Escuela Politécnica Superior de Linares

Grado en Ingeniería de Tecnologías de la Telecomunicación

UNIVERSIDAD DE JAÉN
Escuela Politécnica Superior de Linares

Trabajo Fin de Grado

SEPARACIÓN DE LA VOZ CANTADA EN SEÑALES MUSICALES MONOCANAL, UTILIZANDO TÉCNICAS NMF APLICADAS A ESPECTROGRAMAS DE DIFERENTE RESOLUCIÓN

Alumno: **Francisco Jesús Moreno Infantes**

Tutor: Francisco Jesús Cañadas Quesada
Depto.: Ingeniería de Telecomunicación

Junio, 2016

UNIVERSIDAD DE JAÉN
Escuela Politécnica Superior de Linares

Trabajo Fin de Grado

SEPARACIÓN DE LA VOZ CANTADA EN SEÑALES MUSICALES MONOCANAL, UTILIZANDO TÉCNICAS NMF APLICADAS A ESPECTROGRAMAS DE DIFERENTE RESOLUCIÓN.

RESUMEN

Este Trabajo Fin de Grado está centrado en el procesado de audio, concretamente en la separación de fuentes sonoras. El alumno deberá implementar un sistema que sea capaz de separar la señal vocal (singing-voice) procedente del cantante del resto del acompañamiento musical (instrumentos musicales). Para tal fin, se utilizarán técnicas de descomposición de matrices no negativas (NMF) aplicadas al espectrograma de la señal musical con diferente resolución tiempo-frecuencia. Además, se implementará un método para discernir continuidad espectral y temporal lo cual permitirá descomponer la señal instrumental en armónica y percusiva. El proyecto se implementará utilizando el entorno MATLAB.

UNIVERSIDAD DE JAÉN
Escuela Politécnica Superior de Linares

Trabajo Fin de Grado

SEPARACIÓN DE LA VOZ CANTADA EN SEÑALES MUSICALES MONOCANAL, UTILIZANDO TÉCNICAS NMF APLICADAS A ESPECTROGRAMAS DE DIFERENTE RESOLUCIÓN.

ABSTRACT

This thesis is focused on audio processing, namely in the sound sources separation. The student must implement a system which will be able to separate the speech signal (singing voice) originating from the singer, from the rest of the musical accompaniment (musical instruments). To this end, non-negative matrix factorization techniques (NMF) will be used, applying them to the music signal spectrogram with different time-frequency resolution. In addition, a method to discern spectral and temporal continuity will be implemented, which will allow to separate the instrumental signal into pitched and percussive signal. The project will be implemented using the MATLAB environment.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

El sistema auditivo humano tiene la capacidad de realizar un procesamiento y una discriminación de los sonidos presentes a su alrededor. También tiene la capacidad de distinguir la naturaleza de un sonido, ya que identifica su origen y detecta si consiste en una fuente de voz, en una fuente instrumental o si por el contrario, es simplemente ruido. Un ser humano, que escucha un concierto de una orquesta, puede identificar al instrumento que lleva la melodía y seguirla mentalmente, o bien, puede elegir los instrumentos acompañantes sin tener en cuenta la melodía, incluso puede centrar su atención sólo en un tipo de instrumento y seguir las notas tocadas. Todas estas funciones se realizan de forma subconsciente, es decir, 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, tampoco son capaces de identificar los diferentes instrumentos que suenan en una canción, ni por supuesto, pueden entender la letra que acompaña a la melodía interpretada por un cantante. Este hecho representa un gran inconveniente para el desarrollo de aplicaciones que buscan automatizar sistemas que permitan la extracción de la melodía. Ya sean aplicaciones que requieran la extracción de la voz cantada, como pueda ser un karaoke o simplemente para identificar el cantante o asociar una determinada pista a un género musical determinado. Incluso aplicaciones que permitan extraer determinados sonidos para volverlos a emplear en nuevas piezas musicales.

La separación de fuentes musicales se ha abordado en los últimos años mediante métodos de descomposición de señal, junto con una clasificación de dichas componentes y finalmente con la síntesis de las fuentes por separado. El problema de la separación de fuentes se tratará sobre señales monoaurales, siendo éste el escenario más complejo que existe si se atiende al número de canales de información para la separación de fuentes.

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.

En este Trabajo Fin de Grado se va a realizar el algoritmo de separación de señales monoaurales en sus correspondientes señales de armónicos y percusivos, además se obtendrá también la señal vocal separada de las dos anteriores.

1.1. Introducción a la teoría musical

El primer concepto que hay conocer es la definición de música. ¿Qué es la música? Dependiendo de la persona a la que se le pregunte y de su área de especialización, las definiciones pueden ser muy diversas. Una definición aceptada debido a su objetividad es: "La música es el arte de combinar sonidos agradablemente al oído según las leyes que lo rigen". [1]

Es interesante analizar la definición dada. "La música es el arte…", la música es una forma de comunicación entre el interior de una persona (el músico en este caso) y el resto de personas (los oyentes). La música se define como arte porque es un don con el que se nace, la música no es una materia que se pueda enseñar en una academia o en una universidad. Esto es así debido a que dos personas no son capaces de interpretar exactamente de la misma forma una misma composición musical, cada individuo transmite su personalidad cuando están interpretando o componiendo. "…combinar sonidos…", en la naturaleza que nos rodea hay infinidad de sonidos, pero eso no significa que cualquier sonido pueda ser considerado como música. Para ello los sonidos tienen que ser combinados en los tiempos precisos, con unas duraciones concretas y unas intensidades adecuadas, para que cuando sean escuchados por el oído se produzca una sensación de agrado.

Algunos de los conceptos básicos que forman la notación musical tradicional son los siguientes:

- **Evento [27]:** es la unidad básica musical que representa un sonido con un determinado pitch o frecuencia fundamental (f_0), que se encuentra activo durante un determinado intervalo temporal. Las unidades básicas de la música son las notas musicales, que se agrupan en escalas, las cuales están formadas por 12 notas musicales. Estas 12 notas se dividen en dos grupos, 7 notas sin alteración (Do, Re, Mi, Fa, Sol, La y Si, que están ordenadas de más grave a más aguda); y 5 notas con alteración, ya sea con un sostenido (#), que incrementa un semitono la nota a la que acompaña, o con un bemol (♭), que disminuye un semitono la nota.
- **Octava [28]:** se define como la serie diatónica en la que se incluyen los siete sonidos constitutivos de una escala y la repetición del primero de ellos, es decir, una octava estaría formada por las notas Do, Re, Mi, Fa, Sol, La, Si y Do.
- **Onset [3]:** es el instante de tiempo en el que se inicia un evento, es decir, el momento en el que se inicia una nota musical.
- **Acorde [4]:** es una combinación de dos o más notas sonando de manera simultánea o casi simultánea. Los acordes más frecuentes son las tríadas, que son

acordes de tres notas. Un acorde puede ser consonante o disonante dependiendo de la relación entre las frecuencias de los parciales que componen dichos sonidos armónicos. Un acorde consonante provoca una sensación agradable cuando se escucha, mientras que un acorde disonante provoca una sensación molesta.

- **Pentagrama [1]:** está formado por cinco líneas horizontales y cuatro espacios que se usan para escribir sobre ellos las notas musicales (figura 1). Por convenio se sigue la misma dirección que en la escritura del lenguaje natural, es decir, de izquierda a derecha, por tanto, el pentagrama es una línea de tiempo en la que las notas situadas más a la izquierda se tocan antes que las que se encuentren más a la derecha.

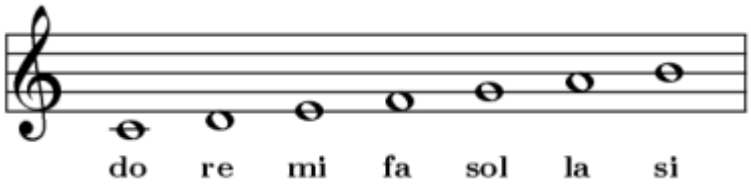

Figura 1: Pentagrama en clave de Sol con las notas musicales ordenadas [29].

- **Clave [1]:** es un símbolo musical que se usa para indicar el pitch o frecuencia de las notas escritas sobre el pentagrama. Se sitúa sobre una de las líneas al comienzo del pentagrama e indica el nombre y pitch de las notas que se sitúen sobre dicha línea. Existen tres claves, clave de Sol, la clave de Fa y la clave de Do (figura 2).

Figura 2: Tipos de claves que existen [30].

- **Duración [1]:** la duración de cada nota se representa mediante figuras musicales (figura 3). Cada figura tiene una duración relativa respecto a las demás. Se establece como referencia la duración de la figura negra (1 segundo), por lo que las duraciones de las figuras respecto de la negra serían: redonda (4), blanca (2), negra (1), corchea (1/2), semicorchea (1/4), fusa (1/8) y semifusa (1/16).

NOMBRE	FIGURA	SILENCIO	VALOR/PULSOS
Redonda			4 Tiempos
Blanca			2 Tiempos
Negra			1 Tiempo
Corchea			1 / 2 Tiempo
Semicorchea			1 / 4 Tiempo
Fusa			1/8 Tiempo
Semifusa			1/16 Tiempo

Figura 3: Clasificación de las figuras musicales según su duración [31].

1.2. El sonido

El sonido es la sensación percibida por el sentido del oído como resultado de la energía mecánica transportada por ondas longitudinales de presión en un medio material elástico, ya sea sólido, liquido o gaseoso (madera, agua o aire, respectivamente) [2]. Las ondas sonoras son denominadas ondas longitudinales porque el campo de presión se propaga en la dirección de propagación de onda, al contrario de las ondas electromagnéticas cuyos campos eléctrico y magnético son perpendiculares a la dirección de propagación, por lo que se denominan ondas transversales. Las ondas acústicas son ondas mecánicas y al hablarse de ondas sonoras está refiriéndose a ondas acústicas producidas en el rango audible del ser humano. Una clasificación según este criterio es:

- **Infrasonidos [2]:** son sonidos cuya frecuencia es inferior a los 15 Hz y no suelen ser percibidos por el oído humano, aunque eventualmente es posible percibir las vibraciones en los tejidos blandos del cuerpo.
- **Sonido audible [2]:** son los sonidos cuya frecuencia está comprendida entre los 15 Hz y los 20 KHz. La máxima frecuencia sonora que es capaz de percibir el oído humano depende de diversos factores, entre ellos la edad y, en tanto que un niño puede percibir frecuencias cercanas a los 20 KHz, una persona de más de 60 años sólo percibe frecuencias de hasta unos 10 o 12 KHz.
- **Ultrasonidos [2]:** son sonidos cuya frecuencia es superior a los 20 KHz y pueden ser percibidos por algunos animales como los perros. No hay realmente un límite superior de frecuencia para lo que se designa como ultrasonido.

