro set de muestras, esta identificación mejoró a la obtenida por los descriptores separados (a excepción de Zero Crossing Rate, que dio igual). Esta combinación también mejoró la identificación de flautas, con el hecho notable de que por separado ningún descriptor había mostrado buena precisión. También mejoró la identificación de violines, aunque la precisión sigue siendo baja (por debajo del 50%).

En el resto de los instrumentos vimos caer la precisión entre combinar los descriptores y utilizarlos por separado. No podemos afirmar que estos descriptores sean igual de determinantes para cada instrumento, ni para cada familia de instrumento, tampoco que mientras más descriptores combinemos, mejores resultados obtengamos. Es posible que los descriptores elegidos sean relevantes para los instrumentos percusivos inarmónicos (como los platillos) y no para los armónicos o melódicos (el resto). Una continuación de este trabajo debería constatar estos resultados con otros instrumentos percusivos e incluir descriptores nuevos para probar si pueden distinguir aquellos que producen sonidos armónicos entre sí.

Asi, podemos ver como queda la matriz de confusion resultante:

Instrumento / Instrumento	Violin	Flauta	Trombon	Guitarra	Clash Symbals	Trompeta
Violin	3	0	0	3	1	0
Flauta	2	6	0	0	0	0
Trombon	3	0	0	0	0	0
Guitarra	0	7	0	0	0	0
Clash Symbals	0	0	0	0	4	0
Trompeta	0	1	0	2	0	4

Figure 2: Matriz de Confusion entrenando con todos los features

Por lo que se muestra en la misma tiene bastante concentración en la diagonal para la flauta y clash cymbals mientras que para el resto de los instrumentos no hubo un gran desempeño. En particular todos los sonidos provenientes de las guitarras los clasificó como flautas pero se debe a que no analizamos tantos features. Con respecto a los violines su performance fue bastante buena ya que la gran mayoría de los sonidos los clasificó como violín o guitarra. Es decir las confusiones fueron sobre instrumentos de cuerda.

3 Aislando Instrumentos

Analizamos la posibilidad de separar audios de distintos instrumentos por medio de la factorización NMF. Como se vio antes esta factorización nos descompone el sonido en una matriz tiene una base de vectores y en otra matriz de activación de michos vectores. Realizamos dos experimentos, uno combinando una guitarra con una flauta y luego otro combinando una trompeta con un violín. Notar que en ambos casos mezclamos una instrumento de viento con uno de cuerda.

3.1 Modelo

El modelo es similar el propuesto en la sección anterior 2.2 . Utilizamos la factorización NMF y luego reconstruimos el sonido con dicha factorización.

3.2 Experimentación

3.2.1 Guitarra y flauta

Experimentamos anulando algunos de estos vectores e intentamos reconstruir los audios originales por separado. Tomamos un audio de Guitarra (nota LA en 4ta octava) y un audio de Flauta (nota RE en 5ta octava), los mezclamos en un mismo audio superponiendo el sonido.

