Escuela Politécnica Superior de Linares
Grado en Tec. De Telecomunicaciones

UNIVERSIDAD DE JAÉN
Escuela Politécnica Superior de Linares

Trabajo Fin de Grado

SEPARACIÓN PERCUSIVA, ARMÓNICA Y VOCAL DE SEÑALES

Alumno: **Jiménez Romero, Miguel**

Tutor: Cañadas Quesada, Francisco J.
Depto.: Ingeniería Telecomunicaciones

Septiembre, 2014

UNIVERSIDAD DE JAÉN
Escuela Politécnica Superior de Linares

Trabajo Fin de Grado

SEPARACIÓN PERCUSIVA, ARMÓNICA Y VOCAL DE SEÑALES

ABSTRACT

Este trabajo fin de grado está centrado en la separación de fuentes sonoras. Concretamente, el alumno deberá implementar un sistema que sea capaz de separar una señal musical estéreo en tres pistas musicales diferentes: percusiva (instrumentos que presentan características percusivas, por ejemplo, baterías), armónica (instrumentos que presentan características armónicas, por ejemplo, piano, bajo...) y vocal (singing-voice o voz cantada). En una primera etapa, se realizará la separación vocal-instrumental y en una segunda etapa se realizará la separación percusiva-armónica a partir de la parte instrumental separada en la etapa previa. El enfoque de dicha separación se basará en el filtrado de mediana del espectrograma de audio. El proyecto se implementará utilizando el entorno MATLAB.

1. INTRODUCCIÓN

El sistema auditivo del ser humano, es un sistema capaz de realizar una serie de tareas complejas e increíbles, con el uso exclusivo de dos flujos de entrada (oído izquierdo y derecho). Es capaz de detectar la naturaleza de un sonido, identificando su origen, si consiste en una fuente de voz, instrumental o si por el contrario es simplemente ruido. Concretando para las fuentes de voz, las posibilidades son infinitas, siendo capaz de distinguir palabras habladas en una composición musical, dando una compresión semántica a la melodía. Es capaz de centrar la atención en una única fuente independiente de si es predominante o no. Todas estás funciones se realizan de forma subconsciente, en un segundo plano sin que realmente llegue a representar un esfuerzo para la persona.

Sin embargo los ordenadores por el contrario no son capaces de realizar este tipo de acciones de forma "subconsciente", no pueden separar las diferentes fuentes de una melodía, identificar los diferentes instrumentos que suenan, ni por supuesto entender la letra que acompaña a la melodía interpretada por un cantante.

Todo ello representa un problema, para el desarrollo de aplicaciones que buscan automatizar sistemas que permitan la extracción de la melodía, generando un pentagrama con las notas musicales correspondiente. Aplicaciones que requieran de la extracción de la voz cantada (Singing Voice), como pueda ser un karaoke o simplemente para identificar el cantante, o asociar una determinada pista a un genero determinado. Incluso aplicaciones que permitan extraer determinados sonidos para volverlos a emplear en nuevas piezas musicales.

Luego parece necesario el desarrollo de un sistema capaz de dotar a un ordenador de la capacidad de separar una pista musical en sus diferentes componentes musicales y vocales, incluyendo la separación de la música en sus correspondientes armónicos y percusivos.

1.1 INTRODUCCIÓN A LA TEORÍA MUSICAL

Quizás la mejor forma de empezar sería respondiendo a la pregunta ¿Qué es la música?. La respuesta puede parecer algo trivial, y estará influenciada por el área de especialización de la persona a la que le preguntemos. Una definición aceptada desde una perspectiva puramente musical es:

"La música es el arte de combinar sonidos agradablemente al oído según las leyes que lo rigen." [1]

Antes de centrarnos en la perspectiva puramente matemática, es interesante analizar la definición dada. "La música es el arte…" es una arte, puesto que se considera una forma de expresión de comunicación, así como no hay dos personas iguales, la percepción y la interpretación de una melodía es única para cada individuo. "…combinar sonidos agradablemente al oído…", en la naturaleza disponemos de diversidad de sonidos, y no por combinarlos ya podemos decir que tenemos música, tendríamos ruido, la verdadera música es la unión de todas esas variables o sonidos que disponemos, en los tiempos precisos y con la duración adecuada que permitan una transición melódica entre los diferentes sonidos que resulten agradables al oído.

Aunque en principio cualquier objeto puede servir como instrumento, podemos clasificarlos según la naturaleza de la acción que genera el sonido:

1. Cuerda: Aquellos instrumentos capaces de generar un sonido mediante la puesta en vibración de cuerdas.

Figura 1: Guitarra

Figura 2: Violín

2. Viento: Aquellos instrumentos capaces de generar un sonido mediante la puesta en vibración de una columna de aire.

Figura 3: Saxofón

Figura 4:Flauta

3. Percusión: Aquellos instrumentos capaces de generar un sonido a través de golpes a este mismo.

Figura 5:Batería

Figura 6:Bongos

Después de esta breve introducción, si tomamos un tono más científico, el matemático Gottfried Wilhelm Leibniz solía definir la música como un ejercicio inconsciente en la Aritmética. De está expresión se ha derivado la frase “hay matemáticas en la música porque cuando se abre una partitura ésta llena de números” [20]. Esta afirmación quizás se podría justificar sobre la base de que el músico intérprete cuenta los tiempos del compás cuando comienza a estudiar una obra pero después de un tiempo de tocarla, ya no está contando conscientemente sino que deja fluir la “magia” de la música. Sin embargo casi todos “elementos externos” de la música se definen numéricamente:

12 notas por octava, compás de ¾, 5 líneas en el pentagrama, n decibelios, semitono de raíz duodécima de dos, altura (frecuencia fundamental) de 440 Hz… etc. Todos estos números son los que nos permiten caracterizar los sonidos y sus propiedades:

1. Frecuencia fundamental o pitch:

La frecuencia fundamental (f_0) o altura, es la frecuencia más baja a la que resuena un objeto vibrante, por ejemplo un violín. En los instrumentos de música de cuerda por ejemplo, normalmente vibran a frecuencias armónicas de la frecuencia fundamental. Estás frecuencias armónicas no son más que múltiplos enteros de la frecuencia fundamental que se encargan de dar al instrumento su sonido particular, es lo que hace que un sonido en una guitarra suene de forma diferente para la misma frecuencia en una trompeta.

Por ejemplo si se analizase el espectrograma del sonido de un violín, se obtendría algo similar a la figura 7:

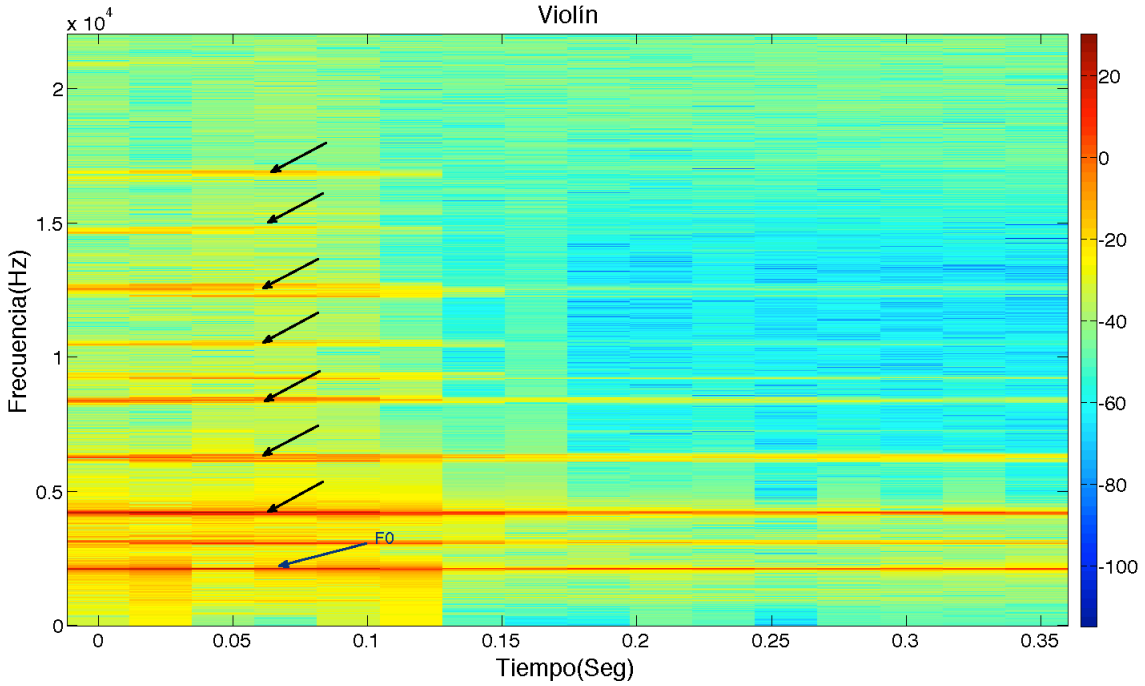

Figura 7: Espectrograma de violín de frecuencia fundamental 2100Hz

Donde la primera flecha (azul) indica la frecuencia fundamental entorno a 2100 Hz, y todas las demás indican sus armónicos (2), notar que la frecuencia fundamental(1) no tiene por que ser la de mayor intensidad, de hecho en la figura 7 podemos ver como el armónico entorno a 4200 Hz, tiene mayor intensidad que la sinusoide de 2100 Hz.

Matemáticamente se define la frecuencia fundamental como la inversa del período fundamental (T_0), es decir el tiempo que tardaría la cuerda en dar una oscilación completa:

$$f_0 = \frac{1}{T_0} \tag{1}$$

$$X_n(\mathrm{n}f_0)\ donde\ n \in \mathbb{Z} \tag{2}$$

En la figura 8 podemos ver el efecto de los armónicos a través de la vibración de una cuerda:

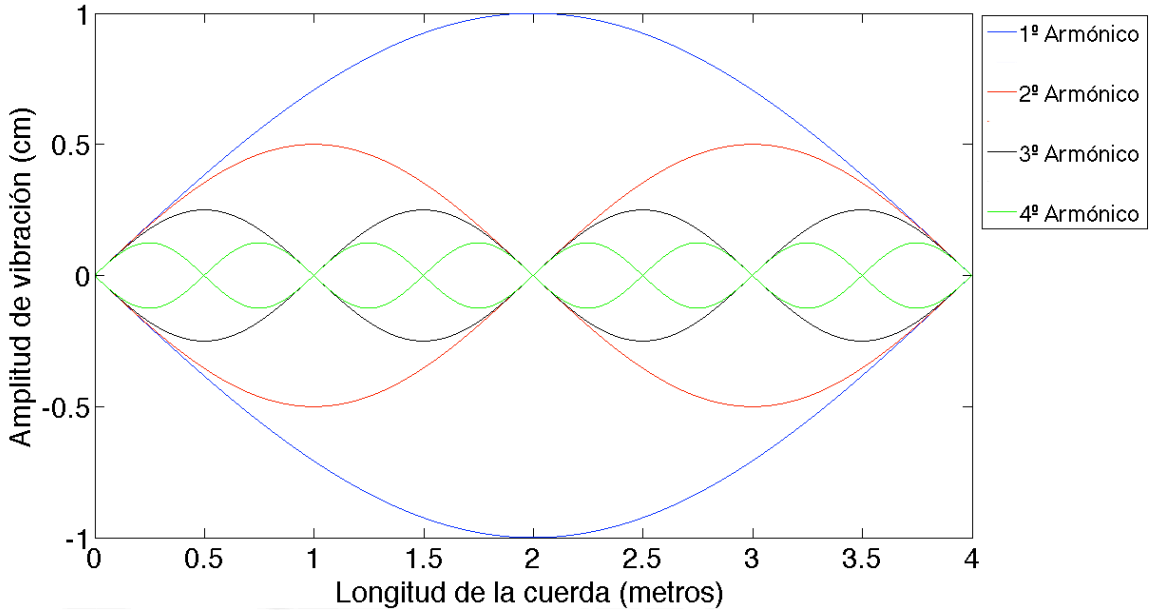

Figura 8: Vibración de una cuerda

Se puede apreciar como el 1º armónico representa la frecuencia de vibración del fundamental y como se producen otras vibraciones en múltiplos de esta frecuencia. Los puntos nulos, son conocidos como nodos, y son puntos en los que no se produce vibración alguna de las cuerdas.

2. Intensidad o sonoridad:

La sonoridad es una medida de la intensidad con la que un sonido es percibido por el oído humano, permitiéndonos ordenar sonidos en una escala de mayor intensidad a menor. La unidad que mide la sonoridad es el fonio.

El fon (o fonio) está definido arbitrariamente como la sonoridad de un sonido senoidal de 1 Khz, con un nivel de presión sonora (intensidad) de 0 dB_{SPL}:

$$S = 10log_{10}(\frac{I}{I_o}) \quad (3)$$

Las primeras curvas de igual sonoridad fueron establecidas por Munson y Fletcher en 1930.

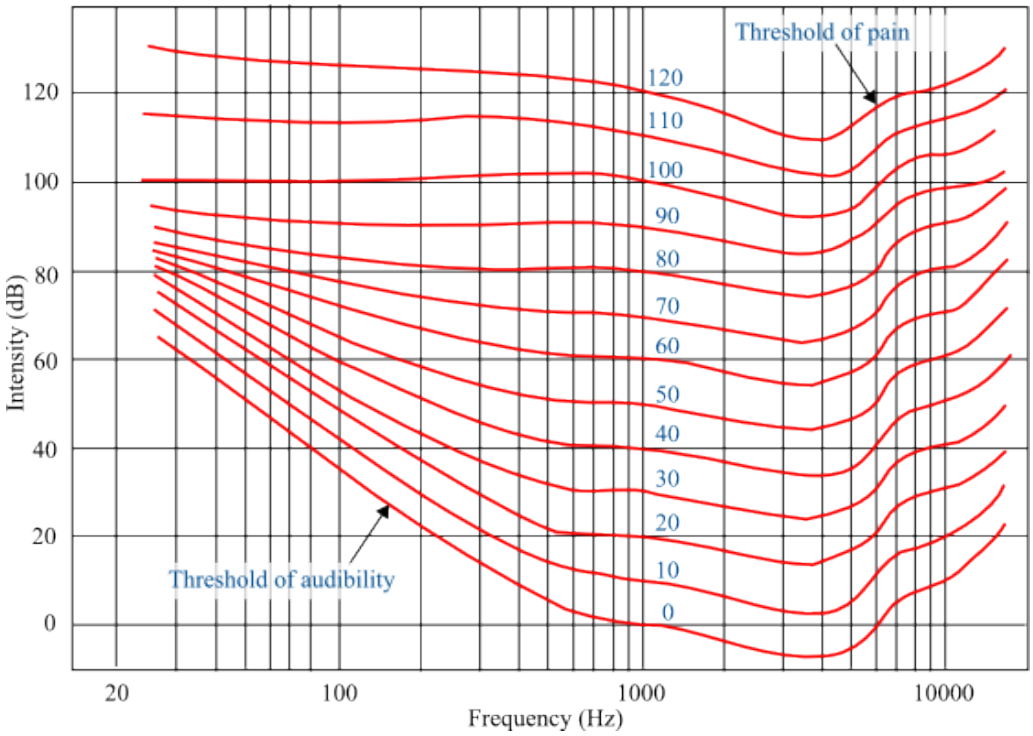

Figura 9: Curvas isofónicas[21]

3. Duración:

Nos informa del espacio temporal que ocupa el sonido desde su aparición hasta su extinción. Este parámetro se relaciona con el ritmo.

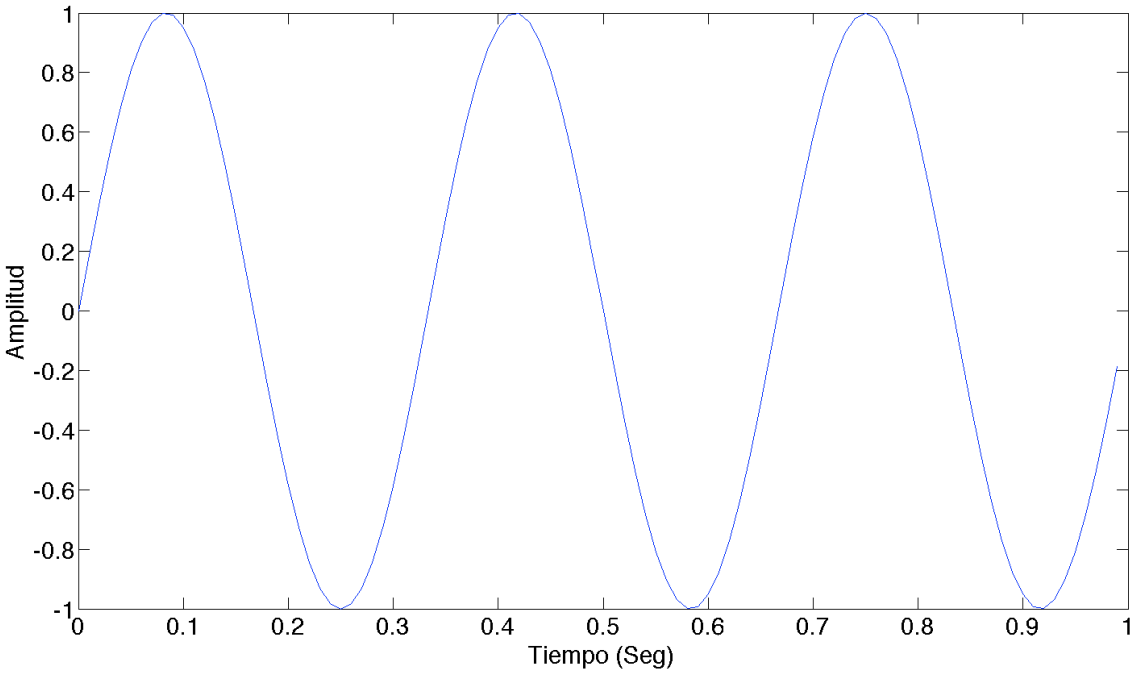

Figura 10: Señal sinusoidal de 3Hz de duración 1 segundo

4. El timbre:

Es la característica o atributo que identifica de forma inequívoca a un sonido con igual sonoridad, altura y duración, pero de procedencia diferente. En definitiva nos permite identificar la fuente sonora de la que proviene un sonido.

Podemos hablar de timbre a tres niveles:

1. General: Cuando distingue elementos de diferentes clases, como por ejemplo una guitarra de una flauta.

2. Parcial: Cuando distingue elementos de una misma clase, como por ejemplos diferentes tipos de guitarras.

3. Particular: Cuando distingue las posibilidades de un único elemento dentro de una clase dada, como por ejemplo las diferentes formas de tocar una guitarra pulsando las cuerdas, con o sin púa, golpeando la caja etc.…

Por ejemplo si se analiza el caso parcial, y se supone una guitarra eléctrica y una acústica, con una duración similar de 2,2 segundos, una sonoridad idéntica y un sonido de frecuencia fundamental 527 Hz, al margen del concepto del timbre deberían ser sonidos idénticos, pero lógicamente no lo son:

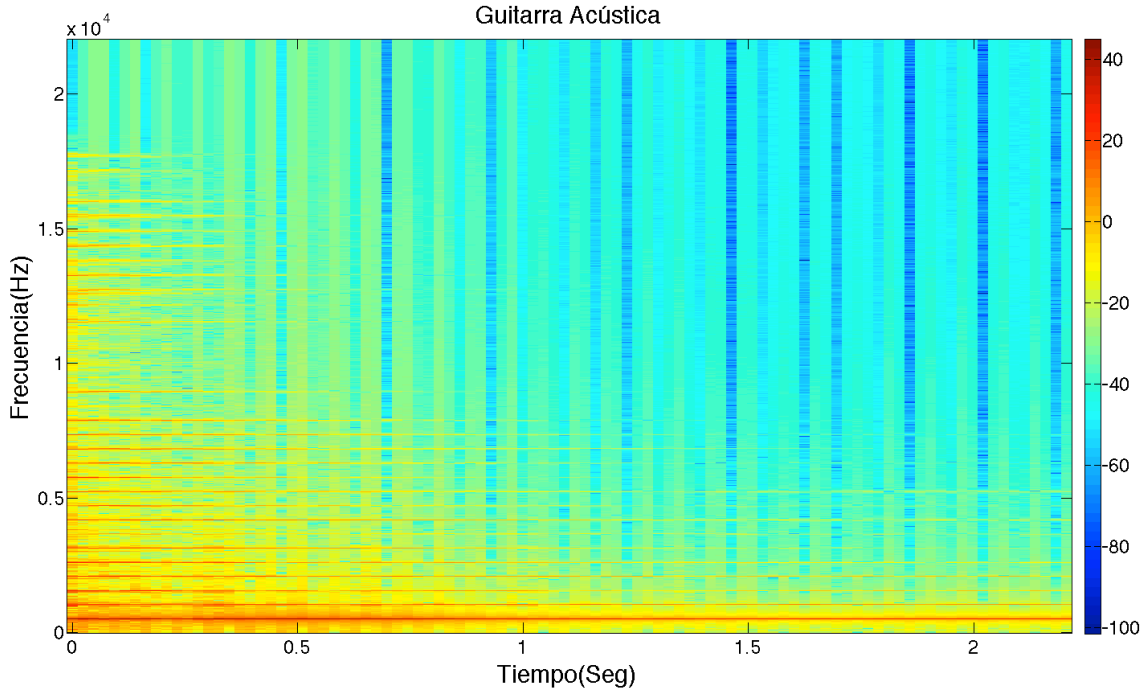

Figura 11:Espectrograma de guitarra acústica a 527 Hz, de duración 2.2 segundos

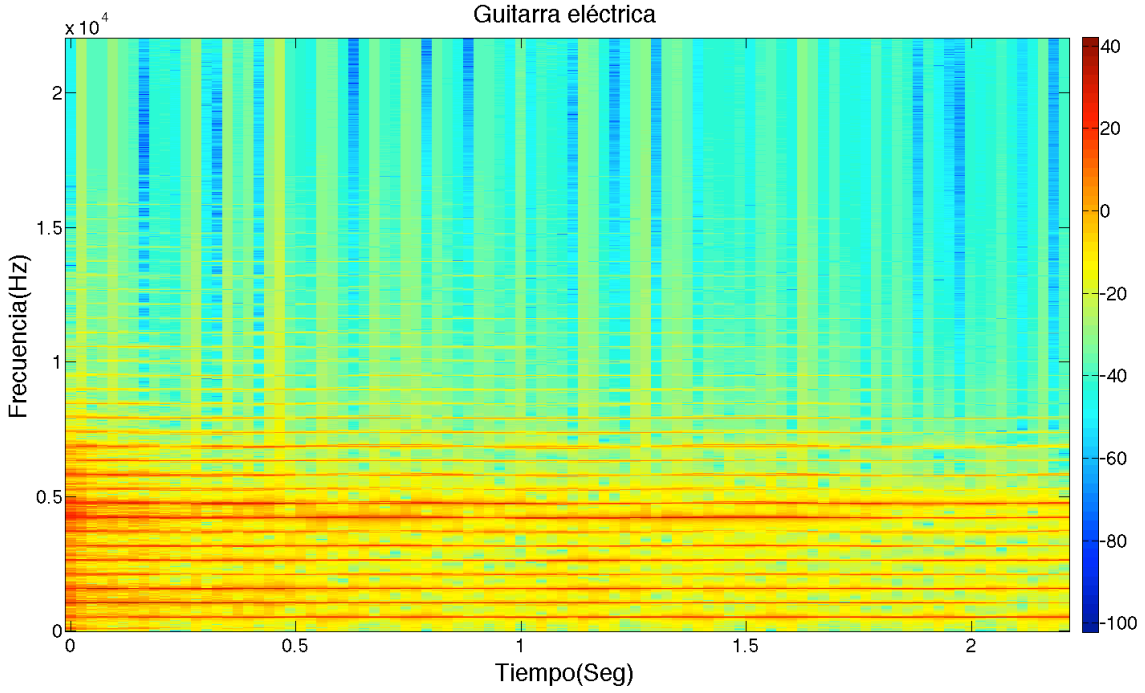

Figura 12:Espectrograma de guitarra eléctrica a 527 Hz, de duración 2.2 segundos

Simplemente con la inspección de la figura 11 y 12, se puede apreciar que son sonidos diferentes, y aunque su frecuencia fundamental es igual, sus armónicos son completamente distintos haciéndolos sonidos completamente diferentes.

Los principales factores que influyen en la determinación del timbre son:

1. La envolvente espectral, es decir la sonoridad relativa de los parciales o armónicos de la señal. Sin analizamos las transformada de Fourier de las señales de la figura 11 y 12:

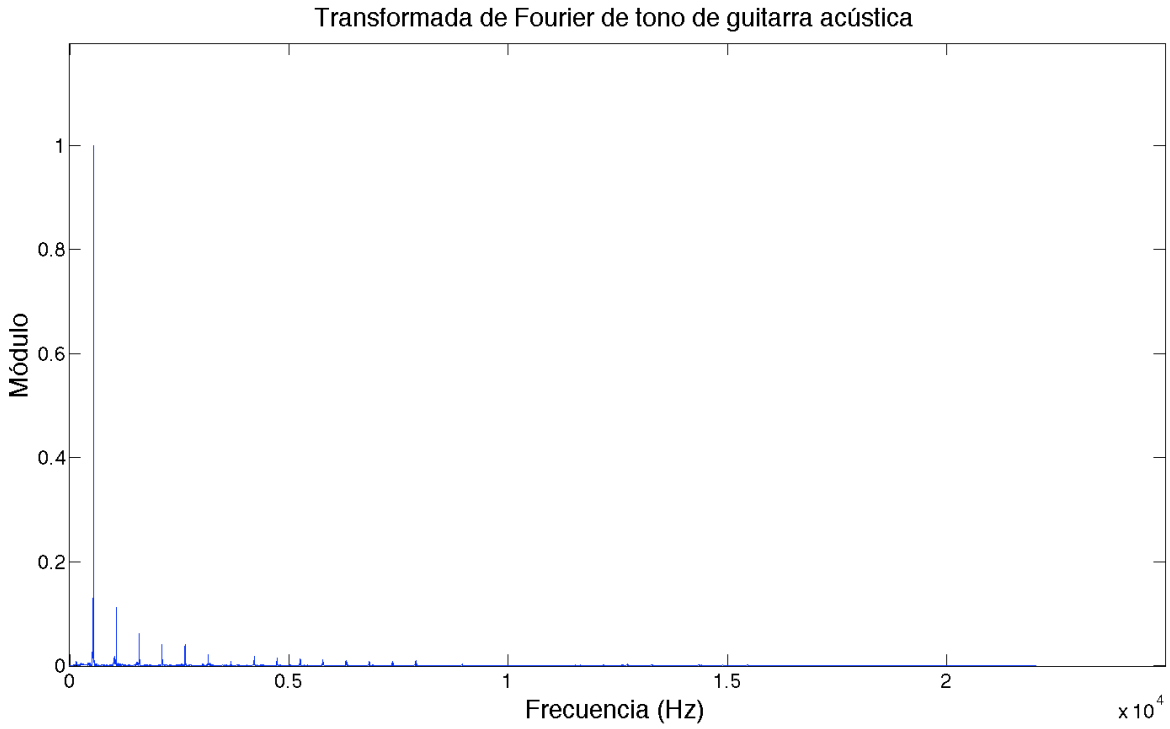

Figura 13: Transformada de Fourier de tono de guitarra acústica a 527 Hz

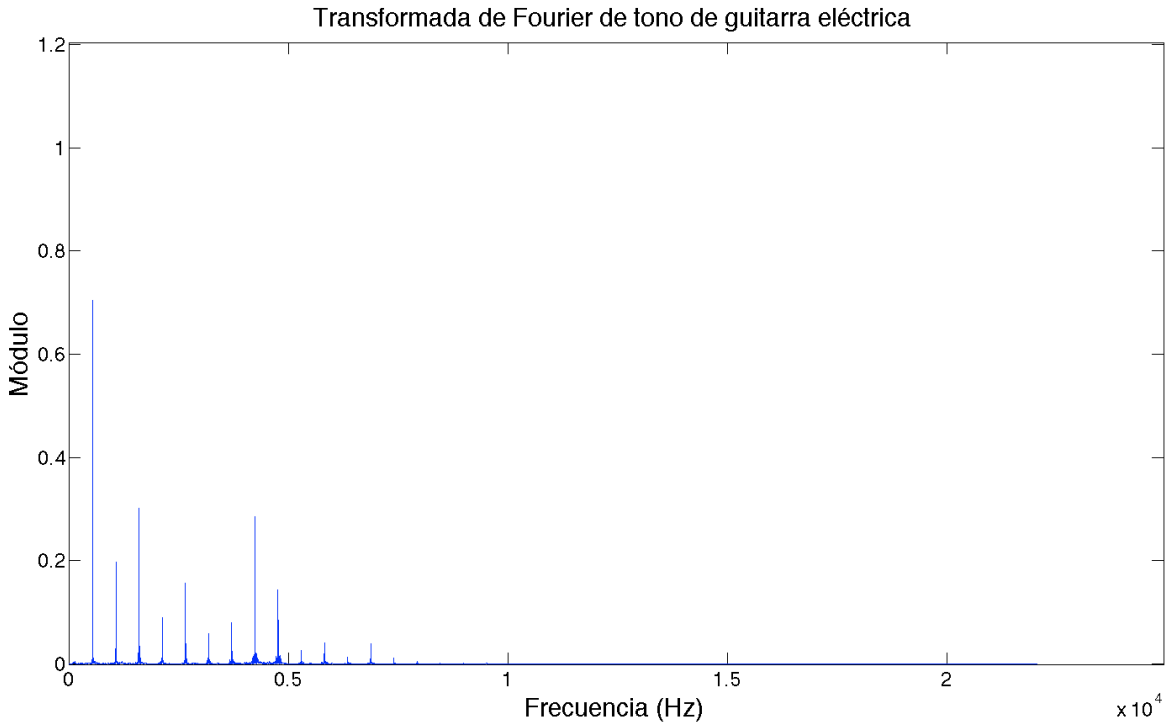

Figura 14: Transformada de Fourier de tono de guitarra eléctrica a 527 Hz

De forma que si se ve la envolvente espectral de cada señal se observa que son completamente diferentes, a pesar de tener la misma frecuencia, misma duración y misma sonoridad, esto se aprecia en la figura 15 y 16:

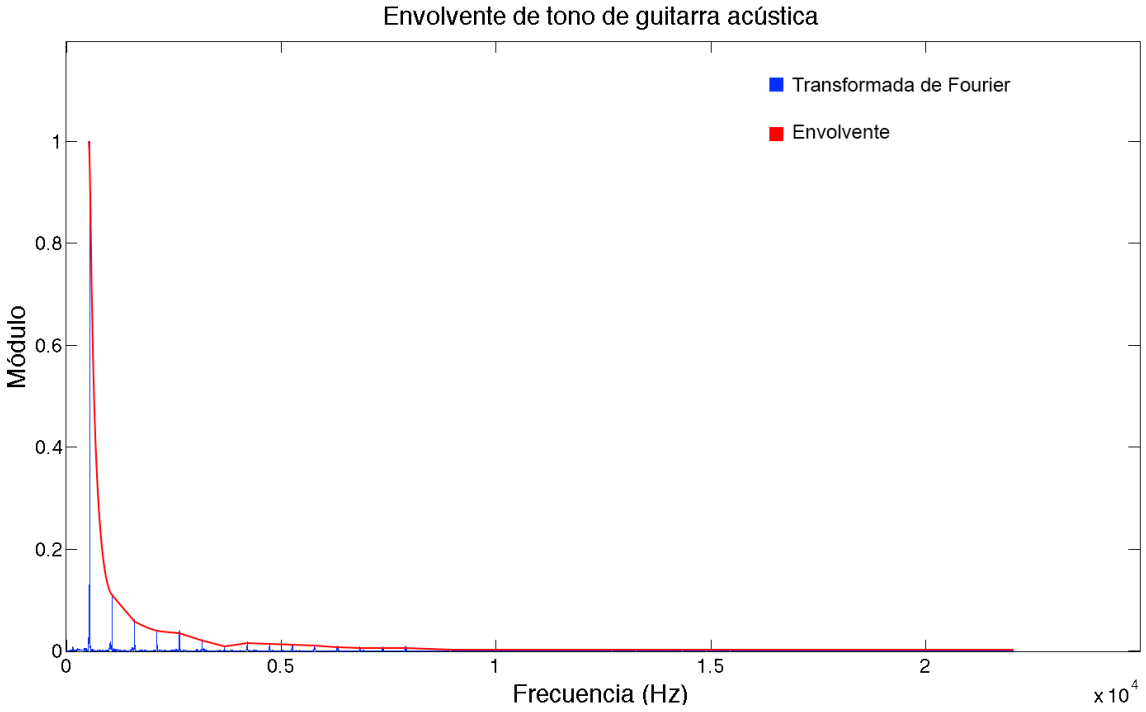

Figura 15: Envolvente de tono de guitarra acústica

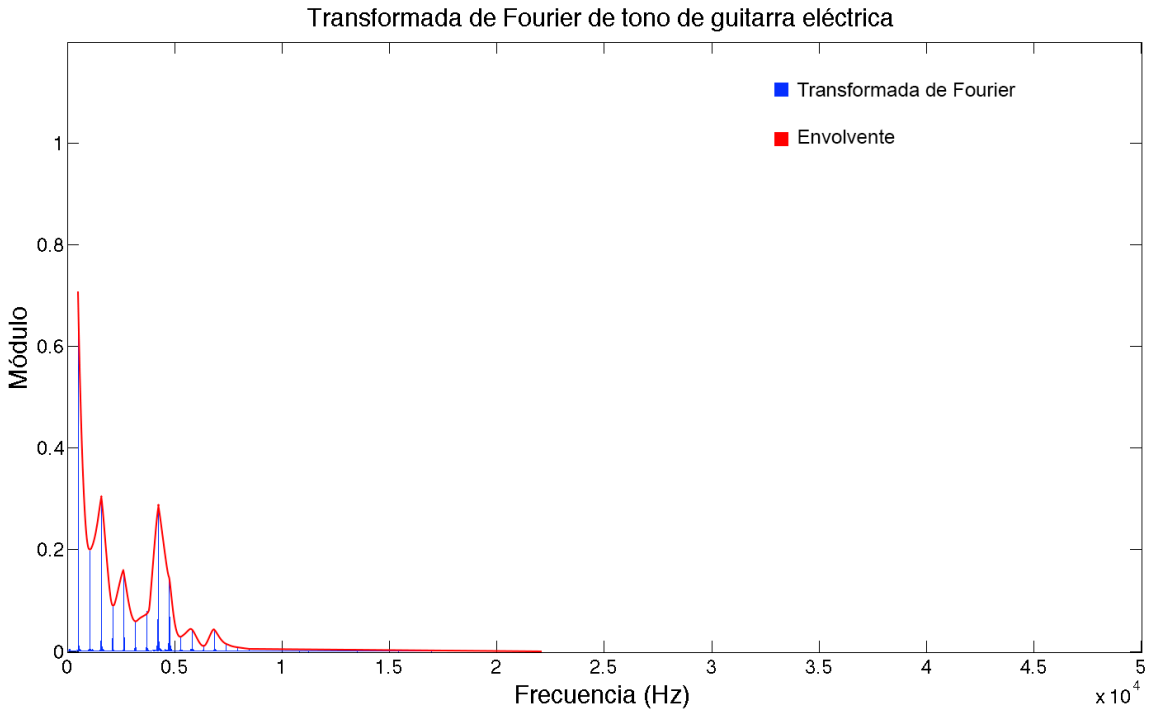

Figura 16: Envolvente de tono de guitarra eléctrica

2. La envolvente dinámica, es la línea imaginaria que une los puntos de amplitud máxima de la onda en el tiempo. En función de la forma de esta envolvente se identifican los diferentes instrumentos, por ejemplo si comparamos la envolvente dinámica de un instrumento de cuerda con uno de percusión, vemos que el de cuerda el ataque es más lento y existe un estado estable mientras que en el de percusión el ataque es muy rápido seguido de una caída sin estado estable.

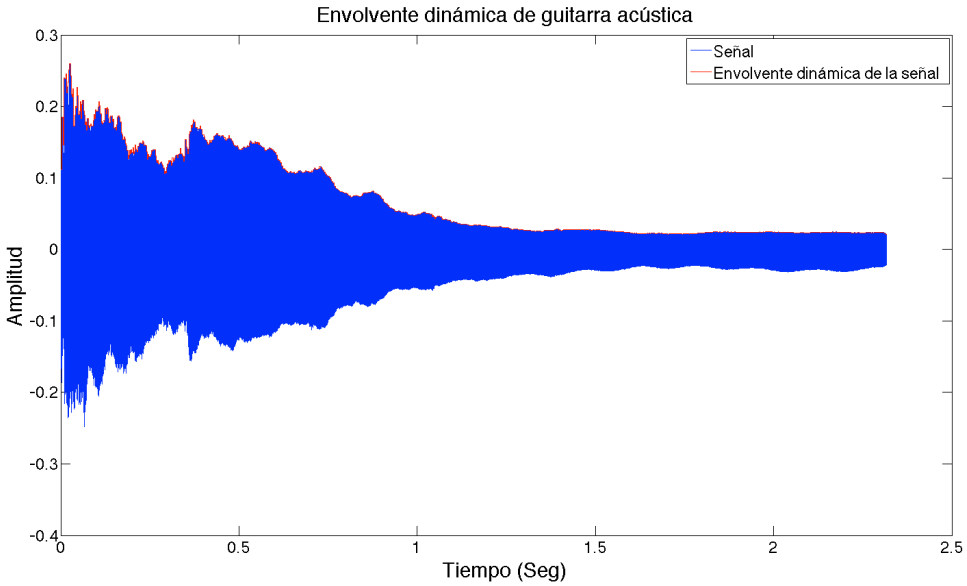

Figura 17: Señal de tono de guitarra acústica con su envolvente dinámica

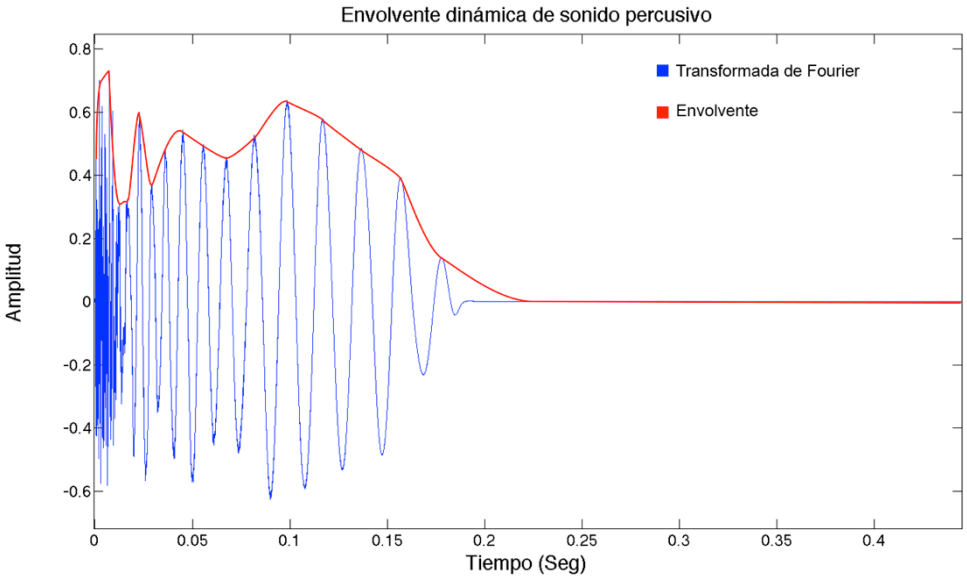

Figura 18: Señal de tono percusivo con su envolvente dinámica.