En la siguiente figura se puede observar esta clasificación de los sonidos de forma gráfica. También se puede ver la clasificación de los sonidos audibles en graves, medios y agudos, y las frecuencias asociadas a cada uno de ellos.

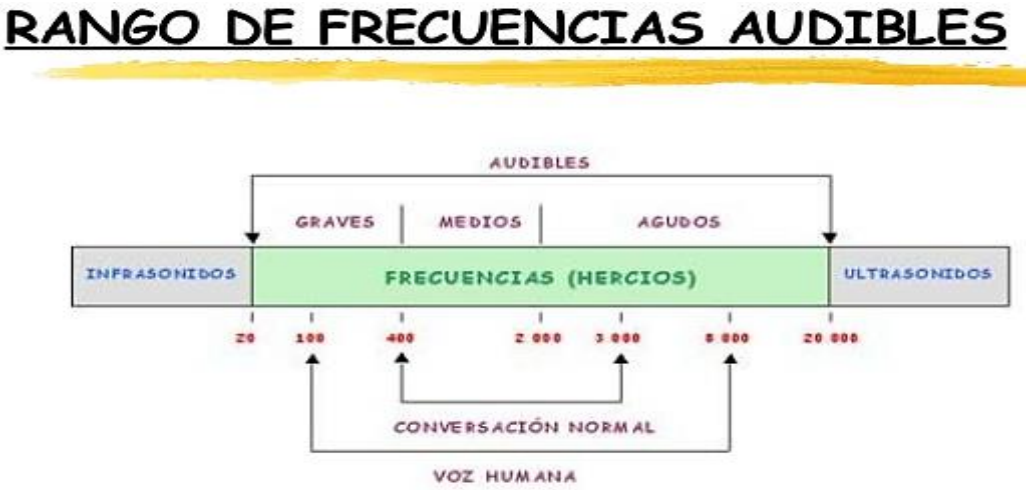

Figura 4: Clasificación de los sonidos en función de su frecuencia [32].

1.2.1. Características del sonido

- **Pitch [4]:** es una característica perceptual del sonido que permite ordenarlos en una escala de frecuencia. El pitch se define como la frecuencia de una determinada sinusoide que se asocia al sonido que se está percibiendo.
- **Frecuencia fundamental [5]:** es la magnitud física que se relaciona con el pitch. Es la frecuencia más baja a la que una nota puede ser tocada. Los armónicos de una nota son aquellas frecuencias que son múltiplos de su frecuencia fundamental. Hay que tener en cuenta que la componente de la frecuencia fundamental no tiene por qué ser la que tenga mayor energía. Algunos autores denominan a la frecuencia de un sonido como su altura [1].
- **Sonoridad [33]:** es la medida subjetiva de la intensidad con la que el oído de una persona percibe un sonido. La unidad que mide esta magnitud es el fonio o fon, que es la sonoridad de un tono con una frecuencia de 1KHz con un nivel de presión sonora de $0dB_{SPL}$.

Las curvas que establecen la relación entre la frecuencia y el nivel de presión sonora que tienen que tener los sonidos para que sean percibidos con la misma sonoridad se denominan curvas isofónicas (figura 5). Se observa que la curva de 0 fonios es el umbral de audición. Por debajo de esa sonoridad el ser humano no es capaz de percibir los sonidos. Mientras que la curva de 120 fonios delimita superiormente el rango de los sonidos audibles por las personas. Si los sonidos superan ese umbral conocido como umbral del dolor, podríamos sufrir daños irreversibles en nuestros oídos.

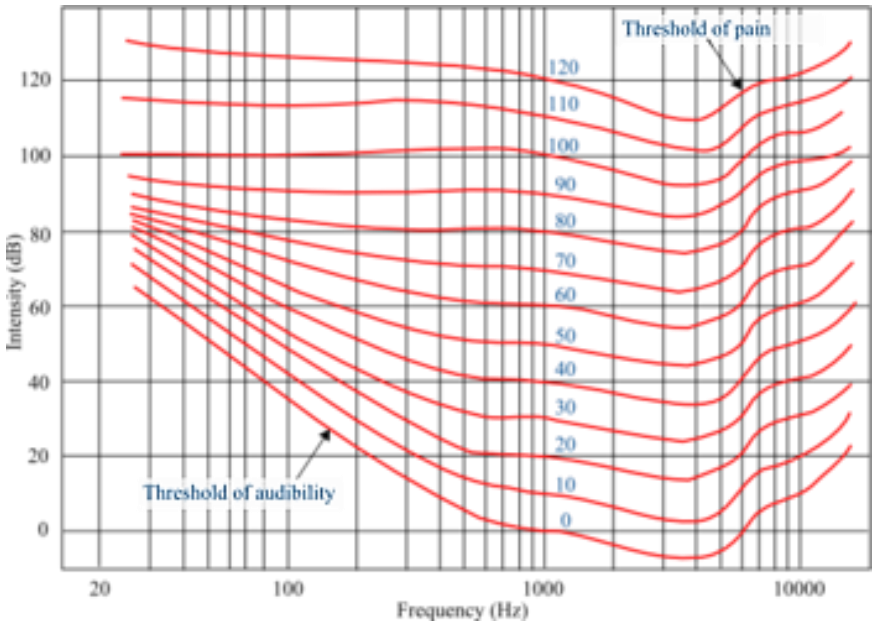

Figura 5: Curvas isofónicas de Fletcher y Munson [34].

- **Duración [1]:** informa del espacio temporal que ocupa un determinado sonido, es decir, el período de tiempo comprendido desde su aparición hasta su extinción.
- **Timbre [5]:** es una característica o atributo del sonido que permite saber de qué fuente proceden sonidos con igual sonoridad, altura y duración.

Se puede hablar de timbres a tres niveles:

- **General:** cuando se distinguen elementos de diferentes clases, por ejemplo, diferenciar una guitarra de una flauta.
- **Parcial:** cuando se distinguen elementos de la misma clase, por ejemplo, diferenciar varios tipos de guitarra.
- **Particular:** cuando se distinguen las diversas posibilidades de un único elemento de una clase específica, por ejemplo, las diferentes formas de tocar una guitarra (con púa, sin púa, golpeando la caja, etc.).

- **Reverberación [2]:** es el efecto que se produce en los recintos cerrados, a cusa de los diversos mecanismos de propagación sonora, principalmente reflexiones múltiples y difusas, que dan lugar a que el sonido tarde un cierto tiempo en “apagarse” aun cuando la fuente sonora haya dejado de emitir. Este tiempo, denominado tiempo de reverberación, se define como el tiempo en el que el sonido reverberante alcanza un nivel de presión de -60 dB 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 de las paredes, del mobiliario, de las personas o de los objetos presentes en el recinto. Por tanto, la constante de atenuación es un fenómeno de gran importancia en los diseños de estudios de grabación.

1.2.2. Clasificación de los sonidos

Los sonidos pueden clasificarse atendiendo a diferentes criterios. A continuación se van a explicar los distintos tipos de sonidos según el número de canales (1.2.2.1.), según el número de eventos que se producen de forma simultánea (1.2.2.2.), según el número de instrumentos musicales (1.2.2.3.) y según el grado de periodicidad (1.2.2.4.).

1.2.2.1. Monoaurales, estéreo y multicanal

- Los sonidos monoaurales [7] son más conocidos como sonidos mono. Se caracterizan porque no muestran información espacial de ningún tipo y tampoco permiten saber el número de fuentes que se han empleado, ya que este tipo de sonidos están compuestos por la suma de todas las fuentes que se han utilizado para

su grabación. Todas las fuentes son grabadas por un único canal, de ahí su denominación como señales monocanal.

- Los sonidos estéreo [8] son aquellos que están grabados por dos canales, es decir, los sonidos son recogidos por dos micrófonos independientes. Este hecho permite obtener información espacial acerca de la localización de las fuentes musicales. Esta forma de grabación es muy usual en la grabación de la música comercial actual.
- Los sonidos multicanal [27] son también conocidos como sonidos surround. Son aquellas señales que se generan con al menos cuatro canales de audio diferentes. Con esta forma de grabación se obtiene una gran cantidad de información espacial sobre las fuentes que se están utilizando. Uno de los sistemas de audio multicanal más conocido y empleado el sistema 5.1, este sistema emplea cinco altavoces para la reproducción de los sonidos, dos de ellos son para fuentes provenientes de la parte derecha del oyente, otros dos son para los sonidos provenientes de la parte izquierda y un altavoz central que se emplea para los sonidos vocales. El ".1" es un altavoz adicional conocido como subwoofer, que se utiliza para la reproducción de sonidos de muy baja frecuencia (los efectos especiales de las películas o los sonidos de los instrumentos de percusión de las canciones musicales).

1.2.2.2. Monofónicos y polifónicos

- Un sonido es monofónico [9] si en un instante de tiempo un único instrumento toca únicamente una nota, dicho de otro modo, un sonido es monofónico si en cada instante temporal aparece una única frecuencia fundamental activa.
- Un sonido es polifónico [9] si en un instante de tiempo se producen dos o más notas a la vez, ya sean tocadas por un único instrumento o por varios de ellos, es decir, en un instante de tiempo aparecen varias frecuencias fundamentales activas.

1.2.2.3. Monotímbricos y multitímbricos

- Un sonido es monotímbrico [27] si todos sus eventos tienen la misma envolvente espectral, es decir, si tienen el mismo timbre. Lo que quiere decir que han sido interpretados por el mismo instrumento.
- Un sonido es multitímbrico [27] cuando sus eventos tienen envolventes espectrales diferentes, es decir, son tocados por distintos instrumentos los cuales poseen timbres diferentes.

1.2.2.4. Armónicos, percusivos y vocales

- Un sonido es armónico o periódico si está formado por una o varias sinusoides de frecuencias armónicamente relacionadas entre sí. En el espectrograma de este tipo de sonidos aparecen componentes espectrales que son múltiplos de la frecuencia fundamental. Los instrumentos musicales que producen sonidos armónicos son los instrumentos de viento y los de cuerda.
- Un sonido es percusivo si no muestra periodicidad en el tiempo y por tanto en frecuencia no presenta ninguna relación armónica entre sus componentes. Los instrumentos musicales que producen estos sonidos son los de percusión, excepto el xilófono, cuyos sonidos si tienen una cierta periodicidad.
- Los sonidos vocales son aquellos que son producidos por personas, ya estén hablando o cantando. Tienen características tanto de los sonidos armónicos como de los percusivos.