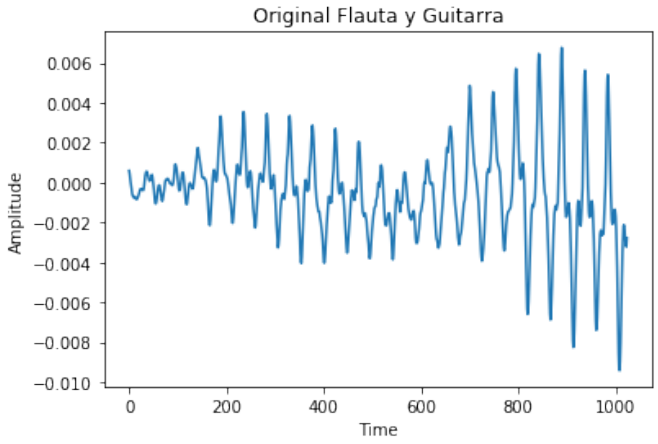
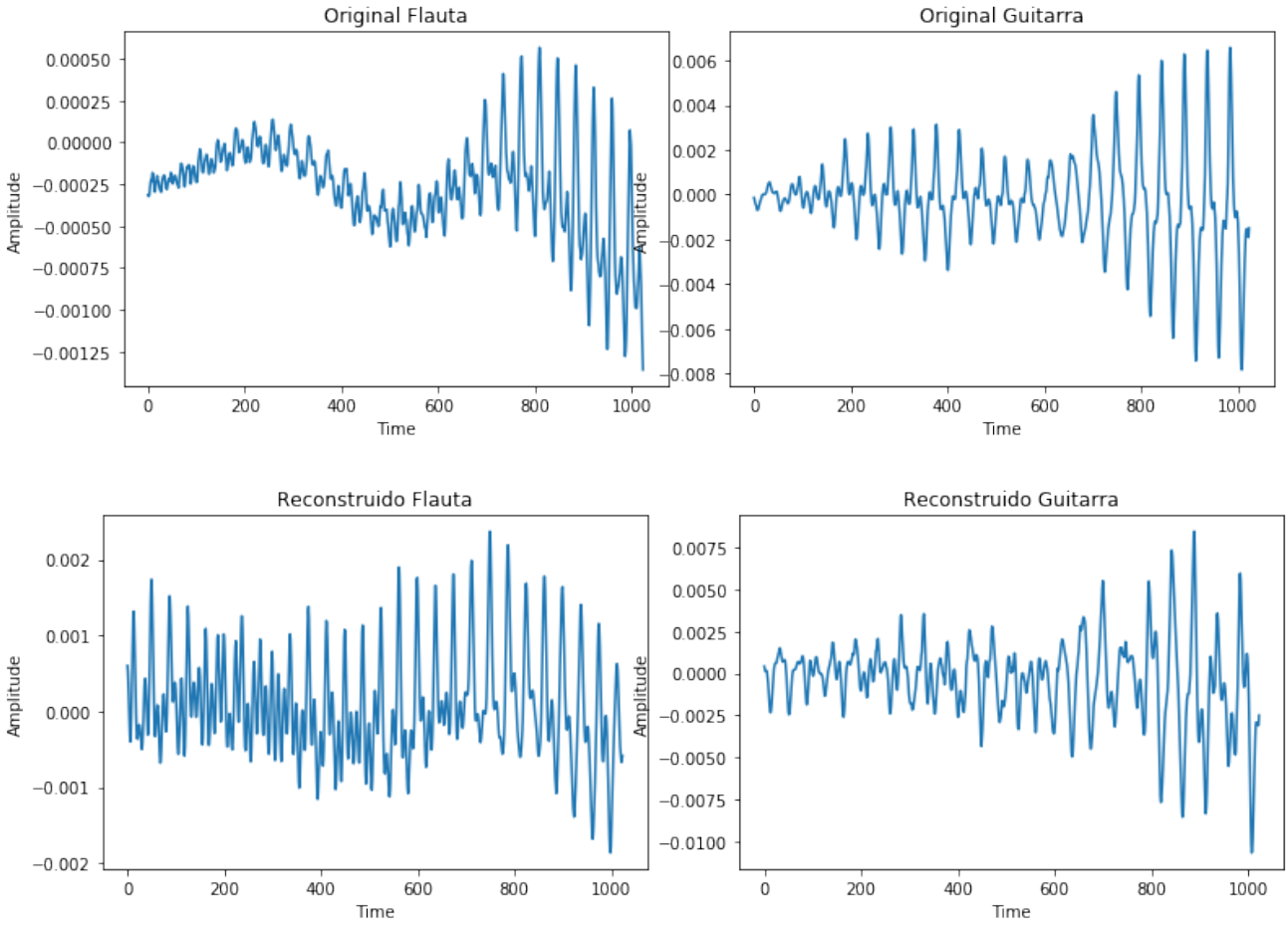


Figure 3: Grafico del audio donde se mezclan Guitarra y Flauta

Luego separamos el audio en varios componentes usando la librería librosa para generar la matriz NMF. Con la matriz de factorización fuimos anulando distintos vectores y reconstruyendo el audio con cada una de las nuevas matrices para luego quedarnos con aquellas matrices que tienen un mayor grado de correlación entre la matriz de factorización de los audios por separado.