Para entender la diferencia entre sonidos armónicos y percusivos, basta con inspeccionar los espectrograma, en la figura 12, representábamos el espectrograma de

un sonido armónico y como se observa se trata de sonidos que se extienden en el tiempo, pero de banda relativamente estrecha, sin embargo si observamos el espectrograma de la figura 18:

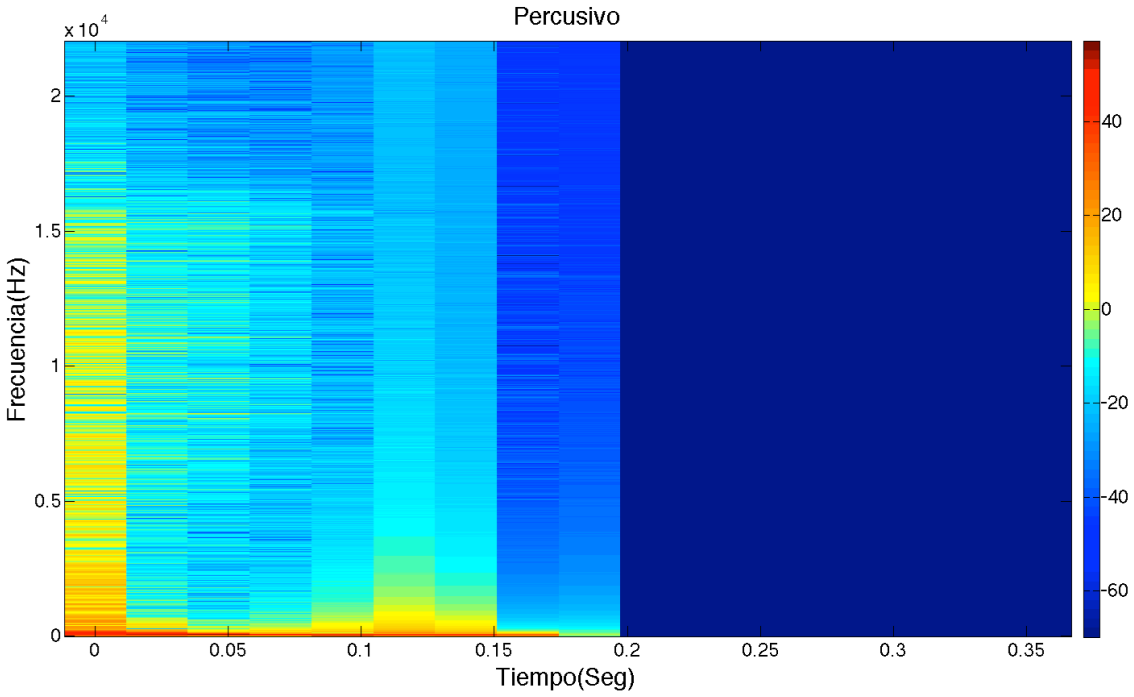

Figura 19:Espectrograma de un sonido percusivo

Vemos que consiste en un sonido de banda ancha, pero de una duración muy corta en el tiempo.

1.2 EL SONIDO

En la sección anterior se ha hablado de cómo la música está compuesta por diferentes tipos de sonidos, pero ¿Qué es el sonido?, desde un punto de vista físico:

"El sonido es una vibración o perturbación que se propaga en un medio elástico (sólido, líquido o gaseoso)" [2].

Como toda perturbación que se propaga por un medio el modelo de análisis es el de una onda. La magnitud física cuya perturbación se propaga en el medio se expresa como una función tanto de la posición del como del tiempo $\psi(\vec{r},t)$. Matemáticamente se dice que dicha función es una onda si verifica la ecuación de ondas:

$$\nabla^2\psi(\vec{r},t) = \frac{1}{v^2}\frac{\delta^2\psi}{\partial t^2}(\vec{r},t) \tag{4}$$

Donde v es la velocidad de propagación de la onda. Una de las perturbaciones que verifica la ecuación 4 se la conoce como el sonido.

Las ondas se pueden clasificar atendiendo a diferentes criterios:

- Atendiendo al medio por el que se propagan:
    1. Ondas mecánicas: Necesitan de un medio elástico (sólido, líquido o gaseoso) para propagarse.
    2. Ondas electromagnéticas: No requieren de un medio para propagarse, se propagan a través del espacio y por tanto a diferencia de las ondas mecánicas pueden propagarse por el vacío.
    3. Ondas gravitacionales: Las ondas gravitacionales son perturbaciones que alteran la geometría misma del espacio-tiempo y aunque es común representarlas viajando en el vacío, técnicamente no podemos afirmar que se desplacen por ningún espacio, sino que en sí mismas son alteraciones del espacio.
- Atendiendo al movimiento de sus partículas:
    1. Onda longitudinal: Las partículas del medio se mueven de forma paralela a la dirección de propagación.
    2. Onda trasversales: Las partículas del medio vibran de forma perpendicular a la dirección de propagación.

En base a la definición del sonido de la página anterior, podemos clasificar el sonido como una onda mecánica, que fluye por el medio como una onda longitudinal de presión.

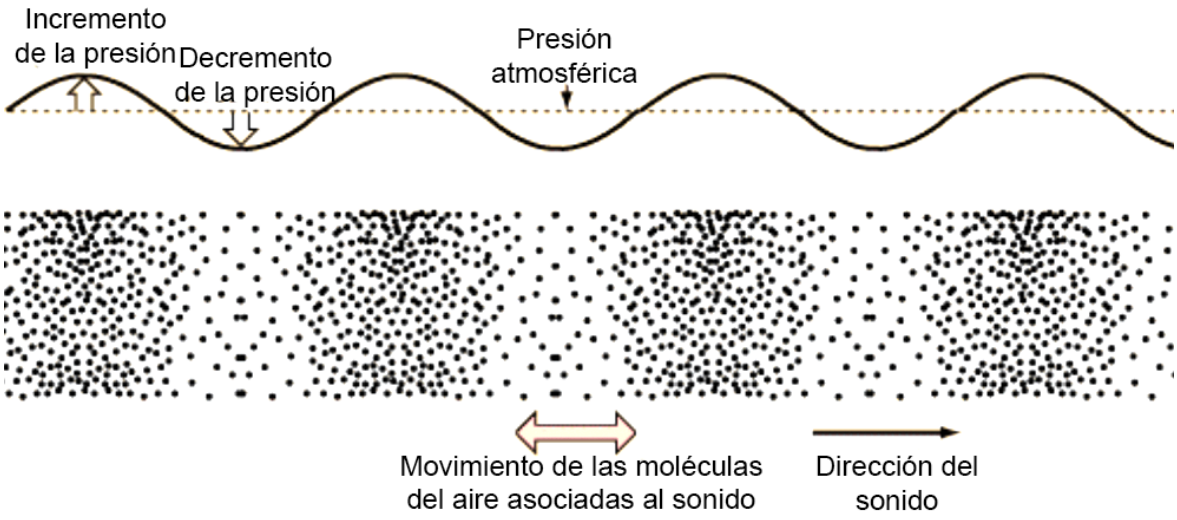

Figura 20: Representación de la variación de presión debida a la propagación del sonido a través de un medio elástico como el aire[2].

En la figura 20 se puede apreciar como varían las partículas de un medio como el aire debido al incremento de la presión por parte de las ondas mecánicas producidas por el sonido. Notar como la vibración se produce en el sentido de propagación como una onda longitudinal.

Como todo movimiento ondulatorio, el sonido puede representarse mediante la transformada de Fourier como una suma de curvas sinusoides. Lo que permite descomponer el sonido en longitud de onda (λ), frecuencia (f) o período (T), amplitud y fase. Esta descomposición simplifica el estudio de sonidos complejos ya que nos permite estudiar cada componente frecuencial independientemente.

Entre los fenómenos de interés que se dan en el sonido, podemos destacar 4 características o efectos[2]:

1. Reverberación: Es el efecto producido en recintos cerrados, por los diversos mecanismos de propagación sonora, principalmente reflexiones múltiples y difusas, que dan lugar a que el sonido tarde un cierto tiempo en "apagarse" aún cuando la fuente sonora haya dejado de emitir. Este tiempo, designado tiempo de reverberación, se define como el tiempo en que el sonido reverberante alcanza un nivel de presión de -60dB respecto al nivel de presión del sonido inicial. El tiempo de reverberación depende de diversos factores como el volumen del recinto, la velocidad del sonido, la absorción en paredes, mobiliario, personas u objetos en el recinto y la constante de atenuación, es un fenómeno de gran importancia en los diseño de estudio de grabación.

2. Directividad: Igual que pasa con las antenas, las fuentes sonoras no suelen radiar de forma omnidireccional y sino que lo hacen de forma direccional, dando unos máximos en una dirección y dando otros más bajos para otra, podemos definirla como:

$$\mathrm{Q} = \frac{(Presión\ axial\ en\ la\ dirección\ de\ máxima\ radiación)^2}{(Presión\ sonora\ media)^2} \tag{5}$$

Cuando la longitud de onda es mucho más grande que la de la fuente sonora, podemos aproximar la radiación como que radia por igual en todas las direcciones.

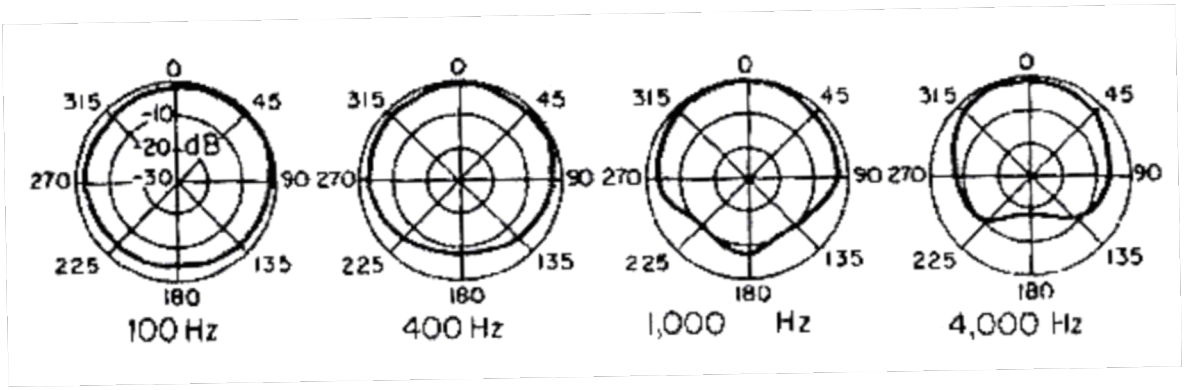

Figura 21: Direccionalidad de la voz a diferentes frecuencias[2]

3. Resonancia: fenómeno que se produce cuando dos cuerpos tienen la misma frecuencia de vibración, uno de los cuales empieza a vibrar al recibir las ondas sonoras emitidas por el otro.

4. Nivel de intensidad sonora(I): Es el flujo de energía sonora que atraviesa una unidad de área, en un punto dado a una dirección perpendicular respecto a la fuente.

$$I = \frac{\rho^2}{Z_A} = \frac{\rho^2}{\rho_0 \, v_s} \; watt/m^2 \tag{6}$$

Donde ρ es la presión, ρ_0 la densidad del medio y v_s la velocidad de propagación. Aunque es más frecuente verlo expresado en unidades logarítmicas (dB) como:

$$I(dB) = 10 log_{10}\left(\frac{I}{I_{ref}}\right) \; dB \tag{7}$$

1.2.2 CLASIFICACIÓN DEL SONIDO

En las secciones anteriores se ha introducido a las definiciones básicas de la música, y de cómo esta se compone por sonidos, así como de la forma de

caracterizarlos. Sin embargo para el estudio de la música comercial es importante distinguir entre otro tipo de caracterización de la música.

1.2.2.1 MONOCANAL Y ESTÉREO

Otra de las formas de caracterizar un sonido hace referencia a si contiene o no información espacial.

Los sistemas monocanales, son sistemas en los que el sonido está definido por un único canal, donde todas las señales de audio se suman, originando un sonido semejante al que se escucharía con un solo oído. Lógicamente, es un sistema que carece de información espacial y aunque es un sistema que ha sido sustituido en su gran mayoría, todavía sigue empleándose en comunicaciones telefónicas, y algunos sistemas de radio, debido a que al transmitir en sonido monoaural se posibilita tener una mayor fuerza de la señal frente a la estereofónica de la misma potencia, cubriendo así una mayor área.

Por el contrario los sistemas estéreo, consiste en sistemas de dos o más canales, aunque en la música comercial se emplean principalmente dos canales, el izquierdo y el derecho, es cierto que en sistemas de alta calidad de sonido para el entretenimiento, como pueden ser cines, teatros y sistemas de cine domésticos, si emplean tres canales de audio. La principal diferencia que presentan frente a los sistemas de un solo canal es que nos permiten obtener información espacial. A día de hoy es el sistema que se emplea en toda la industria de la música comercial y será una de las bases para alguno de los sistemas que se emplean en este trabajo.

1.2.3 LA VOZ

Como se ha comentado en la naturaleza, existen una variedad casi infinita de sonidos. Uno de los de mayor importancia de estudio para este trabajo, son los sonidos capaces de ser reproducido por la voz humana, concretamente la voz cantada (Singing Voice).

Sin adentrarse mucho en la parte fisiológica de la voz, se puede decir que la voz como todo sonido, se ha visto que necesita de un medio elástico para propagarse en este caso es el aire que generan los pulmones, por otro lado el ser humano dispone de la “cuerdas vocales” aunque realmente no son cuerdas sino pliegos vocales, donde los extremos están unido a los músculos de la laringe, y los bordes internos son libre, permitiendo regular el volumen de aire que dejan pasar.

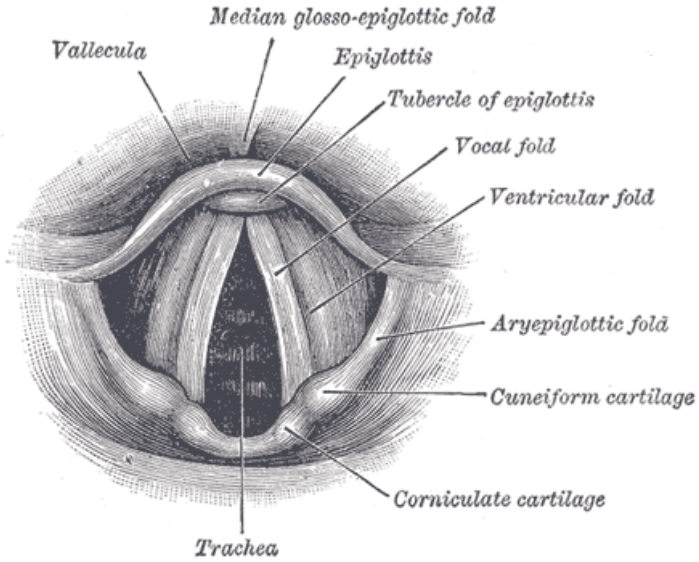

Figura 22:Ilustración de cuerdas vocales[22]

Aunque realmente sean pliegos, y consista en un mecanismo que deja o no pasar una mayor cantidad de aire, la comparación con unas cuerdas es interesante, ya que como se vio en la sección 1, el sonido generado por las cuerdas de una guitarra, generaba una serie de componentes armónicas en torno a la frecuencia fundamental. Una de las características de la voz es su alto contenido armónico, es por ello que resulta propio el nombre de “cuerdas vocales”.

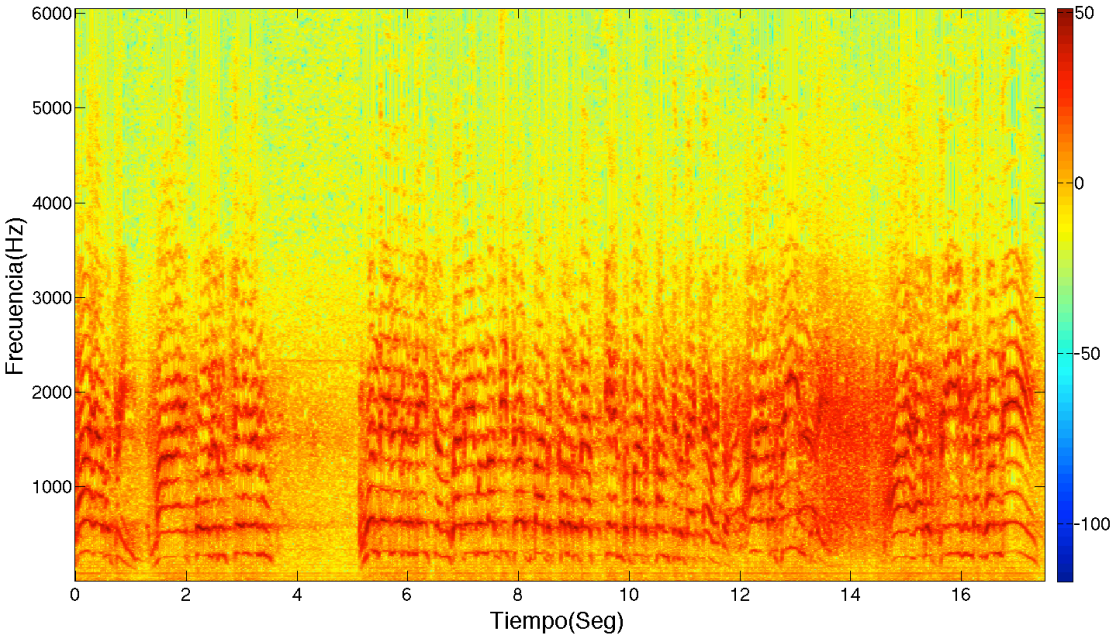

Figura 23:Espectrograma de un fragmento del discurso "I have a dream" de Martin Luther King [23]

En la figura 23, se puede apreciar el alto contenido armónico de la voz, esto se debe principalmente a la resonancia del tracto vocal.

Una segunda característica de la voz son los formantes[3], es el pico de intensidad en el espectro de un sonido, consiste en la concentración de energía que se da a una determinada frecuencia. En otras palabras son bandas de frecuencia en la que se concentra la mayor parte de la energía sonora de un sonido.

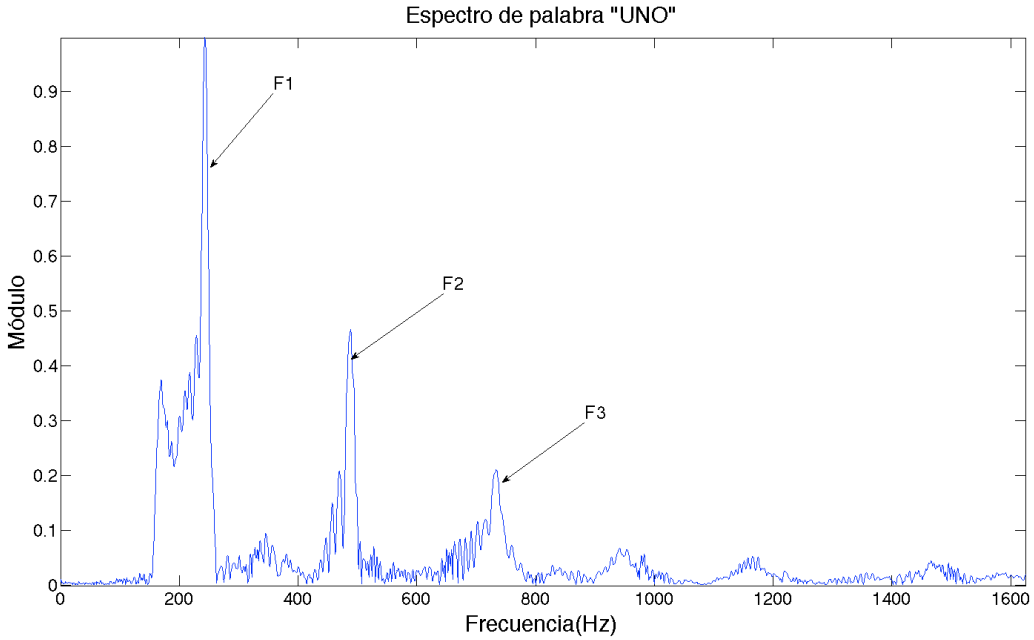

Figura 24: Espectro de la palabra "UNO" donde se identifican los 3 formantes

Esta características es de gran importancia a la hora de identificar sonidos en el habla, principalmente vocales, como se puede ver en la figura 23 se aprecian 3 tipos de sonidos en torno a 3 frecuencias, correspondientes a /U/, /N/ y /O/. Cada uno de estos sonidos tiene varios formantes, que se aprecian mejor en el espectrograma:

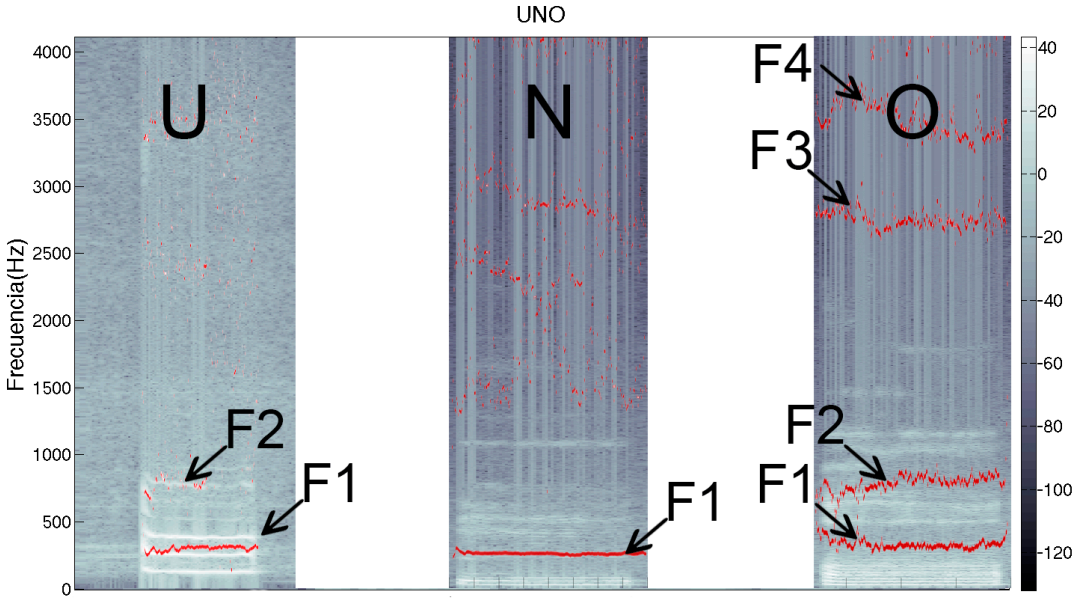

Figura 25: Estimación de los formantes hasta 4 Khz para los fonemas /U/, /N/ y /O/

En la figura 24, podemos ver una estimación de la localización de los fonemas /U/, /N/ y /O/, la línea roja de puntos representa la localización de los formantes, como se ha comentado las vocales son más sencillas de localizar. La dispersión de los puntos indica la veracidad de la estimación siendo continuos para una buena estimación y muy dispersos para una mala estimación. Vemos como los resultados de las dos vocales coinciden con los datos establecidos en la tabla 1, de la siguiente página. Si nos fijamos en el formante 1 tanto de la U como de la O, podemos ver que se encuentran en las frecuencias asociadas al espectro de la figura 23, siendo el orden de aparición /U/, /O/ y /N/, aunque /N/ la estimación se encuentra muy por debajo de la de la figura, lo que indica que la estimación de la consonante no es correcta. Las estimaciones se han realizado con el software libre PRAAT [27].

Existen tantos formantes como resonadores posee el tracto vocal. Sin embargo se considera que sólo los tres primeros, asociados a la cavidad oral, bucal y nasal respectivamente proporcionan la suficiente información para poder diferenciar tipos de sonidos. Si nos centramos en el Singing Voice, hay que destacar un cuarto formante [3], conocido como *"Singing formant",* localizado en frecuencias en torno a 3000 Hz. Este formante ayuda al cantante a permanecer por encima del acompañamiento instrumental.

En la tabla 1 se muestran las frecuencias para el primer formante de las diferentes vocales:

Formantes vocálicos	
Vocal	Región principal formántica
/u/	200 a 400 Hz
/o/	400 a 600 Hz
/a/	800 a 1200 Hz
/e/	400 a 600 y 2200 a 2600 Hz
/i/	200 a 400 y 3000 a 3500 Hz

Tabla 1:Formantes vocálicos[25]

La tercera característica interesante es el rango de frecuencias en el que la voz trabaja, en términos de frecuencia fundamental, en la figura 22 los armónicos alcanzan frecuencias de 1000 y 2000 Hz, pero realmente el fundamental está a frecuencias muy inferiores.

En la figura 26, podemos ver como el rango general para la frecuencia fundamental de los hombres está comprendido entre 100 y 200 Hz, mientras que el de las mujeres entre 150 y 300 Hz. Esto se debe principalmente a la diferencia en la longitud de los pliegos del sistema vocal entre 17 y 25 mm para los hombres y 12.5 a 17.5 mm para las mujeres.

Límite de edades (hombres)	*Nº*	*Frecuencia fundamental media*	*Investigadores*
20-29	175	120	Hollien y Shipp (1972)
	27	119	Hanley (1951)
	157	128	Hollien y Jackson (1973)
	24	132	Philhour (1948)
	6	132	Pronovost (1942)
	103	138	Majewski et al. (1972)
30-39	175	112	Hollien y Shipp (1972)
40-49	175	107	Hollien y Shipp (1972)
	39	113	Mysak (1959)
50-59	175	118	Hollien y Shipp (1972)
60-69	175	112	Hollien y Shipp (1972)
70-79	175	132	Hollien y Shipp (1972)
	39	124	Mysak (1959)
80-89	175	146	Hollien y Shipp (1972)
	39	141	Mysak (1959)

Límite de edades (mujeres)	*Nº*	*Frecuencia fundamental media*
20-29	10	227
30-29	10	214
40-49	10	214
50-59	10	214
60-69	10	209
70-79	10	206
80-90	10	197

Figura 26: Resultado de estudios de la frecuencia fundamental de la voz a diferentes edades para hombres y mujeres[24]

Por último otra características de la voz que también se puede apreciar en la figura 23, es el vibrato y el trémolo [3]. El vibrato vocal físicamente corresponde a una modulación periódica sinusoidal de frecuencia fundamental de la fonación. Es decir el sonido que en el espectrograma debería ser una única línea recta y continua fija en la frecuencia fundamental del sonido en cuestión, pasa a modularse como una sinusoide de frecuencia igual a la frecuencia fundamental que el sonido que representa. Acústicamente el vibrato hace que la voz suene agradable, viva, excitante, en un tono menos plana, en pocas palabras es lo que da a la música “sentimiento”. Por otro lado el trémolo es una

modulación similar a la del vibrato pero en el nivel de sonoridad de la señal, dicho de otra forma el sonido pasa a ser más fuerte y después más suave, más fuerte y después más suave, así sucesivamente.

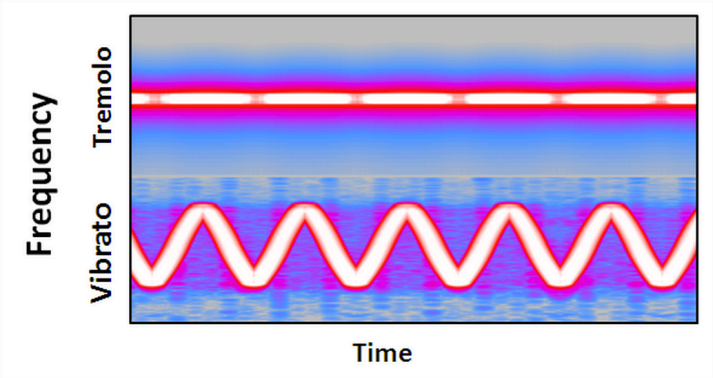

Figura 27: Representación del efecto de trémolo y el vibrato

En la figura 27 se puede observar la representación de ambos efectos, y como estos quedan modulados en forma de sinusoide, el vibrato lo hace en frecuencia (modulación FM) mientras que el trémolo en amplitud (modulación AM). Mientras que en la figura 28 podemos ver el efecto en un caso real no ideal. Observándose las oscilaciones en frecuencia entorno a un sonido así como la ligera variación de amplitud.

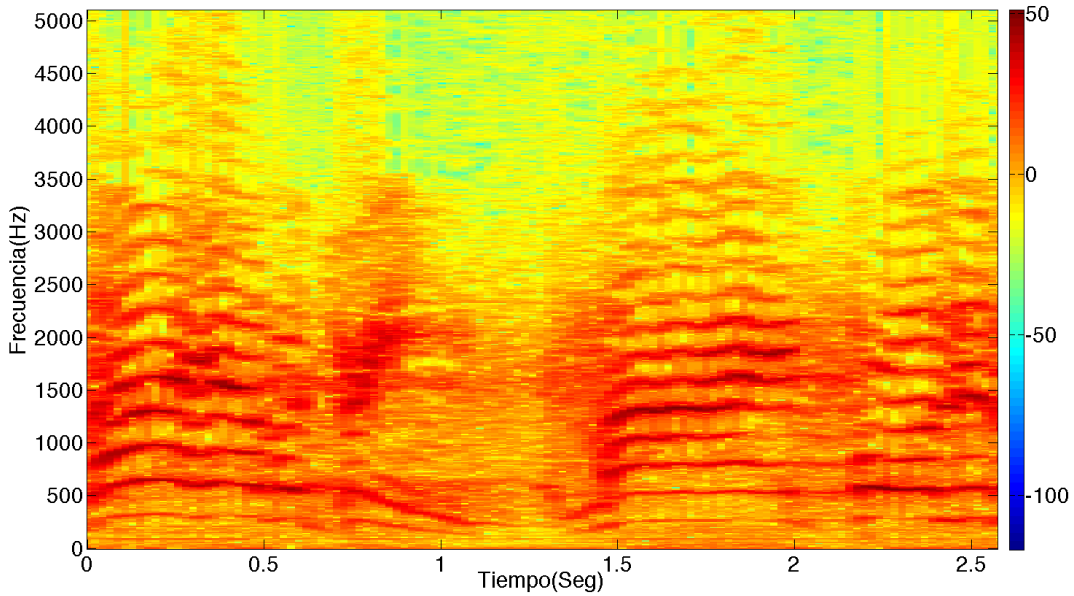

Figura 28:Ampliación de la figura 22, donde se representa un caso real de efectos de trémolo y vibrato

Como se ve ambos efectos están íntimamente relacionados y se pueden describir cada uno por dos características: sus frecuencias (rango del vibrato/trémolo) y sus amplitudes (extensión del vibrato/trémolo). Para la voz el rango de oscilación está

alrededor de 6 Hz y se incrementa exponencialmente durante la duración del sonido. La media de extensión es de 0.6 a 2 semitonos para el cantante y de 0.2 a 0.35 semitonos para instrumentos de cuerda. Donde un semitono se define como cada una de las dos partes, iguales o desiguales, en que se divide el intervalo de un tono.

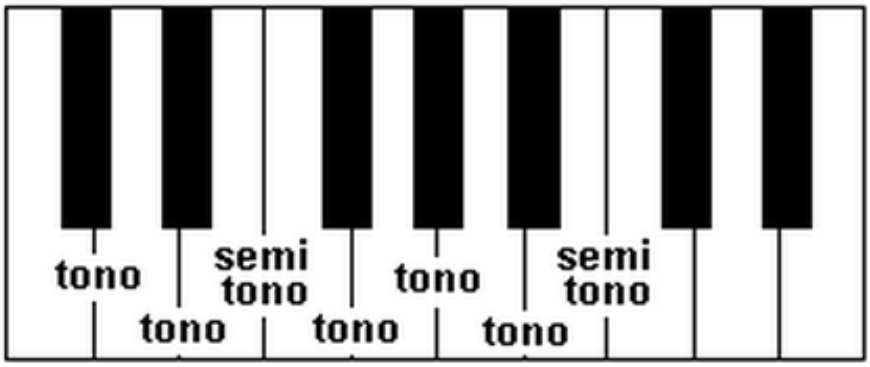

Figura 29: Representación de los semitonos contenidos en una octava[26]

La principal ventaja que ofrece esta característica [3] es que debido a los aspectos de la forma en la que se produce la voz humana, contiene ambos efecto simultáneamente, esto se aprecia en la figura 28, sin embargo son muy poco los instrumentos que pueden producir simultáneamente ambos efectos, generalmente instrumentos de viento se modulan como una modulación AM, mientras que los de cuerda suelen hacerlo mediante una modulación FM, esto se puede ver en la figura 30 y 31 respectivamente.

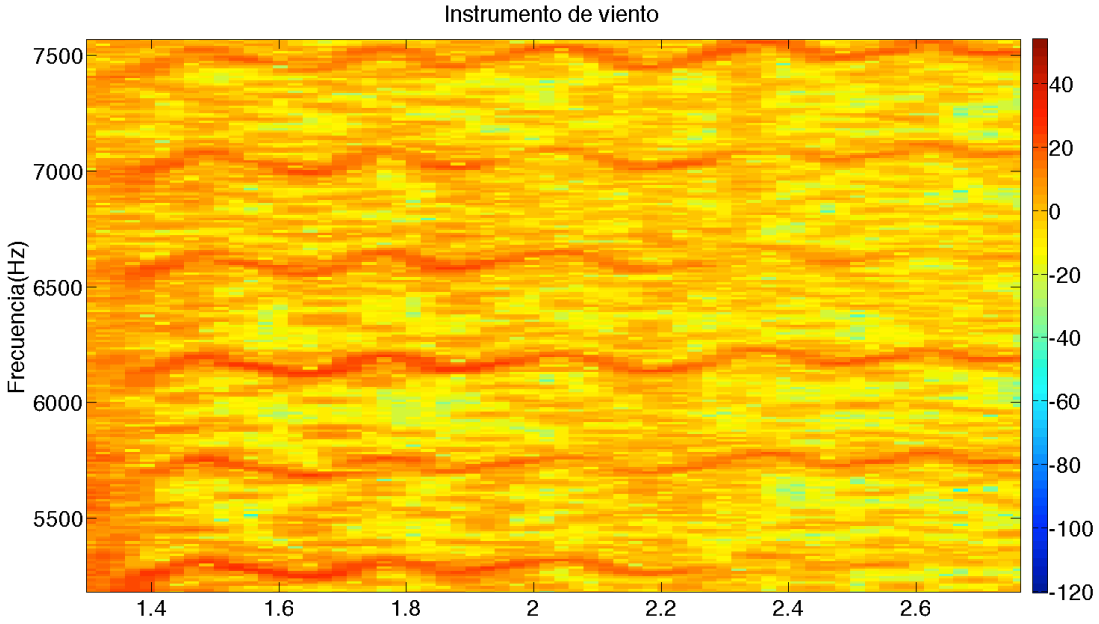

Figura 30: Instrumento de viento donde se aprecia la modulación AM

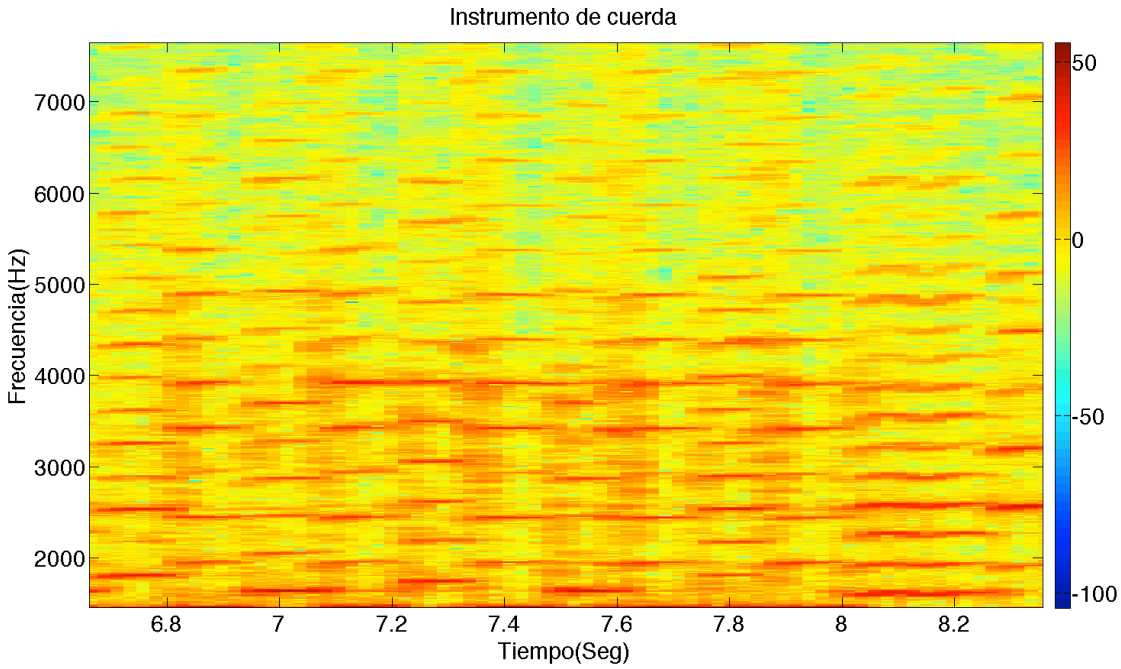

Figura 31: Instrumento de cuerda donde se aprecia la modulación FM

1.2.3.1 FONEMAS SONOROS Y SORDOS

En la sección anterior se ha hablado mucho acerca de la generación de sonidos por parte de los seres humanos, mediante la voz. Se han explicado términos y características de la misma, pero existe otras características asociadas a la fisiología de la voz, que permite otro tipo de agrupaciones diferentes, pero antes de ello es necesario introducir el concepto de fonema:

"Cada una de las unidades fonológicas mínimas que en el sistema de una lengua pueden oponerse a otras en contraste significativo. [28]"

En otras palabras las unidades teóricas básicas postuladas para estudiar el nivel fónico-fonológico de una lengua humana.

La forma de clasificar los fonemas atiende principalmente a la fisiología de la voz, es decir a la forma y posición de los músculos para emitir un sonido, pero previo a esa clasificación podemos distinguir tres grupos más genéricos:

1. Series: Poseen un mismo rasgo particular común (p/t/k).

2. Órdenes: Cuando el rasgo común es el punto de articulación (p/b/m).

3. Correlaciones: Cuando dos series se diferencian por medio de un rasgo distintivo (p/t/k) vs (b/d/g).

Una vez realizada esta agrupación o tipología genérica, podemos clasificar los fonemas como en función de las partes implicadas en la generación del sonido:

1. Vocales/Consonantes: No existe contrición ni interrupción de la corriente de aire en los vocales mientras que en los consonantes existe constricción total o parcial de la cavidad oral.

2. Orales/Nasales: En las orales el velo del paladar esta pegado a la pared de la faringe mientras que en las nasales se encuentra caído y el aire puede salir por las cavidades nasales.

3. El modo de articulación: Oclusiva, fricativa, aproximante, nasal, lateral y vibrante.

4. Órdenes: El lugar de articulación (bilabial, interdental, alveolar…).

5. Series: sonoridad, O bien existe presencia de vibraciones perceptibles en las cuerdas vocales (sonoras) o no existe vibraciones (sordas).