En los siguientes apartados se explicaran con detalle todas las características de estos tres tipos de sonidos.

1.3. Sonidos armónicos

En 1807, el matemático francés Jean-Baptiste Joseph Fourier publicó su teorema sobre las series de Fourier. Este teorema demuestra que cualquier señal continua que se repita en el tiempo, es decir, que tenga un periodo T y por tanto una frecuencia f, puede ser descompuesta en una suma de señales sinusoidales de frecuencias más simples que la de la señal a descomponer. Estas sinusoides son armónicas, ya que sus frecuencias están relacionadas entre sí mediante múltiplos enteros. Esto se demuestra matemáticamente en la siguiente ecuación:

$$f(t) \approx \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos\left(\frac{2\pi n t}{T}\right) + b_n \sin\left(\frac{2\pi n t}{T}\right) \right) \quad (1)$$

Donde *T* es el periodo de la señal original $f(t)$, a_n y b_n son los coeficientes y se calculan de la siguiente manera:

$$a_n = \frac{2}{T}\int_0^T f(t)\cos\left(\frac{2\pi nt}{T}\right)dt \qquad \text{n=0, 1, 2,…} \qquad (2)$$

$$b_n = \frac{2}{T}\int_0^T f(t)\sin\left(\frac{2\pi nt}{T}\right)dt \qquad \text{n=1, 2, 3…} \qquad (3)$$

Esta afirmación que matemáticamente parece compleja y difícil de entender, se puede comprender fácilmente con un ejemplo básico. Cualquier nota tocada por un instrumento de viento o de cuerda se puede descomponer según el desarrollo en series de Fourier. En este caso tenemos una nota musical tocada con un piano:

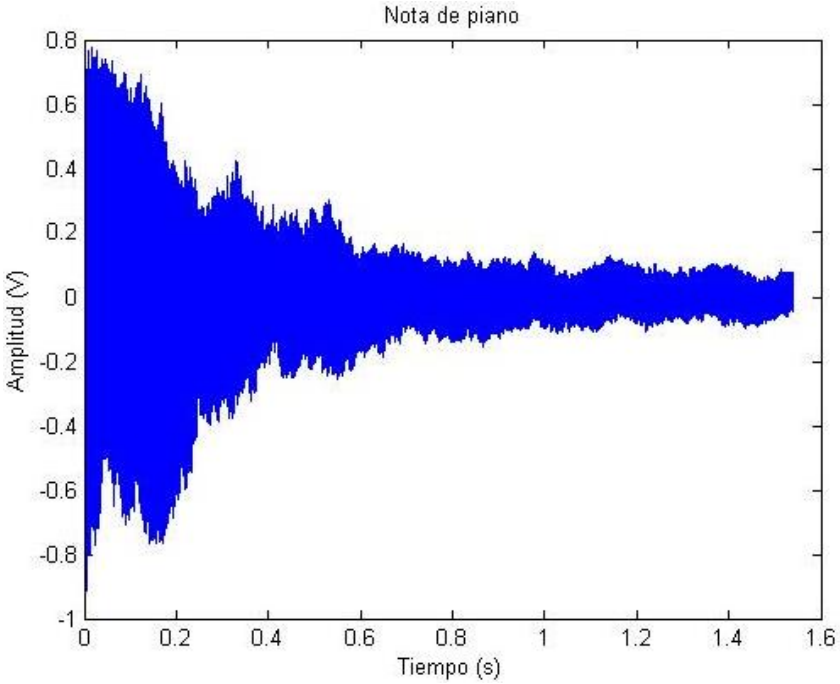

Figura 6: Representación de una nota de piano en el tiempo.

Si se calcula su espectrograma y se representa (figura 7), se observa lo que se había demostrado numéricamente en la ecuación 1, es decir, que el sonido producido al tocar una tecla de piano es armónico. Cada línea horizontal que se ve en el eje de la frecuencia se corresponde con una sinusoide, la primera línea se corresponde con la sinusoide de frecuencia fundamental f0 (720 Hz), que es la frecuencia de la nota que ha sido tocada. Las otras líneas por encima se corresponden con los múltiplos enteros de la frecuencia fundamental (2f0, 3f0, 4f0,…). En este caso la componente de mayor energía es la de la

frecuencia fundamental, pero esto no sucede en todos los sonidos, ya que se puede dar el caso de que la componente con mayor energía sea algún armónico.

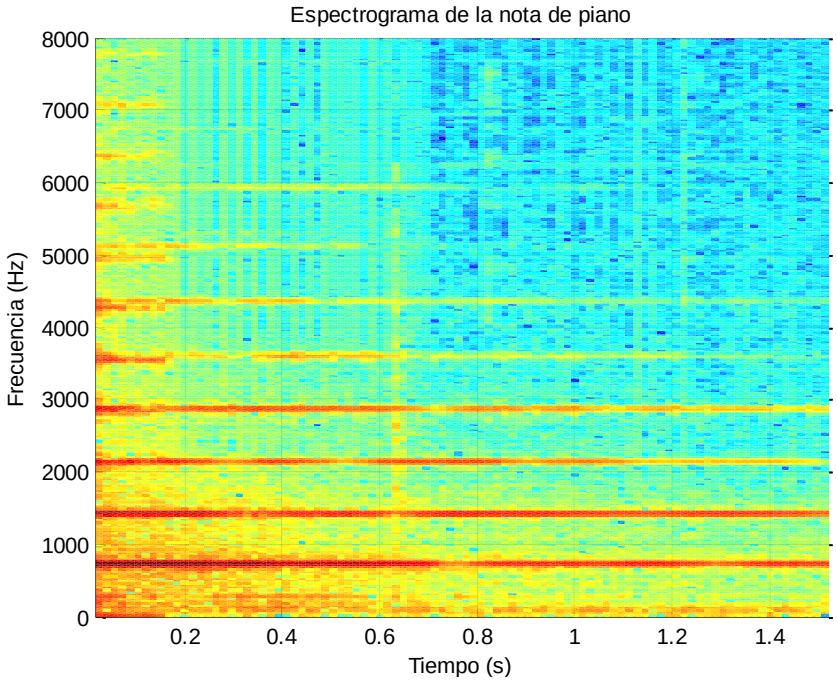

Figura 7: Espectrograma de la nota de piano representada en la figura 6.

Además, al observar el espectrograma se puede apreciar otra característica de los sonidos armónicos [12]. Esta característica es que las componentes espectrales de los sonidos armónicos aparecen como una señal continua a lo largo del tiempo. En cambio en el eje de la frecuencia estas componentes no son continuas, sino que solo aparecen componentes en f_0 y en sus múltiplos, dando lugar a sus componentes armónicas. Este comportamiento de los sonidos armónicos será la base para el proceso de separación que se llevará a cabo en este Trabajo Fin de Grado.

1.4. Sonidos percusivos

Los sonidos percusivos, al contrario que los sonidos armónicos, no se rigen por el Teorema de Fourier, ya que no presentan ni continuidad ni periodicidad. Por lo tanto sus características son totalmente las opuestas. Puesto que los sonidos percusivos no pueden ser descompuestos en una suma de sinusoides, se llega a la conclusión de que no presentan periodicidad en frecuencia. Pero por el contrario si presentan periodicidad en el tiempo.

Con un ejemplo se ve claramente estas características [12]. Se tiene una señal que son tres golpes de tambor (figura 8).

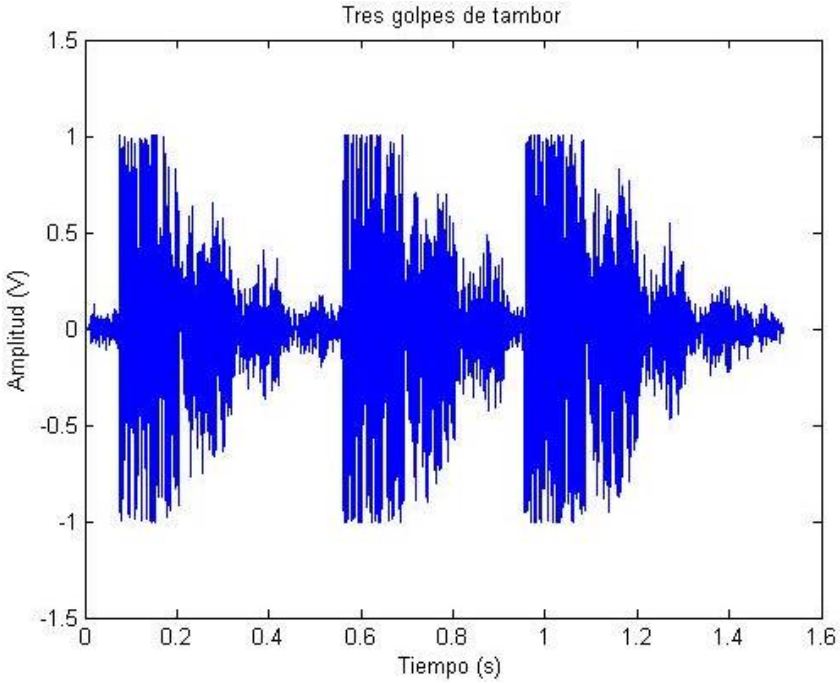

Figura 8: Representación de tres golpes de tambor en el tiempo.