Con las nuevas matrices que tienen algunos vectores anulados pero que guardan mayor correlación con los sonidos originales reconstruímos el audio y los volvimos a graficar.

Como se puede observar, los audios reconstruidos mantienen un grado de similitud con los audios originales por separado. El sonido resultante de la reconstrucción también tienen un gran parecido con su audio original por separado a diferencia de otras tecnicas donde el audio reconstruido no guardaba mucha relación con el audio original.



Esta técnica podría ser util para la caracterización de instrumentos como también separación de audios. Con esta técnica no logramos separar audio en algunas de nuestras prueba indicando que esta técnica por si misma no es capaz de lograrlo. Sin embargo creemos que podría ser parte de técnicas más avanzadas de caracterización de sonido usando por ejemplo machine learning para detectar algunos aspectos del sonido.

3.2.2 Violín y trompeta

Al igual que en la sección anterior superpusimos los sonidos de dos instrumentos tocando una nota distinta. En este caso tomamos un violín y una trompeta. La nota correspondiente al sonido del violín es una D4, es decir octava prima de Re, en tanto la nota de la trompeta es una A4 lo que es lo mismo que una octava prima de La. La señal correspondiente a la mezcla se la puede ver en la siguiente figura:

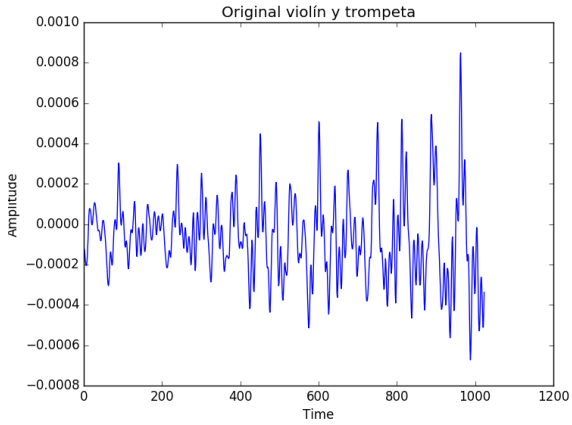
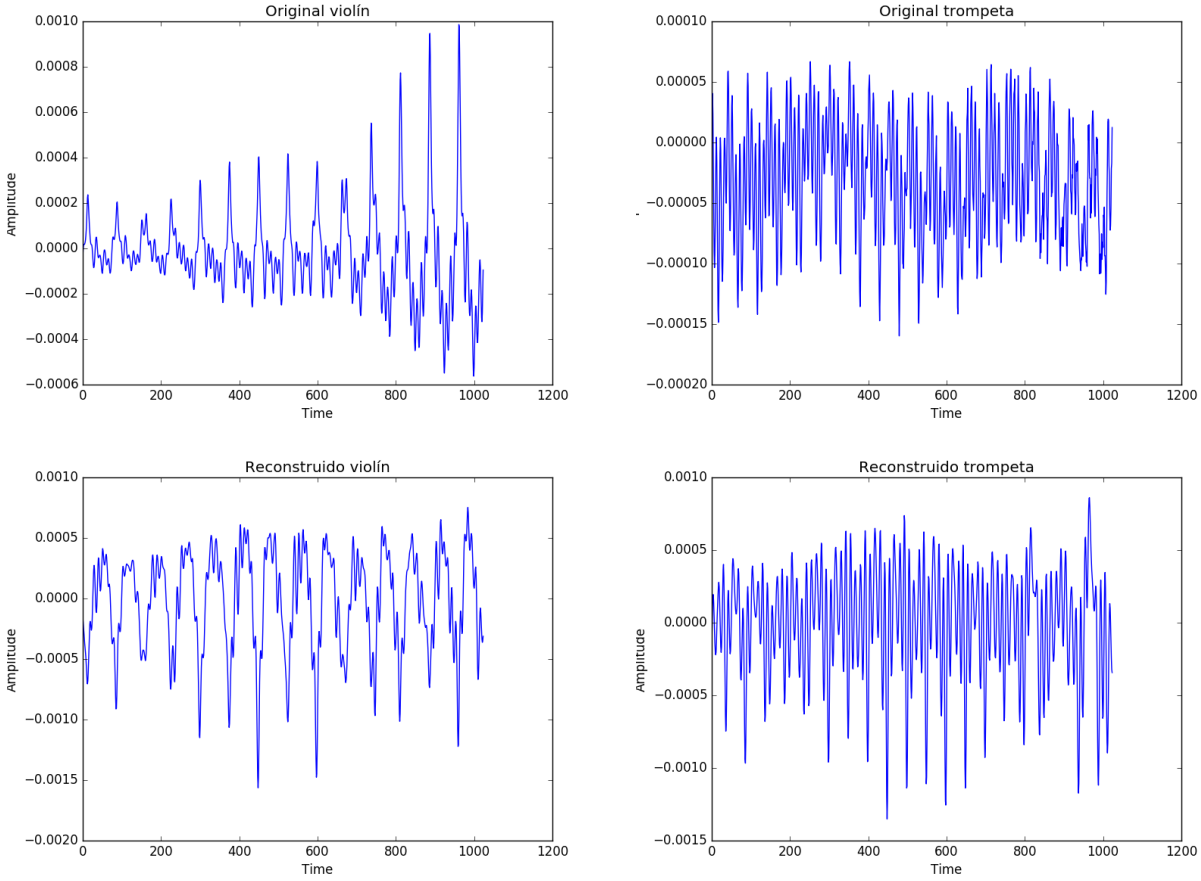