	Bilabial		**Labiodental**		**Interdental**		**Dental**		**Alveolar**		**Palatal**		**Velar**	
	D	*S*	*D*	*S*	*D*	*S*	*D*	*S*	*D*	*S*	*D*	*S*	*D*	*S*
Oclusivas	p	b					t	d					k	g
Nasal		m					s	n						
Vibrante Simple														
Vibrante Multiple								r						
Fricativa			f	v			s	z						
Lateral								l						

TABLA 2: Clasificación de fonemas D(Sordos) y S(Sonoros) [Ref. 5]

Un tipo de fonema de especial interés para este trabajo, son los fonemas sordos, como /t/, /p/ y /k/, principalmente por la forma que presenta su espectrograma:

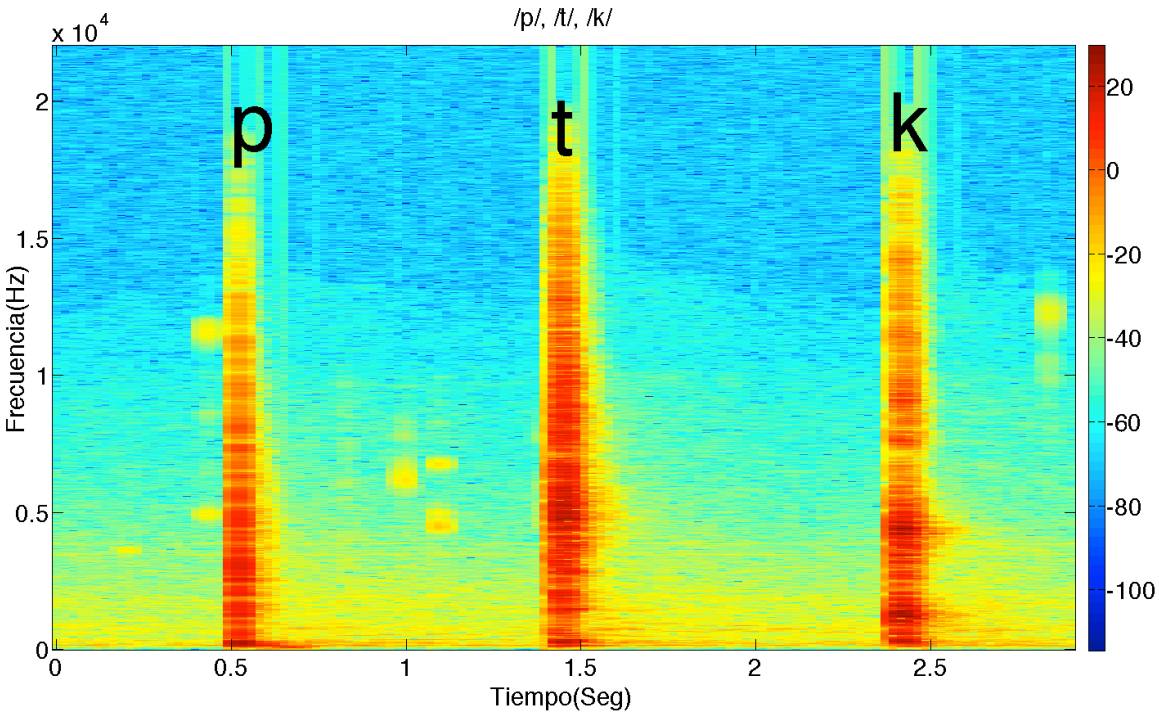

Figura 32: Espectrograma de fonemas sordos

En la figura 32 podemos observar el espectrograma de 3 fonemas sordos, y como se puede ver se trata de fonemas que se producen en breves instantes de tiempo y tienen un gran ancho de banda, muy parecido al espectrograma de la figura 19, es decir son sonidos muy similares a los percusivos desde un punto de vista del espectrograma, lo que hace que identificarlos sea un tanto complejo. Por el contrario los fonemas sonoros, son más parecidos a los armónicos:

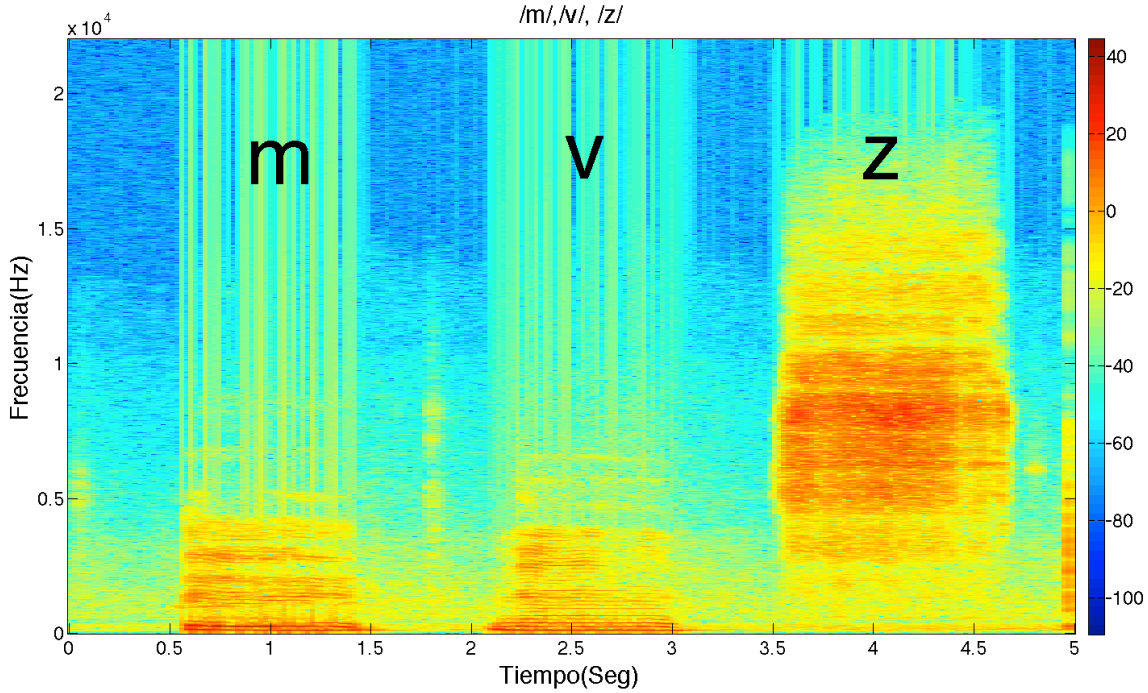

Figura 33: Espectrograma de fonemas sonoros

A pesar de su estructura similar a los armónicos como se muestra en la figura 33, no resultan tan complicado de diferenciar de los armónicos como los fonemas sordos de los percusivos.

2. OBJETIVOS

Objetivo principal

El objetivo principal de este trabajo fin de grado, es el desarrollo de un software capaz de separar pistas de música comercial en sus compontes armónico-percusivo y vocal, apoyándose en el software MatLab, en un principio no estaría pensado para trabajar en tiempo real. Una vez realizado comprobar la calidad de los resultados obtenidos y compararlos con sistemas actuales que se encuentran en funcionamiento.

Objetivo específico

Para cumplir el objetivo principal, se han desarrollado una serie de objetivos específicos o sub-objetivos:

1. Realizar una primera separación instrumental-vocal, empleando para ello algunos algoritmos y parte de la teoría explicada en las secciones anteriores.

2. Realizar una separación exclusiva para la voz y los posibles restos de instrumental que puedan quedar en la misma.

3. Realizar una separación armónico-percusivo al conjunto de música instrumental obtenida de los otros procesos.

4. Realizar una medición del nivel de calidad de los resultados a través de medidas subjetivas (test de Mushra) a un número determinado de personas.

5. Procesar los resultados obtenidos a través de unas funciones matemáticas para obtener valores numéricos y cuantificados (SIR,SDR) y poder compararlos con otros algoritmos ya implementados que se encuentran en la actualidad.

3. ESTADO DEL ARTE

La separación de voz cantada (Singing Voice Separation (SVS)), puede ser definida como el proceso de extraer elementos vocales de una pista grabada de una canción [10]. La importancia de este campo reside principalmente en la posibilidad de automatizar aplicaciones de música capaces de recuperar información de la canción, como reconocimiento e la letra, identificación del cantante o simplemente la extracción de la parte instrumental de una canción.

Se trata de un área que lleva muchos años de estudio, y distingue claramente varios campos de trabajo en función de las técnicas utilizadas:

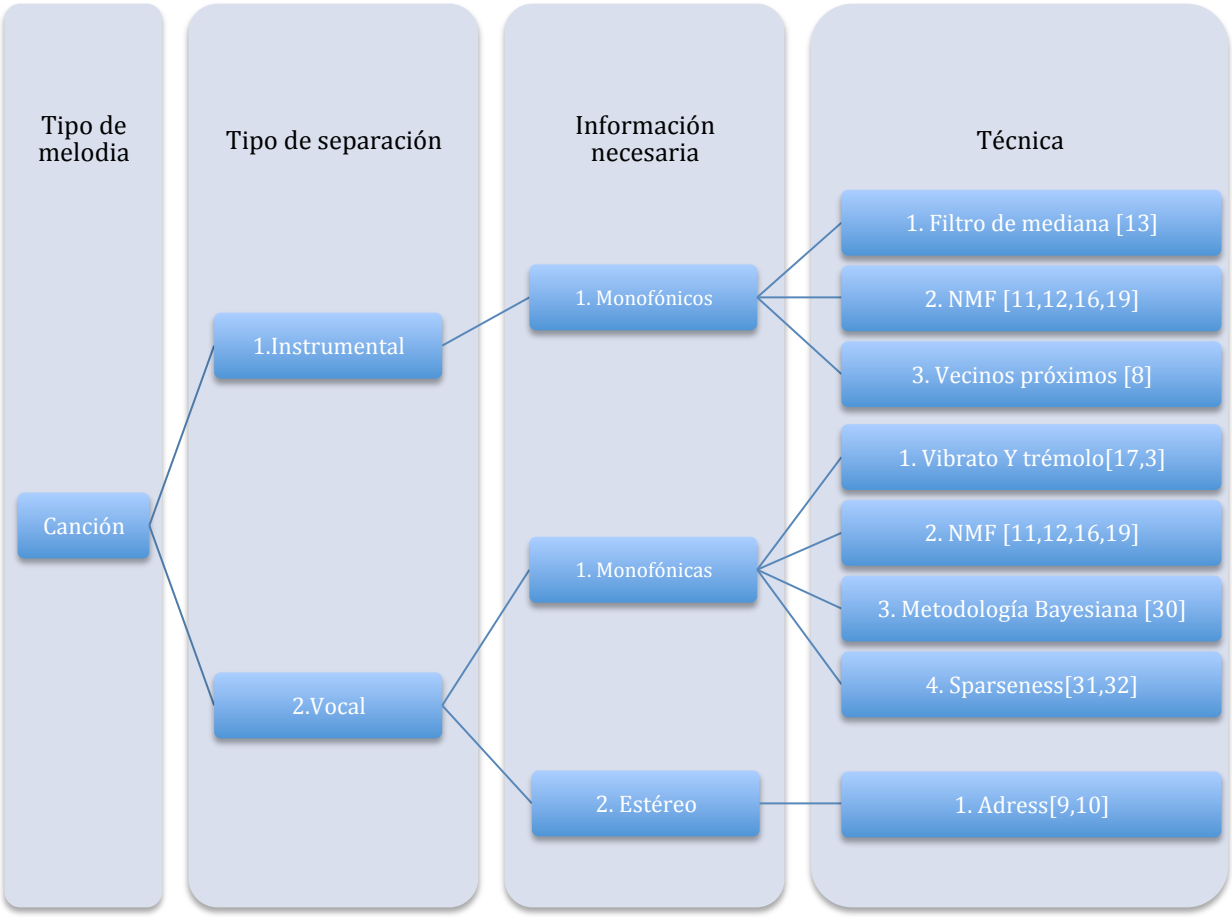

Figura 34: Breve clasificación de algunos sistemas de SVS

En la figura 34, se pueden apreciar algunos de los sistemas o técnicas que existen en la actualidad para SVS. Existen muchos más pero estos son algunos de los más significativos.

1. Grupo Instrumental: Dentro de este grupo encontramos diferentes técnicas, en base a la información necesaria, y aunque existen técnicas que requieren de la

información que proporciona una pista estéreo, casi todos los algoritmos que se emplean son monofónicos.

1.1.1 Filtro de mediana[13]: Consiste en un algoritmo que nos permite separar la parte percusiva de la armónica. Realmente no es un sistema de procesado de audio entra más en sistemas de procesado de imagen ya que lo que realiza es una búsqueda de picos y llanos. Es decir si se recuerda de la sección 1.1, las figuras 12 y 19 representaban el espectrograma de un sonido armónico y percusivo respectivamente. Como se explicó, los percusivos tienen una energía de banda ancha y de poca duración en el tiempo, pudiendo ser identificados como líneas verticales en el espectrograma, mientras que los armónicos todo lo contrario se pueden identificar como líneas horizontales en el espectrograma. Dentro de este algoritmo existen varias variantes, definidas en base a la función de pesos que se emplee. Principalmente se trabaja con dos funciones de peso, distinguiendo así entre los dos algoritmos. Por un lado tenemos la distancia Euclidea:

$$D = \sqrt{x^2 - y^2} \tag{8}$$

Y la Kullback-Leibler:

$$D = \mathrm{x} \cdot log_{10}\left(\frac{x}{y}\right) - x + y \tag{9}$$

La principal diferencia entre ambos métodos radica en la importancia que le dan a las muestras grande y pequeñas, mientras que (9) no hace distinción y da la misma importancia a todas las muestras la (8) da mayor peso a las muestras parecidas que a las que son muy diferente. Se trata de un algoritmo de bajo coste computacional que no requiere de muestras futuras de la canción lo que lo hace ideal para la implementación en tiempo real.

1.1.2 NMF[11,12,16,19]: Non-Negative Matrix Factorization, es un sistema que funciona tanto para separar la voz de la parte instrumental como para separar las componente instrumentales. Su uso no se limita al análisis de audio, sino que se extiende a campos completamente diferentes como el procesado de imágenes. Desde un punto de vista matemático consiste en la descomposición de matrices, en donde la única condición necesaria es que los valores de estas no sean negativos, lo que se adapta perfectamente para el módulo de los espectrogramas que se suelen emplear. La idea es simple se trata de descomponer la matriz original en dos matrices una matriz de bases y otra ganancia a

lo largo del tiempo, en otras palabras, una matriz que diga que ocurre y otra que diga cuando ocurre.

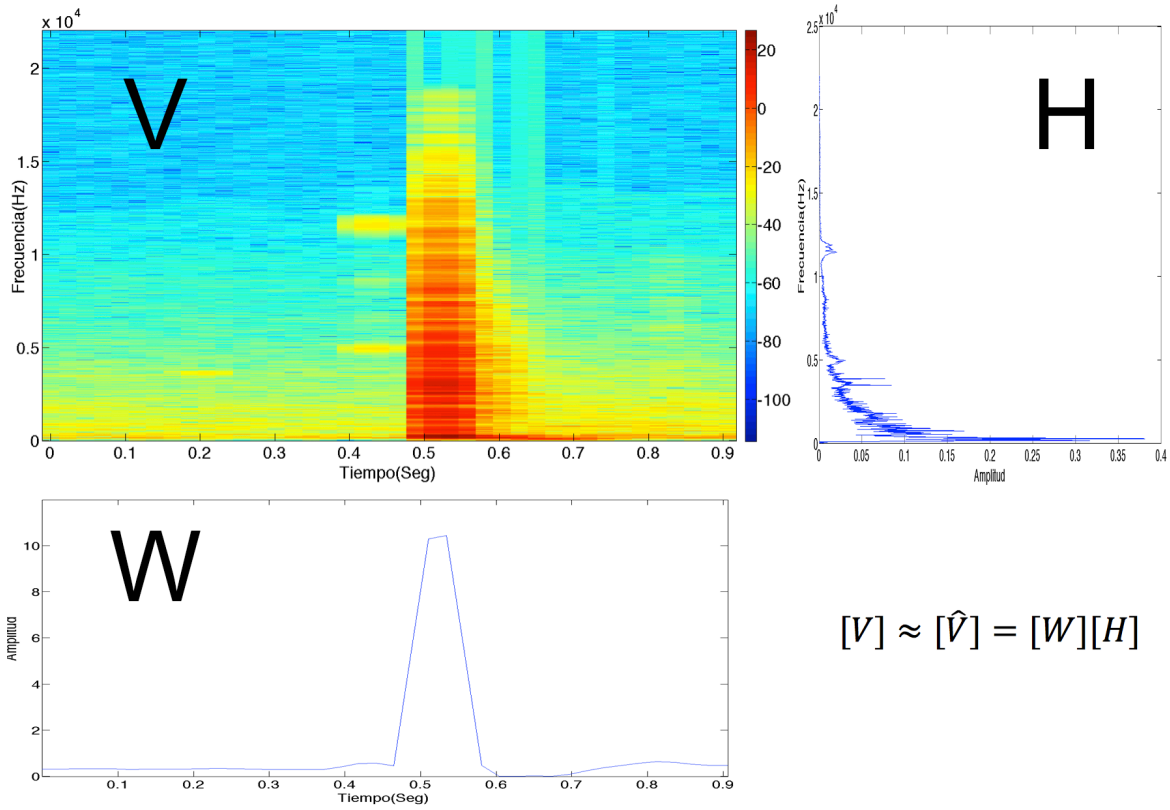

Figura 35: Ejemplo de la descomposición de un sonido mediante el uso de NMF

En la figura 35, se puede apreciar como H representa que es lo que ocurre para cada frecuencia, y W en que momento ocurre. De esta forma podemos "enseñar" al ordenador que forma tiene por ejemplo un percusivo o que duración tienen de media determinados fonemas, de forma que por simple comparación podríamos identificar las componente de interés de una melodía. Luego el NMF se compone de dos pasos un primer paso de entrenamiento almacenándolo en una base de datos y otro segundo de comparación, estas dos etapas es lo que nos permite clasificar los diferentes tipos de NMF en base al tipo de entrenamiento y en base a la función de peso o método empleado para la comparación:

1. En base al tipo de entrenamiento [16]:

1.1 Supervisado: Consiste en entrenar completamente todas las componentes que se van a analizar, armónicos, percusivo o voz.

1.2 Semi-supervisado: Consiste en entrenar alguna parte del conjunto que se desea identificar, por ejemplo solo los percusivos.

1.3 No supervisado: Es el más complejo de todos, no se entrena ninguna base de datos, sino que el ordenador “aprende” solo a identificar las componentes y a reconocerlas.

2. En base al tipo de comparación[29]:

2.1 Divergencia Kullback-Liebler.

2.2 Divergencia Itakuro-Saito.

2.3 Distancia Euclidea.

1.1.3 Vecinos más próximos [8]:

Consiste en un método para la separación de la parte instrumental de la parte cantada en la melodía. Se basa en la idea de que en la música comercial actual, la melodía de instrumental de fondo que acompaña al cantante, sigue unos patrones repetitivos, mientras que la voz del cantante va variando continuamente salvo estribillos y demás y por tanto podemos realizar una comparación de a lo largo de toda la canción y ver que zonas del espectrograma son más parecidas a otras de esta forma obtenemos las zonas instrumentales de la canción. Para realizar la comparación entre las diferentes zonas de una melodía se emplea como viene siendo habitual una función de peso concretamente la misma que hemos visto en el filtro de mediana (8).

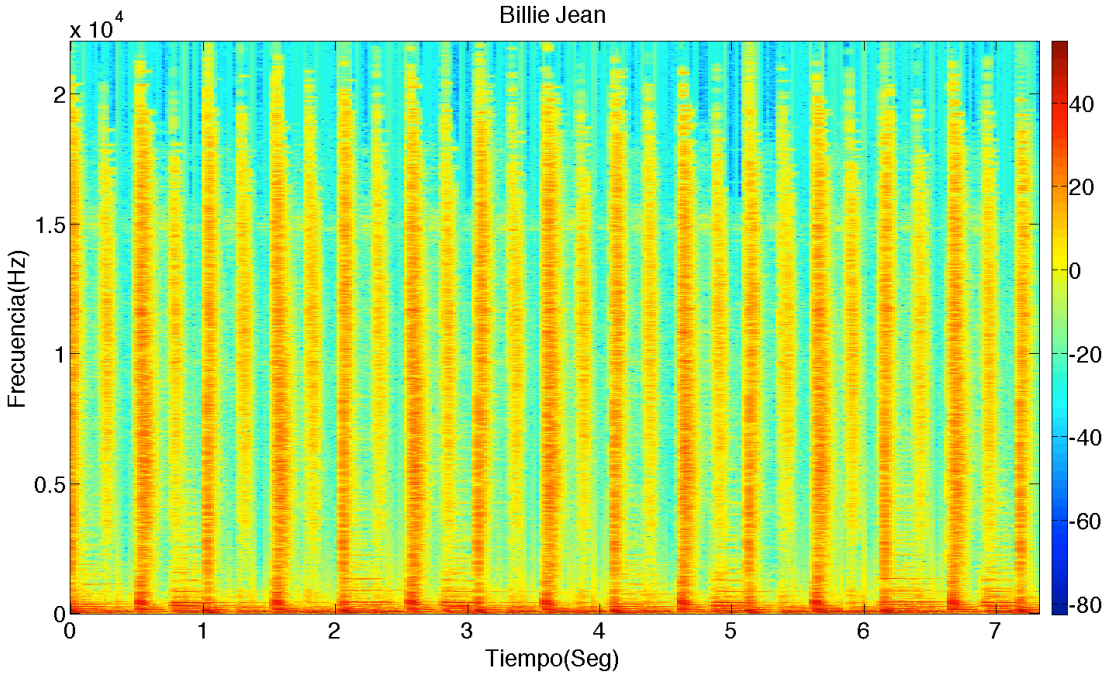

Figura 36: Espectrograma de un fragmente de la canción "Billie Jean" de Michael Jackson

En la figura 36, se puede ver un fragmento de la canción de “Billie Jean” de Michael Jackson, claramente se observa un patrón repetitivo cada medio segundo aproximadamente, esa repetición es lo que aprovecha este método para separar las pistas de música en instrumental y voz. El principal inconveniente de este método es que dependen de muestras futuras, lo cuál representa un problema para implementarlo en tiempo real, se necesita llenar un buffer de unos 5 o 10 segundos dependiendo de la canción.

2. Grupo Voz: Dentro de este grupo encontramos otra gran variedad de técnicas, algunas son similares a las ya vistas para la separación instrumental, mientras que otras son exclusivas para la detección de la voz.

2.1.1 Vibrato y trémolo [17,3]: Se basa en las dos características explicadas en la sección 1.2.3. Permite identificar la voz y algunos instrumentos, así como descartar otros, pudiendo estimar la frecuencia fundamental de la voz del cantante para los diferentes instantes de tiempo. Se trata de un proceso de localización y selección de una serie de regiones en el dominio del tiempo-frecuencia donde el pitch del cantante esté presente. Para ello se buscan oscilaciones que se extienden en torno a los 6 Hz en los diferentes parciales o armónicos, para identificarlos como vibrato, lógicamente se ha de elegir un umbral que permitirá el paso de parciales que no nos interesen. De forma similar se eligen los parciales con tremolo, solo que esta vez nos centramos en las amplitudes y finalmente eligiendo un segundo criterio que permite estimar la voz, basado en el número de parciales que se producen en un instante de tiempo[17]. De esta forma se consigue identificar los parciales que producen ambos efectos y que con casi toda seguridad pertenecen a la voz.

2.1.2 NMF[11,12,16,19]: El principio es el mismo que el del caso de separación instrumental, simplemente que las bases que se entran aquí están orientadas a identificar patrones de voz.

2.1.3 Metodología Bayesiana [30]: Es uno de los sistemas mas complejos que hay, se trata de una solución al problema de separación de fuentes ciegas (*“Blind source separation”*). Este tipo de problemas se definen cuando a priori se desconoce o se sabe muy poco acerca de la situación física de la fuente. Entonces se supone que la mezcla de fuentes es lineal, y que pueden ser descritas mediante algún tipo función de densidad de probabilidad. Independientemente de la información inicial, rara vez esta es suficiente para dar con una solución única, por ello se recurre a un procedimiento de razonamiento

inductivo, tratando este tipo de problemas a través de la metodología bayesiana. La técnica general consiste en formar un modelo que describe un problema de separación de fuentes en particular. Los parámetros que describen este modelo pueden ser tan complejos como se desee, el conjunto de señales de origen, una matriz de cómo están mezcladas o puede incluir más detalles como la posición y orientación de las fuentes o sus interacciones dinámicas. Una vez se tiene construido el modelo que describe todas las características relevantes del problema de separación de fuentes, se puede calcular la probabilidad de que valor particular de esos parámetros proporcionan una descripción precisa de la situación real. Lógicamente a mayor complejidad del modelo se requerirá un mayor tiempo de calculo y no necesariamente se obtendrán mejores resultados.

2.1.4 Sparseness [31,32]: Realmente no es un sistema independiente completo, sino que se suele emplear como complemente al NMF para separar mejor las fuentes. Por ejemplo cuando el espectro de una fuente (ej: Pedal de batería) cubre parcialmente el espectro de otra (ej. Tambor), la fuente más retardada podría ser modelada como una suma del sonido de la primera fuente y algún residuo. El uso del sparse facilita una representación donde solo un solo espectro se use para modelar la fuente más retardada. Matemáticamente son unos términos añadidos a la función de peso del NMF. En otras palabras su objetivo es la representación de la señal utilizando la mínima cantidad de datos al tiempo que conserva la máxima cantidad de información.

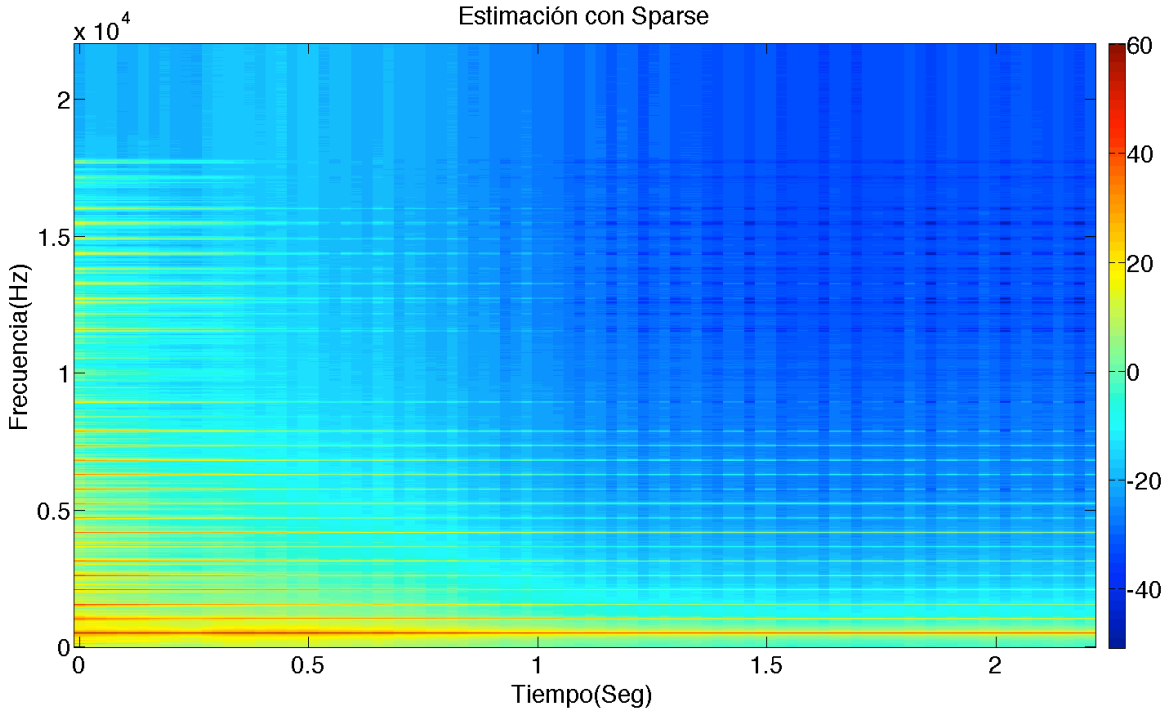

Figura 37: Ejemplo de una estimación de la figura 11 con Sparse

En la figura 37 se puede apreciar como se representa la figura 11, pero con muchos menos datos, pero la estructura general de la señal sigue estando.

2.2.1 ADRess [9,10]: El proceso de *"Azimuth Discrimination and re-synthesis"* se trata de uno de los algoritmos más utilizados en pistas estéreo, se basa en el principio de que la voz se encuentra centrada en los dos canales en el momento de grabar la pista, es decir aprovecha la información espacial que ofrecen las pistas en estéreo. Dicho de otra forma el algoritmo del ADRess se aprovecha de la ventaja de la variación de intensidades entre los dos canales izquierdo y derecho para crear lo que se llama azimugrama. El azimugrama es una representación de una fuente individual de la señal de entrada, en este caso la señal estéreo y siendo la fuente la voz, que permiten ver que porcentaje de dicha fuente se encuentra en cada canal. En la figura 38, se puede ver la distribución de los diferentes instrumentos a la hora de grabar una canción estando tanto el cantante como los percusivos en la zona central.

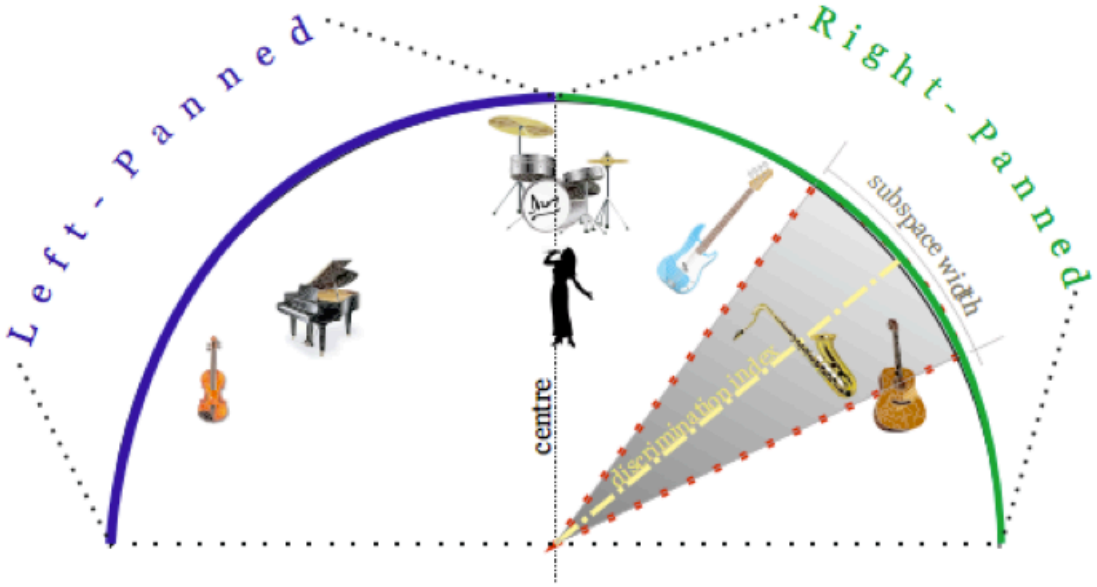

Figura 38: Representación de la localización de los diferentes instrumentos en una orquesta [10]

Mientras que en la figura 39, se observa un ejemplo del concepto de azimugrama, en el que se detectan la proporción de dos tipos de fuentes diferentes en el canal izquierdo.

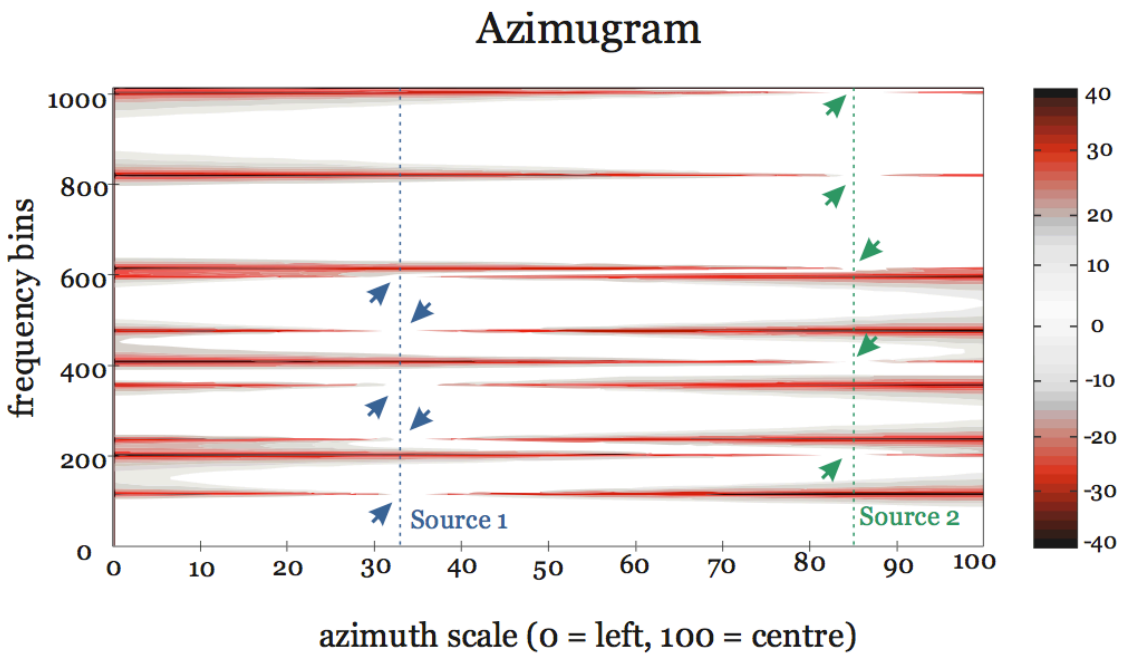

Figura 39: Ejemplo de azimugrama con detección de dos fuentes [10]

Como se ha visto el campo del SVS es un campo muy amplio, como muchas áreas de estudio y algoritmos tan importantes como el NMF que son usados en muchos campos y no sólo se limitan a la música. En la tabla 2, se ofrece un breve resumen de los algoritmos comentados en esta sección y sus referencias para interés del lector.

Algoritmo	Clasificación		Referencia
	Monofónico	Estéreo	
Filtro de mediana	☑	☐	[13]
NMF	☑	☑	[11,12,16,19]
Vecinos próximos	☑	☐	[8]
Vibrato y trémelo	☑	☐	[17,3]
Metodología Bayesiana	☑	☐	[30]
Sparseness	☑	☐	[31,32]
ADRess	☐	☑	[9,10]

Tabla 2: Comparativa de distintos métodos existente para el SVS

4. IMPLEMENTACIÓN

A continuación se presentan las diferentes etapas que se han utilizado, así como la descripción de su funcionamiento y el proceso hasta completar un programa capaz de cumplir los objetivos de la sección 2.

4.1 DESCRIPCIÓN DEL PROYECTO

La implementación consta de 4 etapas, que se producen en serie, es necesario que para el buen funcionamiento, todos la pista sean estéreo, aunque como veremos más adelante se analicen en su mayor parte cada canal por separado.

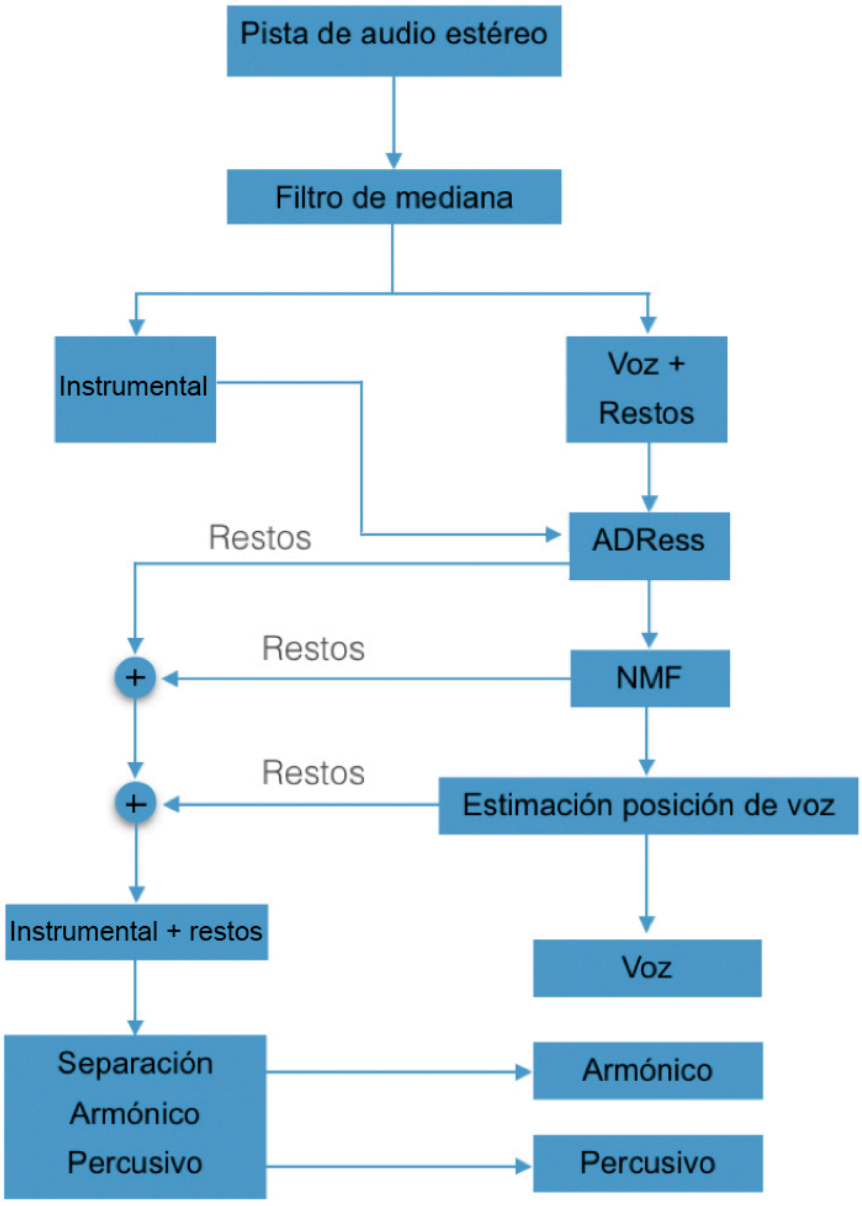

Figura 40: Diagrama completo del módulo

En la figura 40 se puede apreciar el diagrama del módulo que se irá describiendo en los siguientes apartados. Una vez se obtiene la voz con los restos se entra en un proceso en serie y por tanto el error que se cometa en cada etapa repercutirá en el resultado de la siguiente. Explicando brevemente las diferentes etapas:

En una primera etapa, se divide la melodía en los dos grupos principales, instrumental y voz. En esta etapa no es necesario ser muy estrictos con la selección de voz, es decir no resulta muy problemático que en la parte de voz se introduzca un poco de instrumental, es más es preferible eso a que en la parte instrumental se introduzca voz, debido a que a la parte instrumental solo se le aplica una etapa más para separar armónicos de percusivos.

Siguiendo por el camino de la derecha según la figura 40, en una segunda etapa, emplearíamos el algoritmo del ADRess, para separar la gran parte de instrumental que no se consigue separar en la primera etapa, de igual forma y paralela a esta etapa, se hace lo mismo con la parte instrumental por si tuviese restos de voz, de forma que podamos reagruparlos cada uno en su pista correspondiente.