A continuación se calcula su espectrograma:

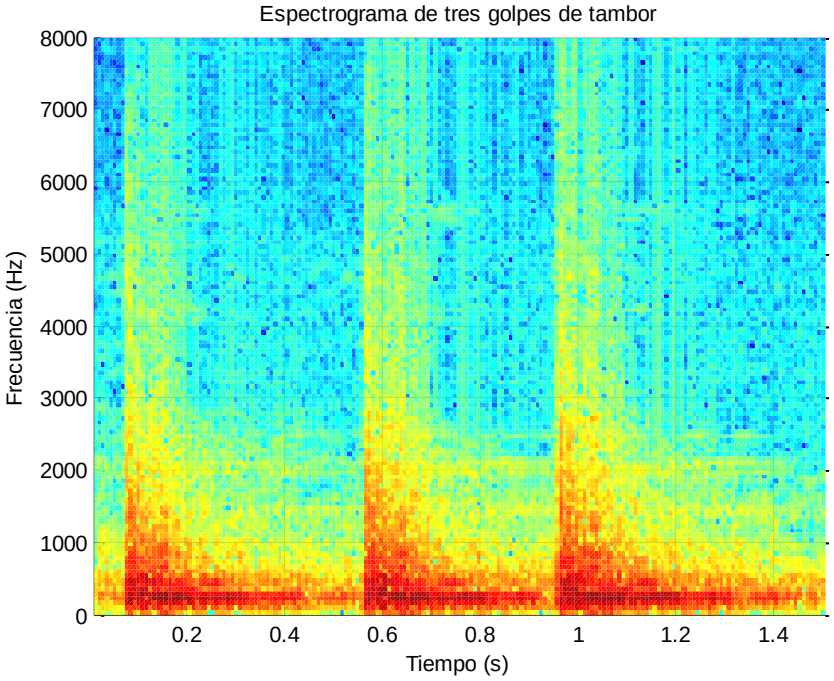

Figura 9: Espectrograma de la señal de tambor de la figura 8.

Se observa claramente que hay tres golpes de tambor. Se ve que en el eje de frecuencia la señal es continua (los sonidos percusivos no están formados por sinusoides), característica por la cual a los sonidos percusivos se les denomina como sonidos de banda ancha. También se puede observar que en el eje del tiempo sí se diferencian los tres golpes

de tambor, por esta característica se dice que los sonidos percusivos son discontinuos en el tiempo.

1.5. La voz

1.5.1. Aparato fonador

En la producción de la voz humana intervienen tres órganos que tienen funciones claramente diferentes. Estos son los pulmones, la laringe y el tracto vocal [10]. La primera etapa de la voz humana son los pulmones. Los pulmones realizan una función poco compleja, ya que se encargan de almacenar el aire que se coge del exterior en el proceso de inhalación. Una vez almacenado, los pulmones se convierten en una fuente de aire, expulsándolo de nuevo hacia el exterior en el proceso de exhalación. Para que el sonido sea más alto es necesario que los pulmones expulsen el aire con mayor presión. Los pulmones tienen la capacidad de superar la presión atmosférica entre un 0,4% y un 2%.

Cuando el aire sale de los pulmones pasa por la tráquea hacia la laringe, en este instante comienza la segunda etapa del proceso del habla. La laringe es un conducto formado principalmente por cartílagos. Un conjunto de estos cartílagos son las cuerdas vocales, que tienen una función vital en el proceso del habla. Las cuerdas vocales, al contrario de lo que su nombre indica, no son cuerdas, sino que son pliegues de ligamentos que se extienden desde el cartílago del tiroides, que se encuentra en la parte delantera del cuello, hasta el cartílago del aritenoides, que se encuentra en la parte trasera. Las cuerdas vocales se disponen en forma de V y el espacio entre ellas se denomina glotis. La función de las cuerdas vocales es modular el flujo de aire que pasa a través de ellas, mediante la apertura y el cierre de estas. Estos movimientos se realizan de forma automática produciendo la vibración de las cuerdas, lo que da a lugar a la creación de las vocales y de las consonantes.

En las personas adultas, las cuerdas vocales de los hombres son más largas y pesadas que las de las mujeres, como consecuencia, las de los hombres tienen una capacidad de vibración más baja. Las frecuencias típicas de vibración de las cuerdas en las personas son de 110 Hz para los hombres, de 220 Hz para las mujeres y de 300 Hz para los niños, aunque dependiendo de la persona la frecuencia de vibración varía.

Figura 10: Esquema del aparato fonador [35].

La tercera etapa es el tracto vocal (figura 10), que se encarga de transformar los zumbidos y vibraciones que provienen de las cuerdas vocales en los sonidos que producimos al hablar. El tracto vocal se considera como el conducto que va desde las cuerdas vocales hasta los labios, pero en realidad está formado por la faringe, la cavidad bocal y la cavidad nasal. El velo de paladar se encarga de regular la cantidad de aire que pasa de la faringe a la cavidad nasal y a la boca. La boca es la parte más importante del tracto vocal, e incluso del proceso del habla, ya que está formada por el paladar, la lengua, los dientes y los labios. Todas estas partes se pueden mover para variar la forma y el tamaño de esta cavidad. Esta característica del tracto vocal permite que cada persona genere sonidos distintos, es decir, en cada persona la movilidad del tracto vocal hace que la voz tenga un timbre diferente y por tanto hablen de forma distinta.

1.5.2. Características de la voz

Como se ha explicado en el apartado anterior, una parte fundamental en el proceso del habla son las cuerdas vocales, que aunque realmente sean pliegues de ligamentos, se les denomina cuerdas como un símil con las cuerdas de los instrumentos de cuerda. Esto es así porque al vibrar las cuerdas vocales en el proceso del habla, generan componentes armónicas al igual que dichos instrumentos. En la figura 11 se puede ver el espectrograma de un fragmento de un discurso de Steve Jobs en la Universidad de Stanford, donde se aprecia claramente esta característica, la armonicidad de la voz:

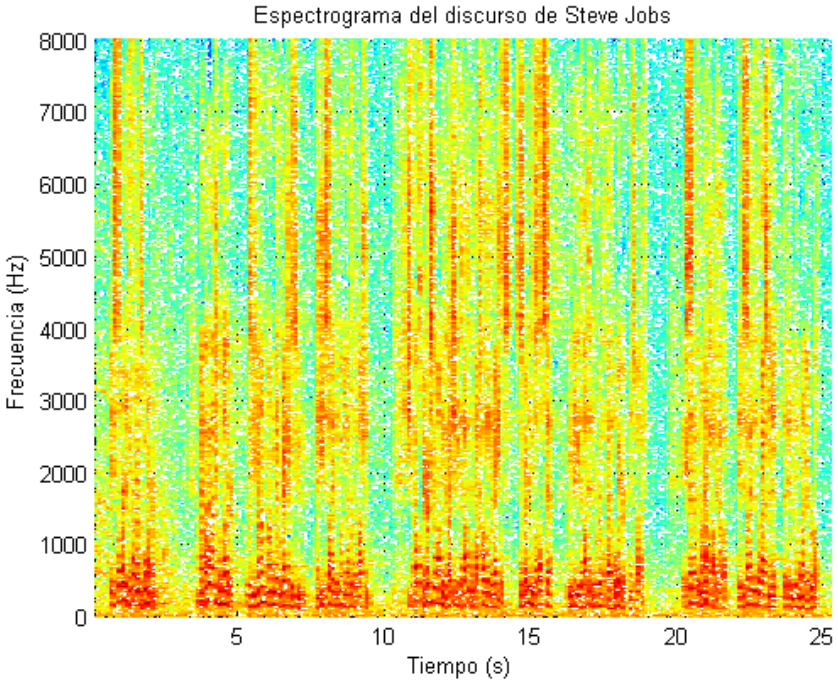

Figura 11: Espectrograma del discurso de Steve Jobs en la Universidad de Stanford [36].

En esta figura se puede ver que la voz tiene un alto contenido armónico, debido a la resonancia que tiene el tracto vocal, que como se ha visto en el apartado anterior, está formado por muchas partes móviles (boca, lengua, dientes y velo del paladar), que permiten la formación final de los sonidos. Esta característica de la voz será de gran utilidad para el desarrollo de este Trabajo Fin de Grado.

Otra característica muy importante de la voz son los formantes [10] [11], que permiten diferenciar los sonidos vocálicos entre sí. En la producción del habla las cuerdas vocales vibran a unas determinadas frecuencias, y es en el tracto vocal donde unas frecuencias se refuerzan y otras se atenúan. A cada conjunto de frecuencias o armónicos que se realzan se les denomina formantes. Esta propiedad sirve para reconocer los sonidos vocálicos. Los tres primeros formantes son los más importantes, porque dan información de la posición de la boca, de la lengua y del velo del paladar.

El primer formante (F1) está relacionado con la apertura de la cavidad bucal, cuanto más abierta se encuentre la boca, mayor frecuencia tendrá F1. Así pues, como la vocal a es la más abierta, la frecuencia de su F1 será la mayor de las 5 vocales, mientras la frecuencia del F1 de la vocal u será la más baja, ya que es la vocal más cerrada.

El segundo formante (F2) está relacionado con el retroceso de la lengua y el redondeo labial. Cuanto más anterior sea la vocal, es decir, cuanto más cerca esté la lengua de los dientes al pronunciar el sonido, mayor será la frecuencia del F2. Además, cuanto menos cerrada y redonda esté la boca, mayor será la frecuencia del F2. Por tanto,

la vocal i tendrá la frecuencia más alta para el F2, mientras que la vocal u tendrá la frecuencia más baja de las 5 vocales.

Por último, el tercer formante (F3) está relacionado con el descenso del velo del paladar y la elevación de la punta de la lengua. Cuanto más arriba esté el velo del paladar y más abajo esté la punta de la lengua, mayor será la frecuencia del F3 (sonido más agudo). Cuanto más arriba este la punta de la lengua, menor será la frecuencia del F3.

A continuación se va a realizar la transformada de Fourier de cada una de las cinco vocales (figura 13). El objetivo es ver las frecuencias a las que aparecen los distintos formantes y ver si se corresponden con las frecuencias medias que se establecen en la carta de formantes (figura 12).

Al observar ambas figuras, se puede comprobar que las frecuencias de los dos primeros formantes de las cinco vocales coinciden. Y además se demuestra las propiedades que van asociadas a cada formante, que han sido explicadas en los párrafos anteriores.

Por ejemplo si se coge la vocal i, en relación al F1 su frecuencia debería de ser la más baja junto con la de la vocal u, por ser las dos vocales más cerradas. Y efectivamente esto es así, ya que su frecuencia está en torno a los 220 Hz, al igual que en la vocal u. En relación al formante F2 también se cumple lo esperado, la vocal i tiene el F2 más alto en torno a los 2650 Hz, mientras que la vocal u lo tiene en torno a 650 Hz.