Figure 6: Grafico del audio donde se mezclan violín y trompeta

Luego ejecutamos el script para separar la señal en las dos señales correspondientes a los instrumentos originales. Vale la pena aclarar que el script es exactamente el mismo que el que utilizamos para el experimento anterior.



Como se puede apreciar en las figuras correspondientes a los audios originales la diferencia entre las señales es muy marcada pero en las versiones reconstruidas son similares. Además la trompeta se reconstruyó de manera satisfactoria mientras que el violín tiene bastantes diferencias. Por lo que podemos concluir que logró aislar correctamente la trompeta.

4 Conclusiones

Podemos ver que caracterizar instrumentos y separar audio presesntan un desafio complejo para el cual existen una gamma de técnicas muy diversas para resolverlo. Ademas, la mayoria suele tener resultados parcialmente exitosos, de manera que resulta dificil resolver el problema de manera exacta y la manera de encararlo resulta ser estimando una solución relativamente buena.

En particular, la factorización NMF nos fue muy util para descomponer el sonido en una cantidad arbitraria de componentes distintos. Esta factorización es usada en distintas tecnicas de manipulación de audio pero usualmente resulta ser una pequena parte de un proceso mas grande para separar el audio.

Podemos ver que caracterizar instrumentos de manera "manual" en el sentido de intentar usar una tecnica particular resulta ser dificil para caracterizarlos, mas que nada debido a la cantidad de variables distintas a tener en cuenta sobre la senal, esto nos lleva a pensar que una solucion de ML puede llegar a ser la mas razonable para este tipo de problemas. Aun más, nuestros resultados muestran como un modelo relativamente simple tiene una performance masomenos aceptable sobre un conjunto de datos, dejando abierta la pregunta de si se puede llegar a usar nuestro modelo como un modulo individual de un modelo mas grande.

Con respecto a la separación de audios, en nuestra experimentación logramos obtener casos donde dicha factorización nos permitio separar alguna caracteristica del audio y relacionarla con el original, dando la idea de que este método podría llegar a ser util como un modulo de un modelo mas grande, al igual que el modelo de ML.

References

[1] University of Iowa Musical Instrument Sample Database, http://theremin.music.uiowa.edu/index.html.

[2] Emmanouil Benetos, Margarita Kotti and Constantine Kotropoulos. Musical instrument classification using non-negative matrix factorization algorithms. In *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE.*