Tras la segunda etapa, se observó que junto a la voz centrada suelen situarse físicamente instrumentos percusivos, como la batería y por tanto era común que quedasen algunos restos de instrumentos percusivos en el resultado a la salida de esta etapa, por ello se pensó en una tercera etapa que se centrase en la selección de percusivos únicamente, decidiéndose finalmente implementar una etapa de NMF, semi-supervisado donde se entrenase únicamente con instrumentos percusivos. Para mejorar los resultados de esta etapa, se decidió también incluir junto al NMF, la restricción sparseness.

Por último en la línea de separación de voz, se aportó una pequeña contribución propia, ya que a la salida de la etapa anterior, si bien se notaba claramente una disminución de la intensidad sonora de los percusivos frente a la entrada de esa etapa, estos no terminaban de atenuarse lo suficiente para se imperceptibles, y por ello se ideó un sistema que tomase la zonas de la canción donde se registrasen niveles de baja intensidad sonora y se supusiesen percusivos, de esta forma todas las zonas de la canción donde no participa el cantante quedan “eliminadas”.

Finalmente, para la separación de instrumental en armónico percusivo, se incluyó una etapa que se basa en el principio de la forma espectral de los armónicos y de los percusivos, principalmente por su simpleza y sus buenos resultados.

Con todas estas etapas se cerraría el módulo propuesto para este trabajo fin de grado, como se ha comentado al principio, es un proceso muy serie, es decir los errores de las primeras etapas serán arrastrados hasta el final del proceso.

4.2 ETAPA 1: VECINOS MÁS PRÓXIMOS Y FILTRO DE MEDIANA

La técnica de separación que se propone en este etapa, parte inicialmente de una señal mezclada monocanal, es decir no aprovecha la información espacial de la canción de entrada. Se basa en encontrar los valores más similares del momento actual con respeto a los momentos siguientes de la canción. Tras aplicar a la pista la transformada de tiempo corto de Fourier, la idea es que los frames más cercanos a cualquier frame no vuelven a ocurrir en breves espacios de tiempo, sino que ocurren en momentos separados a lo largo de la canción. Obteniendo de esta forma un camino para estimar la parte instrumental que acompaña al cantante, ya que los frames que más se parezcan indicarán fragmentos que se repiten de la melodía.

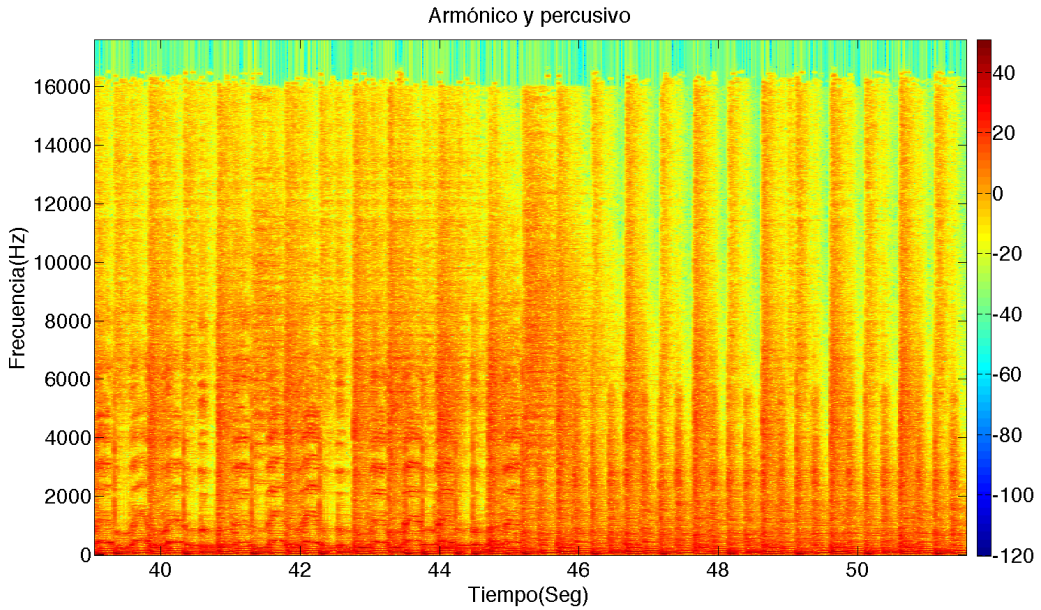

Figura 41:Espectrograma de los armónicos y percusivos de un fragmento de la canción 'Livin'On a prayer' de Bon Jovi

En la figura 41 se puede apreciar un fragmento de la parte instrumental de la canción *'Livin'on a prayer'* de Bon Jovi, en ella se ve claramente como existen ciertos patrones que se repiten (a corto plazo) tanto a nivel armónico como a nivel percusivo.

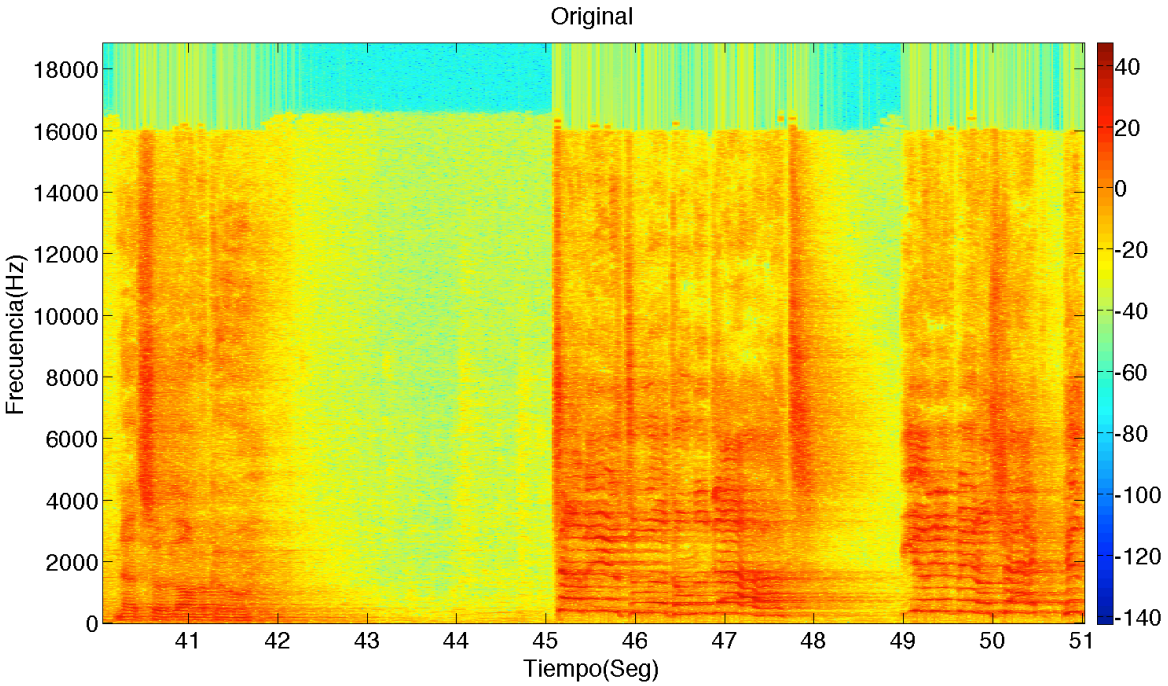

Figura 42: Espectrograma de la voz de un fragmento de la canción 'Livin'On a prayer' de Bon Jovi

Sin embargo como se observa en el espectrograma para el mismo momento de tiempo de la parte vocal en la figura 42, la parte de voz de una canción carece de esos patrones repetitivos.

Bajo estas premisas, se entiende que la música de fondo predominará cuando se calcule la distancia entre dos frames, ya que solo en un pequeño número de bins en los que la energía de la voz esté presente no coincidirán. Para el calculo de la distancia se ha decidido emplear la distancia Euclidea al cuadrado [8], principalmente por su rapidez de calculo, se define como:

$$D_{k,l} = \sum (X_k - X_l)^2 \qquad (10)$$

Donde X es la magnitud del espectrograma de la mezcla de tamaño NxM, donde N representa los bins de frecuencia y M los frames de tiempo. Mientras que X_k es

el k-esimo frame del espectrograma y X_l el frame actual, $D_{k,l}$ representa la distancia Euclidea al cuadrado entre el frame k y el frame l y sumados sobre todos los N bins de frecuencia.

Una vez se tiene la matriz D de dimensiones MxM, se pasa a ordenarla de forma ascendente por filas, aunque indiferentemente puede ser por columnas ya que en este momento D es una matriz simétrica de diagonal 0, de esta forma se consigue que las muestras que más se parecen se encuentren en las primeras posiciones, luego tomamos un valor de p vecinos más próximos, de esta forma se obtiene una matriz de Mxp, a la que llamaremos P. Entonces podemos estimar un umbral para la música de fondo para el k-esimo frame como:

$$Y_k = M(P) \tag{11}$$

Donde Y_k denota el umbral del k-esimo frame del espectrograma para la música de fondo y M denota el operador mediana. Se asume que la música de fondo no tiene mayor energía en un momento y frecuencia dado que la mezcla original con lo que posteriormente se puede realizar una comparativa y eliminar cualquier valor que tenga un valor mayor que el de la muestra original, descartándolo como música de fondo.

$$Y_{f,k} = \min\left(X_{f,k}, Y_{f,k}\right) \tag{12}$$

Donde f denota el f-esimo bin de frecuencia y k el k-esimo frame de tiempo. Llegados a este punto, se puede generar una máscara, que permitiese la separación instrumental de la voz. En aquellos instantes de tiempo y para cada frecuencia en la que no exista voz del cantante $X_{f,k} \approx Y_{f,k}$, mientras que para los instantes en los que si exista voz del cantante $X_{f,k} \gg Y_{f,k}$.

Podemos elegir entre dos tipos de máscaras, la máscara binaria que presenta el problema que en ocasiones introduce artefactos audibles en la resíntesis, por ello se

ha decidido emplear mejor una más suavizada [8] Wiener, basada en una función Gaussiana:

$$W_{f,k} = exp(-\frac{(log_{10}(X_{f,k}) - log_{10}(Y_{f,k}))^2}{2\lambda^2}) \quad (13)$$

Donde W es una máscara suavizada que se aplica sobre el espectrograma de la señal original y λ es un parámetro de tolerancia que sirve para controlar el peso de la máscara.

El valor complejo de la música del espectrograma B se puede estimar como:

$$B = W \otimes R \quad (14)$$

Donde R denota el espectrograma complejo de la señal original y $\otimes$ denota el producto de Hadamard. Finalmente la música de fondo se puede recuperar vía transformada inversa de tiempo de Fourier, tomando la fase de la señal de entrada.

De forma similar podemos se puede recuperar el espectrograma complejo de la voz del cantante V estimado como:

$$V = (1 - W) \otimes R \quad (15)$$

Para concluir se puede realizar un pequeño post-procesado de la señal que mejora el resultado basado en la teoría explicada en la sección 1.2.3, aplicando un filtro paso bajo que descarte todas las frecuencias por debajo de los 100 Hz, ya que como se vio, la voz humana no genera frecuencias por debajo de los 100 Hz, y por tanto se puede suponer que todos ellos serán instrumental.

Analizando esta primera etapa, resulta evidente uno de los principales problemas que presenta, para implementar el algoritmo en tiempo real, se trata de un

proceso que depende de muestras futuras, y por tanto su implementación en tiempo real requerirá de un buffer de al menos 5 segundos. Estos 5 segundos son relativos y dependerán de la canción lógicamente, pero tras escuchar varias canciones, en un promedio de 5 segundos se encuentran fragmentos similares, que permitan una separación relativamente óptima. Por esto se trata de un buen algoritmo, pero no está pensado para trabajar en tiempo real.

Ejemplo de separación

En este apartado se va a demostrar la teoría de la etapa aplicada a un caso práctico, concretamente tomamos la canción “Wouldn’t it be nice” de los Beach Boys.

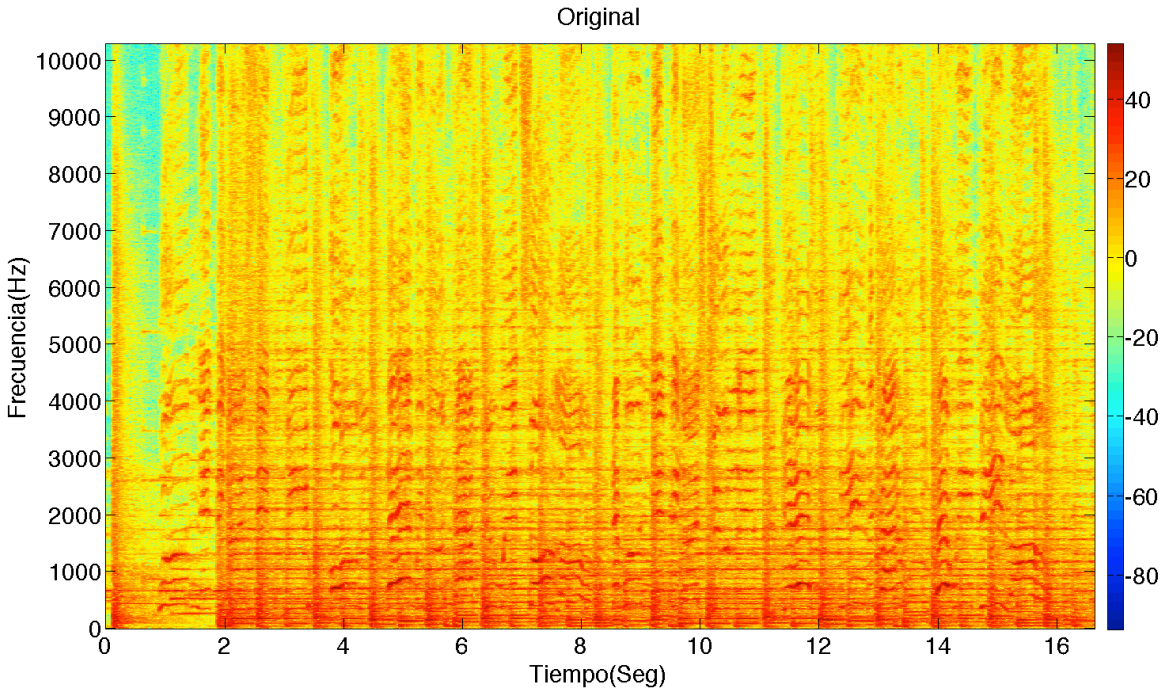

Figura 43:Espectrograma de un fragmento de la canción 'Wouldn't be nice' de los Beach Boys

La figura 43 muestra el espectrograma de la señal original de entrada a esta primera etapa de filtro de mediana, se considerarán valores óptimos para esta etapa (8), $p = 80$ y $\lambda = 1$, donde p recordemos que es el número de vecinos próximos que se va a tomar, la frecuencia de muestro con la que se trabaja es de 44100 Hz. La matriz está compuesta por 717 frames de tiempo y 2049 bins de frecuencia.

Aplicando la ecuación (10), obtenemos una matriz D de 717x717 simétrica con una diagonal de cero, esto se puede observar en la figura 44. La diagonal 0 se da porque lógicamente al comparar un frame consigo mismo la distancia es 0.

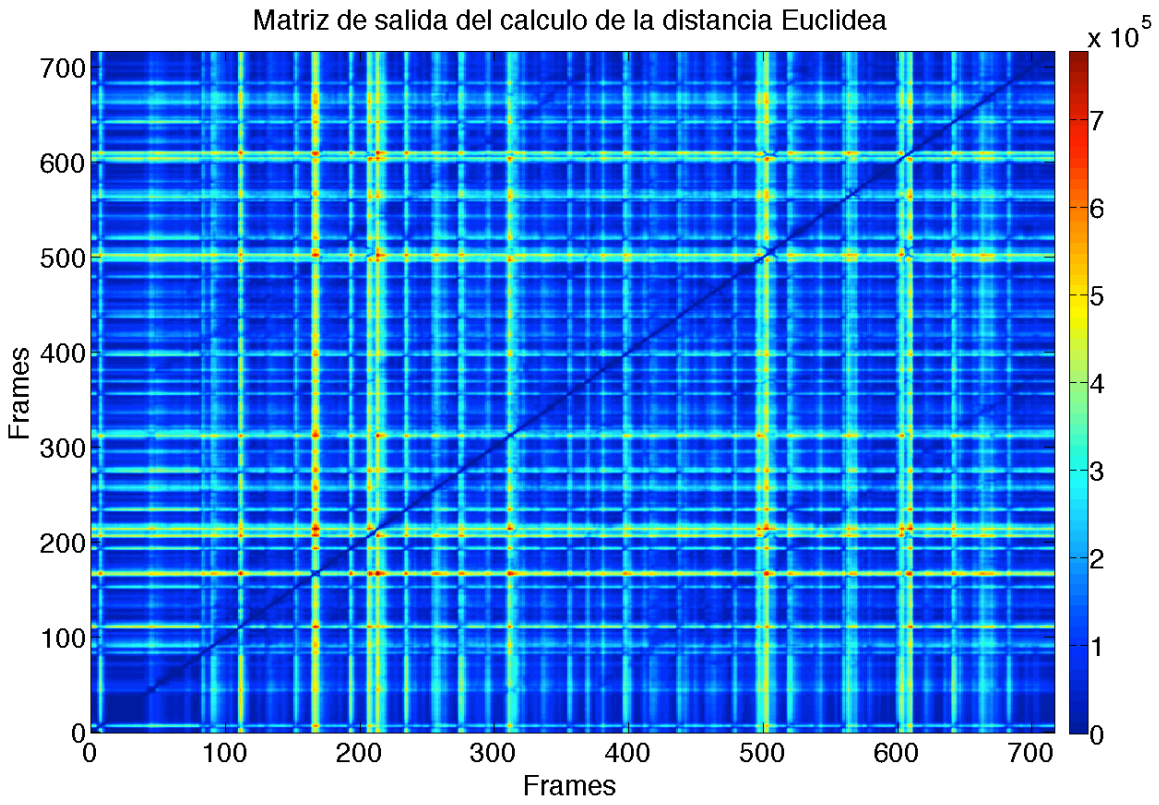

Figura 44: Matriz de salida tras calcular la distancia Euclidea

Si por ejemplo se observa en la figura 45 donde se ve la distancia del primer frame con respecto a los demás:

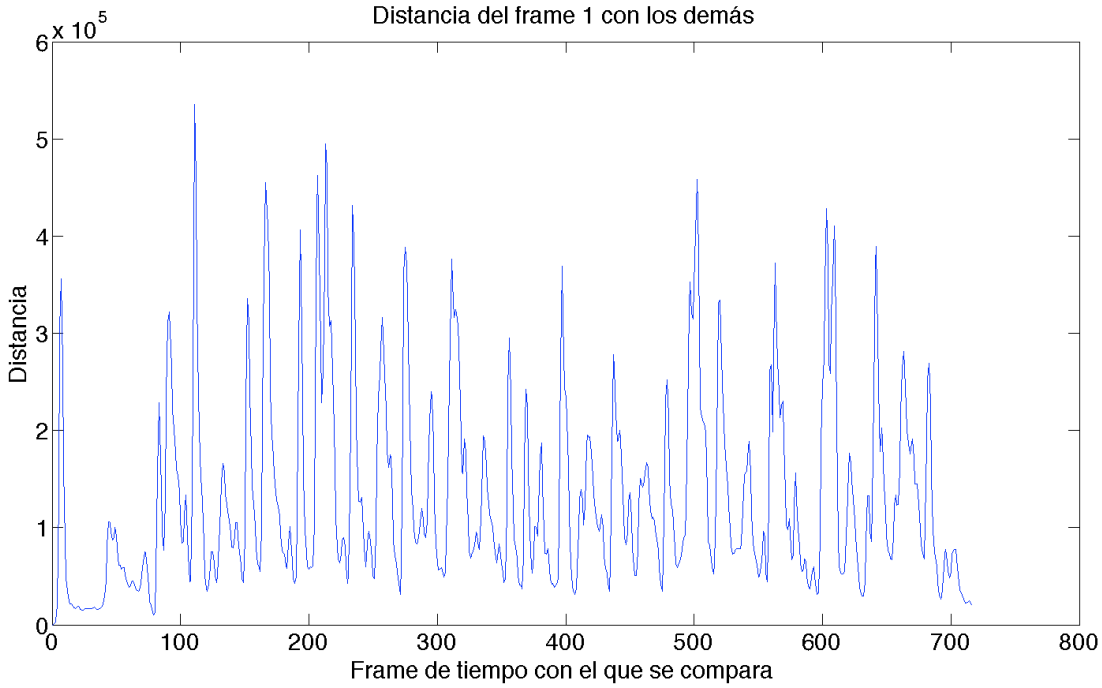

Figura 45:Distancia del frame 1 con respecto a todos los demás

Claramente se observan unos picos muy por encima de la mayoría que representan la zona donde los frames son completamente distintos, pero por lo general se observa un umbral medio en el que existe cierto parecido. Si ordenamos D de forma ascendente, y tomamos los vecinos más próximos con $p = 80$. Esto se puede ver en la figura 46:

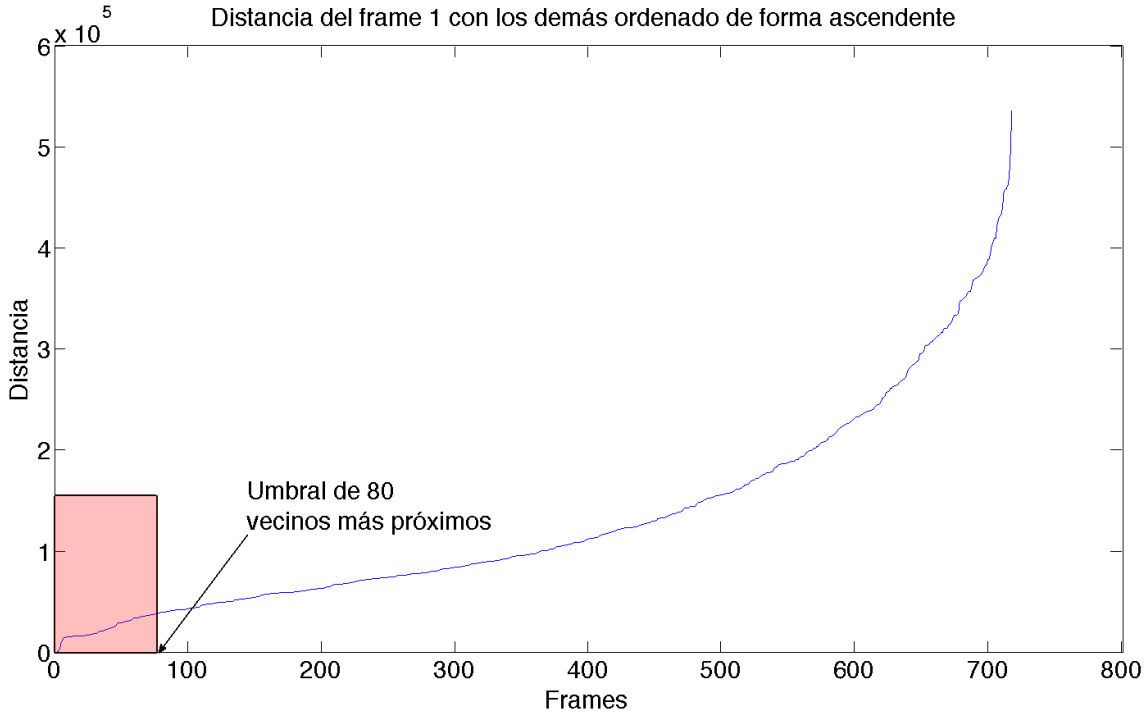

Figura 46: Matriz D ordenada de forma ascendente seleccionando los p vecinos más próximos

Tras aplicar esta selección de 80 muestras por frame y a todos los bins de frecuencia por cada frame obtenemos una matriz de 2049x80x717, en otras palabras los valores de la señal original que corresponden a esas 80 muestras.

Finalmente y aplicando las ecuaciones (11) y (12), obtenemos matriz Y de dimensiones 2049x717 en las que se tienen muestras de energía menor o igual que la señal original. Por último solo faltaría aplicar la ecuación (13) para generar la máscara W y de esta forma poder separar la parte instrumental de la vocal.

Los resultados de la separación se muestran en las figuras 47 y 48:

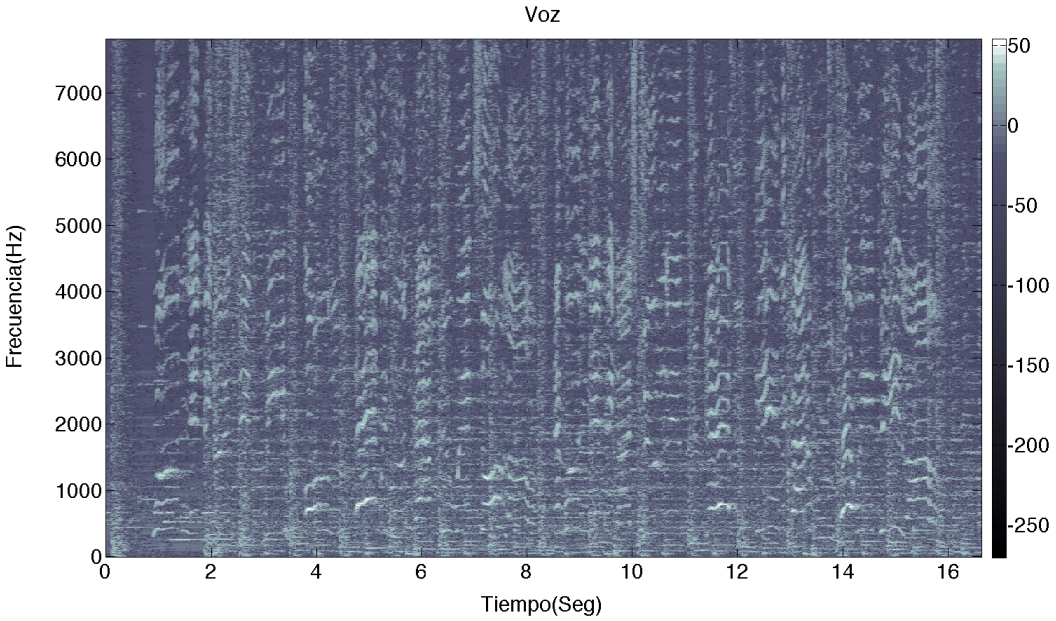

Figura 47: Espectrograma de la voz tras separación de filtro de mediana

En la figura 47 se pueden observar claramente formas propias del espectrograma de la voz, aunque lógicamente quedan algunos restos de la parte percusiva y armónica.

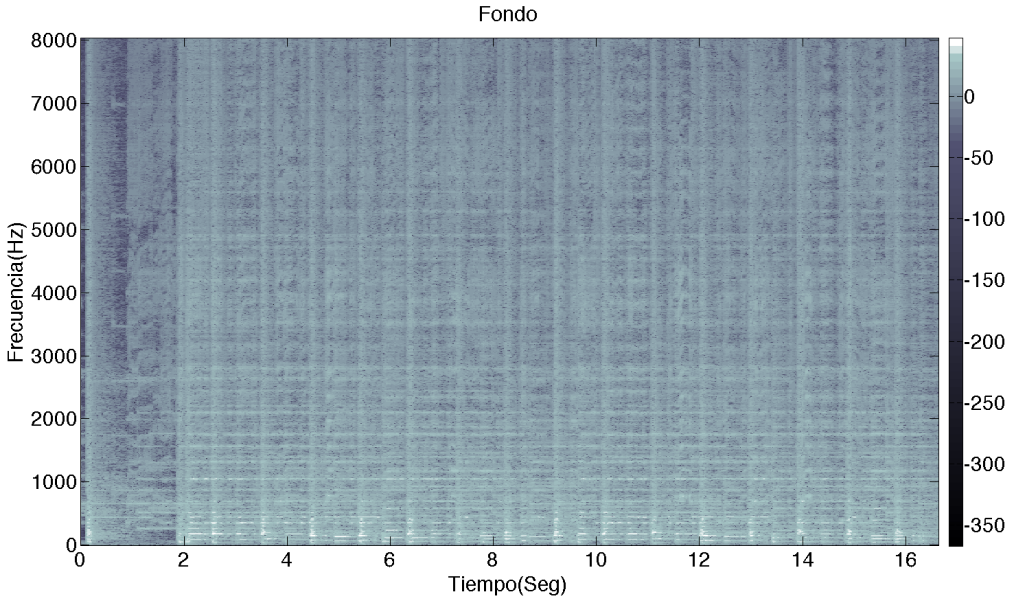

Figura 48: Espectrograma de la música de fondo tras separación del filtro de mediana

En la figura 48, claramente se observan formas propias de los armónicos y de los percusivos sin que apenas se aprecien restos de voz. Aunque si escuchamos los resultados en ambas pistas quedan restos tanto de voz en la pista instrumental como de

armónico y percusivos en la pista de voz, pero esos restos se afrontarán en las siguiente etapas.

Por último es interesante ver en la figura 49 como el filtro paso bajo explicado en la teoría del algoritmo limpia completamente las bajas frecuencias de la pista de voz, en las que se asume que no se puede producir ningún tipo de sonido por parte del cantante:

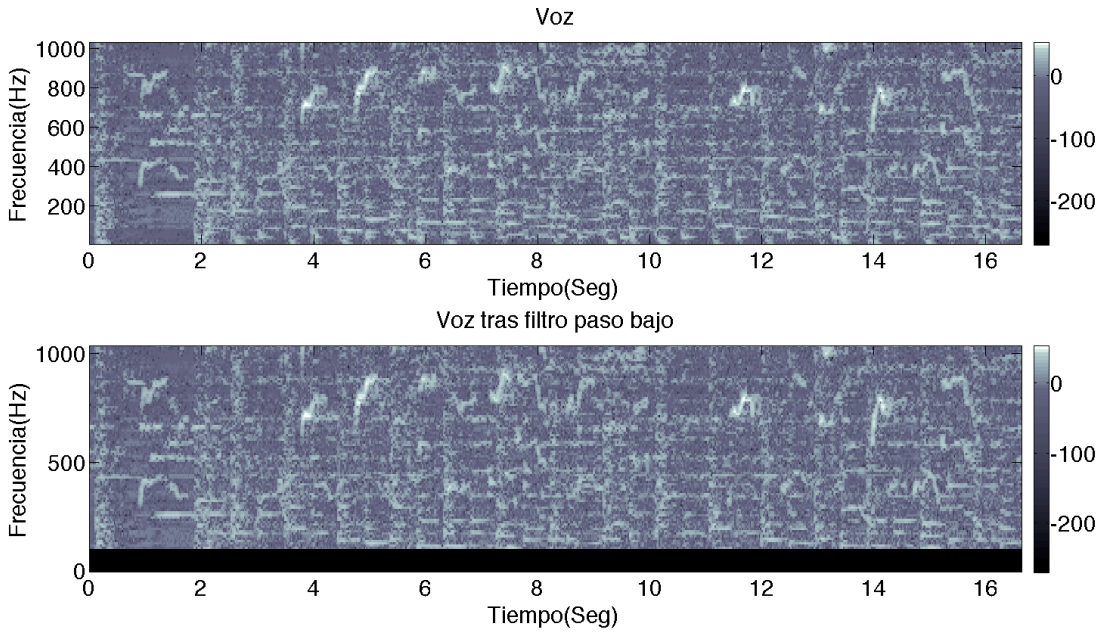

Figura 49: Comparativa de espectrogramas de la voz tras aplicar el filtro paso bajo a 100 Hz

4.3 ETAPA 2: ADRESS(AZIMUTH DISCRIMINATION AND RE-SYNTHESIS)

Se ha visto como en la primera etapa, se consigue implantar las bases para una buena separación de la parte instrumental de la voz, a continuación y en las siguientes etapas hasta la separación de armónico y percusivo, se aplicarán técnicas con intención de mejorar la selección de voz frente a la instrumental. Para situarnos en el diagrama del módulo (figura 40), se desarrollan las etapas de la parte derecha.

Sobre el año 1960, se realizó la primera producción en estéreo de música comercial de la historia para el grupo The Beatles [10]. En la actualidad por lo general y salvo casos muy concretos todas las grabaciones se realizan en pistas estéreo, esto nos permite aprovechar la información espacial de este conjunto de señales. Con el tiempo y la experimentación, se tendió a localizar la voz junto a los bajos, en el centro de la grabación, fundamentalmente debido a que la voz es el sonido principal de una canción y

necesita existir un equilibrio entre la zona derecha e izquierda. Por otra parte las frecuencias bajas debido a su gran longitud de onda, no sonaban bien si se encontraban con cierta desviación respecto al centro.

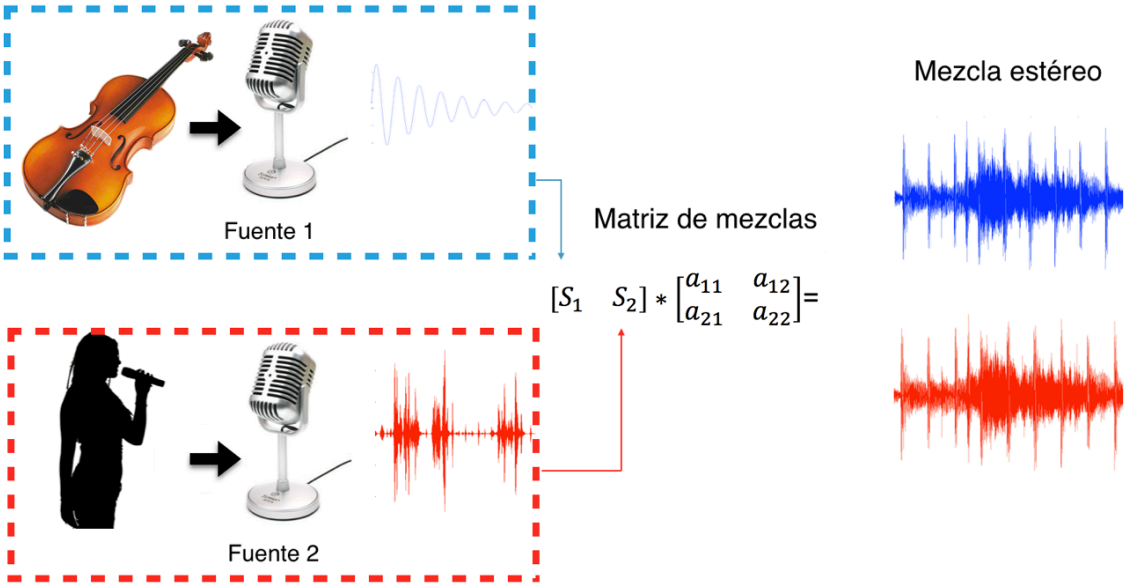

Figura 50: Mezcla estéreo de dos fuentes independientes

En la figura 50 se puede ver la forma en la que se mezclan dos fuentes, aunque por lo general en una pista de audio musical se encuentran muchas más fuentes sonoras.

Bajo estas premisas se describe el método del ADRess, el cuál explota la información espacial que ofrecen las pistas estéreo. Esta etapa permitirá ser más selectivo a la hora de seleccionar la voz en la salida de la primera etapa descartando gran parte de los instrumentos que acompañen al cantante.

En la práctica se basa en generar un azimugrama. Un azimugrama consiste en la recreación de un espacio que nos permite modelar la posición de las distintas fuentes, de forma que seleccionando un sub-espacio del mismo podemos seleccionar la fuente deseada. La localización espacial (‘panning’) entre dos canales (izquierdo y derecho) es posible gracias a un análisis panorámico de la potencia de las señales. La idea es asociar el porcentaje de intensidad sonora que llega de cada fuente a cada canal, constituyendo así lo que definimos como el campo estéreo. Este campo tiene una representación geométrica (semicírculo) que facilita el objetivo del algoritmo, aislar un ángulo centrado de este semicírculo. En la figura 51, se puede observar el concepto de campo estéreo y de cómo se introduce el término de índice de discriminación que

permitirá generar un sub-espacio de forma que podamos elegir un subconjunto de fuentes en base a su posición:

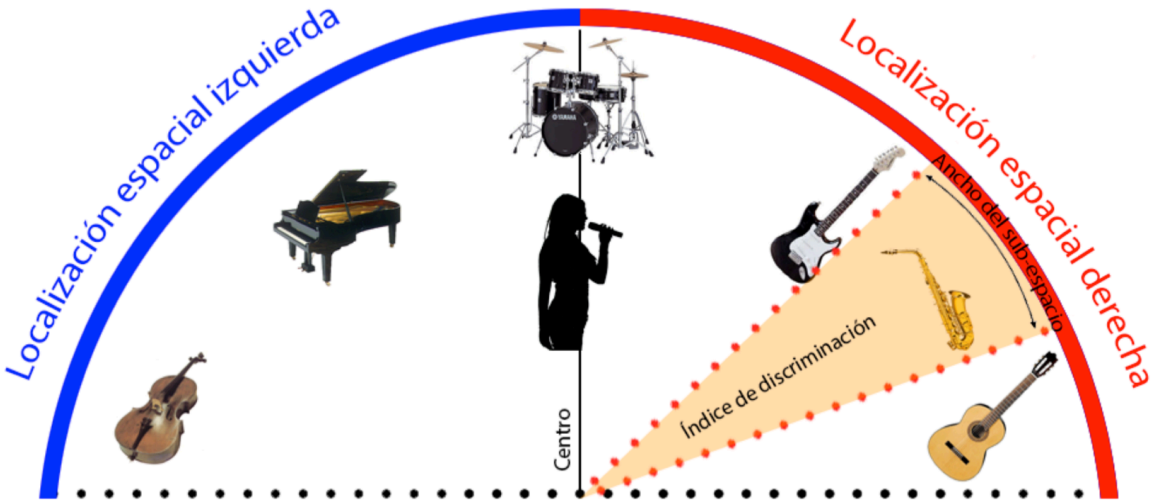

Figura 51: Ejemplo de campo geométrico generado en la etapa del ADRess

En la figura 51, se puede ver los dos parámetros principales de esta etapa, por un lado se tiene el índice de discriminación que como se verá más adelante corresponde al parámetro d y el ancho del sub-espacio que será el parámetro H. Estos parámetros son introducidos por el usuario en base a la selección de las fuentes que deseen, ya que ADDress no es capaz de detectar por si mismo las fuentes que son objeto de análisis, por ellos y para el análisis supondremos que nuestra fuente se encuentra centrada.

De la figura 51, también podemos extraer otra conclusión, es cierto que con solo dos parámetros se puede estimar la posición de las fuentes de interés, pero que pasa si como en la figura 51, se centra el índice de discriminación en cero apuntando hacia la cantante, el ancho del sub-espacio necesario para tomar la zona de voz también incluiría la batería que se encuentra detrás, en otras palabras algunas de las frecuencias de los instrumentos que físicamente se encuentren cerca del cantante se incluirán en la selección. La solución reducir el ancho del sub-espacio, asegurándonos una mejor selección de la posición de la voz, pero está restricción si bien eliminaría mejor los instrumentos cercanos, podría hacernos perder parte de información de la voz, por ellos y como casi todo en el mundo del SVS, deberemos tener un compromiso de ancho de sub-espacio, si se prefiere que la voz incluya algunos instrumentos cercanos o si se prefiere ser más selectivo y renunciar a zonas de la voz. Para este proyecto, se usa el primer criterio, se intentará recoger toda la información de la voz posible, aunque no se eliminen en su mayor parte algunos instrumentos. En otras etapas posteriores nos encargaremos

de perfeccionar esa selección ya que como se ha visto, la mayor parte de esos instrumentos serán percusivos de baja frecuencia.