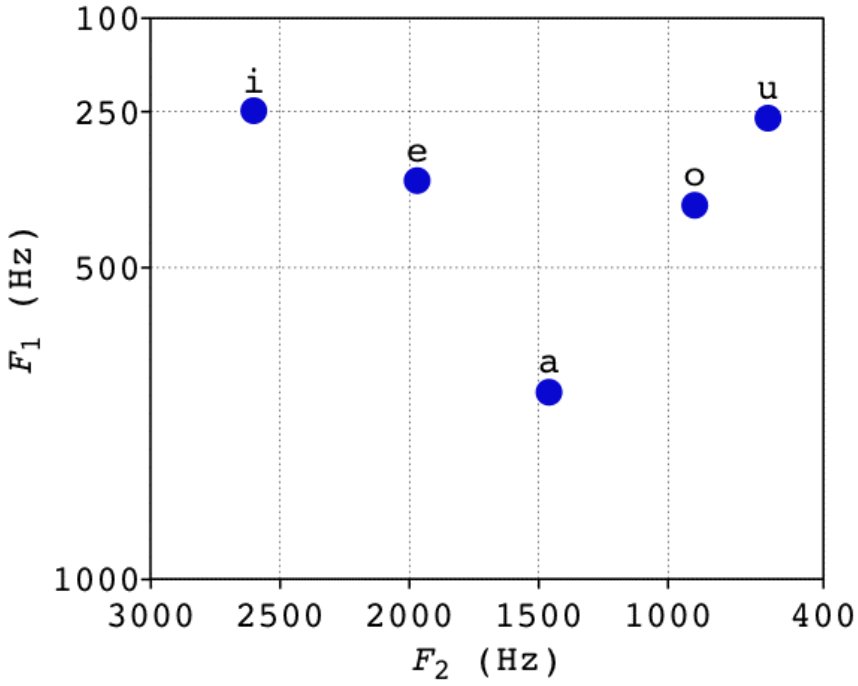

Figura 12: Carta de formantes [37].

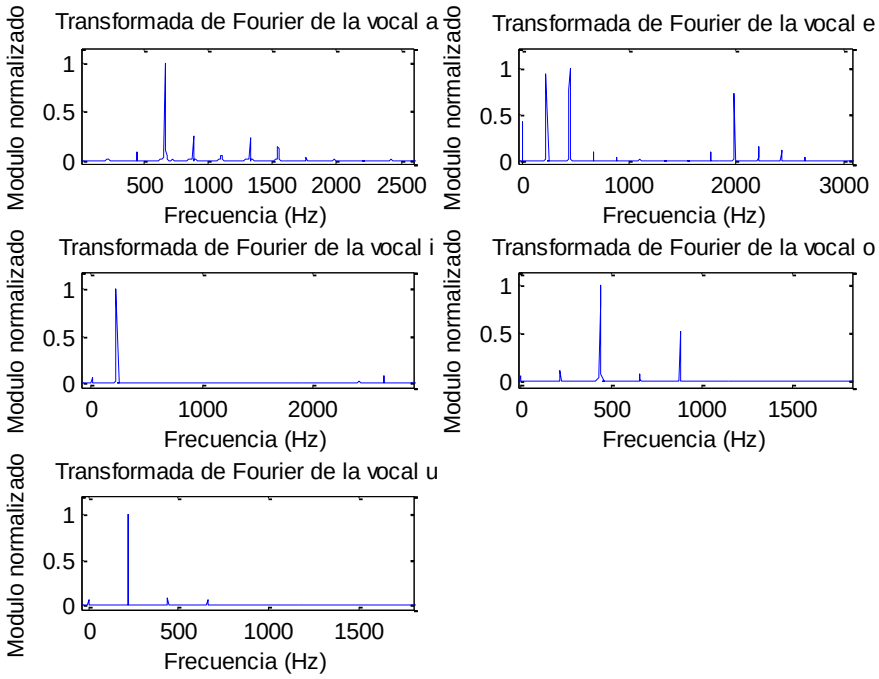

Figura 13: Módulo de la transformada de Fourier de las cinco vocales.

Otra característica de la voz que es muy utilizada es el vibrato y el trémolo [11]. Aunque están estrechamente relacionadas, no son exactamente el mismo efecto. El vibrato consiste en la oscilación regular de una nota, es decir, consiste en una modulación de la frecuencia de una nota (modulación FM). Mientras que el trémolo consiste en la variación de la intensidad o volumen de una nota, es decir, consiste en una modulación de la amplitud de una nota (modulación AM). Estos dos términos están profundamente relacionados, porque en la mayoría de las ocasiones los cantantes realizan estos fenómenos indistintamente y llegan a mezclarlos. El rango de oscilación de frecuencia para la voz de los cantantes es de unos 6 Hz, mientras que el rango de amplitud esta entre 0,6 y 2 semitonos.

1.5.3. Clasificación de los fonemas

Según el modo de articulación de los órganos del tracto vocal, las consonantes se clasifican en [10]:

- Oclusivas: se bloquea el paso de aire durante una fracción de tiempo en alguna parte del tracto vocal (normalmente se cierra la boca). Cuando la presión de la columna de aire logra vencer la oclusión, éste sale bruscamente, originando un ruido o chasquido característico, que se denomina explosión. Las consonantes de este tipo son p, b, d, t, k y g.

- Fricativas: los órganos de la articulación se acercan sin llegar a cerrarse, dejando entre ellos un estrecho canal por el que la columna de aire, al pasar, produce un ruido turbulento característico conocido como fricción. Las consonantes de este tipo son: f, v, s, z y h.

- Nasales: se producen moviendo el velo del paladar para conectar la cavidad nasal con la faringe dejando bloqueada la cavidad bucal. Las consonantes de este tipo son: m y n.

- Semivocales: se posiciona el tracto vocal como si se fuera a pronunciar una vocal y luego se cambia la forma, pronunciando la vocal que le sigue, este tipo de consonantes siempre van seguidas de una vocal. Las consonantes de este tipo son: w e y.

- Líquidas: la punta de la lengua se eleva y la cavidad bucal se estrecha. Estas consonantes son: l y r.

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC) © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Figura 14: Clasificación de los fonemas del alfabeto internacional [38].

Según el lugar de articulación de los órganos del tracto vocal, las consonantes se clasifican en:

- Bilabiales: el labio inferior toca o se aproxima al labio superior. Son: p, b y m.

- Labiodentales: el labio inferior toca o se acerca a los incisivos superiores. Es la f.

- Dentales: la punta de la lengua se coloca tras los incisivos superiores. Son: t y d.

- Alveolares: la lengua toca o se aproxima a los alvéolos. Son: s, n, l y r.

- Palatales: el dorso de la lengua toca o se aproxima al paladar duro. Es la ñ.

- Velar: el dorso de la lengua toca o se acerca al velo del paladar. Son: k, g y x.

- Glotales: las cuerdas vocales participan en la articulación, bien uniéndose fuertemente sin vibrar (lo que se suele llamar golpe de glotis) o bien acercándose. Es la letra h.

Además existe otro criterio de clasificación, que va a ser de vital importancia para la realización de este Trabajo Fin de Grado. Según la vibración de las cuerdas vocales los fonemas se clasifican en:

- Sonoros: se caracterizan porque las cuerdas vocales vibran para producir el sonido. Los fonemas sonoros son: b, d, g, v, z, l, m, n y r.

- Sordos: se caracterizan porque las cuerdas vocales no vibran para producir el sonido. Los fonemas sordos son: p, t, k, f, ɵ, s y h.

El comportamiento espectral de los fonemas sordos y sonoros es una de las bases de este Trabajo Fin de Grado, por lo que se ha realizado el espectrograma de tres fonemas sordos (t, p y k) (figura 15) y de tres fonemas sonoros (p, d y g) (figura 16).

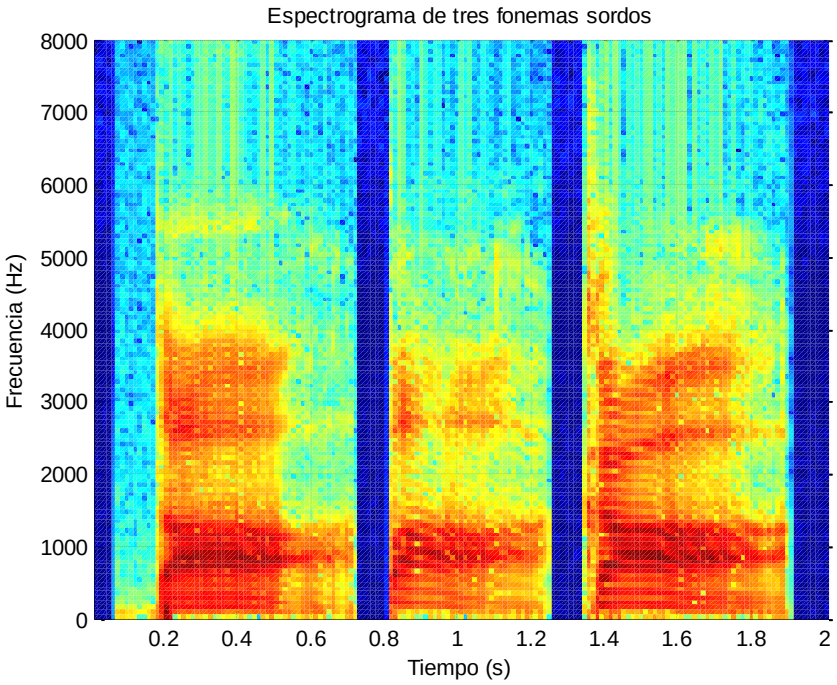

Figura 15: Espectrograma de los fonemas sordos t, p y k.

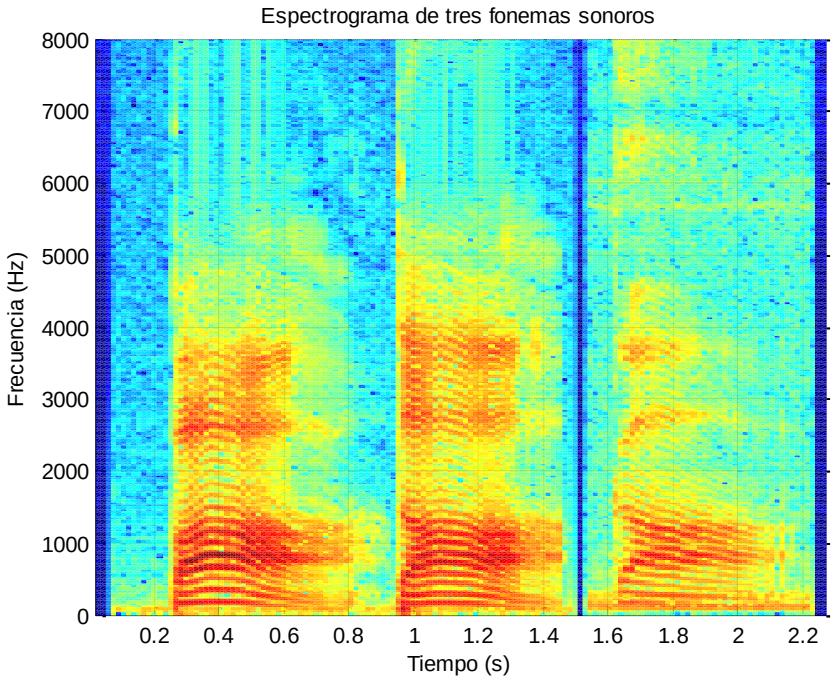

Figura 16: Espectrograma de los fonemas sonoros b, d y g.

De las dos figuras anteriores se saca la conclusión de que los fonemas sordos se comportan como los sonidos percusivos, ya que presentan un ancho de banda muy grande, o lo que es lo mismo, son continuos en frecuencia. Mientras que los fonemas sonoros se asemejan más a los sonidos armónicos, ya que se aprecian discontinuidades en frecuencia.

2. OBJETIVOS

El principal objetivo de este Trabajo Fin de Grado es la separación de una señal de audio monocanal en tres pistas musicales distintas, por un lado los sonidos armónicos y por otro lado los sonidos percusivos, siendo los sonidos vocales la señal que queda.

Para alcanzar este objetivo se van a desarrollar una serie de objetivos específicos:

1. Revisión de la bibliografía específica que permita conocer el estado del arte.
2. Implementación de un algoritmo basado en las técnicas de factorización en matrices no negativas (NMF).
3. Estudio de las características de los diferentes tamaños de ventana que se utilizan en el proceso de enventanado.
4. Estudio de las características de los sonidos que permitan discriminar entre continuidad espectral y temporal.
5. Implementación de una interfaz de usuario con el entorno MATLAB.

3. ESTADO DEL ARTE

El objetivo de la separación de fuentes sonoras (Sound Source Separation, SSS) es recuperar las distintas señales de audio que han sido generadas por distintas fuentes, y que se encuentran mezcladas en una única señal (mono) o en varias señales (estéreo o multicanal). La SSS además de realizar la detección de las distintas notas y su asignación a la fuente correspondiente, también se encarga de sintetizar la señal correspondiente a cada fuente.

En la separación de fuentes sonoras, según el tipo de información previa de la que se disponga, se distingue entre separación de fuentes a ciegas (Blind Source Separation, BSS) y separación de fuentes informada (Informed Source Separation, ISS). En la BSS no se dispone de ningún tipo de información espacial o espectral previa. Con este tipo de separación se obtienen buenos resultados para señales multicanal [13], pero por el contrario si se emplea con señales monoaurales los resultados no presentan la calidad deseada [14]. La ISS sí dispone de información espectral [15] y espacial [16] previa. Este método es el que se emplea para que la separación de señales monoaurales tenga una calidad aceptable.

3.1. Independent Component Analysis (ICA)

El análisis de componentes independientes (ICA) [17] es un método que se engloba dentro de la BSS, ya que no emplea información adicional previa y además está diseñado para señales multicanal.

Los algoritmos ICA suponen que cada trama de cada canal $\mathrm{x_c(t)}$, es una mezcla lineal de las señales emitidas por cada fuente $\mathrm{g(t)}$, y lo expresan como en la siguiente ecuación:

$$x_c(\mathrm{t}) = \mathrm{Bg(t)} \tag{4}$$

Donde B es la matriz de mezcla con coeficientes desconocidos.

Basándose, por tanto, en la información de la señal de entrada $\mathrm{x(t)}$, los algoritmos ICA intentan invertir el proceso de mezcla buscando una matriz de separación óptima W, con la que la estimación de las fuentes separadas sería:

$$\hat{g}(t) = Wx(t) \quad (5)$$

Idealmente, la matriz de separación W, sería la inversa de la matriz de mezcla B y las fuentes reconstruidas idénticas a las señales independientes de cada fuente. Por consiguiente, los métodos ICA pretenden buscar una matriz W que reconstruya fuentes las cuales sean lo más independientes posible entre ellas.

Según el teorema central del límite, la suma de dos o más variables aleatorias independientes tiene una distribución que se puede considerar más Gaussiana que cualquiera de las variables aleatorias iniciales por separado [18]. Por tanto, si se reconstruyen las fuentes originales con distribuciones que sean lo menos gaussianas posible, un algoritmo ICA podría, potencialmente, recuperar las fuentes originales. Para medir la no gaussianidad se utilizan dos medidas, la entropía negativa o neguentropía y la curtosis. La entropía negativa se utiliza para medir la distancia a la normalidad de una señal, si una señal es gaussiana el valor de entropía negativa es 0, mientras que si no es gaussiana el valor es positivo. La curtosis es una medida de la forma de la distribución de una señal, la curtosis es positiva para distribuciones con un gran pico en torno a 0 y con amplias colas, mientras que será 0 si la distribución es gaussiana.

De este modo, los algoritmos ICA pueden diseñarse para modificar W de manera que las fuentes reconstruidas obtengan un valor absoluto de curtosis alto hasta que obtengan un conjunto de valores óptimos.

3.2. REpeating Pattern Extraction Technique (REPET)

La técnica de extracción de los patrones de repetición (REPET) [19] se basa en un concepto demostrado por científicos, que consiste en que en muchas piezas musicales el cantante superpone su voz a un acompañamiento instrumental repetitivo. Por tanto, si se identifican los patrones repetitivos de un audio y se aíslan de los no repetitivos, se podrá separar el acompañamiento instrumental de la voz. Este método consta de tres etapas, primero se realiza una identificación del periodo de repetición, luego se realiza un modelado del segmento de repetición y por último se procede a la extracción de los patrones repetitivos.

En la primera etapa se calcula el espectrograma X de la señal de entrada x, se eliminan los elementos negativos del espectrograma ya que se le aplica el valor absoluto, obteniéndose la matriz V y posteriormente la autocorrelación del espectrograma en potencia V^2, obteniéndose la matriz B. La magnitud b indica la similitud de las componentes

y se calcula realizando la media de las filas de la matriz de B. Si la señal original está compuesta por patrones repetitivos, en la matriz b se podrán apreciar picos que se repiten de forma periódica con un periodo p y con diferentes amplitudes.

En la segunda etapa se realiza el modelado de los segmentos que se repiten. El razonamiento que se sigue es que el espectrograma de la voz (patrones no repetitivos) tiene dispersión y varía en relación al espectrograma del acompañamiento musical (patrones repetitivos). Se le realiza la media a los segmentos de cada periodo p de la matriz b obteniendo una matriz S que recoge los segmentos con poca variación (acompañamiento instrumental).

En la tercera etapa se calcula un modelo de espectrograma de repetición W, calculando el mínimo entre cada elemento de S y de V. después se calcula una máscara tiempo-frecuencia suave M, que se calcula normalizando W respecto de V elemento a elemento. Esto hace que los segmentos repetitivos se correspondan con valores cercanos a 1 en la máscara, mientras que los no repetitivos se corresponden con valores cercanos a 0. Por último se aplica la máscara M a X y se pasa al domino del tiempo, obteniendo el acompañamiento instrumental. La voz se obtiene quitando la señal musical resultante a la señal original.

En la siguiente figura se pueden ver las distintas etapas de forma gráfica.

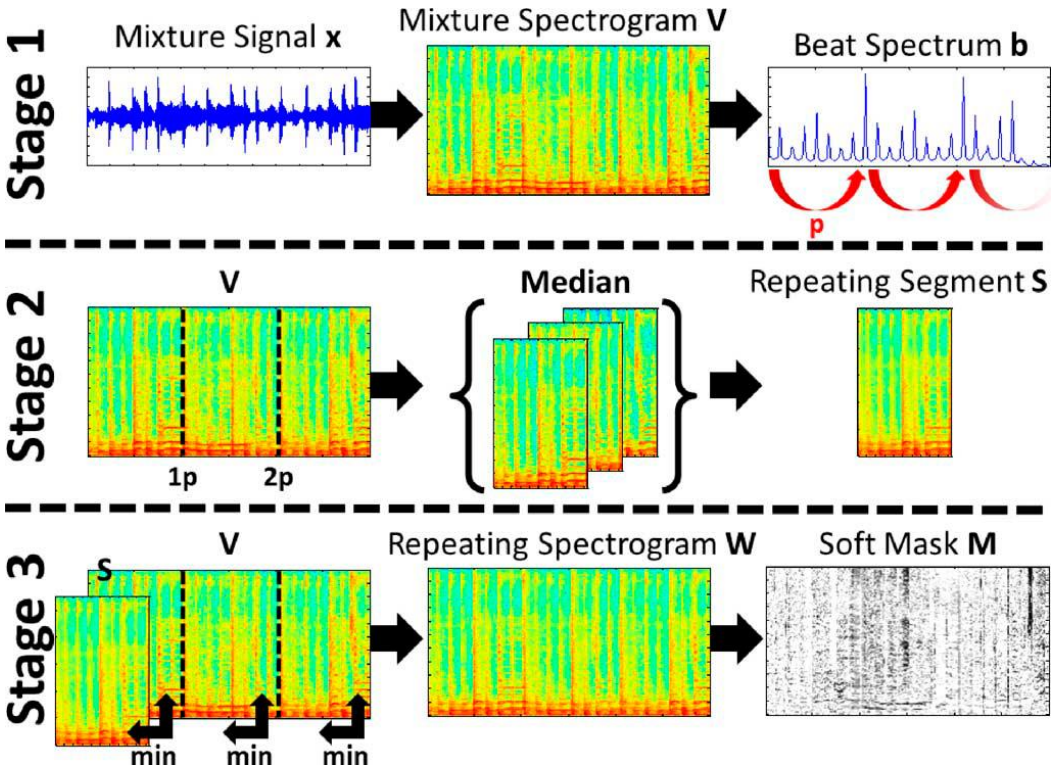

Figura 17: Desarrollo de la técnica de REPET [19].

3.3. Non-negative Matrix Factorization (NMF)

La factorización de matrices no negativas es más conocida por sus siglas en inglés como NMF [20] [21]. Es un conjunto de algoritmos matemáticos que se encargan de realizar la separación de una matriz de entrada X de dimensiones FxT, que es el espectrograma de la señal original x(t), en dos matrices que son multiplicadas entre sí, que son W de dimensiones FxN y H de dimensiones NxT, cumpliéndose la condición de que los elementos de estas dos matrices siempre sean mayores o iguales a cero. Habitualmente los valores de N son tales que se cumple que FxN+NxT<<FxT, de este modo se reduce el tamaño de las matrices a tratar.

$$X \approx V = W * H \tag{6}$$

Las matrices W y H en las que se descompone la señal de entrada también reciben el nombre de matriz de bases y matriz de ganancias o activaciones, respectivamente. La matriz de bases indica el número de fuentes o bases que hay para cada frecuencia, mientras que la matriz de activaciones indica en qué instante temporal está activa cada base.