ADRess explota el principio conocido como intensidad diferencial de canales cruzados IID (inter-channel intensity difference), asumimos que cada fuente puede ser representada en una posición del campo geométrico como una relación entre los dos canales. Por ejemplo, el piano de la figura 51 se podría definir como una fuente donde el 65 % pertenece al canal izquierda y el 35% al derecho, de igual forma que la guitarra acústica se puede expresar como una fuente de 80% derecho y 20% izquierdo, mientras que el cantante es una fuente repartida al 50% entre los dos canales. De esta forma si se multiplica el canal derecho por el 0.25 la guitarra será igualmente distribuida entra los dos canales, permitiéndonos cancelarla. Posteriormente podríamos reconstruir la señal con la guitarra eliminada. Hay dos principios fundamentales que se han comentado en el párrafo anterior pero que si se particulariza para este caso, el algoritmo es incapaz de averiguar ese factor de 0.25 para eliminar la guitarra, luego debe ser introducido por el usuario por prueba y error. El segundo principio es que cada fuente musical consiste en varias componentes en frecuencia que en algunos casos se mezclan entre las fuentes y es por ello que el usuario deberá introducir el parámetro H de ancho del sub-espacio.

Formalmente, la descripción del ADRess para una señal estéreo es: Siendo $L(t)$ y $R(t)$ las señales de audio en el canal izquierdo y derecho de una grabación estéreo respectivamente. Estas pueden ser representadas como [10]:

$$L(t) = \sum_{j=1}^{J} Pl_j S_j(t) \tag{16}$$

$$R(t) = \sum_{j=1}^{J} Pr_j S_j(t) \tag{17}$$

Donde S_j son las *J* fuentes independientes, mientras que Pl_j y Pr_j son los coeficientes de la localización espacial para el canal izquierdo y derecho respectivamente para las fuentes. La relación de intensidad entre $L(t)$ y $R(t)$ para las j^{th} fuentes puede expresarse como:

$$g_j = \begin{cases} \frac{Pl_j}{Pr_j} & si \quad Pr_j > Pl_j \\ \frac{Pr_j}{Pl_j} & si \quad Pl_j < Pr_j \end{cases} \tag{18}$$

Donde podemos ver fácilmente de la ecuación que podemos estimar los coeficientes como $Pr_j = g_j \times Pl_j$ o $Pl_j = g_j \times Pr_j$. Debido a que $L(t)$ y $R(t)$ son combinación lineal de la misma fuente independiente, lo que indica que la fuente podría ser cancelada usando la siguiente expresión:

$$L(t) - g(i) \times R(t) \tag{19}$$

$$R(t) - g(i) \times L(t) \tag{20}$$

La principal diferencia entre ambas ecuaciones radica en que su uso dependerá en que canal se encuentre la j^{th} fuente más predomínate, es decir dependerá de si Pr_j es mayor o meno que Pl_j respectivamente. En caso de que $Pr_j = Pl_j$ será indiferente el uso de cualquiera de las dos ecuaciones. Por otro lado, g_j funciona como un factor de escala, facilitando la extracción de la fuente interés. La importancia de la estimación de g_j determinará la efectividad del algoritmo, pero como contra partida, se trata del factor decisivo a la hora de estudiar la eficiencia del sistema.

Inicialmente se realiza una transformada de Fourier enventanada (STFT), para el caso que nos atañe se trabajará con una ventana Hanning, pero perfectamente se podría trabajar con cualquier otro tipo.

$$Lf(K) = \sum_{t=0}^{N-1} L(t) e^{-\frac{j2\pi}{N} kt}, para\ k = 0,1,2 \dots N-1 \tag{21}$$

$$Rf(K) = \sum_{t=0}^{N-1} R(t)e^{-\frac{j2\pi}{N}kt}, para\ k = 0,1,2 \dots N-1 \quad (22)$$

En las ecuaciones (21) y (22), podemos observar la TF discreta para el canal izquierdo y derecho respectivamente, N representa el número de puntos de la FFT y j la unidad imaginaria.

La forma de determinar g_j se basará en una parametrización de la variable, es decir se tomará un valor β que definirá la resolución del azimugrama, generando un vector $g(i)$ con i comprendido entre 0 y 1, con saltos de $1/\beta$, este vector contendrá los valores de peso que multiplicarán a la STFT de un canal y se restarán posteriormente a la del otro. Como se ha comentado anteriormente, este parámetro es el principal inconveniente en temas de eficiencia por que lógicamente si se elige una β mayor, efectivamente tendremos mayor resolución pero el tiempo de calculo como se verá es más elevado:

$$g(i) = \frac{i}{\beta}, para\ i = 0,1,2,3..\beta \quad (23)$$

En la ecuación (23) se ve la forma de calcular ese vector de valores de peso. Puesto que la voz está centrada, es indiferente si se usa la ecuación (19) o (20), pero en caso de estar interesado en otra fuente no centrada, se debería saber en que canal es más predominante. Finalmente ya se puede definir las ecuaciones de los azimugramas para cada canal, que no es otra que el valor absoluto de las ecuaciones (19) y (20):

$$Az_r(k,i) = |L(t) - g(i)\times R(t)| \quad (24)$$

$$Az_l(k,i) = |R(t) - g(i)\times L(t)| \quad (25)$$

La idea es que por cada frame de cada canal, se genere un azimugrama, que determine las frecuencias donde se producen la anulación de la fuente, de esta forma quedan localizadas simplemente para invertir el espectrograma. Debido a la complejidad de los conceptos, se va a ilustrar un pequeño ejemplo que defina el proceso para dos fuentes muy simple de tonos puros.

1. Fuente 1: Se supone la suma de 5 sinusoides de igual amplitud y frecuencia de 2540 Hz, 5080 Hz, 7620 Hz, 10160 Hz, y 12700 Hz respectivamente (tono con frecuencia fundamental 2540 Hz).

2. Fuente 2: Se supone la suma de 5 sinusoides de igual amplitud y frecuencia de 4350 Hz, 8700 Hz, 13050 Hz, 17400 Hz, y 21750 Hz respectivamente (tono con frecuencia fundamental 4350 Hz).

Notar que es un caso ideal, en el que las muestras no están solapadas y es sencillo identificar los nulos pertenecientes a las diferentes fuentes así como a los diferentes tonos, como ya se ha comentado esto en la realidad no sucede, en principio por lo general tendremos bastantes más fuentes y debido a los armónicos posiblemente muchas fuentes se solapen en frecuencia con otras cercanas.

Si se mezclan la fuente 1 con la 2 en un 75% al canal izquierdo y 25% al derecho y la fuente 2 un 54 % al izquierdo y un 46 % al derecho. En la figura 52 se representa el azimugrama de la ecuación (25), con una $\beta = 100$.

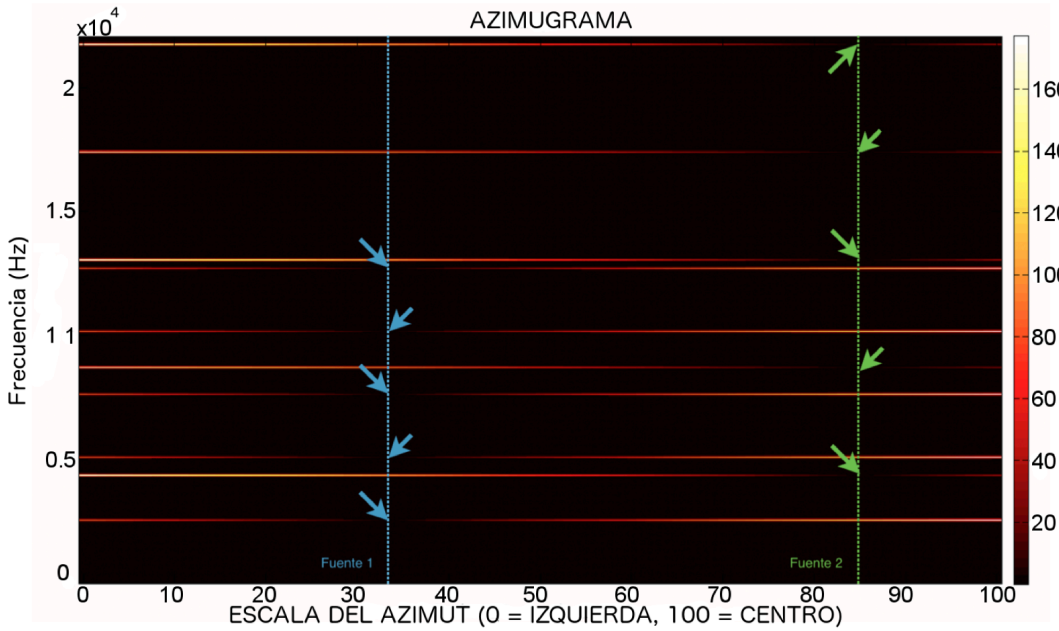

Figura 52: Azimugrama de dos fuentes con tonos puros para el canal izquierdo

En la figura 52, las flechas indican los puntos nulos, es decir donde cada fuente es cancelada. Se puede ver que la fuente 1 sería cancelada si multiplicamos la STFT del canal izquierdo por el valor de $g(i)$para $i = 85$ de la ecuación (23) de esta forma posteriormente podríamos recuperar la fuente 1 del canal derecho. Igualmente pasa con la fuente 2 podríamos cancelarla para $i = 33$.

Después de obtener los nulos, es necesario convertirlos en picos, de forma que permita localizar y recuperar la fuente mediante la inversa de la transformada de tiempo de Fourier (ISTFT). Desafortunadamente, la magnitud de los picos es desconocidas luego será necesaria estimarla:

$$Az_R(k,i) = \begin{cases} Az_R(k)_{max} - Az_R(k)_{min} & si \quad Az_R(k)_{\max} = Az_R(k)_{min} \\ 0 & Cualquier\ otro\ caso \end{cases} \tag{26}$$

$$Az_L(k,i) = \begin{cases} Az_L(k)_{max} - Az_L(k)_{min} & si \quad Az_L(k)_{max} = Az_L(k)_{min} \\ 0 & Cualquier\ otro\ caso \end{cases} \tag{27}$$

$para\ i\ \in \{1,2,3,\dots\beta\}\ y\ k\ \in \{1,2,3\dots N/2\}$

Como se observa en las ecuaciones (26) y (27) son ecuaciones, en las que las condiciones están muy ligadas a la resolución del azimuth (eficiencia del algoritmo). De esta forma conseguimos volver los nulos máximos, haciendo todos los demás bins de frecuencia ceros.

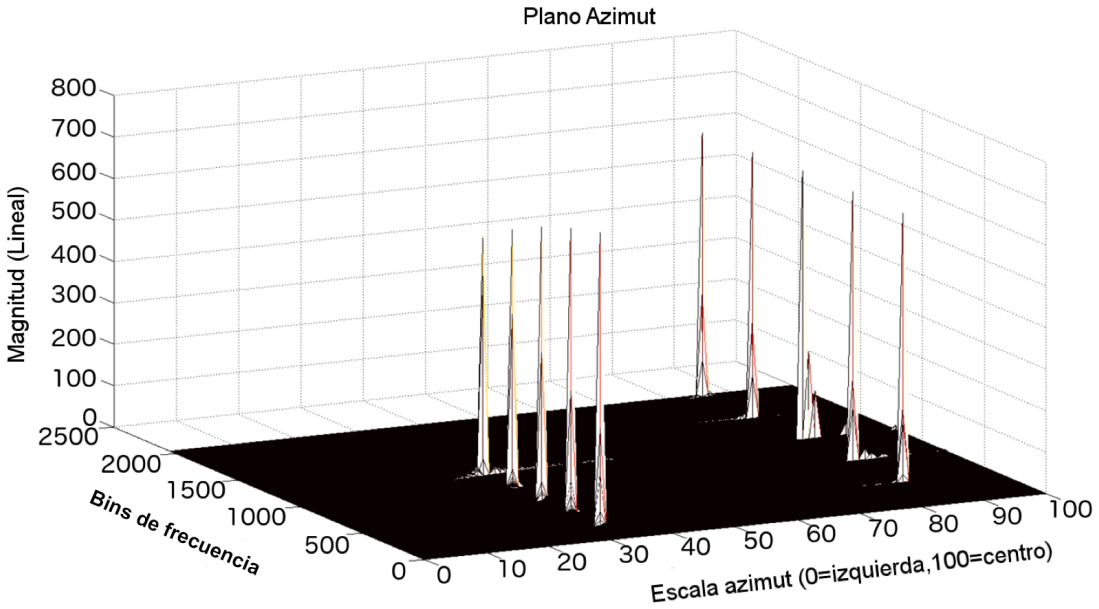

Figura 53: Azimugrama con mínimos invertidos que permite localizar las fuentes

Basándose en las ecuaciones(26) y (27), se puede construir un plano de frecuencia-azimuth, como el que se muestra en la figura 53. No obstante cabe destacar como ya se ha dicho que se trata de un caso aislado, que generalmente no se produce en la música comercial, pues obviamente por simple inspección sabemos que para los valores de $i = 85$ e $i = 33$ recuperaríamos la fuente 1 y 2 respectivamente, con que no se requeriría de un ancho de azimuth que seleccionase un rango de conjunto de picos. El caso normal es aquel en el que existe solapamiento debido principalmente por los armónicos y por tanto se debe usar un ancho de azimuth, que representa el error que asumimos que vamos a cometer para perder el menor número de información perteneciente a la voz. A continuación se presenta un caso muy simple en el que se ha incluido un armónico. Siguiendo en la línea del ejemplo anterior, el canal izquierdo tiene un 75% de la fuente 1 y un 54% de la fuente 2, significará que compartirán un armónico que tendrá una intensidad de $\frac{0.75+0.54}{2} = 0.645$, con lo que se cancelará cuando el canal izquierdo sea multiplicado en la ecuación (25), notar que para ser preciso necesitaríamos una resolución de $\beta = 1000$ para poder obtener el 0.645 necesario.

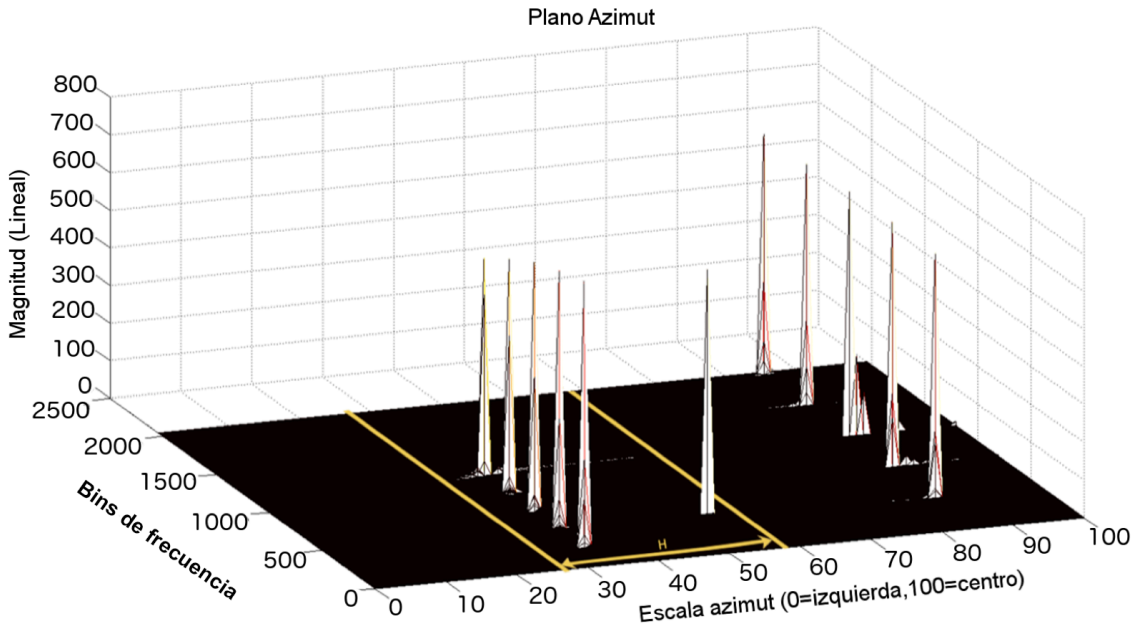

Figura 54: Azimugrama para un caso con armónicos

Como resultado, este pico particular aparece representado en la figura 54 para un valor de $i = 55$. Esto se asemeja más con un caso real que el de la figura 53, por eso es necesario el parámetro H (ancho del azimuth) que nos permite discriminar algunos armónicos, desgraciadamente no siempre será posible. El valor de H estará comprendido entre 1 y β.

Lógicamente un valor de H elevado, hará que se obtenga la fuente deseada bien recuperada, pero también incluirá otras fuentes, complementariamente un valor pequeño, hará que se discrimine mejor entre las fuentes, pero podríamos eliminar información que nos interesa captar. Por ello en la práctica definiremos el índice de discriminación d y nos desplazaremos *-H/2* un lado y *H/2* para el otro, recaerá sobre el usuario determinar que anchura de azimuth le interesa tener, así como el punto de discriminación.

Llegados a este punto tendremos una matriz de BinsxH, por cada frame de tiempo, para poder recuperar la señal podremos estimarla empleando las siguientes ecuaciones:

$$Y_R(K) = \sum_{i=d-H/2}^{d+H/2} Az_R(k,i), para\ i \leq k \geq N \tag{28}$$

$$Y_L(K) = \sum_{i=d-H/2}^{d+H/2} Az_L(k,i), para\ i \leq k \geq N \tag{29}$$

Donde $Y_R(K)$ e $Y_L(K)$ son la magnitud del espectrograma.

Con las ecuaciones (28) y (29), quedaría estimada la magnitud del espectrograma, para la fuente de interés. Por último la fase de la señal también es necesaria, para poder recuperar la señal mediante la ISTFT. Luego deberemos tomar la señal original y añadirlo a nuestra fuente estimada.

Para restaurar la información de fase, obteniendo la forma polar de la forma compleja:

$$\Phi_L(k) = \angle(Lf(k)) \tag{30}$$

$$\Phi_R(k) = \angle(Rf(k)) \tag{31}$$

La parte real e imaginaria del espectrograma de la fuente deseada es estimada como:

$$\hat{S}(K) = \begin{Bmatrix} \Re S(k) = Y(k)\cos\left(\emptyset(k)\right) \\ \Im S(k) = Y(k)\sin(\emptyset(k)) \end{Bmatrix} \quad (32)$$

Donde $\hat{S}(K)$ es el espectrograma complejo. Finalmente cada frame de tiempo es re-sintetizado mediante la IFFT:

$$\hat{s}(t) = \frac{1}{N}\sum_{k=1}^{N} \hat{S}(K)e^{\frac{j2\pi}{N}kt}, para\ t = 1{,}2 \dots N \quad (33)$$

Los frames de tiempo son entonces re-combinados usando un overlap.

Como se ha descrito en esta sección, ADRess es un algoritmo que explota las características de la música estéreo. Pero presenta dos inconvenientes, que el algoritmo es incapaz de identificar las fuentes por sí mismo, y por tanto se requiere de dos parámetros (índice de discriminación y ancho de azimuth) que deben ser introducidos por el usuario. Por otro lado es incapaz de extraer fuentes muy próximas entre sí. Aunque sin embargo presenta otra gran ventaja, y es que no requiere de las muestras futuras como pasaba en la sección en la etapa anterior con el filtro de mediana, es decir limitando los factores como beta y los fragmentos para análisis, podría ser un algoritmo útil para implementación en tiempo real, teniendo en cuenta el parámetro β de resolución y su efecto en el tiempo de computación.

En la figura 55, se presenta el diagrama de bloques para esta etapa en concreto:

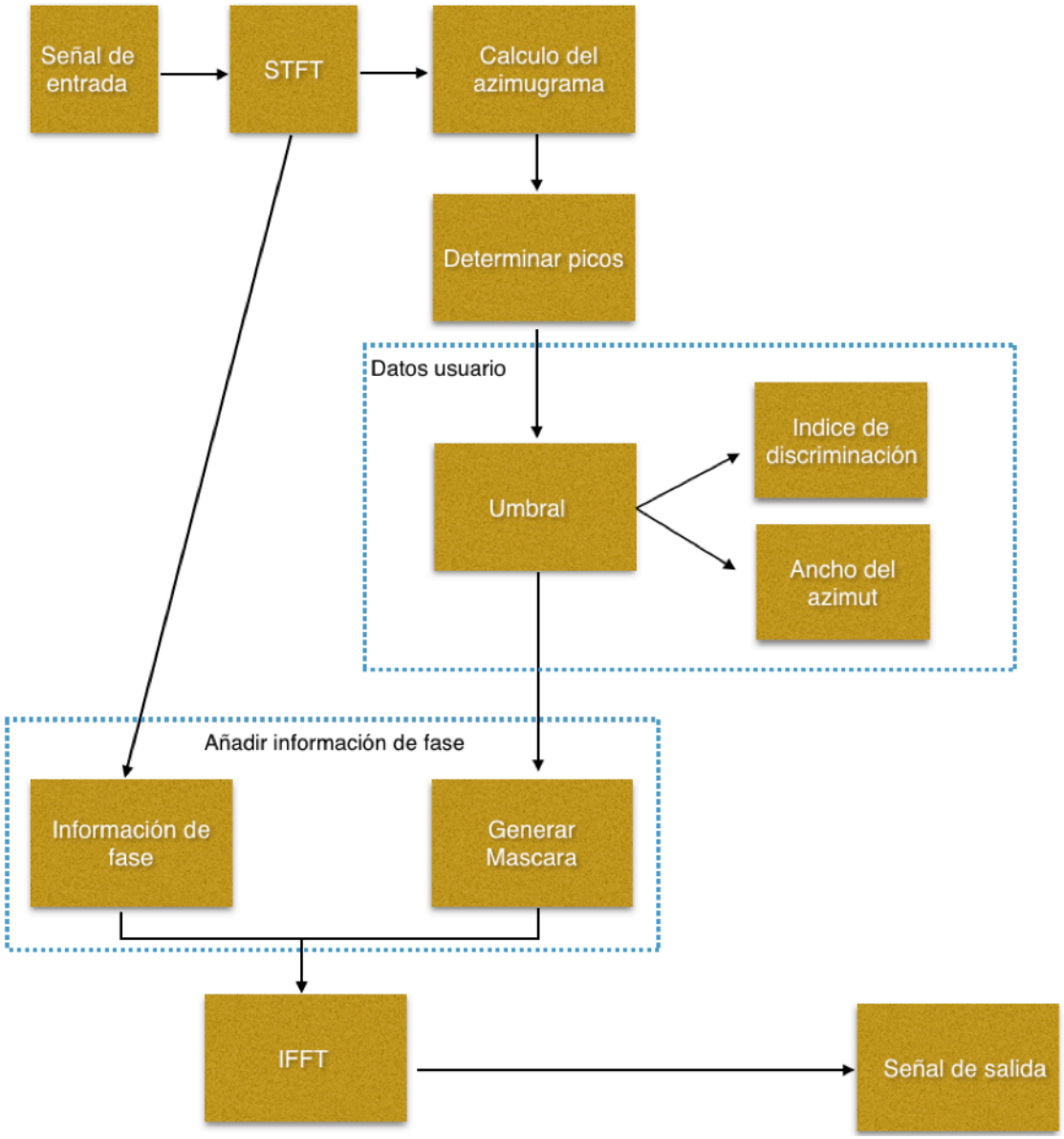

Figura 55: Diagrama de bloques de la etapa ADRess

4.4 ETAPA 3: NMF (NON NEGATIVE MATRIX FACTORIZATION)

Como se ha comentado a la salida de la segunda etapa, es posible que cierta parte de los sonidos percusivos próximos a la voz, no se eliminen correctamente, y por ello es necesario una etapa capaz de filtrar mejor esos restos percusivos, de la señal de salida.

Una de las posibles soluciones, y que está siendo muy empleada hoy día, son las que se basan en el "Training Data", NMF es uno de esos sistemas que se emplean para la separación de fuentes cuando como base del problema se tiene un matriz no negativa, como es el caso de la STFT.

NMF consiste en la factorización de una matriz con el criterio de que todas las matrices en la factorización son no negativas. Se trata de un método para descomponer señales de audio en una combinación ponderada de varios vectores base no negativos. Esta característica facilita que el NMF sea utilizado en la separación de fuentes de audio. Aunque en este caso se emplea una variante no muy utilizada del NMF por lo general, suele emplearse los sistemas de NMF como entrenamientos para las diferentes fuentes ocurrentes en la melodía, pero en nuestro caso la idea es más simple. Solo entrenaremos la base de percusivos y supondremos que todo lo demás es voz, por ellos es importante que todos los sistemas previos hayan funcionado correctamente, puesto que de no ser así es algoritmo será incapaz de funcionar de forma coherente. A este tipo de sistemas se le conoce con el nombre de NMF semi-supervisado [11].

Antes de centrarnos en la etapa en concreto, se explicarán los conceptos y los parámetros de un sistema NMF. Se supone una matriz de datos todos ellos positivo que llamaremos $V,$ se pude descomponer esta matriz V en dos matrices W y $H,$ donde su producto sea aproximadamente la matriz $V,$ esto es:

$$[V] \approx [\hat{V}] = [W][H] \tag{34}$$

Donde V representaría el espectrograma original de la mezcla, W es la matriz de vectores bases y H sus respectivas activaciones, en otras palabras W representaría los patrones espectrales y H el momento en el que pasan.

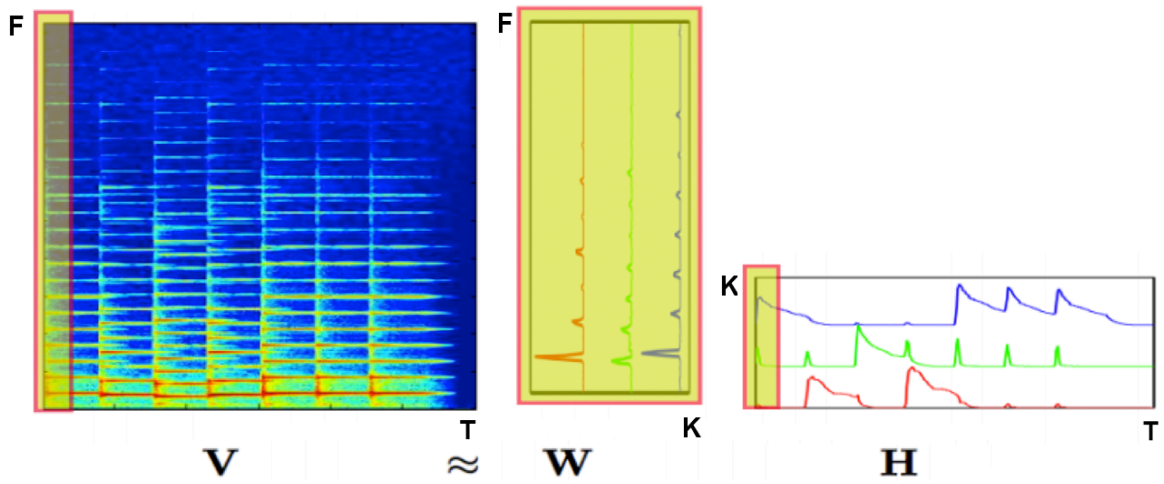

Figura 56: Descomposición NMF [12] para un valor de bases igual a 3

En la figura 56, se puede apreciar un ejemplo de NMF para $k = 3$, k es el primer parámetro de interés de un sistema NMF, nos indica el número de bases que se quieren considerar. En otras palabras un sonido percusivo puede quedar completamente estimado con $k = 2$ o $k = 3$, sin embargo la voz humana tiene muchos más registros o las variaciones en frecuencia son mucho más amplias por ello para la voz un $k = 64$, son valores típicos. En la figura, también se puede ver la descomposición, de la matriz de activación H, que muestra los pesos de esos vectores bases W en cada instante de tiempo. Matricialmente esto se expresa en la ecuación (35):

$$\begin{bmatrix} V_{1x1} & \dots & V_{1xM} \\ \vdots & \ddots & \dots \\ V_{Nx1} & \dots & V_{NxM} \end{bmatrix} \approx \begin{bmatrix} W_{1x1} & \dots & W_{1xM} \\ \vdots & \ddots & \dots \\ W_{kx1} & \dots & W_{kxM} \end{bmatrix} \times \begin{bmatrix} H_{1x1} & \dots & H_{1xk} \\ \vdots & \ddots & \dots \\ H_{Nx1} & \dots & \mathrm{H}_{Nxk} \end{bmatrix} \tag{35}$$

Donde N representa los bins de frecuencia y M los frames de tiempo.

La cuestión es como se puede obtener una matriz W y H en base solo a una matriz V de tal forma que la diferencia entre V y WH sea mínima. Se trata de un problema de optimización y como tal la respuesta es:

$$\min \; D(V \parallel WH) \tag{37}$$

Donde D representa la divergencia.

Luego se necesita una función de peso para el calculo de la divergencia, ya se han comentado y usado algunas como la distancia Euclidea, empleada en la etapa 1:

$$D\left(\hat{V} \parallel V\right) = \sum_{i,j} (V_{ij} - \hat{V}_{ij})^2 \tag{38}$$

Sin embargo para este caso emplearemos otra función de peso conocida como la Kullback-Leibler (KL)[19]:

$$D\left(\hat{V} \parallel V\right) = V log_{10}\left(\frac{V}{\hat{V}}\right) - V + \hat{V} \tag{39}$$

La justificación del uso de la ecuación (39) a favor de la (38), se debe principalmente a que KL presta la misma atención a los cambios grandes de amplitud que a los pequeños mientras que la Euclidea al ser una función cuadrática no tiene da el mismo peso para ambos casos.

Ahora la cuestión es minimizar la función (39), teniendo en cuenta que $\hat{V}$ está representado por WH. Existen múltiples formas de minimizar a función, el método en el que se basa esta etapa consiste en optimizar H para minimizar W y optimizar W para minimizar H, de forma iterativa un número determinado número de veces hasta converger. Como se demuestra [12]:

$$H \leftarrow H \otimes \frac{W^T \frac{V}{WH}}{W^T \cdot 1} \tag{40}$$

$$W \leftarrow W \otimes \frac{\frac{V}{WH} H^T}{1 \cdot W^T} \tag{41}$$

En la ecuación (40) y (41), representan la forma de optimizar H y W respectivamente, donde $\otimes$ representa el producto de Hadamard, $\cdot$ representa el producto de matrices, W^T y H^T representan las matrices traspuesta de W y H respectivamente y 1 es la matriz unidad de dimensiones igual a NxM siendo N el número de bins y M el número de frames.

CASO PARTICULAR

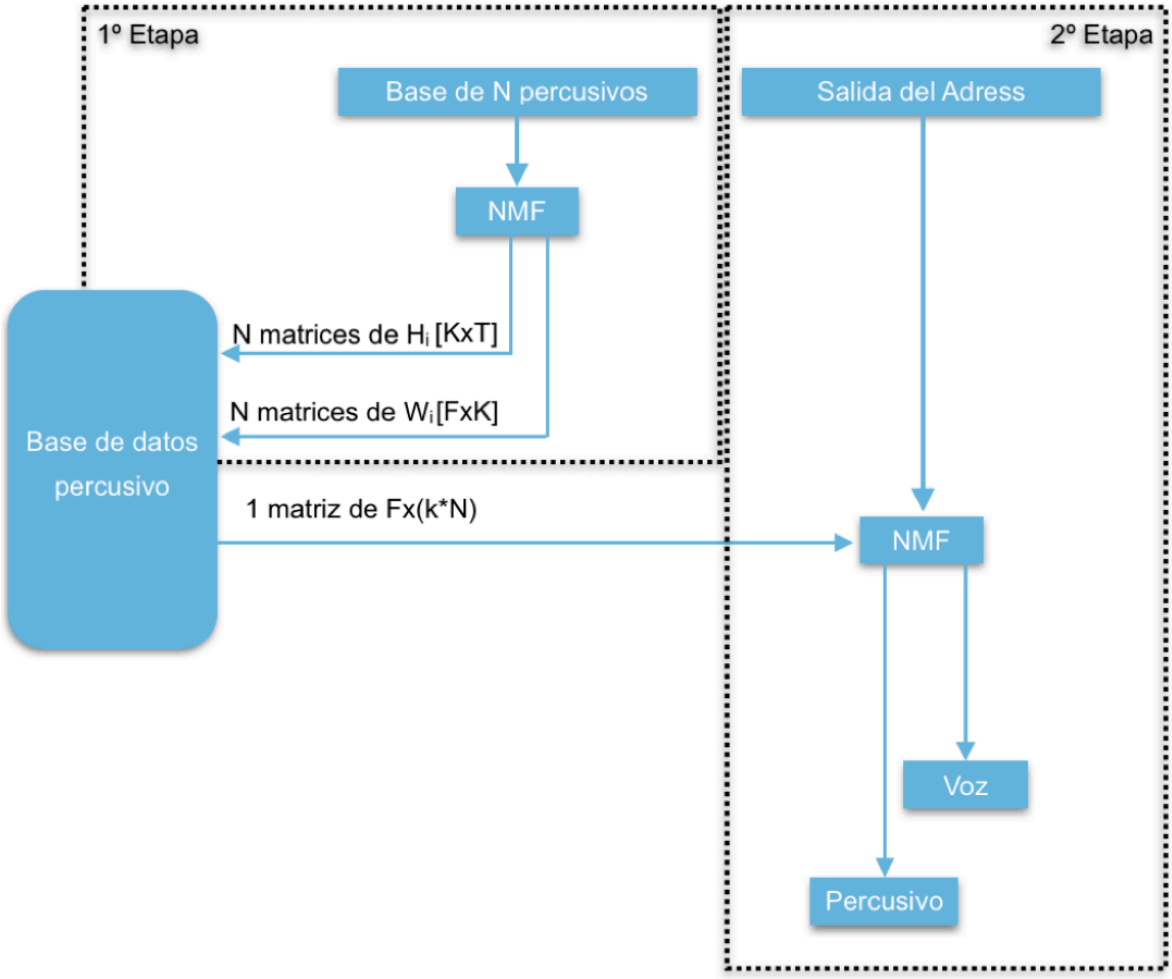

Figura 57: Diagrama de bloques de la etapa 3 de NMF

En la figura 57, se puede observar el diagrama de bloques de la etapa 3 NMF, como se puede ver consta dos sub-etapas, una primera sub-etapa que llamaremos de entrenamiento de percusivos, en la que se "enseña" al sistema el tipo de percusivos que va a identificar, de tal forma que se genera una matriz de dimensiones $NxLk$, donde L es el número de pistas de percusivos con el que se va a entrenar, N el número de bins de frecuencia y k es el número de bases que se considerar en cada pista. Posteriormente con esa matriz *training* se inicia la segunda sub-etapa que comienza con la pista estéreo que sale en la segundo etapa, donde se toma como premisa que es pista solo contendrá

frecuencias correspondiente a la voz y los percusivos de la parte instrumental. La idea en rasgos generales es aplicar dos NMF en esta segunda sub-etapa, uno que estime las activaciones y matriz de bases de la voz y otra que estime la matriz de activación de los percusivos únicamente ya que la matriz *training* representa la matriz de base de los percusivos.

Se supone una base percusiva de entrada con una duración aproximada de 0.3 segundos, cuyo espectrograma es:

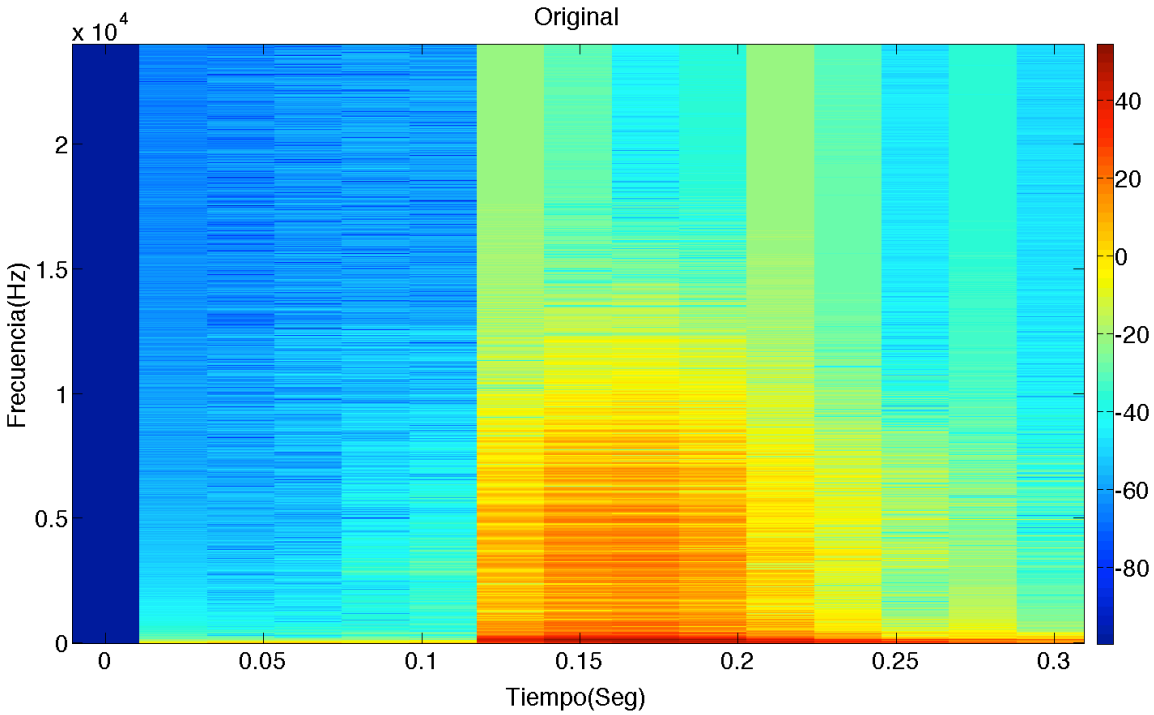

Figura 58: Espectrograma de entrada de NMF para la etapa de entrenamiento

Aunque se comentó que en esta etapa se va a iterar un número de veces determinada para la descomposición, también se podría en este momento establecer un umbral que asegure un error determinado. Para este caso se emplea un número de bases de $k = 3$, con un total de 100 iteraciones, haciendo uso de las ecuaciones (40) y (41), obtenemos las matrices de bases W y la matriz de activaciones H. Esto se ve en la figura 59:

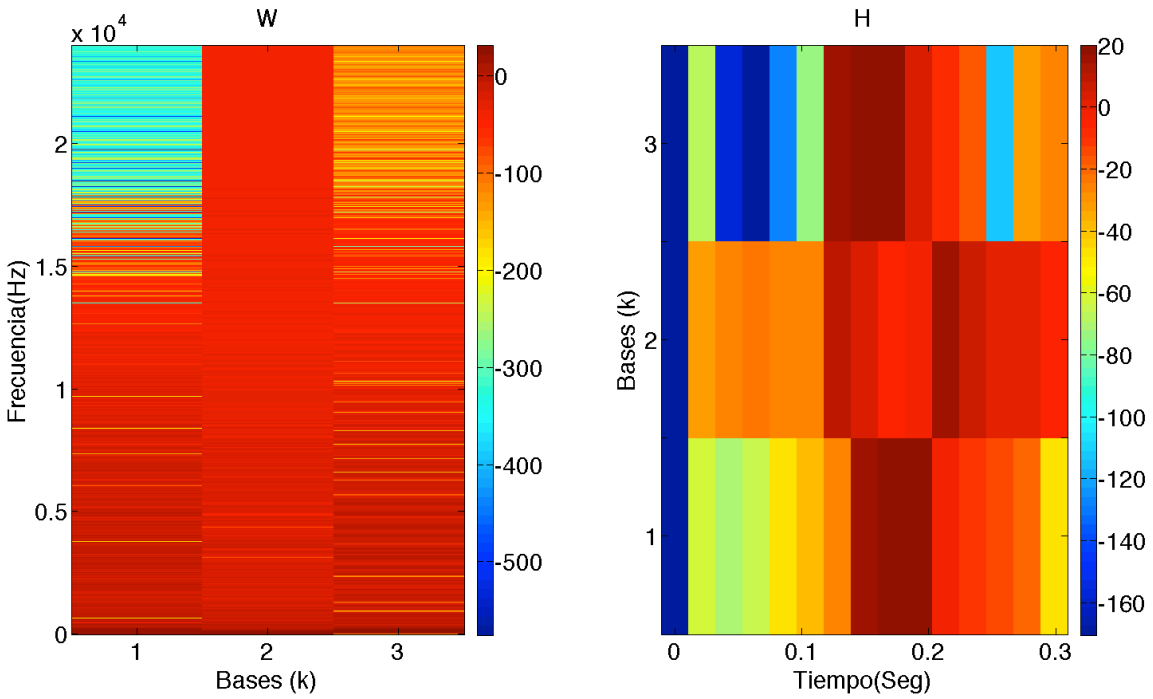

Figura 59: Espectrogramas de la descomposición de la matriz de bases W y de activación H

Para ver que efectivamente la ecuación (39) cumple con la convergencia en la figura 60 se puede ver como tras 100 iteraciones el error es mínimo, es más a partir de la iteración 20 la diferencia es mínima, pero como el número de iteraciones es un valor fijo quizás para otras pistas de entrenamiento no sean suficiente 20 iteraciones, por ello se toma un valor bastante por encima del que se estima necesario.