Para una señal sencilla como pueden ser cuatro tonos consecutivos, se va a calcular su espectrograma X y su descomposición en las matrices W y H.

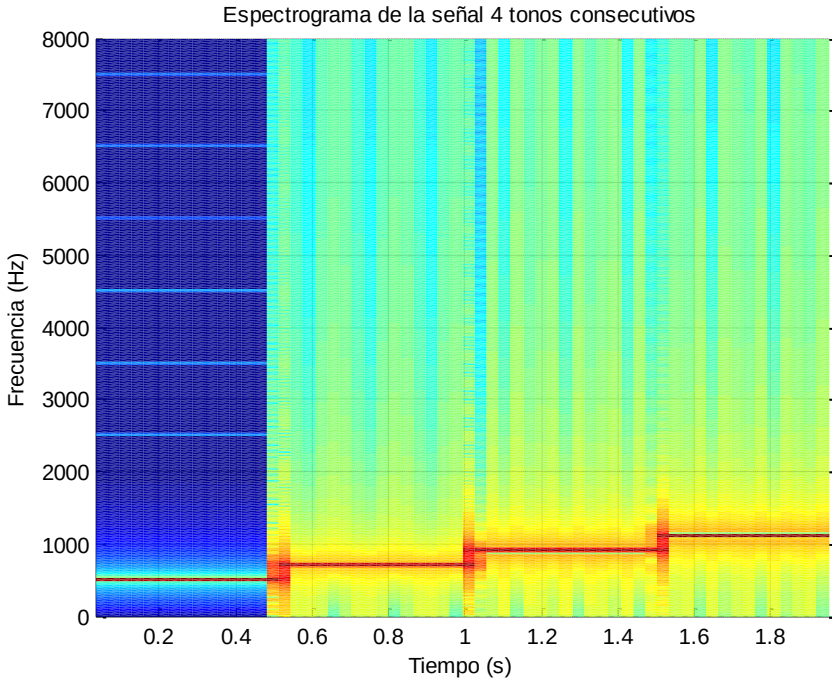

Figura 18: Espectrograma de cuatro tonos consecutivos.

Como se puede observar, la señal está formada por cuatro tonos de duración 0,5 segundos cada uno. Las líneas de color rojo oscuro horizontales son las componentes de cada tono y muestran que frecuencia tiene cada uno y el instante temporal en el que se producen:

- El primer tono tiene una frecuencia de 500 Hz y se produce en el intervalo de 0 a 0,5 segundos.
- El segundo tono tiene una frecuencia de 700 Hz y se produce en el intervalo de 0,5 a 1 segundos.
- El tercer tono tiene una frecuencia de 900 Hz y se produce en el intervalo de 1 a 1,5 segundos.
- El tercer tono tiene una frecuencia de 1100 Hz y se produce en el intervalo de 1,5 a 2 segundos.

A continuación se calculan las matrices W y H mediante unas reglas de actualización que se explicarán más adelante.

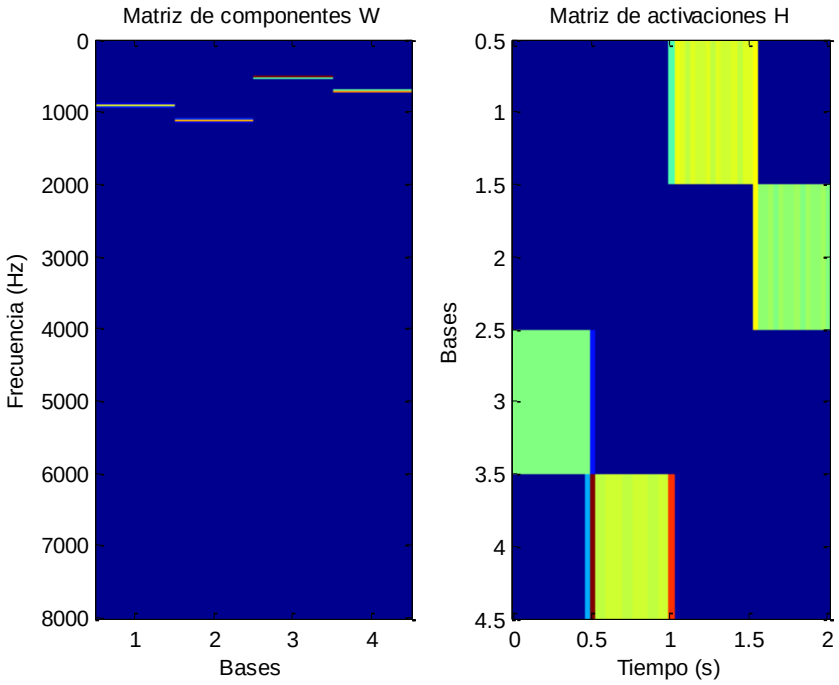

Figura 19: Separación de la señal en sus matrices W y H.

Se puede observar claramente que el espectrograma X ha sido descompuesto en cuatro bases, una para cada tono. En la matriz de bases W se observa que la primera base se corresponde con el tono de 900 Hz, la segunda base se corresponde con el tono de 1100 Hz, la tercera base se corresponde con el tono de 500 Hz y por último, la cuarta base se corresponde con el tono de 700 Hz. Mientras que si se analiza la matriz de activaciones H, se tiene la información de cuándo están activas cada una de las bases. Por tanto la base número 3, que se corresponde con el primer tono, debería de estar activa en el intervalo

de 0 a 0,5 segundos y efectivamente, en la matriz H se comprueba que la tercera base está activa en ese período de tiempo. Con el resto de bases tiene que ocurrir lo mismo, la base número 4 que es la que se corresponde con el segundo tono (700 Hz), se ve que está activa en el intervalo de 0,5 a 1 segundos. La base número 1, la correspondiente al tercer tono (900 Hz), se comprueba que está activa en el intervalo de 1 a 1,5 segundos. Para finalizar, la base número 2 (tono de 1100 Hz) también aparece activa en su intervalo de tiempo correcto (de 1,5 a 2 segundos).

Gracias a este ejemplo se ha demostrado la gran cantidad información que se puede obtener de la descomposición de un espectrograma con esta técnica.

Dependiendo de los procesos que se lleven a cabo para realizar la separación, los algoritmos NMF se dividen en tres [22]:

- Método supervisado: en este método existe una fase de entrenamiento y una fase de separación. En la fase de entrenamiento se entrenan los patrones espectrales de todas las fuentes que forman la señal de entrada, es decir, se calcula la matriz de bases W. Una vez que se tiene la matriz W, en la fase de separación solo es necesario calcular la matriz de activaciones H.

- Método semisupervisado: al igual que en el método anterior, existe una fase de entrenamiento y otra de separación. En este caso, se entrena la matriz de bases de un único tipo de sonido (por ejemplo los sonidos percusivos únicamente). De nuevo en la fase de separación hay que calcular la matriz de activaciones H, pero además hay que calcular la matriz de componentes de los sonidos que no han sido entrenados.

- Método a ciegas o no supervisado: en esta ocasión no hay fase de entrenamiento, por lo que en la fase de separación no se tendrá ninguna información de los patrones espectrales de los sonidos. La fase de separación se realizará atendiendo a una serie de restricciones para una correcta separación. Estas restricciones pueden ser discontinuidad espectral, discontinuidad temporal o dispersión (sparse).

Para realizar la separación del espectrograma X de la señal de entrada x en las dos matrices correspondientes W y H, es necesario definir unas funciones de coste [22] [16] que permitan cuantificar la calidad de la separación. Dichas funciones de coste pueden ser construidas en base a la medida de la distancia entre dos matrices no negativas. La función más sencilla es la distancia Euclídea:

$$d_{EUC}(x|y) = \frac{1}{2}(x-y)^2 \quad (7)$$

Claramente se puede observar que esta función está delimitada por el 0, ya que el mínimo valor que puede adoptar es 0 y se produciría cuando x=y.

Otra función de coste muy utilizada es la de Kullback-Leibler, que se calcula con la siguiente ecuación:

$$d_{KL}(x|y) = x * \log\frac{x}{y} - x + y \quad (8)$$

Esta función también es positiva y al igual que en la función de la distancia Euclídea, cuando las dos matrices son iguales la ecuación vale 0. Pero en este caso no se le puede denominar distancia, ya que no existe simetría en x e y, a esta medida se le denomina divergencia desde x hasta y.

Existe una tercera función que es también ampliamente utilizada, es la función de coste de Itakura-Saito:

$$d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1 \quad (9)$$

Esta función de coste fue definida a partir de la estimación de máxima probabilidad del espectro de la voz con modelado autorregresivo. Esta función mide el parecido entre dos espectros. Su característica principal es la invarianza de escala, es decir, a las componentes de baja energía en x se les asigna el mismo peso relativo que a las que poseen una energía superior. Esta característica es muy importante para los espectros de las señales de audio, en los que el rango dinámico de la señal es muy amplio.

A continuación, se muestra una gráfica con las tres distancias para x=1, en la que se observa que la distancia Euclídea es convexa para todo $\mathbb{R}$, la divergencia de Kullback-Leibler es una línea recta con una pendiente constante, y por último, la distancia de Itakura-Saito para valores de y superiores a 1 en valor absoluto permanece constante en torno a cero, mientras que para valores inferiores cae asintóticamente a 0 (valores entre -1 y 0) o crece asintóticamente a 0 (valores entre 0 y 1). Hay que tener en cuenta que para la

realización de la gráfica no se ha cogido el valor y=0, ya que esto produciría una asíntota creciente hacia infinito a la izquierda y a la derecha del 0, no pudiéndose apreciar el comportamiento real de las funciones de coste.