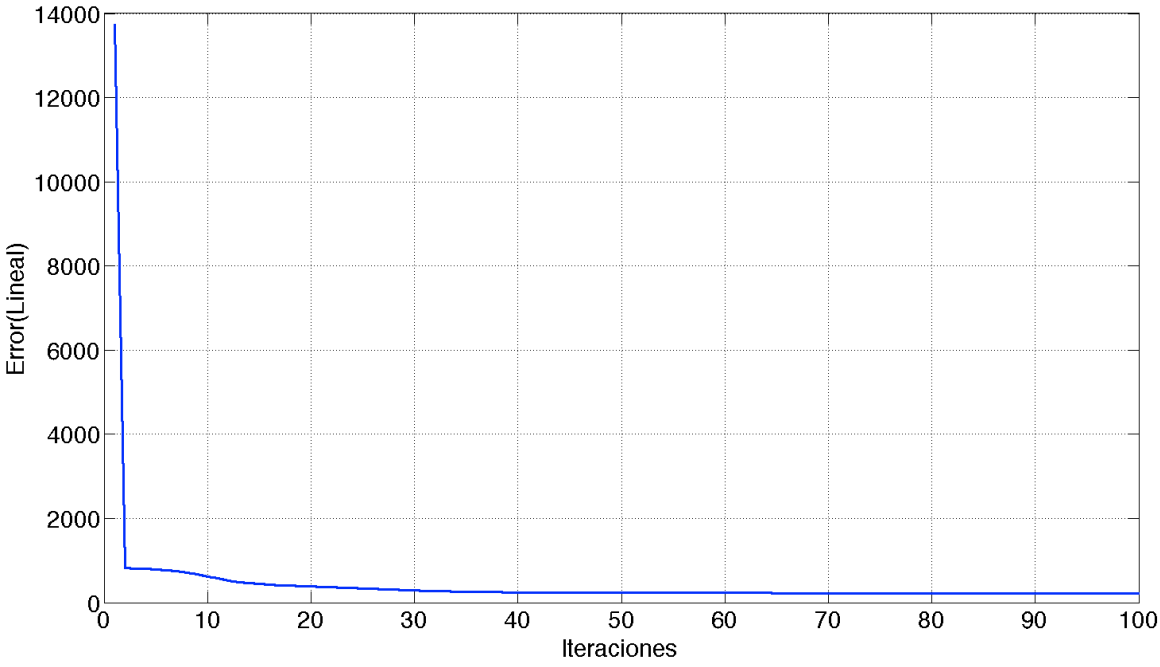

Figura 60: Convergencia de la KL para 100 iteraciones

Como se observa en la figura 60, el error es apenas de 140, un error asumible teniendo en cuenta que se trata de una señal de 30735 muestras. En la figura 61, se aprecia una comparativa entre la señal original con la señal reconstruida:

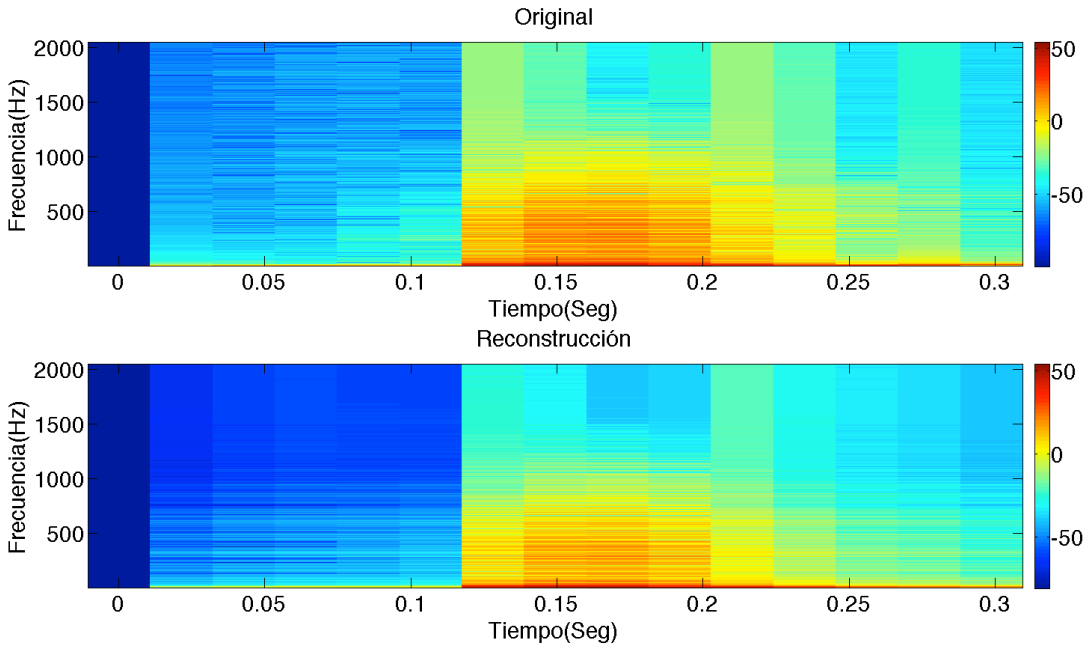

Figura 61: Comparativa de la señal original y la estimación tras NMF, para 100 iteraciones y k=3

Con la figura 61, queda demostrado que los cálculos expuesto en la etapa 3, son consistentes y generan resultados prácticos válidos.

Resulta curioso analizar el caso extremo en el que $k = 1$, y para un número mucho menor de iteraciones unas 10 aproximadamente de forma que el error sea bastante grande, vemos como apenas distinguimos una única fuente:

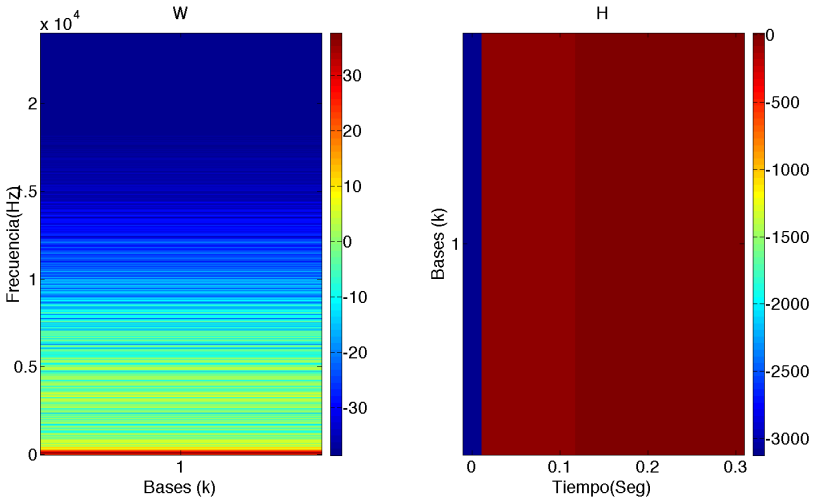

Figura 62: Descomposición de NMF para un valor de k=1

Creando conflictos a la hora de reconstruir la señal:

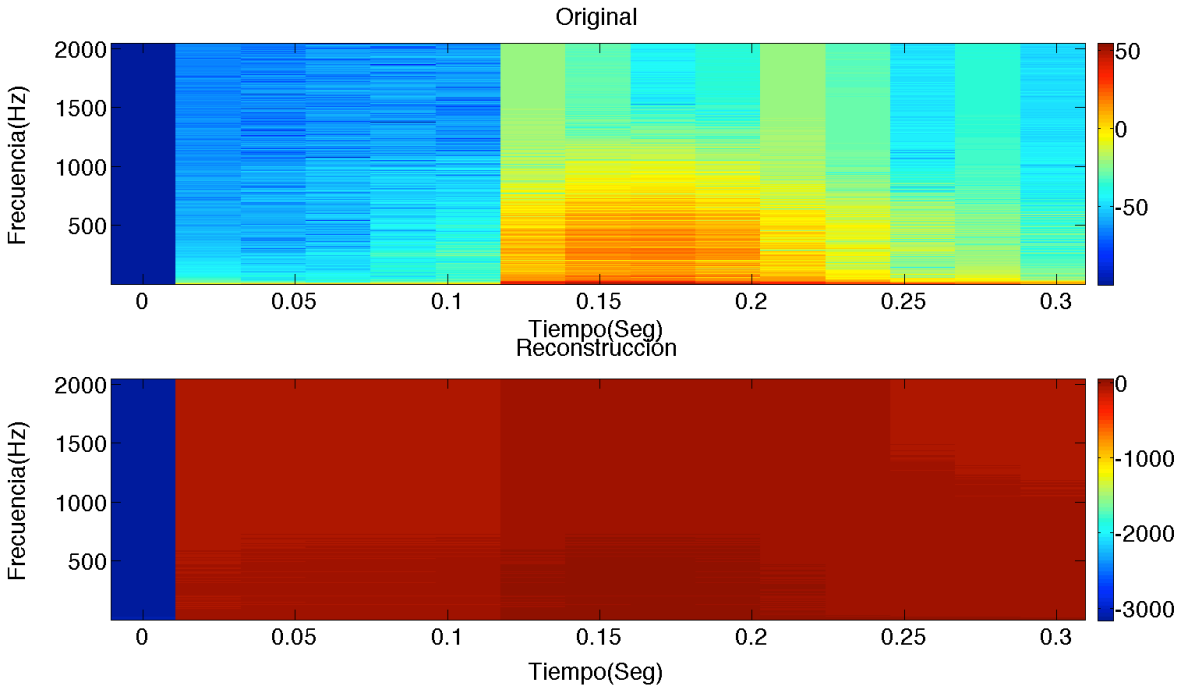

Figura 63: Comparativa entre la señal original y la estimada para k=1 y 10 iteraciones

Donde prácticamente se hace irreconocible la señal, basta con fijarse en la magnitud de las dos señales, por ello será importante tanto la elección de unas iteraciones lo suficientemente elevadas que permitan minimizar el error al máximo, así como un número de bases percusivas aceptable, el número de bases percusivas suele estar en torno a 2 ó 3, aunque lógicamente depende mucho de la señal que se esté analizando.

Sin embargo si se vuelve a 100 iteraciones pero se mantiene $k = 1$, se observa como la señal reconstruida es muy similar a la original. En la figura se aprecia una buena reconstrucción aunque quizás por debajo de su valor en intensidad, se puede ver que los tiempos están bien reconstruidos y bien proporcionados, pero el error cometido hace que se pierda energía en las frecuencias altas de la señal.

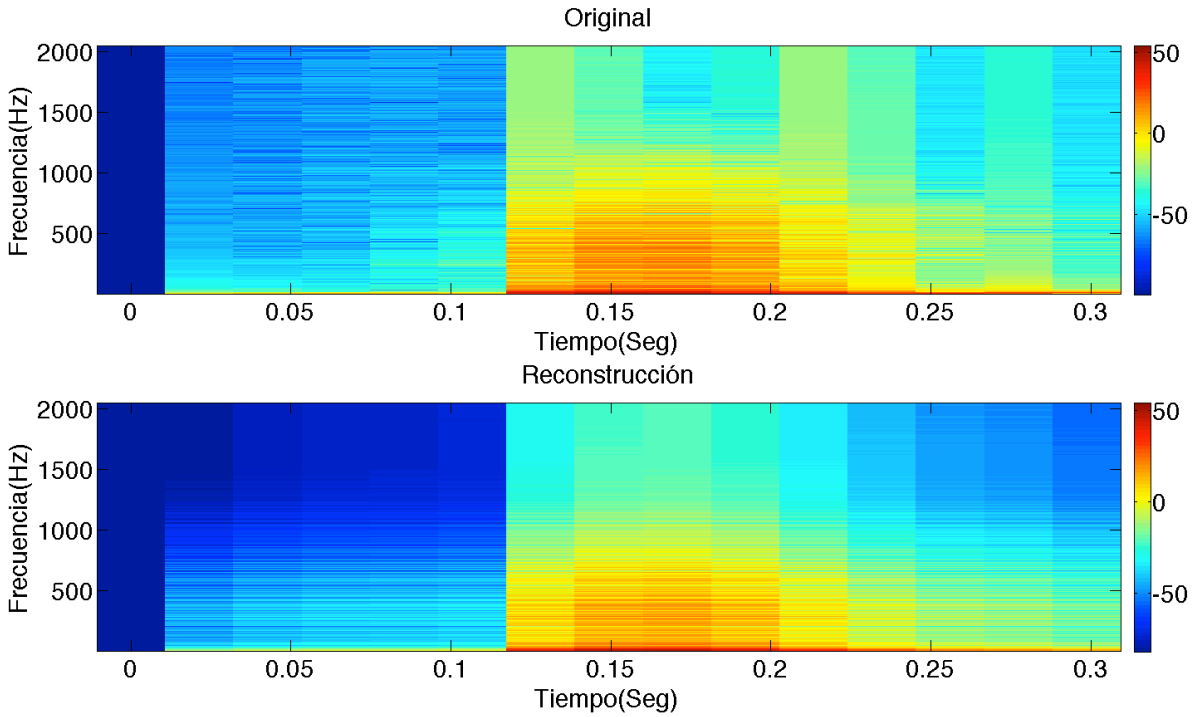

Figura 64: Comparativa entre la señal original y la reconstrucción para k=1 y 100 iteraciones

La conclusión a la que se llega es que el aumentar las bases k no implica una mejora directa de la reconstrucción de la señal, ya que la figura 64, es prácticamente similar a la 61, y de igual forma en la figura 60 vemos que a partir de un cierto número de iteraciones el error es prácticamente constante, con lo que habrá que buscar un compromiso entre iteraciones y valor óptimo de bases.

En este punto se completaría la primera sub-etapa de entrenamiento, simplemente se debería repetir para cada pista que se desee entrenar. La matriz que es de interés de esta etapa es W, de forma que después de L pistas de entrenamiento se pueda montar la matriz *training* W_T*:*

$$[W_T] = [W_1 \quad \dots \quad W_L] \tag{42}$$

En la segunda sub-etapa, la forma de realizar la estimación del NMF, se realiza de una forma un poco diferente a la primera sub-etapa, hasta ahora se había usado la ecuación (34), pero ahora empleamos una ecuación que agrupa dos ecuaciones como la (34):

$$[V] = [S] + [P] \tag{43}$$

Donde $[V]$ es la matriz del espectrograma de entrada, $[S]$ representa la matriz de la parte de la pista que pertenece al cantante y $[P]$ la matriz de la parte de la pista que pertenece al percusivo. De igual forma que en la ecuación (34), se puede descomponer ambas partes en:

$$[S] = [W_S][H_S] \tag{44}$$

$$[P] = [W_P][H_P] \tag{45}$$

En donde W_S son los vectores bases de la voz y H_S sus respectivas activaciones, mientras W_S que W_P son los vectores bases del percusivo y H_P sus respectivas activaciones. Resultando entonces:

$$[\hat{V}] = [W_S][H_S] + [W_P][H_P] \tag{46}$$

Llegando a la conclusión de que realmente se está planteando dos NMF independientes de las que una función de una de ellas es conocida, entrenada previamente. W_P representaría la matriz entrenada, es decir W_T (41) De forma que únicamente necesitaríamos estimar W_S, H_S y H_P. El proceso de estimación para cada una es exactamente el mismo que el explicado con anterioridad, simplemente teniendo en cuenta que se debe emplear la reconstrucción a través (45).

Otro punto a tener en cuenta es que en la parte de entrenamiento hemos supuesto que $k = 3$, ya que se trataba de entrenar fuentes percusivas, donde claramente estamos limitados a un número pequeño de sonidos, pero cuando se estima W_S, no se debe olvidar que se trata de estimación de voz, donde el número de fuentes sonoras, o de variables del sonido que pueden producirse son mayores luego el valor por lo general estará comprendido entre 16 y 128 fuentes sonoras para voz. Como es lógico todo este

rango de valores es muy relativo y dependerá en gran medida de nuestras bases, así como de la canción a tratar.

Una vez obtenido los diferentes valores de $[W_s]$, $[H_s]$, $[W_P]$ y $[H_P]$ se puede ser capaz de generar un par de mascaras suavizadas que permitan extraer de la señal original los fragmentos de voz y de percusivo.

$$mascara_P = \frac{P}{P+S} \tag{47}$$

$$mascara_s = \frac{S}{P+S} \tag{48}$$

Donde (47), representa la máscara de percusivo y (48) la de voz, siendo S y P en las ecuaciones (45) y (44) respectivamente.

ETAPA 3.1: SPARSENESS

En un intento de mejorar el sistema de NMF, se emplea una técnica que aprovecha una de las propiedades de la voz humana. Como se ha visto en la sección 1.2.3 la voz humana es monofónica, esto quiere decir que solo produce un sonido en un instante dado. Esta propiedad es útil si tenemos en cuenta que en el NMF somos capaces de separar voz y percusión durante el algoritmo.

Sparseness [31], es un complemente añadido a los sistemas NMF, que busca obtener una representación de la señal con el menor número de datos posibles, de forma que se pierda la menor cantidad de información, de esta forma se consigue que el solapamiento que se produce debido a los armónicos en las diferentes frecuencias se limite. Conceptualmente consiste en añadir un valor de restricción λ que ponderado entre 0 y 1, siendo 1 el más restrictivo.

La forma de implementación sobre NMF es sencilla, teniendo en cuenta que se requiere emplear de una normalización respecto a FT, siendo F los bins de frecuencia y T los frames de tiempo.

$$X_{FT} = \frac{X_{FT}}{\sigma} \; con \; \sigma = \frac{\sum_{FT} X_{FT}}{FT} \tag{49}$$

En la ecuación (46), se puede ver un factor de normalización necesario sobre la señal de entrada. Si se sigue con la función de peso de KL, si se añade la restricción quedaría:

$$D\left(\hat{V} \parallel V\right) = V log_{10}\left(\frac{V}{\hat{V}}\right) - V + \hat{V} + \lambda_V \sum\sum H_V + \lambda_P \sum\sum H_P \tag{50}$$

Se observa como la ecuación (50), es similar a la (39), pero con el añadido de los valores restrictivos, donde $\lambda_V \sum\sum H_V$, es la restricción aplicada a la voz y $\lambda_P \sum\sum H_P$ es la restricción para la parte instrumental de percusivos. Como se observa la restricción solo afecta a la matriz de activación, λ_V y λ_P, son los valores de restricción para cada parte vocal y percusiva respectivamente, esto lógicamente afecta a la función de optimización de la matriz H_V y H_P, que se recalcularía como:

$$H \leftarrow H \otimes \frac{W^T \frac{V}{\overline{WH}}}{W^T \cdot 1 + \lambda C} \tag{51}$$

El término C, es un valor de peso que hace que todas las restricciones tengan la misma importancia, es decir como se ha visto en la ecuación (50), se han añadido dos nuevos términos, de tal forma que si analizamos la cantidad de datos que aporta cada término en la ecuación (50):

$$V log_{10}\left(\frac{V}{\hat{V}}\right) - V + \hat{V} \rightarrow FT \tag{52}$$

$$\lambda_V \sum \sum H_V \to KT \tag{53}$$

$$\lambda_P \sum \sum H_P \to KT \tag{54}$$

Donde K, representa el número de bases, T el número de frames y F el número de bins. Lógicamente $FT > KT$, en cuanto a número de datos, luego necesitamos multiplicar los dos nuevos términos por un factor de ponderación que permita tener el mismo peso que el primer término, es decir C:

$$C = \frac{F}{K} \tag{55}$$

Por último y para ver los efectos de la estimación con y sin sparseness, supongamos una tono percusivo cuyo espectrograma es el de la figura 65:

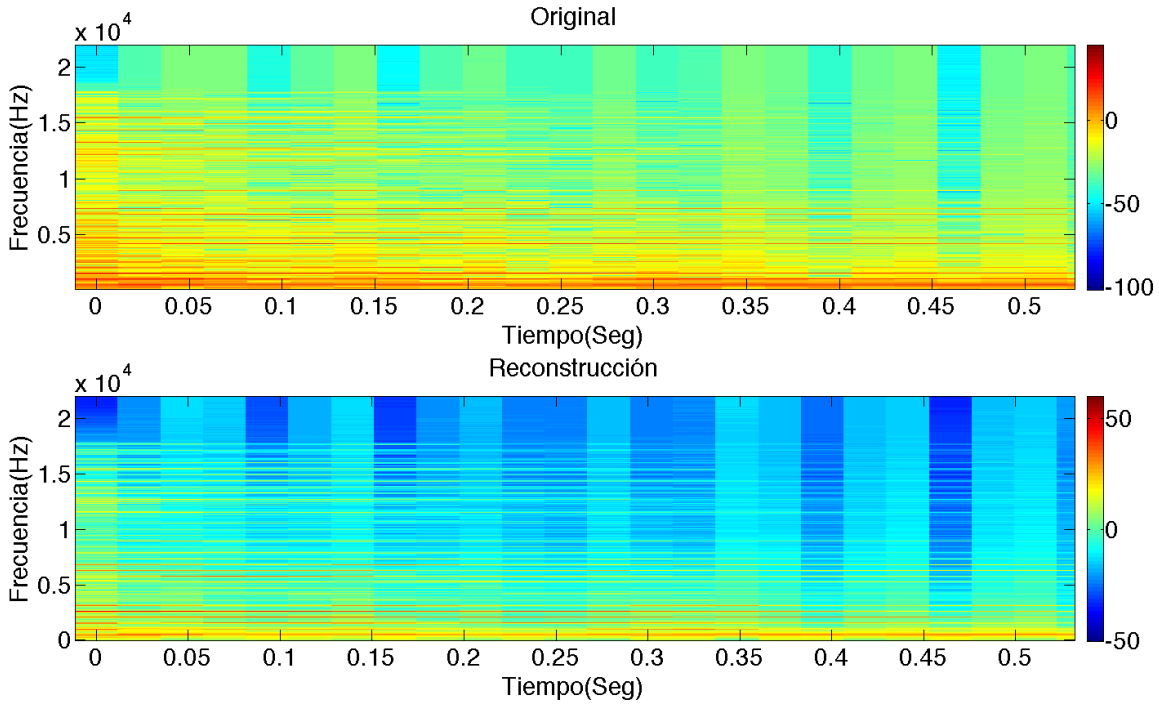

Figura 65: Espectrograma de señal a estimar sin y con sparseness

Supongamos que se toman las restricciones comentada y tomamos $\lambda = 0$ y $\lambda = 1$, las dos situaciones extremo, si vemos las funciones de error de ambas, vemos que para el tono instrumental de la figura 65, la estimación se realiza con un menor error sin la restricción:

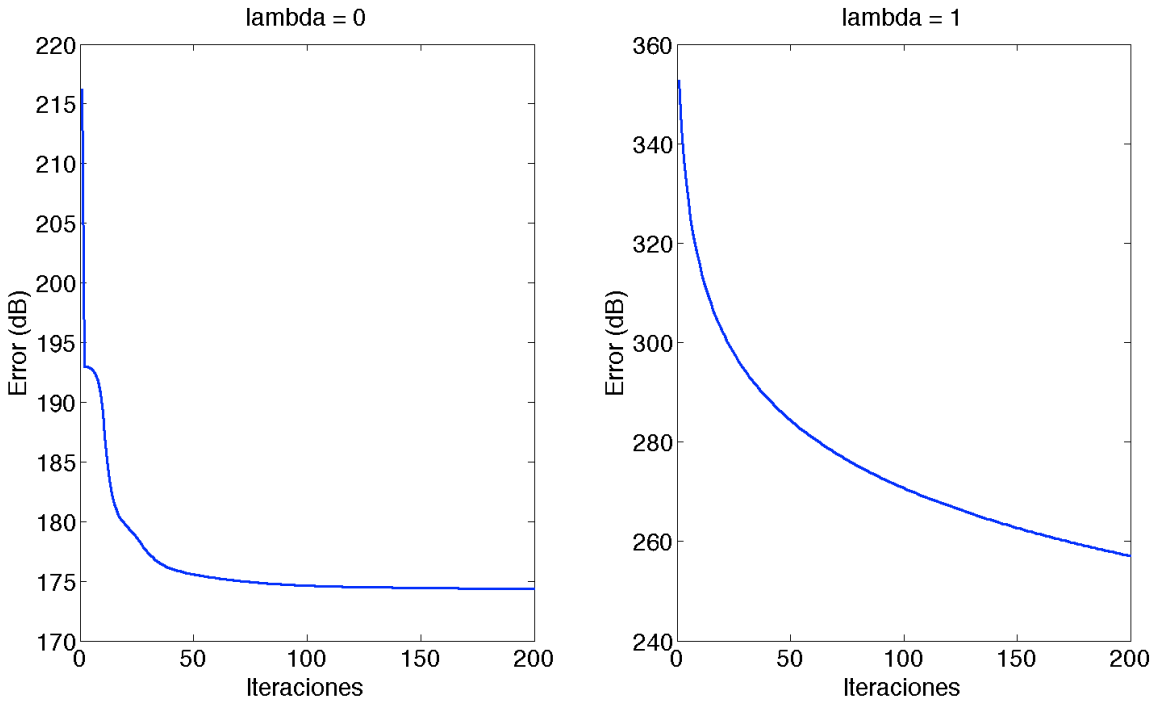

Figura 66: Comparativa del error que se comete en la aproximación para distintos valores de λ

4.6 ETAPA 4: ESTIMACIÓN DE ZONAS DE VOZ

Como se ha podido ver en la etapa 3, NMF consigue identificar bien los restos de percusivos que quedaban entorno a la voz. Esta última etapa, se aporta como método novedoso y completamente nuevo. El problema del NMF, es que actúa por igual en toda la canción, es decir rebajas los percusivos tanto en la parte cantada como en la que no, pero la parte en la que no hay voz podría ser eliminada por completo, esa es la idea de esta cuarta etapa. Supongamos la figura 67 como pista a analizar:

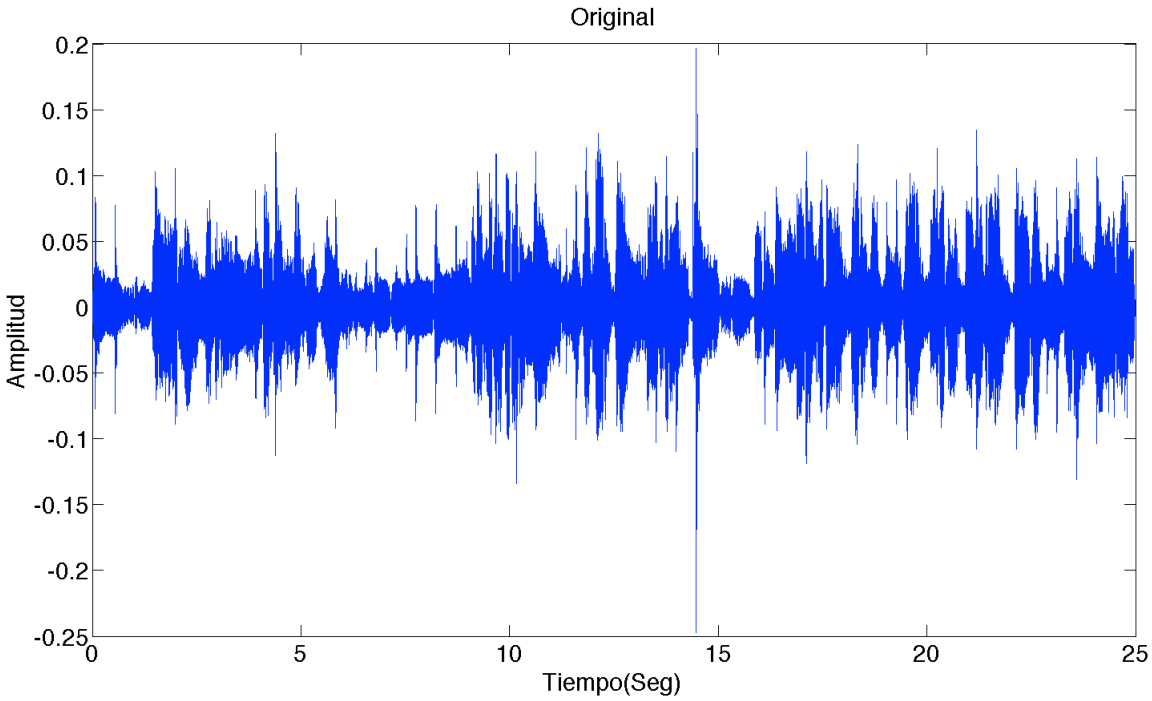

Figura 67: Señal de entrada para analizar

El efecto que produce esta etapa, se puede ver en la figura 68, donde se muestra en la parte superior la salida del NMF y donde las flechas indican las zonas que están rebajadas con respecto a la original que ha identificado como percusivo, por simple inspección se puede ver que el nivel de intensidad de la voz está muy por encima de esta:

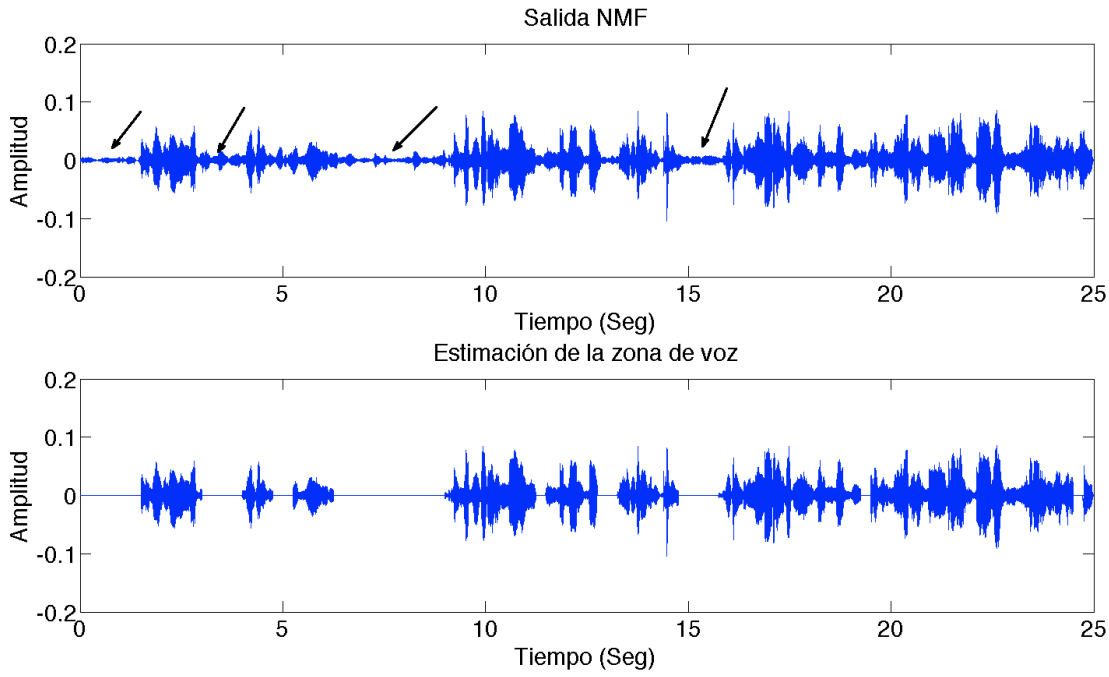

Figura 68: Salida del NMF (superior) y salida de la estimación tras la etapa 4 (inferior)

Como es lógico se necesitará de un umbral que diga que partes son percusivos y que partes voz, donde tras varias pruebas se tomó como umbral el valor de medio del conjunto de muestras de la pista. Este umbral esta basado en el tipo de canción que se presupone de antemano que se va a analizar y como se ha comentado varias veces son canciones repetitivas donde la voz del cantante no tiene grandes oscilaciones de intensidades, es decir música tipo pop, country…etc.

El funcionamiento es muy sencillo, se basa en analizar fragmentos de 0.25 segundos, debido a que es la duración aproximada de un sonido percusivo, donde se estima un valor para el sub-conjunto de datos tomados mediante un filtro de mediana, y si se encuentra por debajo del valor umbral X_{umbral}, se tomará como percusivo y será descartado de la pista de voz. En la figura 69, se ilustra este proceso:

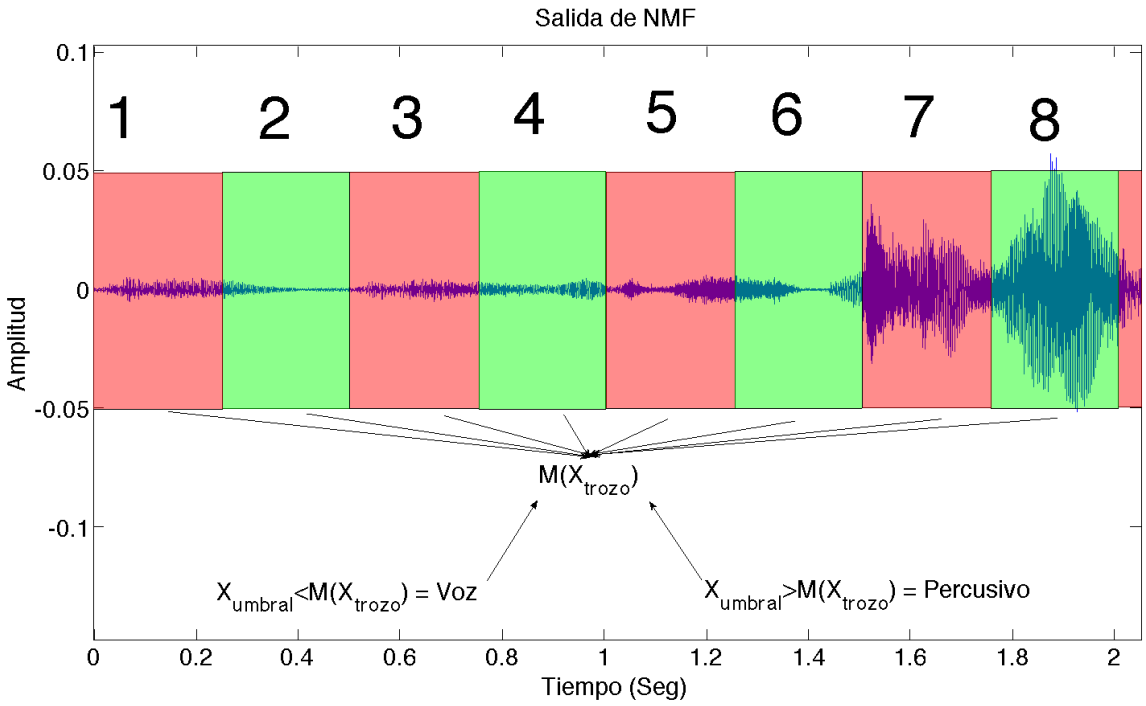

Figura 69: Ilustración del proceso de estimación de zonas de voz

En la figura 69, se ilustra el proceso selección de la voz, como se puede ver los fragmentos 1,2,3,4,5 claramente tiene mucha menos intensidad que fragmentos como el 7 o el 8 por ello serán detectados como umbral. Donde $M(X_{trozo})$ representa el operador mediana del subconjunto de muestras.

Como se viene comentando, lo efectivo que resulten las etapas anteriores influencian fuertemente el resultado de esta etapa, ya que está basada en que NMF es capaz de identificar bien los percusivos con su entrenamiento y aunque no los elimine si sea capaz de atenuarlos.

4.7 ETAPA 5: SEPARACIÓN ARMÓNICO PERCUSIVO

Con la etapa 4, se concluye la separación de la voz, con lo que si se vuelve a la figura 40 del diagrama del módulo, retomaríamos el camino izquierdo donde simplemente se aplica una única etapa con el objetivo de separar armónicos de percusivos.

Esta técnica hace uso de un filtro de mediana sobre el espectro de la señal de audio, este filtro analizará la forma del espectro de la señal y con estos resultados se podrán generar las mascaras que procesarán la señal original. En la práctica es más un algoritmo de análisis de imagen más que de audio [13].

Esto se puede apreciar en la siguiente imagen, donde claramente se observa la forma horizontal de los percusivos:

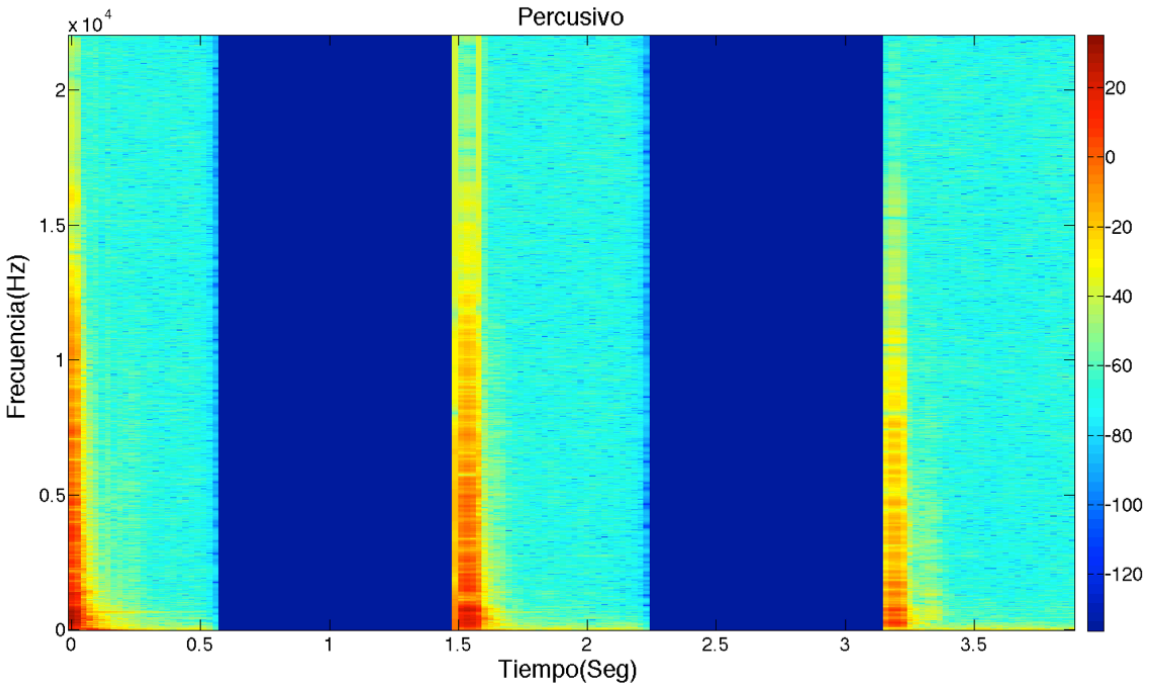

Figura 70: Representación de tono percusivo

Por contra si nos fijamos en una muestra de armónicos de una melodía, podemos comprobar como la energía se distribuye de forma horizontal:

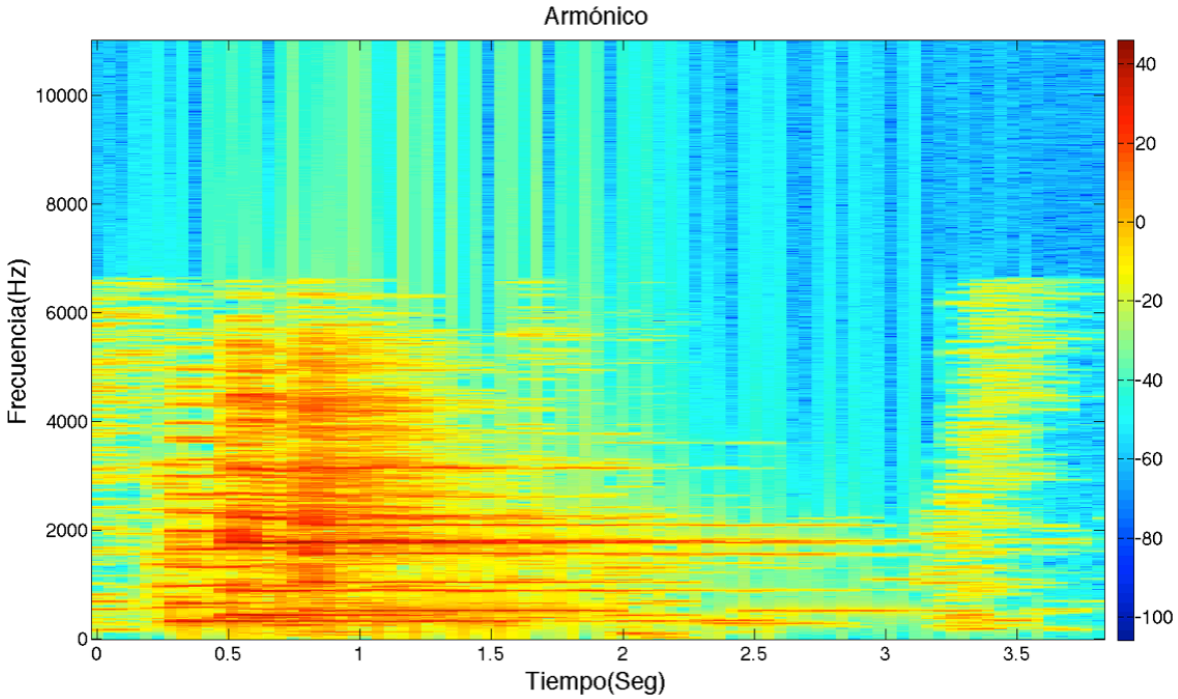

Figura 71: Representación de tonos armónicos

En definitiva, se aprovechan las características propias de armónicos y de percusivos que se han comentado en la sección 1.2, donde se indicó que los percusivos tenía una estructura corta en el tiempo pero de banda ancha mientras que los armónicos todo lo contrario. En otras palabras más de procesado de imagen, los percusivos son entendidos como eventos verticales en el espectrograma y los armónicos como eventos horizontales.

Los filtros de mediana han sido muy usados en el procesado de imágenes, para eliminar ruido de puntos, o de sal y pimienta. El funcionamiento del filtro de mediana es sencillo, se le pasa una muestra de la señal original, tomada mediante un enventanado entorno a la muestra de interés. Dando un vector de entrada $x(n)$ y otro de salida $y(n)$, la longitud del filtro l vendrá determinada por el número de muestras sobre las cuales actúa el filtro. Donde si l es impar, podemos definir el filtro de mediana como:

$$y(n) = mediana\{x(n-k:n+k)\} \quad con \quad k = (l-1)/2 \tag{56}$$

Como se puede observar, la muestra original es reemplazada con la mediana obtenida de la lista ordenada de las muestras en lo vecinos de la muestra original. En

caso de que l sea par, la mediana se obtiene como la media de las dos muestras en el medio de la lista ordenada. De esta forma conseguimos eliminar picos de ruido ya que ellos no siguen el patrón de los valores típicos de la región de análisis.

Dado un espectrograma S, y denotando el i^{th} frame de tiempo como S_i, y el h^{th} bin de frecuencia como S_h, se puede obtener un espectrograma mejorado de percusión P_i, generado un filtro de mediana sobre S_i:

$$P_i = M\{S_i, lperc\} \tag{57}$$

Donde M denota el filtrado de mediana y $lperc$ es la longitud del filtro, de esta forma podemos combinar P_i para producir un espectrograma de percusión P. De forma similar actuamos sobre los armónicos, desplazándonos sobre S_h:

$$H_i = M\{S_h, lharm\} \tag{58}$$

Donde $lharm$ representa la longitud del filtro.

Los resultados de los espectrogramas, tanto armónico como percusivo se pueden usar para generar dos máscaras las cuales posteriormente son aplicadas sobre el espectrograma original. Existen dos tipos de mascaras que se pueden emplear. La primera de estas es una mascara binaria, donde se asume que cada bin de frecuencia pertenece a la fuente percusiva o armónica. En esta caso la mascara se define como:

$$M_{H_{h,i}} = \begin{cases} 1 & si\ H_{h,i} > P_{h,i} \\ 0 & En\ otro\ caso \end{cases} \tag{59}$$

$$M_{P_{h,i}} = \begin{cases} 1 & si\ P_{h,i} > H_{h,i} \\ 0 & En\ otro\ caso \end{cases} \tag{60}$$

Pero como se comentó en la etapa 1, este tipo de máscaras suelen dejar un tipo de ruido de fondo audible por ello es referible usar máscaras suavizadas como la Wiener, que se definen como:

$$M_{H_{h,i}} = \frac{H_{h,i}^{p}}{H_{h,i}^{p} + P_{h,i}^{p}} \tag{61}$$

$$M_{P_{h,i}} = \frac{P_{h,i}^{p}}{H_{h,i}^{p} + P_{h,i}^{p}} \tag{62}$$

Donde p denota la potencia a la cual cada elemento individual del espectrograma es elevado. Por último podemos recuperar el espectrograma complejo de cada señal como:

$$\widehat{H} = S \otimes M_H \tag{63}$$

$$\hat{P} = S \otimes M_P \tag{64}$$

Donde $\otimes$ representa el producto de Hadamard y donde S denota el valor original del espectrograma. Para recuperar la señal, bastará con invertir el espectrograma al dominio del tiempo y se obtendrá la forma de onda armónico y percusivo respectivamente.

Como se puede ver el algoritmo solo requiere de dos iteraciones, lo que lo convierte en una etapa muy eficiente y además al no requerir de muestras futuras al momento de análisis, es un buen algoritmo para emplearlo en tiempo real.

EJEMPLO

Si se emplea el siguiente código con un fragmento de audio a modo de ejemplo como “Billie Jean” de Michael Jackson, el espectrograma de entrada será:

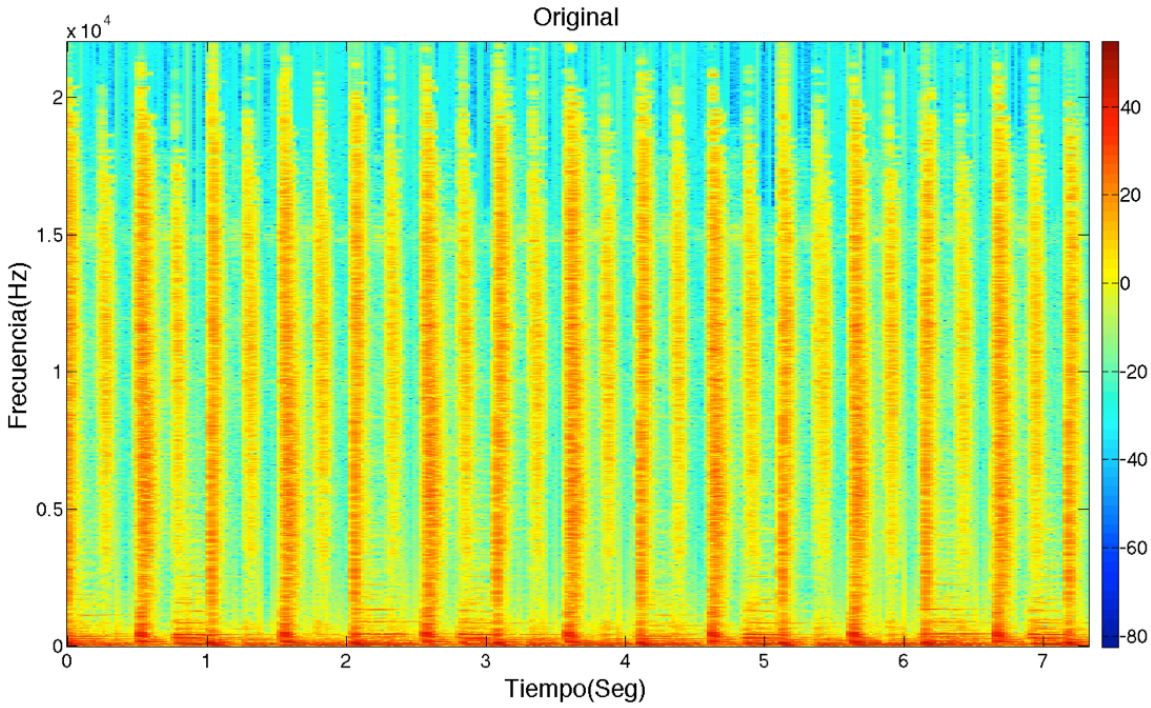

Figura 72: Espectrograma de la señal de entrada a la etapa 5

Claramente se puede ver que el fragmente contiene tanto zonas de armónicos como de percusivos, según el autor [13] tras varios análisis la longitud que mejor funciona para el filtro es de 17, tanto para armónico como percusivo, ya que las variaciones no son tan drásticas como para ser perceptibles, por tanto será lo longitud que se emplee. Por otro lado dice que el valor de p optimo que mejores resultados da para la separación es $p = 2$, luego también se tomará este valor.

Aplicando los dos análisis de mediana sobre el espectrograma de la figura 72, y generando las máscaras de (61) y (62), obtenemos los resultados de la figura 73 para armónico y 74 para percusivo:

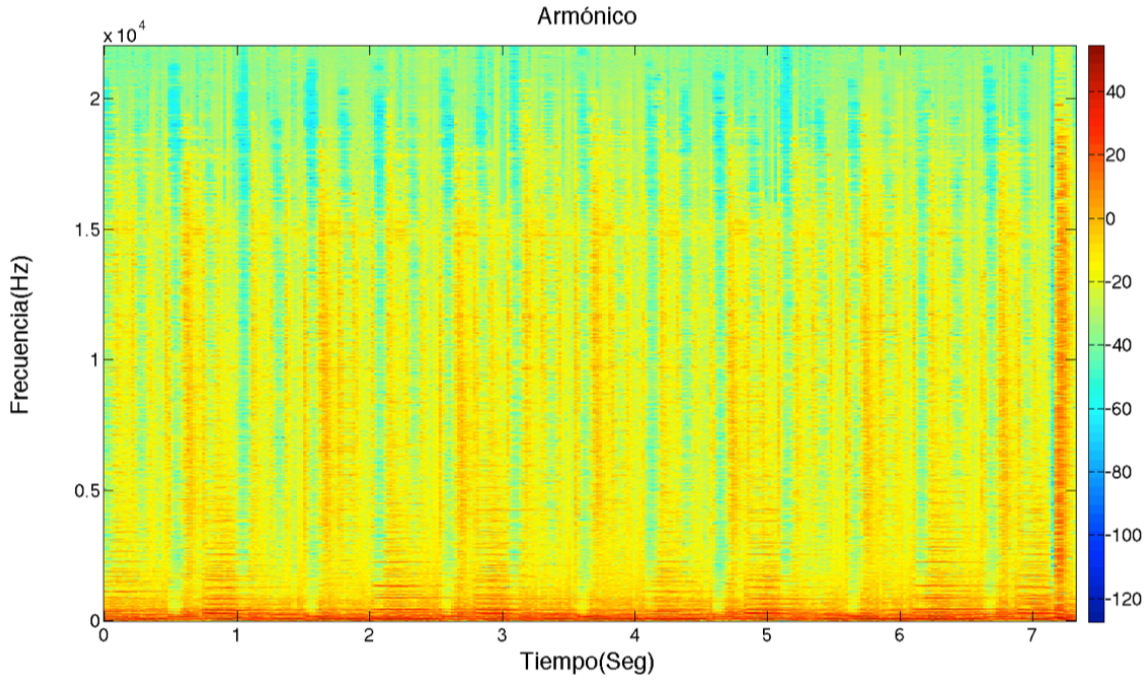

Figura 73: Resultados de la separación de la pista de armónicos

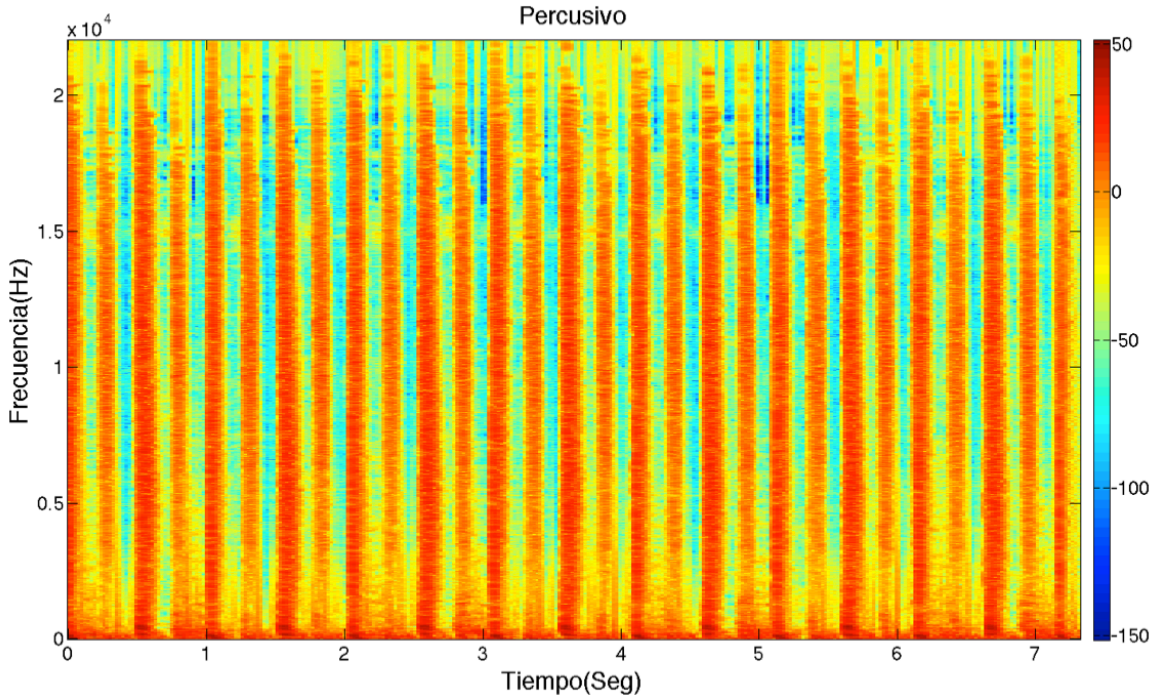

Figura 74: Resultados de la separación de la pista de percusivos

Se puede ver como claramente se han identificado los parámetros armónicos y percusivos, de tal forma que si escuchamos las pistas de audio separadas vemos que se obtienen unos resultados bastante buenos.

5. EVALUACIÓN

Como se ha visto en la sección anterior se dispone de un sistema solido, que permite la separación de pistas en armónicos, percusivos y voz, en la presente sección se va a intentar cuantificar la calidad del sistema a través de una serie de test, tanto subjetivos como cuantificados.

5.1 BASES DE DATOS

Para el análisis tenemos que distinguir entre 3 bases de datos:

1. Por un lado se tiene la base de percusivos, con las que se realiza el entrenamiento del sistema NMF, esta base de datos no entra dentro de la evaluación, sino simplemente su mejor o peor elección hará que algunas canciones se separen mejor o peor. Para este programa se ha empleado un total de 60 bases percusiva para el entrenamiento del NMF.

2. Por otro lado se tiene una base de melodías de 7 canciones con las que se va a optimizar los parámetros del sistema, es decir el número de bases k, la resolución del azimugrama β…etc.

Estas 7 pistas son:

1. Another dreamer – the ones we love
2. Fort minor – remember the name
3. Ultimate nz tour
4. Bon Jovi - Livin' On A Prayer
5. Fleetwood mac – Go our own way
6. Los lobos – La bamba
7. LostProphets – Rooftops

Donde las 3 primeras se han extraído de la página web del SISEC [15]. El SISEC es una competición a nivel mundial, en la que multitud de personas interesadas en el estudio de este campo, desarrollan sus algoritmos y los comparan unos con otros en las mismas condiciones. El resto de pistas han sido extraídas del juego *Guitar Heros.*

Como se puede observar son un número bastante elevado de pistas que como se verá en la sección de set-up se hará un análisis híper-paramétrico con un total de 11 variables que terminarán generando un total de 4320 pistas de audio entre percusivos, armónico y voz, alcanzando más de 25 Gigabytes de datos de audio y 12960 valores numéricos. No obstante como se trata de ver como mejora la extracción de voz, solo los datos de la voz serán representativos en la sección de análisis.

3. Por último tendremos una base de 2 pistas de audio de testeo, es decir con los parámetros ya optimizados que solución se obtiene directamente sin haber tratado con esas pistas con anterioridad.

Estas 2 pistas son:

1. Tamy- Que pena tanto faz
2. Bearlin

5.2 SET-UP

En esta sección se habla de los parámetros que se emplean en el sistema para la híper-parametrización de las variables, así como los empleados en las transformadas.

Para la elección de la STFT, se han tomado los valores típicos que se han observados en la bibliografía[8][9][13]:

Tipo de ventana	Hamming
Número de muestras	4096
Saltos	1024

Tabla 3: Parámetros de la STFT usados

Como ya se indicó en sus respectivas etapas existen parámetros que no se van a analizar, ya que como se vio en sus respectivas referencias ya fueron optimizados, por tanto los parámetros que se van a optimizar son principalmente los de la etapa ADRess, NMF y Sparseness.

Filtro de median	• $\lambda = 1$ • $p = 80$
Separación armónico de percusivos	• $l = 17\ muestras$ • $P = 2\ (suavizado\ de\ la\ máscara)$
ADRess	• $\beta = [30 \quad 100]$ • $H = [20]$
NMF	• $K_v = [16 \quad 32 \quad 64 \quad 128]$ • $K_p = [1 \quad 2 \quad 3 \quad 4 \quad 5]$ • $Iteraciones = [100]$
Sparseness	• $\lambda_v = [0 \quad 0.2 \quad 0.4 \quad 0.6 \quad 0.8 \quad 1]$ • $\lambda_p = [0 \quad 0.2 \quad 0.4 \quad 0.6 \quad 0.8 \quad 1]$

Tabla 4: Parámetros tomados para el análisis híper-paramétrico de los diferentes sistemas

5.3 METRÍCAS

Los métodos de evaluación serán de dos tipos, por un lado se realizará un test Mushra, ya que al tratarse de análisis de audio, la percepción es subjetiva y única de cada individuo, por ello se realizará un test a diferentes personas con intención de obtener según sus criterios lo valores de las variables para los que mejor quedaría separada una pista. Por otra parte necesitaremos una cuantificación para los resultados de forma que se pueda comparar con otros algoritmos, por ello como se comentó a comienzos de la sección nos compararemos con los resultados del SISEC [15]. La forma de compararse será a través de los parámetros SDR, SIR, SAR, que son generados por la función *bss_eval_images_nosort* incluida en la toolbox de MatLab BSS_EVAL [14].

El principal objetivo de utilizar esta toolbox, es descomponer una fuente $\hat{S}(t)$ como una suma $S_i(t)$:

$$\hat{S}(t) = S_{target}(t) + e_{interf}(t) + e_{noise}(t) + e_{artif}(t) \quad (64)$$

Donde $S_{target}(t)$ es una aproximación de la fuente de interés $S_i(t)$, $e_{interf}(t)$ representa las interferencias no deseadas que se encuentran en nuestra señal, $e_{noise}(t)$, es una deformación del propio ruido que no afecta a la fuente, por último $e_{artif}(t)$ es un término artificial que corresponde a los 'artificios' creados como consecuencia de los algoritmos de separación, como musicales, ruido…etc.

Existen diferentes formas de descomponer la señales en los términos de la ecuación (47) la función que se emplea [14] esta basada en una descomposición mediante filtrados, descomponiendo la fuente estimada en las 4 componentes que representan la distorsión, interferencias y artefactos. Aunque solo nos centraremos en dos de ellas:

1. La relación de distorsión entre fuentes:

$$SDR = 10\ log_{10}\left(\frac{\|S_{target}(t)\|^2}{\|e_{interf}(t)+e_{noise}(t)+e_{artif}(t)\|^2}\right) \quad (65)$$

Da información de como se parece la forma de onda de la señal original, a la de la fuente estimada.

2. La relación de fuentes interferentes:

$$SIR = 10\ log_{10}\left(\frac{\|S_{target}(t)\|^2}{\|e_{interf}(t)\|^2}\right) \quad (65)$$

Da información de como afectan las interferencias debido a una mala separación de las fuentes, por ejemplo que en la voz se introduzcan percusivos.

5.3 RESULTADOS

La forma de exponer los resultados se dividirá en 3 etapas, una primera etapa mostrará los resultados de la híper-parametrización, analizando primero los parámetros del ADRess para todas las canciones y obteniendo la media, tomando el mejor para los siguientes valores y así sucesivamente.

Resultados para $\beta = [30 \quad 100]$

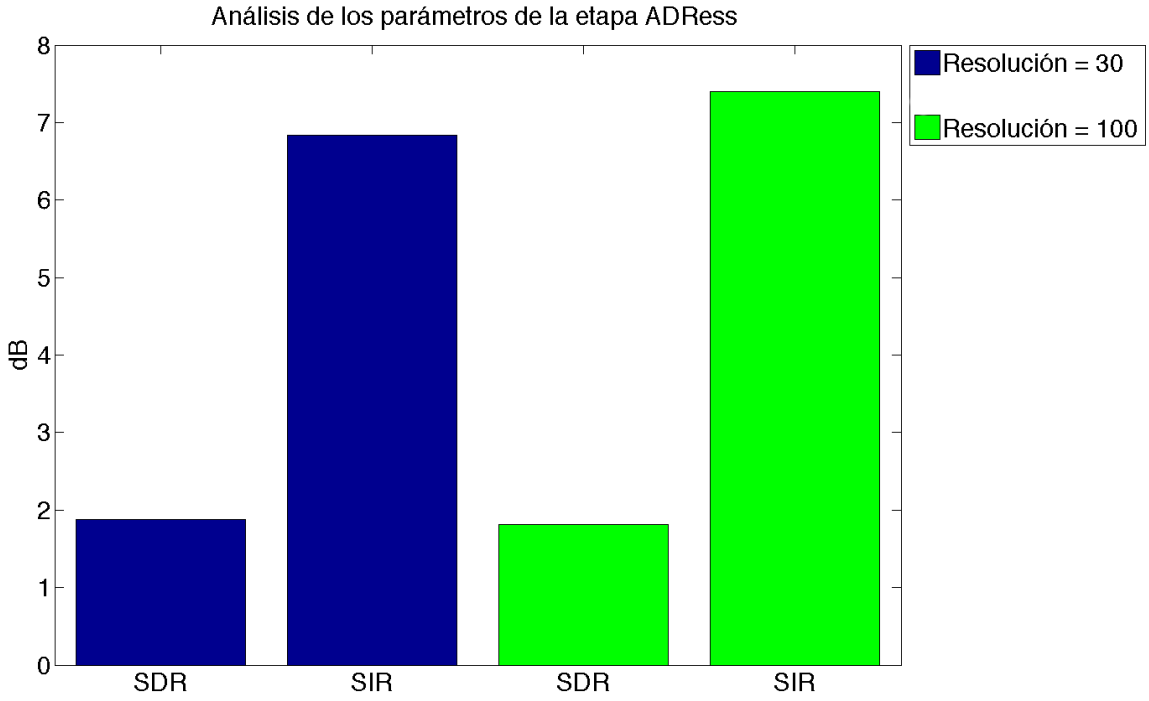

Figura 75:Análisis de los parámetros de la etapa ADRess

Bajo la premisa de que el principal parámetro de interés es el SIR ya que representa las interferencias de señales como consecuencia de una mala separación de fuentes. En la figura 75 se puede apreciar como era de esperar que una mayor resolución implica una mejor separación de fuentes, por esto las siguientes gráficas se realizarán en base a esa resolución es decir $\beta = 100$.

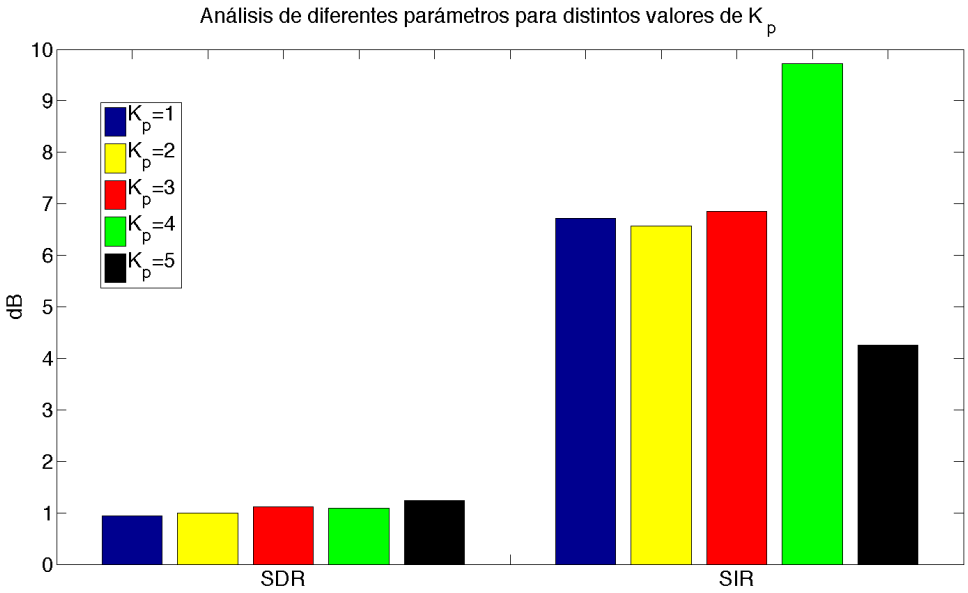

Figura 76:Análisis de los diferentes parámetros para distintos valores de K_p

En la figura 76, se observa el análisis para diferentes valores de K_p para un valor de $K_v = 16$, se puede ver que en base al criterio del SIR el mejor valor es para un valor de $K_p = 4$, mientras que para el valor SDR, prácticamente son muy similares todos los resultados por ello se decide tomar el valor de $K_p = 4$.

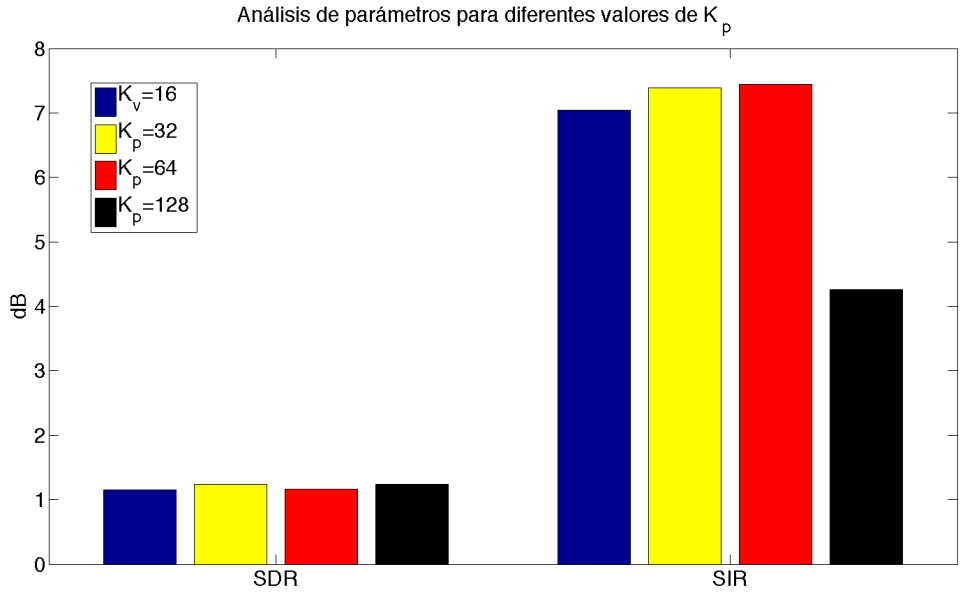

Figura 77: Análisis de parámetros para diferentes valores de K_v

En la figura 77, se puede ver el análisis para diferentes parámetros de K_v, este parámetro es muy dependiente la canción y del cantante lógicamente, por eso los datos son muy relativos, escuchando los resultados, se puede apreciar una mejor separación

en algunas canciones para 16 bases y sin embargo otras con 128 bases suenan mejor, pero como se trata de un análisis cuantificado, para continuar el análisis de los parámetros del sparseness, se tomará como valor óptimo un valor de bases $K_v = 64$ bases vocales.

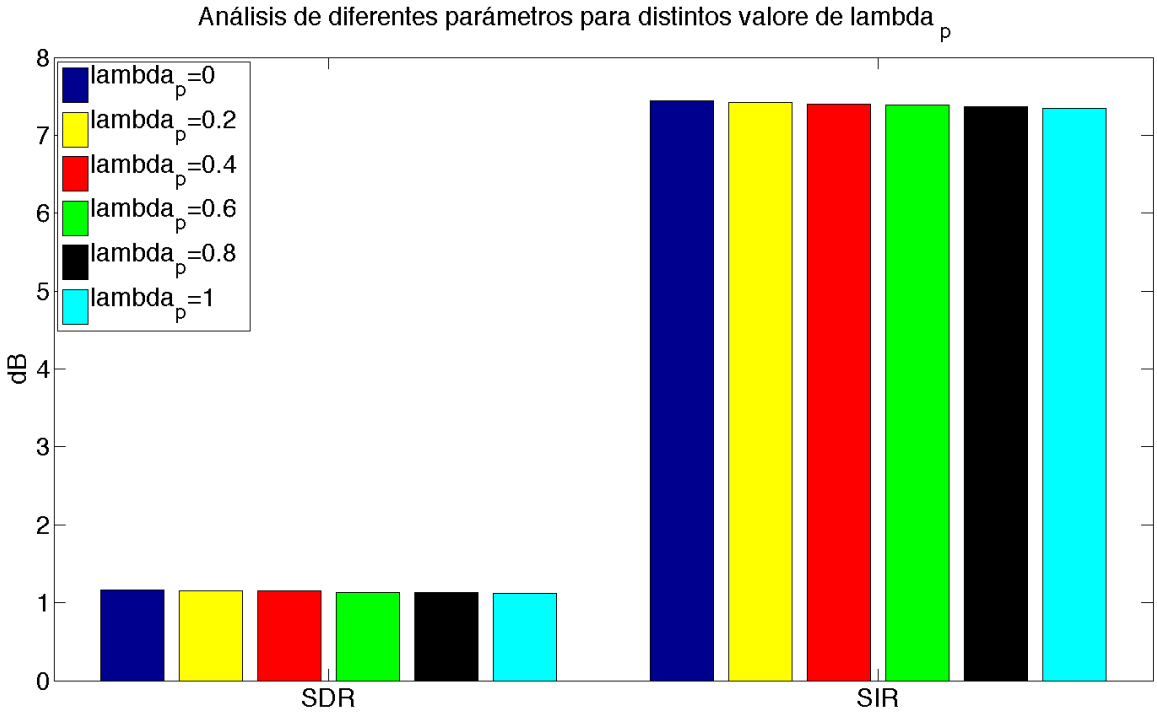

Figura 78: Análisis de diferentes parámetros para distintos valores de λ_p

En la figura 78 se puede observar el análisis para diferentes valores de λ_p, aunque acústicamente no representa una mejora significativa ni perceptible, se puede observar en el gráfico de barras que las variaciones son mínimos obteniendo los mejores resultados para $\lambda_p = 0$, aunque al ser medidas medias de un conjunto de canciones es posible que debamos admitir un margen llegando a la conclusión que los mejores valores estarán entorno a 0, 0.1 ó 0.3.

Finalmente en la figura 79, se puede observar el análisis para el último parámetro λ_v. En la figura se puede ver como igual que pasaba con λ_p, las variaciones son mínimas, lo que se refleja en que acústicamente son imperceptibles, no obstante se ve que conforme nos acercamos a $\lambda_v = 1$, se obtienen mejores resultados en la separación y por tanto llegamos a la conclusión de que entorno a 0.8, 0.9 ó 1 serán los valores óptimos para las restricciones del sparseness en la voz.

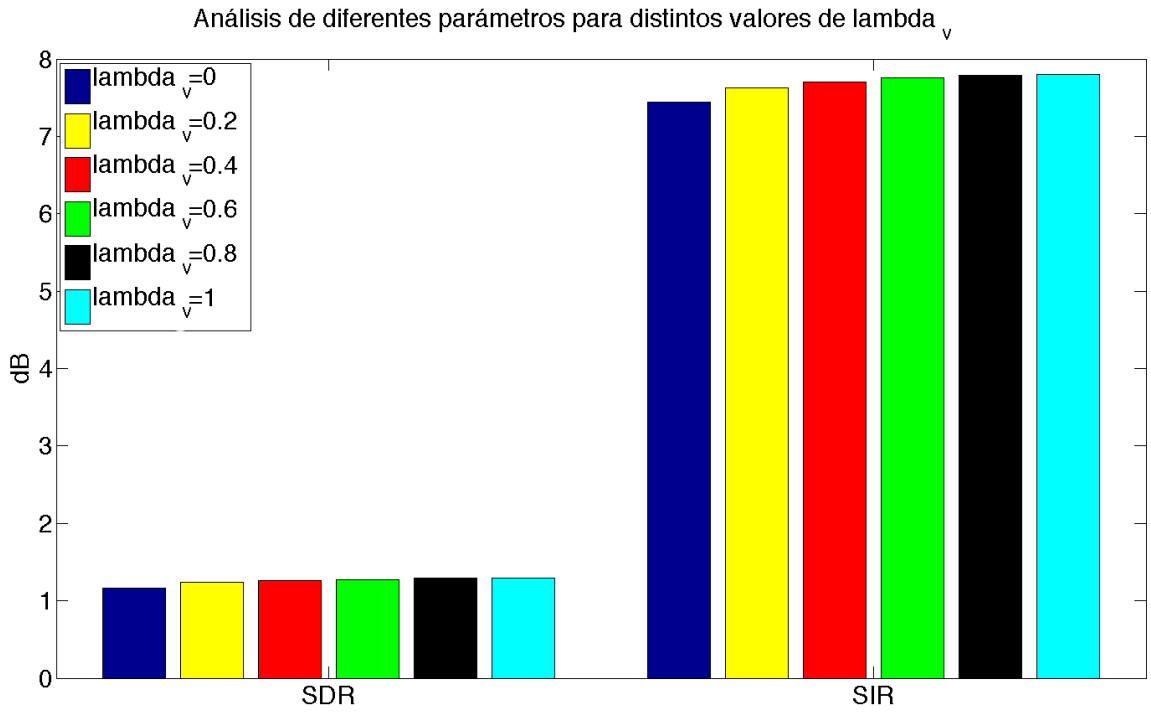

Figura 79: Análisis de diferentes parámetros para distintos valores de λ_v

Finalmente se puede concluir que después de la optimización para la evaluación en la etapa de test, se emplearán los siguientes valores:

Filtro de median	• $\lambda = 1$ • $p = 80$
Separación armónico de percusivos	• $l = 17\ muestras$ • $P = 2\ (suavizado\ de\ la\ máscara)$
ADRess	• $\beta = [100]$ • $H = [20]$
NMF	• $K_v = [64]$ • $K_p = [4]$ • $Iteraciones = [100]$
Sparseness	• $\lambda_v = [0]$ • $\lambda_p = [1]$

Tabla 5: Resolución de los valores optimizados

Como se ha dicho al margen de los resultados obtenidos, la música y la percepción que cada persona tiene de ella es diferente ya que influyen muchos factores, así como para alguien, puede resultar que una pista donde en la separación la voz queda con algo de instrumental puede opinar que suena mejor, que alguien que no desea nada de instrumental en su pista de voz y prefiere que algunos fonemas no se oigan

correctamente. Por ello se ha pensado al margen de las pruebas de optimización y de comparación, esta última se verá más adelante, resultaba interesante realizar un test de Mushra [33]. El test de Mushra es un método para realizar test de audio, mediante una evaluación subjetiva de la calidad de audio y así cuantificar la calidad percibida de las señales que han sufrido un procesado de audio. El test ha sido realizado por diez personas al azar completamente independientes a este proyecto, donde primero se ha reproducido la pista con toda la información, instrumental y vocal, posteriormente se han ido reproduciendo la correspondiente separación para diferentes valores que se han analizado:

• Filtro de median	• $\lambda = 1$ • $p = 80$
• Separación armónico de percusivos	• $l = 17\ muestras$ • $P = 2\ (suavizado\ de\ la\ máscara)$
• ADRess	• $\beta = [100]$ • $H = [20]$
• NMF	• $K_v = [32 \quad 64]$ • $K_p = [2 \quad 4]$ • $Iteraciones = [100]$
• Sparseness	• $\lambda_v = [0 \quad 1]$ • $\lambda_p = [0 \quad 1]$

Tabla 6: Valores empleados en el test de Mushra

Se le pidió a los participantes que puntuasen las pistas en base a su propio criterio de lo bien que le sonaba la separación del 1 al 5, siendo 1 una mala separación y 5 una muy buena separación. En la tabla 7, se pueden apreciar los valores medio de cada resultado que han aportado. La numeración de las canciones corresponde a:

1. Bon Jovi - Livin' On A Prayer
2. Another dreamer – the ones we love
3. Fort minor – remember the name
4. Ultimate nz tour
5. Los lobos – La bamba
6. LostProphets – Rooftops

	Canción 1		Canción 2		Canción 3		Canción 4		Canción 5		Canción 6	
	K=2	K=4	K=2	K=4	K=2	K=4	K=2	K=4	K=2	K=4	K = 2	K=4
Alphav = 0 alphap = 0 bases vocales = 16	2.57	3.14	3	3	2.85	2.57	2.42	1.71	2.57	2.85	2.42	2.28
Alphav = 0 alphap = 0 bases vocales = 32	2.71	3.14	3.71	2.57	2.85	3.57	1.85	1.71	3.14	3.28	2.42	2.71
Alphav = 0 alphap = 0 bases vocales = 64	3.14	3.28	3.14	3.28	2.85	3.28	1.85	1.71	2.85	3	2	2.42
Alphav = 0 alphap = 0 bases vocales = 128	2.28	3	4	3.57	2.71	3	1.71	2.28	3	2.57	2	2.14
Alphav = 1 alphap = 0 bases vocales = 16	3.57	2.85	2.57	2.71	2.48	2.28	2.85	2.42	2.85	3.71	2.71	3.14
Alphav = 1 alphap = 0 bases vocales = 32	2.85	2.85	2.71	3	2.42	2.85	2.14	2.14	2.57	3.42	2.57	2.57
Alphav = 1 alphap = 0 bases vocales = 64	2.85	2.85	3.14	3	2.28	2.85	1.85	2	3	3	2.42	2.42
Alphav = 1 alphap = 0 bases vocales = 128	2.14	2.57	3.57	3.71	2.85	2.28	1.57	2.28	3.28	2.57	2.14	2.42
Archivo de Voz ADRess	2.85		3.14		2.71		1.71		3		2.14	

Tabla 7: Resultados del test de Mushra

Lógicamente estos resultado no son representativos de los anteriores que se han cuantificado, sin embargo se observan similitudes, principalmente en la elección de bases, vemos que una elección de 4 bases percusivas es mejor frente a las de 2, estos resultados ya se habían visto numéricamente, pero ahora se puede observar que acústicamente se cumplen esos resultados, por el contrario la variación de los valores de sparseness si se cumplía en el análisis numérico, sin embargo auditivamente no son cambios tan perceptible, dando resultados contradictorios.

A continuación se va a ver el tercer análisis donde se compara los resultados que se obtienen para esos valores optimizados, para dos pistas independientes a todas las etapas de análisis anterior y se compararan con los resultados de otros algoritmos. Como existen muchos algoritmos diferentes y el texto se extendería demasiado si los comparamos con todos, en la sección de referencias se incluyen algunos enlaces [34] y [35] que corresponden a la página donde se recogen los resultados para esa pista de audio en concreto:

1. Tamy – Que pena tanto faz [34]:

Algoritmo	SDR (Voz)	SIR (Voz)
TFG	4.4 dB	4.5 dB
Algoritmo 1	3.6 dB	5.5 dB
Algoritmo 2	5.1 dB	6.9 dB

Tabla 8: Comparativa de la canción Tamy con diferentes algoritmos

2. Bearlin [34] y [35]:

Algoritmo	SDR (Voz)	SIR (Voz)
TFG	4.2 dB	15.2 dB
Algoritmo 1	6.3 dB	12.8 dB
Algoritmo 2	7.9 dB	16.6 dB

Tabla 9: Comparativa de la canción Bearlin con diferentes algoritmos

Como se puede apreciar los resultados no son tan diferentes a otros algoritmos que se encuentran ya presente de otros autores, es más en algunos casos llega a superarlos. Cabe destacar que la canción de Tamy, es una canción sin apenas percusivo, y por tanto la etapa del NMF y las que le siguen no serían realmente necesarias pues estropean la calidad de la señal, ya que originalmente con la primera separación se obtenían 11 dB de SIR y 2.3 dB de SDR, esto se debe a que los datos optimizados para el ADRess, y las otras etapas no se adapten bien a esta canción, es por esto que en el programa final se recogen los resultados de la salida de cada etapa, siendo en este caso innecesario aplicar NMF o sparseness. En cambio la canción de Bearlin vemos como los resultados son bastante buenos viendo que la mayoría de algoritmos que se han visto [34], tienen un SDR de 6 de media y un SIR de 16 de media con algunas excepciones que llegan a SIR de hasta 26 dB. La conclusión es que se muestran unos resultados válidos y coherentes con otros algoritmos actuales.

6. CONCLUSIONES

A lo largo de este proyecto, se ha desarrollado una serie de algoritmos, con la intención de separar fuentes instrumentales de vocales y las instrumentales en armónicos y percusivos. Para ello se desarrolló un módulo que implementase todos estos algoritmos en serie, y fuese capaz de dar a la salida una solución al problema que se planteaba. Finalmente se han desarrollado un total de 5 etapas:

1. Una etapa para separar instrumental de voz.
    a. Filtro de mediana.
2. Una etapa para separar armónicos de percusivos.
    a. Separación armónico y percusivo
3. Tres etapas para perfeccionar la separación de la voz con la instrumental.
    a. ADRess
    b. NMF (con Sparseness)
    c. Estimación de zonas de voz

En base a estas etapas, se ha cumplido el objetivo principal de este TFG, que buscaba separar fuentes instrumentales en sus armónicos y percusivos correspondiente y fuentes de voz, partiendo de una pista con todas estas fuentes mezcladas. Con el desarrollo de cada etapa se han ido cumpliendo los sub-objetivos y finalmente se han evaluado los resultados comparándolos con resultados de algoritmos reales que están hoy en funcionamiento y viendo como no son tan diferentes los que se obtienen en este proyecto como los que hay, mejorando incluso alguno de ellos.

Tras finalizar el TFG, se puede ver que en base a los resultados obtenidos en la sección 5, se han cumplido los objetivos planteados al comienzo del trabajo. Se consigue una buena separación de las diferentes fuentes en la mayoría de los casos, no muy diferentes a otros resultados actuales, con algunas excepciones. Se ha demostrado como cada etapa que se ha ido añadiendo ha conseguido una mejora significativa en la separación de los restos percusivos que quedaban tras elADRess, con lo que se concluye que los algoritmos son válidos, incluso se ha aportado una mejora que limpia la señala de voz en las zonas donde no canta el cantante. Aunque quizás no sean los

algoritmos más eficientes pero desde un principio no se pensó en desarrollar el módulo para que funcionase en tiempo real.

7. LINEAS DE INVETIGACIÓN FUTURAS

De cara a continuar investigando con las bases plantadas de este proyecto, son 3 campos principalmente que sería interesante a la hora de continuar el estudio.

Un primer campo, que buscase realizar la separación instrumental de voz, mediante técnicas que aprovechen las características propias de la voz y de los instrumentos, si se recuerda aquí se emplea una etapa que estudia la similitud de la forma del espectrograma, lo que lo hace muy dependiente del tipo de música mientras que si el estudio se enfoca desde un punto de vista de las características de la voz como el vibrato o el trémolo o incluso los diferentes formantes de la voz, se ampliaría el rango de tipo de música a la que aplicar el algoritmo.

Un segundo campo, que mejore alguno de los algoritmos presentes como el ADRess, o el de filtro de mediana, para hacerlos más eficientes de cara a poder realizar un programa que se ejecute en tiempo real, que es lo verdaderamente interesante de este terreno. Si fuese necesario por imposibilidad de mejorar la eficiencia de algunos algoritmos reemplazarlos por otros buscando un compromiso entre calidad y eficiencia.

Un tercer campo que busque ir más allá de los que en este TFG se plantea, una vez se tiene la separación instrumental en armónicos y percusivos, buscar técnicas que identifiquen los instrumentos y sean capaz de mostrar las notas musicales de cada uno de ellos para la melodía, de forma similar con la voz, capaz de extraer los fonemas y representar la letra de la canción.

Con estos tres campos se ve que aunque la base de este TFG es sólida, queda bastante por investigar en esta área, así como nuevos campos de investigación derivados de esta base.

8. BIBLIOGRAFÍA

[1]	Juan Sebastián Guevara Sanin. Teoría de la música. 2010
[2]	Constantino Pérez Vega. El sonido y audición. Universidad de Cantabria.
[3]	L. Regnier, G. Peeters. Singing voice detection in music tracks using direct voice vibrato detection. Acoustics, Speech and Signal Processing, 2009, 1685 - 1688
[4]	MI Uzcanga Lacabe et al. Voz cantada. Clínica Universitaria. Facultad de Medicina. Universidad de Navarra.
[5]	Beatriz Gallardo Paúls. Presentación del tema de fonología. Universidad de Valencia.
[6]	Diego Gómez Fernández. La teoría universalista de Jakobson y el orden de adquisición de los fonemas en la lengua española. Universidad de Sevilla.
[7]	Demestre, J., Llisterri, J., Riera, M. y Soler, O. La percepción del lenguaje. Psicologia del lenguaje. En O. Soler (Ed.),Barcelona: Editorial UOC. 2006
[8]	Derry FitzGerald. Vocal Separation using Nearest Neigthbours and MedianFiltering. Signals and Systems Conference, IET Irish, 2012, pp 1-5
[9]	Derry FitzGerald. Stereo vocal extraction using adress and nearest neighbors median filtering. Signals and Systems Conference, 16th IET Irish, 2013,
[10]	Stratis Sofianos. Singing Voice extration from stereophonic recordings. Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference on, 2010, pp 233 - 236
[11]	Emad M. Grais and Hakan Erdogan. Single channel speech music separation using nonnegative matrix factorization and spectral mask. Digital Signal Processing (DSP), 17th International Conference on, 2011, 1 - 6
[12]	Nicholas Bryan, Dennis Sun. Introduction to Non-Negative Matrix Factorization. Standford University. 2013
[13]	Derry FitzGerald. Harmonic/Percussive Separation Using Median Filtering. 13th International Conference on Digital Audio Effects (DAFX10), 2010
[14]	C.Févotte, R. Gribonval, E. Vincent. BSS EVAL ToolBox user guide, revision 2. IRISA. 2005.
[15]	https://sisec2011.wiki.irisa.fr/tiki-index.php?page=Professionally+produced+music+recordings
[16]	Youngmin Cho · Lawrence K. Saul. Nonnegative Matrix Factorization for Semi-supervised Dimensionality Reduction. University of California.

[17]	Chao-Ling Hsu, Jyh-Shing Roger Jang. Singing pitch extraction by voice vibrato/tremolo estimation and instrument partial deletion. 11th International Society for Music Information Retrieval Conference, 2010, pp 525
[18]	http://es.wikipedia.org/wiki/Vibrato
[19]	Jingu Kim and Haesun Park. Sparse Nonnegative Matrix Factorization for Clustering
[20]	¿Matemáticas en la Música? Miscelánea Matemática. Núm.27. Soc. Mat. Mex. (1999) pp. 15-27.
[21]	http://es.wikipedia.org/wiki/Sonoridad_(sicoac%C3%BAstica)
[22]	http://es.wikipedia.org/wiki/Cuerdas_vocales
[23]	http://en.wikipedia.org/wiki/I_Have_a_Dream
[24]	Arauz Benvenuto, Guevara Jackson, Sapaly Tosi. La voz normal. Panamerica. pp 168-170.
[25]	http://es.wikipedia.org/wiki/Formante
[26]	http://www.teoria.com/referencia/t/tonos-semi.php
[27]	http://www.fon.hum.uva.nl/praat/
[28]	http://lema.rae.es/drae/?val=fonema
[29]	Cédric Févotte, Nancy Bertin, Jean-Louis Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Département Traitement du signal et des Images Groupe AAO : Audio, Acoustique et Ondes. Mayo 2008
[30]	Daniel B. Rowe, “A Bayesian approach to blind source separation,”2002.

[31]	Tuomas Virtanen. Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria. Audio, Speech, and Language Processing, IEEE Transactions on (Volume:15 , issue: 3). Marzo 2007. Pp 1066-1074
[32]	Alexey Castrodad, Zhengming Xing, Lawrence Carin et al. Learning discriminative sparse models for source separation and mapping of hyperspectral imagery. University of Minnesota. Octubre 2010
[33]	http://c4dm.eecs.qmul.ac.uk/downloads/#mushram
[34]	http://www.irisa.fr/metiss/SiSEC08/SiSEC_professional/
[35]	http://www.durrieu.ch/phd/eusipco09/

ANEXO 1: MANUAL DE USUARIO

Para la implementación del software con interfaz gráfico se ha hecho uso del programa MatLab. Uno de los principales problemas de MatLab, es que no resulta para nada eficiente a la hora de trabajar con audio en tiempo real, pero aún así se ha decidido emplear este programa debido a la gran cantidad de funciones que trae implementada y por su facilidad de uso.

La GUI (Graphics User Interface), es un entorno de desarrollo para crear aplicaciones dentro del programa desarrollado por MathWork, permitiendo incorporar botones, gráficos…etc.

En este programa (recordemos la figura 40), se dispone de los diferentes algoritmos de los que se ha hablado, siendo necesario una pista en estéreo para poder usar todos por completo. Nada más ejecutar el programa se muestra la siguiente ventana:

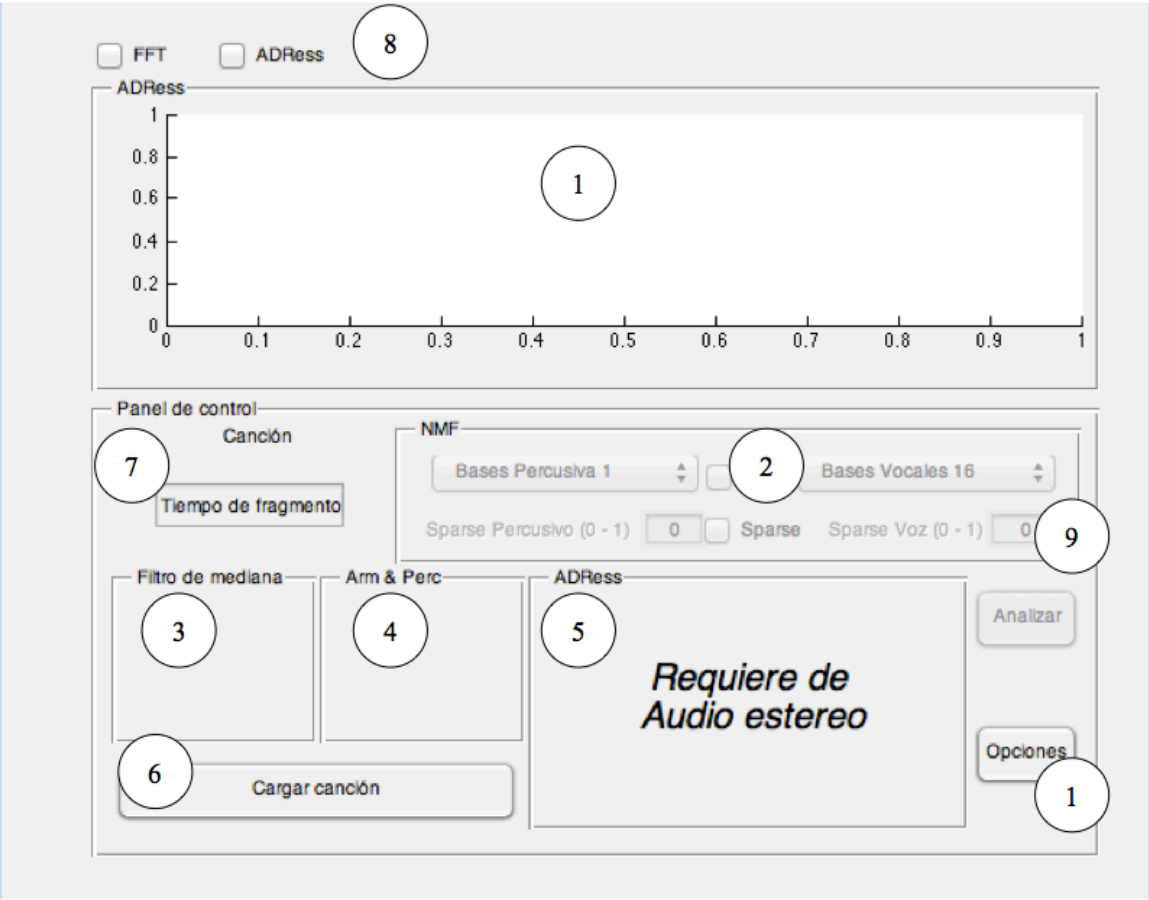

Figura 80: Interfaz del programa

1) Zona dedicada a presentar las gráficas necesarias.

2) Zona para la configuración del algoritmo de NMF, bloqueada si la pista no es estéreo,

3) Zona para la configuración del filtro de mediana

4) Zona para la configuración de la separación de armónico/percusivo.

5) Zona para la configuración del Adress.

6) Botón dedicado para cargar en el sistema una pista.

7) Por eficiencia del sistema las canciones son fragmentadas en segundos, por defecto 5 segundos.

8) Ticks para visualizar en la zona de gráficas la información deseada.

9) Botón de analizar, queda desbloqueado una vez cargada una pista.

10) Opciones que amplia la ventana para reproducir los resultados obtenidos tras el análisis.

Se probará a cargar una pista en estéreo de forma que la explicación sea lo más amplia posible, en caso de ser mono, los únicos algoritmos que se permiten usar son el de Filtro de mediana y la separación armónico percusivo.

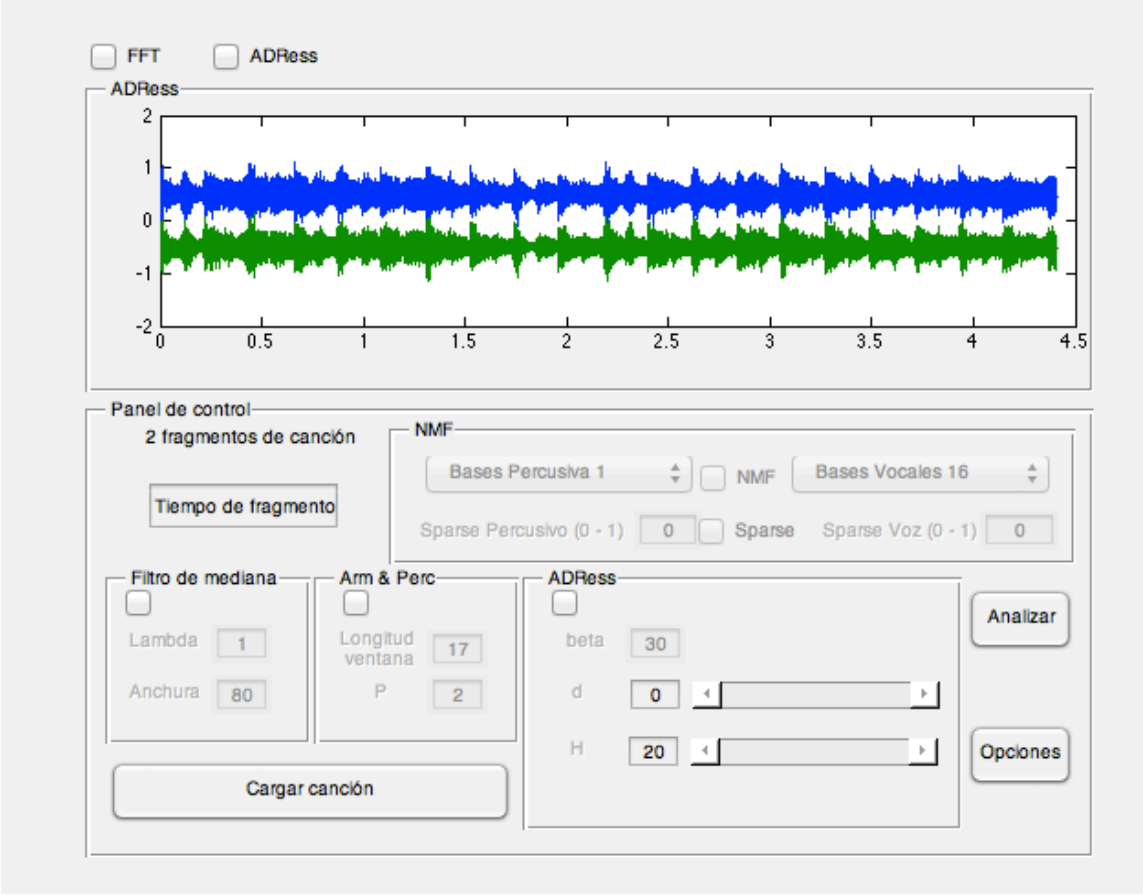

Figura 81: Interfaz del programa

Una vez cargada en memoria la pista se podrá ver como se representan los dos canales en la gráfica, donde el eje X representa el tiempo en segundos, y el eje Y está dividido en los dos canales siendo la zona positiva uno y la negativa otro. Si se desea comprobar que la melodía se ha cargado correctamente podemos pulsar sobre opciones:

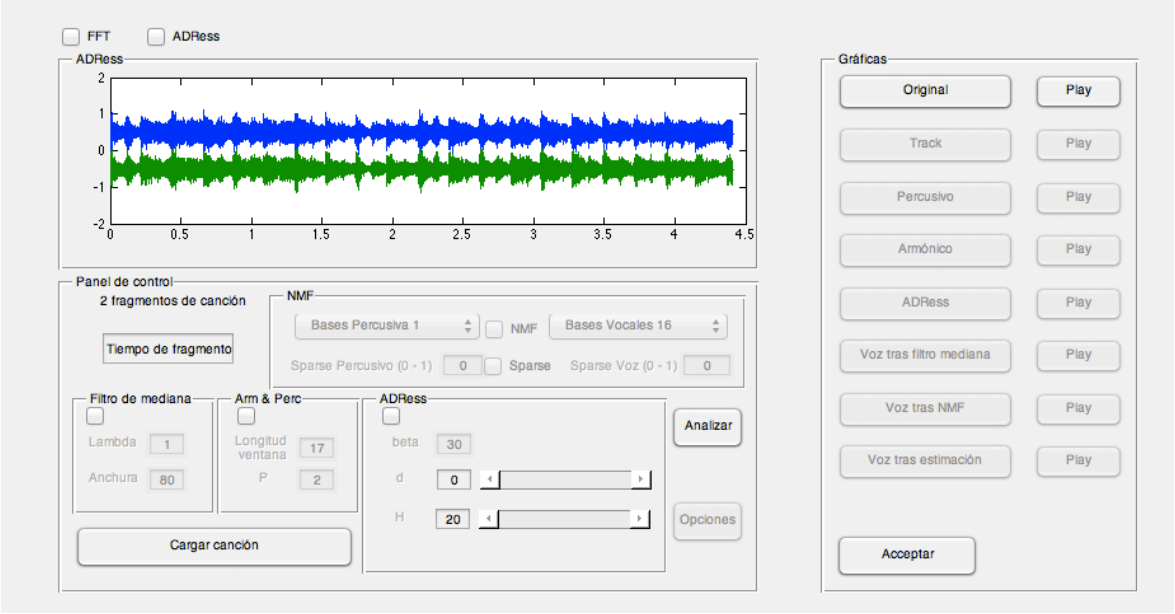

Figura 82: Interfaz del programa

Dejando desplegado el menú de reproducción y gráficas, como es lógico como todavía no se ha realizado ningún calculo, solo podemos escuchar la pista original cargada y ver su espectograma, este se abrirá en una ventana independiente por comodidad por si se desea manipular.

Para configurar los diferentes algoritmos bastará con hacer tick sobre el recuadro para poder empezar a utilizarlo:

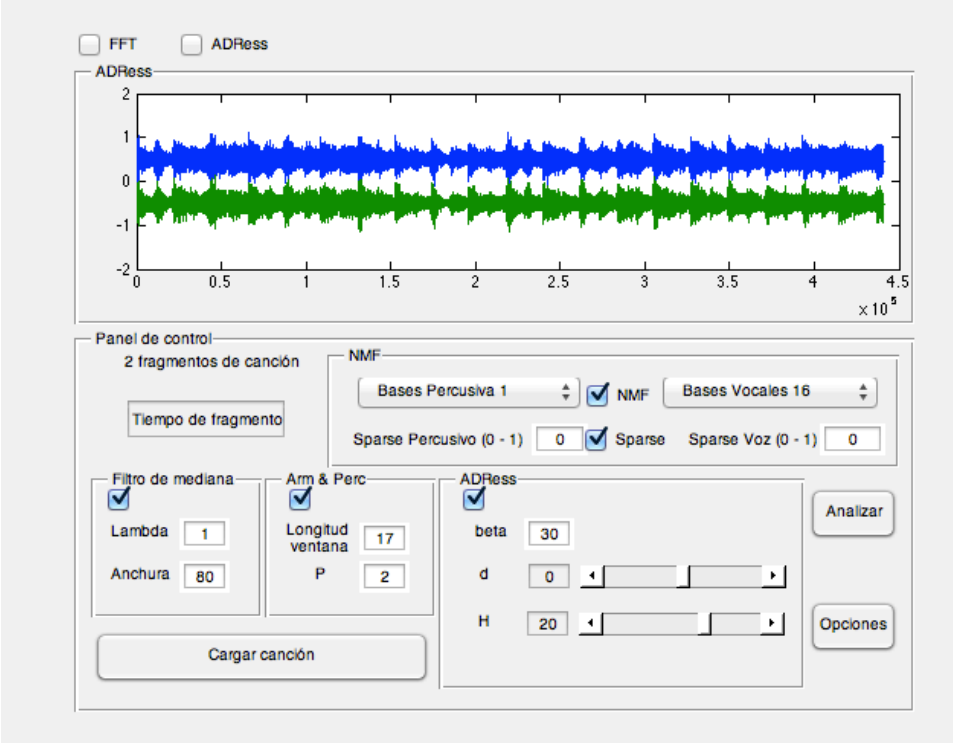

Figura 83:Interfaz del programa

Los dos primeros algoritmos el de filtro de mediana y el de armónico/percusivo, no es recomendable tocarlos, puesto que se encuentran por defecto en sus valores óptimos. Sin embargo el Adress, si incorpora una serie de herramientas para facilitar su uso. Antes de pasar a explicar estas herramientas recordemos cada parámetros. En el filtro de mediana lambda representa la caída de la máscara, y anchura, representa el límite a tomar. En Arm & Perc, el parámetro longitud de la ventana, representa el ancho de señal con el que ir comparando, y P la caída de la máscara. Para el Adress, el parámetro d representa donde queremos centrar el análisis, si sobre el canal izquierdo o el derecho, por defecto está centrado, ya que una de nuestras suposiciones es que la voz del cantante se encuentra centrada entre los dos canales. Por otro lado H representa el ancho que se desea tomar desde la parte centrada recordemos que será H/2 para un lado y H/2 hacía el otro, para cargar la herramienta del Adress en memoria, será necesario dirigirse a los ticks del punto 8 y pulsar sobre Adress.

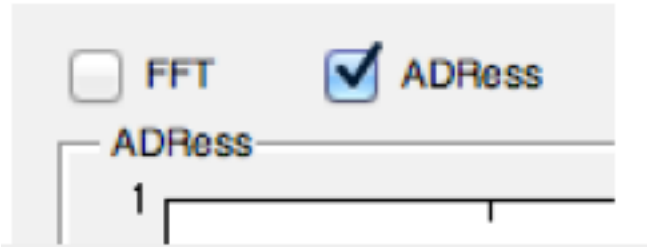

Figura 84:Interfaz del programa

Una vez pulsado reproduciremos la canción desde el botón play, en el menú Gráficas, veremos como automáticamente la zona de gráficas cambias para representar la fuerza de los sonidos que llegan por cada canal. Para poder manipular la ventana en tiempo real, será necesario pulsar el botón Stop que ha sustituido al play, y volver a pulsar el botón play justo después, de esta forma mientras dure la pista, podremos modificar la ventana en tiempo real atendiendo a factores como el ancho que ocupa la voz en los canales:

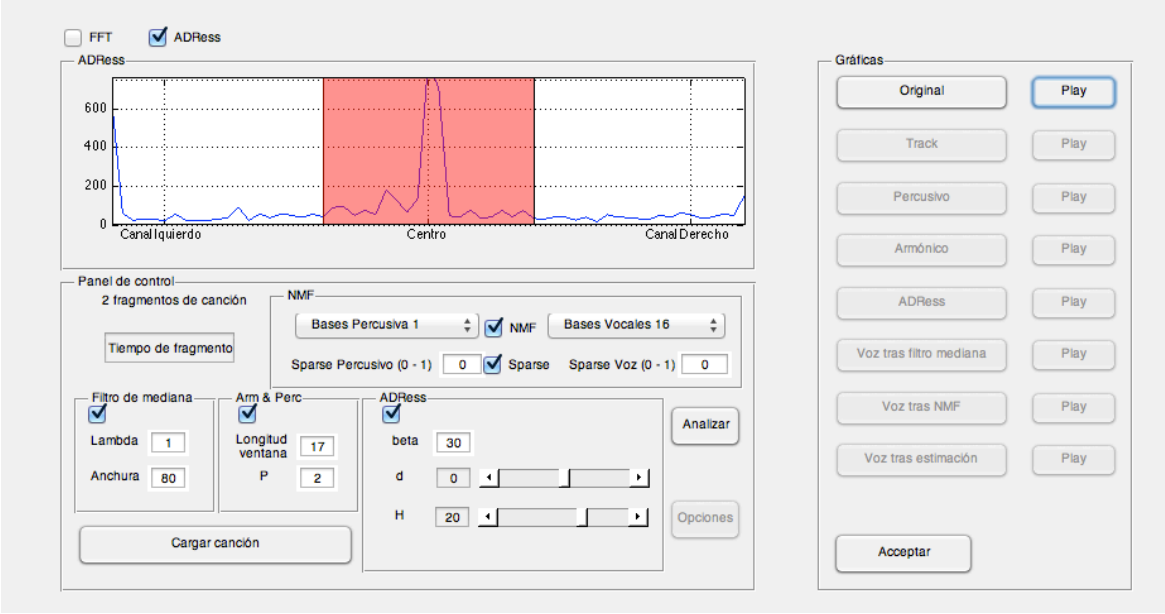

Figura 85:Interfaz del programa

No importa que los valores salgan de la gráfica porque realmente lo que nos interesa son esas “deltas”, centrales que representan la voz, para este caso vemos como el ancho de nuestra venta es ligeramente superior a la zona vocal, luego con casi toda seguridad, estaremos tomando valores pertenecientes al percusivo, para ello vamos a H y rebajamos su valor hasta hacerlo coincidir con las esquinas de la voz, aproximadamente.

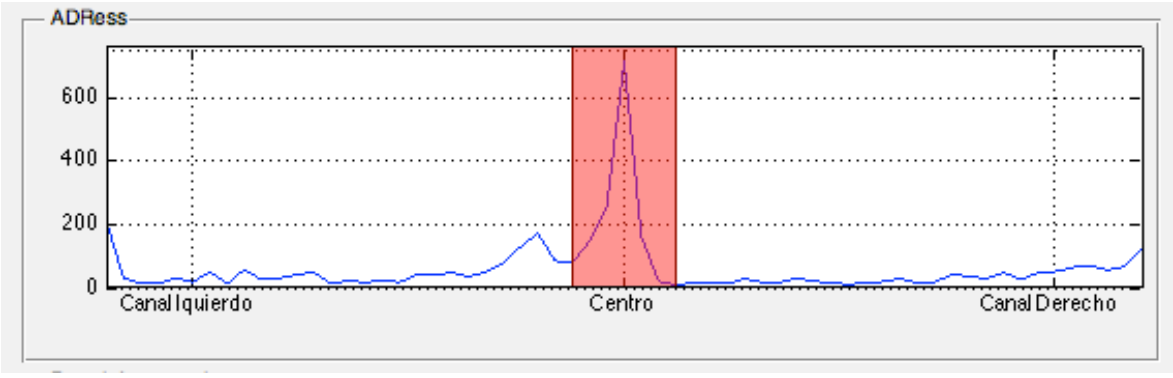

Figura 86:Interfaz del programa

Es importante tener claro, que es un sistema aproximado, ni de lejos es exacto, ya que para poder ejecutarlo en tiempo “real”, se tuvo que muestrear cada cuarto de segundo, lo que genera demasiadas perdidas de información, por ellos es recomendable dejar un margen de seguridad entorno a la voz, ya se encargará el NMF de eliminarlo. El

parámetro beta, es el que nos da la resolución y precisión a la hora de seleccionar una zona u otra, aunque por defecto se encuentre en 30 por temas de eficiencia, puede ponerse sin ningún problema a 100 de forma que como se observa en la siguiente figura, disponemos de mucha más información sobre el gráfico de Adress:

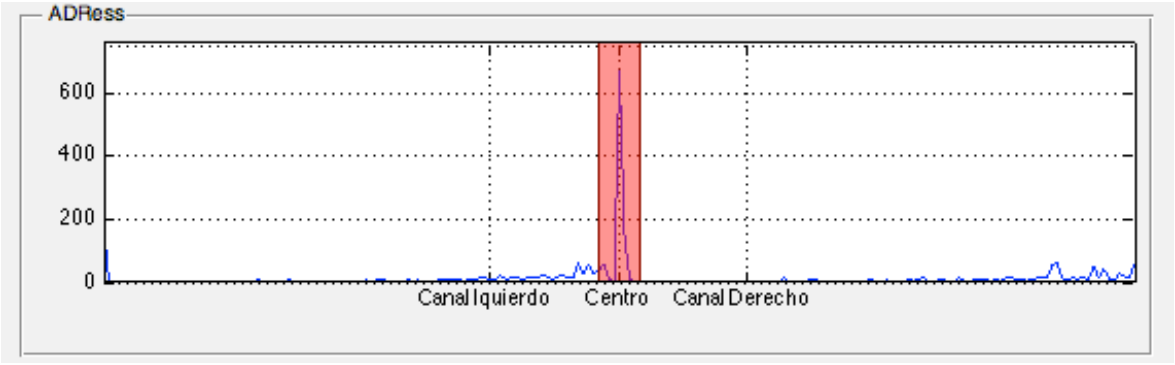

Figura 87:Interfaz del programa

Podemos apreciar como la delta está más definida en el espacio, y nuestra H ahora puede variar entre 1 y 100 en vez de sobre 1 y 30 como anteriormente estaba puesto.

Otra herramienta interesante es poder ver la transformada de Fourier en tiempo real conforme se reproduce la pista, esto nos da una idea de por donde se encuentran la mayor parte de las frecuencias. Por último otra herramienta útil sobre todo para la parte de estimación de la voz, es la reproducción de la señal en el tiempo, en tiempo real, de forma que podamos observar los percusivos atenuados tras el NMF y de esta forma saber si esta estimación de voz funcionará de forma correcta. Estas dos herramientas se encuentran ilustradas en las siguientes figuras:

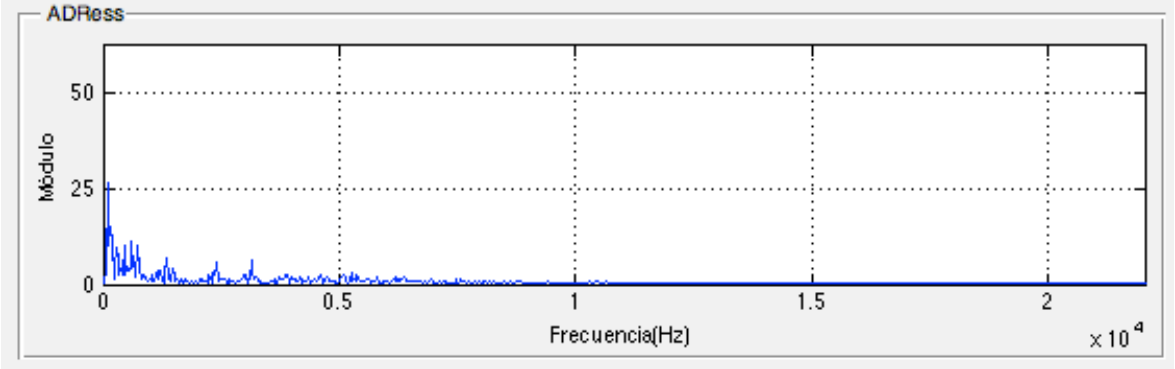

Figura 88:Interfaz del programa

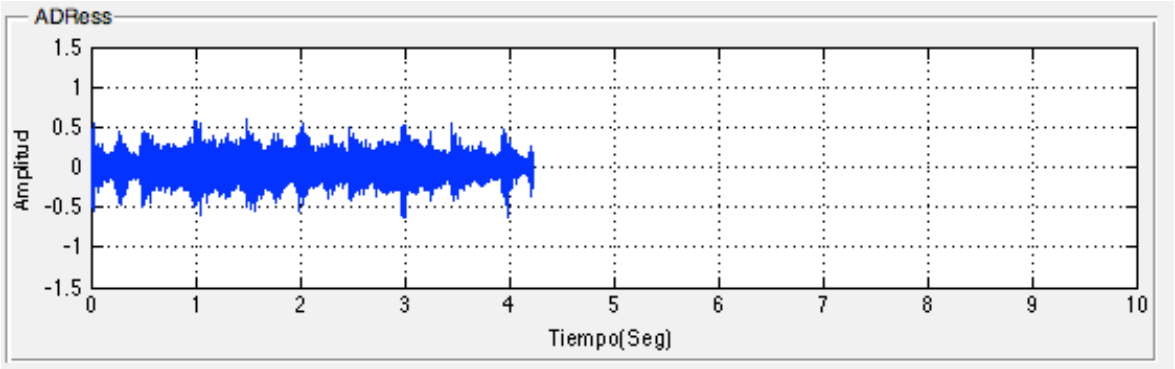

Figura 89:Interfaz del programa

Finalmente, en el NMF, se incluyen las bases entrenadas para las fuentes percusivas (K_p), que se han usado a lo largo de este proyecto, así como las fuentes vocales (K_v) por defecto que se han empleado:

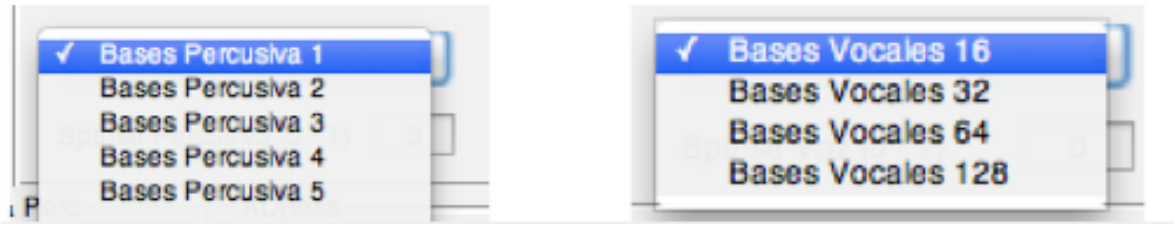

Figura 90: Interfaz del programa

Por último se permite la elección del sparse deseado, tanto para voz como para percusivo, recordemos que estos valores han de estar entre 1 y 0, en caso de emplear valores no válidos se tomará como 0 y en caso de emplear valores mayores que 1 se tomará 1.

El proceso de análisis resulta un poco lento debido a los algoritmos y a la propia naturaleza de MatLab, y aunque tanto adress como armónico/percusivo se realizan en fragmentos de 1 segundo, el algoritmo de filtro de mediana, necesita de al menos 5 segundos o 10 dependiendo de la canción para poder ser procesado correctamente, por ello el valor por defecto es de 5 segundos y en caso de elegir un valor inferior este se corregirá automáticamente a 5 segundos. En caso de analizar canciones enteras, no se recomienda emplear fragmentos muy largos ya que podemos vernos sin memoria suficiente para el calculo del algoritmo.

Durante el proceso en la ventana de trabajo de MatLab podremos ir viendo, los algoritmos que se han ido completando y al finalizar todos los cálculos se mostrará el

tiempo total de duración. También una vez terminado los cálculos deben desbloquearse los botones de la zona de Gráficos de forma que podamos escuchar sobre el propio programas los resultados obtenidos correspondiendo a:

- Track es la suma de armónicos y percusivos
- Percusivo, es la separación de armónico percusivo más los restos que se han conseguido eliminar tras NMF.
- Voz es el resultado final de la voz tras realizarse todo el proceso.
- Armónico es la parte armónica tras la separación de armónico/percusivo.
- Adress representa el resultado tras la salida del Adress.

Figura 91:Interfaz del programa

Por defecto las pistas se guardan en la carpeta del programa, con el nombre que corresponde según el menú Gráficos

ANEXO 2: ÍNDICE DE FIGURAS

ANEXO 3: ÍNDICE DE TABLAS