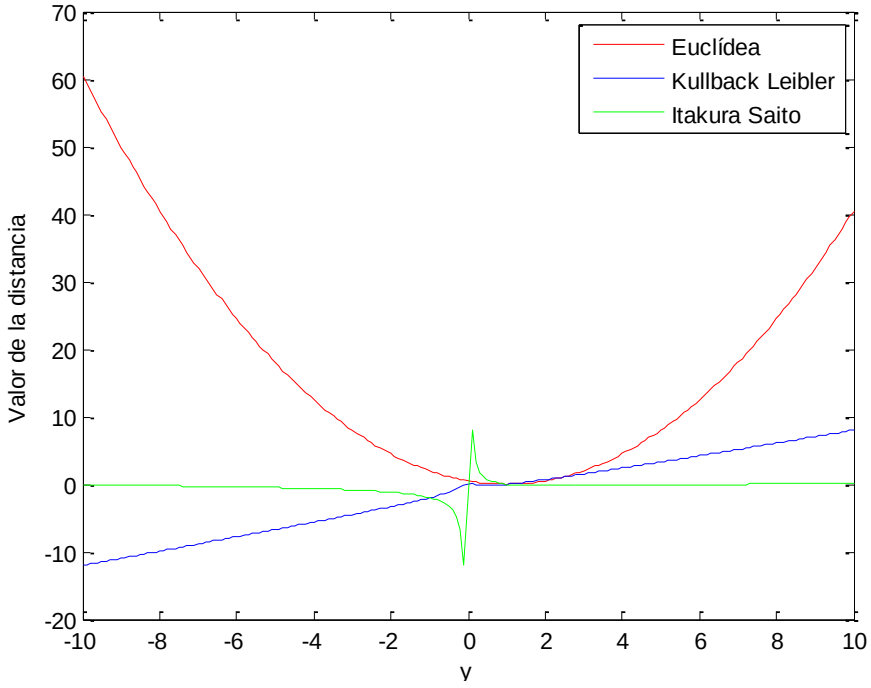

Figura 20: Comportamiento de las tres funciones de coste para x=1.

Una vez que se han definido las funciones de coste, existe el problema de su minimización, ya que la separación será mejor cuanto menores sean los valores de dichas funciones. Para ello se establecen dos problemas [23]:

- **Problema 1:** minimizar la ecuación de la distancia Euclídea con respecto a las matrices W y H, teniendo en cuenta la restricción de que los elementos de estas matrices siempre tienen que ser iguales o mayores que 0. La ecuación (distancia Euclídea) aplicada a la ecuación (separación de X) quedaría de esta forma:

$$d_{EUC}(X|V) = \frac{1}{2}(X - V)^2 = \frac{1}{2}(X - W * H)^2 \qquad (10)$$

- **Problema 2:** minimizar la ecuación de la divergencia de Kullback-Leibler con respecto a las matrices W y H, teniendo en cuenta la misma restricción que en problema anterior, los elementos de estas matrices siempre tienen que ser iguales o mayores que 0. La ecuación 6 aplicada a la ecuación 8 quedaría de la siguiente forma:

$$d_{KL}(X|V) = X * \log\frac{X}{V} - X + V = X * \log\frac{X}{W * H} - X + W * H \quad (11)$$

Tanto la distancia Euclídea como la divergencia de Kullback-Leibler son convexas en W o en H, pero no pueden ser convexas en las dos a la vez. Por lo tanto, no tiene sentido resolver el problema con la finalidad de encontrar un mínimo global, ya que esta tarea sería imposible. Entonces el objetivo es encontrar una técnica de optimización que sea capaz de conseguir mínimos locales.

Para resolver los dos problemas se han definido unas reglas de actualización que cumplen un compromiso entre la velocidad y la facilidad de implementación. Las reglas son las siguientes [20]:

- **Teorema 1:** la distancia Euclídea es no creciente si se aplican las siguientes reglas de actualización:

$$H_{a\mu} \leftarrow H_{a\mu}\frac{(W^T X)_{a\mu}}{(W^T WH)_{a\mu}} \quad (12)$$

$$W_{ia} \leftarrow W_{ia}\frac{(XH^T)_{ia}}{(WHH^T)_{ia}} \quad (13)$$

La distancia Euclídea es invariante bajo estas actualizaciones si y sólo si W y H son un punto estacionario de la distancia.

- **Teorema 2:** la divergencia Kullback-Leibler es no creciente si se aplican las siguientes reglas de actualización:

$$H_{a\mu} \leftarrow H_{a\mu}\frac{\sum_i W_{ia}X_{i\mu}/\,(WH)_{i\mu}}{\sum_k W_{ka}} \quad (14)$$

$$W_{ia} \leftarrow W_{ia}\frac{\sum_\mu H_{a\mu}X_{i\mu}/\,(WH)_{i\mu}}{\sum_u H_{au}} \quad (15)$$

La divergencia de Kullback-Leibler también es invariante bajo estas actualizaciones si y sólo si W y H son un punto estacionario de la distancia.

Para simplificar las reglas de actualización, estas han sido definidas para las tres funciones de coste en una sola ecuación, quedando las reglas de la siguiente manera:

$$H \leftarrow H.* \frac{W^T * (WH)^{(\beta-2)} * X}{W^T * (WH)^{(\beta-1)}} \tag{16}$$

$$W \leftarrow W.* \frac{\left((WH)^{(\beta-2)} * X\right) * H^T}{(WH)^{(\beta-1)} * H^T} \tag{17}$$

Como se puede observar, ambas ecuaciones dependen del parámetro β, por lo que dándole el valor correspondiente se pueden obtener las reglas de actualización para cada tipo de función de coste:

- Para la distancia Euclídea β vale 2, quedando las ecuaciones de esta manera:

$$H \leftarrow H.* \frac{W^T * I * X}{W^T * (W * H)} \tag{18}$$

$$W \leftarrow W.* \frac{(I * X) * H^T}{(W * H) * H^T} \tag{19}$$

Donde I es la matriz identidad de las mismas dimensiones que X.

- Para la distancia Kullback-Leibler β vale 1, quedando las ecuaciones de esta manera:

$$H \leftarrow H.* \frac{W^T * (W * H) * X}{W^T * 1} \tag{20}$$

$$W \leftarrow W.* \frac{((W * H)^{-1} * X) * H^T}{1 * H^T} \tag{21}$$

- Para la distancia Itakura Saito β vale 0, quedando las ecuaciones de esta manera:

$$H \leftarrow H.* \frac{W^T * (W * H)^{-2} * X}{W^T * (W * H)^{-1}} \tag{22}$$

$$W \leftarrow W.* \frac{((W * H)^{-2} * X) * H^T}{(W * H)^{-1} * H^T} \tag{23}$$

En las ecuaciones 20 y 21, 1 es una matriz de las mismas dimensiones que la matriz X y con todos sus elementos con valor 1. Todas las divisiones de matrices son divisiones elemento a elemento, al igual que el producto ($.*$). Mientras que el producto ($*$) es el producto de matrices.

Una vez definidas las reglas de actualización, ya se podrá separar la señal de entrada X, en las dos matrices W y H. Para saber qué bases de la matriz W se corresponden a cada sonido (armónico, percusivo o vocal) se pueden usar varios criterios:

- **Umbralización en frecuencia [12]:** también se denomina como discontinuidad espectral. Este criterio se utiliza para detectar aquellas componentes que tienen un carácter armónico de las que no lo tienen (sonidos percusivos). El fundamento teórico en el que se basa es que los sonidos armónicos al representarlos en un espectrograma, aparecen como discontinuidades en frecuencia, mientras que son continuos en el tiempo.
- **Umbralización en el tiempo [12]:** también se denomina como discontinuidad temporal. Este criterio se utiliza para detectar aquellas componentes que tienen un carácter percusivo de las que no lo tienen (sonidos armónicos). El fundamento teórico en el que se basa es que los sonidos percusivos al representarlos en un espectrograma, aparecen como discontinuidades en el tiempo mientras que son continuos en frecuencia.
- **Sparse [24]:** este método se basa en el principio de que las matrices W y H tienen muchos elementos nulos o cercanos a 0. Este método 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. Matemáticamente consiste en añadir un parámetro en la regla de actualización de la matriz H:

$$H \leftarrow H.* \frac{W^T * (WH)^{(\beta-2)} * X}{W^T * (WH)^{(\beta-1)} + \lambda} \quad (24)$$

Donde λ es el parámetro de sparse, pudiendo tomar valores desde 0 hasta 1.

3.3.1. NMF-KL-FT [12]

Este método es una particularización del conjunto de algoritmos matemáticos conocidos como NMF. Como se puede ver la nomenclatura de este método tiene tres partes:

- **NMF:** estas siglas indican que se trata de un método basado en la factorización de matrices no negativas.
- **KL:** estas siglas indican que se van a utilizar las reglas de actualización de Kullback-Leibler y por consiguiente, la función de coste (divergencia) del mismo nombre.
- **FT:** estas siglas indican que para discernir que bases se corresponden con cada tipo de sonido (armónicos, percusivos y vocales) primero se va a aplicar umbralización en frecuencia para conocer las bases armónicas y posteriormente se va a aplicar umbralización en el tiempo para conocer las bases percusivas.

En este Trabajo Fin de Grado se va a evaluar y a optimizar este método en concreto, el cual está publicado en [12].

4. IMPLEMENTACIÓN

En este apartado se va a explicar el algoritmo en el que se basa este Trabajo Fin de Grado, detallando minuciosamente su funcionamiento y las posibles variantes de cada una de sus etapas. De tal forma que se cumplan los objetivos propuestos en el apartado 2.

4.1. Descripción del Trabajo Fin de Grado

Como ya se ha anunciado en el apartado de los objetivos, la finalidad de este Trabajo Fin de Grado es la separación de una señal de audio monocanal en tres señales distintas, es decir, separar una señal en la que la música y la voz este mezclada, en una señal que contenga a los sonidos armónicos, en otra señal que contenga a los sonidos percusivos y en otra señal que contenga a los sonidos provenientes de la voz. Para ello en primer lugar se va a realizar una etapa de optimización del algoritmo, en la que se realizará la separación de 89 fragmentos de la base de datos (en total son 1000 fragmentos) barriendo todos los parámetros que afectan a la calidad de la separación (número de bases, tamaño de la ventana larga, tamaño de la ventana corta, umbral en frecuencia y umbral en el tiempo). Una vez realizado el proceso de optimización, se podrá saber cuál es la combinación de parámetros con la que mejores resultados se obtienen. Después se realizará la separación de los 911 fragmentos restantes con la combinación óptima, para comprobar que la calidad que se obtiene con dicha combinación es la que se ha calculado en la etapa de optimización.

Tanto para la etapa de optimización como para la etapa de separación posterior, el algoritmo que se va a utilizar es el mismo y se basa en las técnicas NMF. En el apartado anterior se ha explicado que en esta técnica todas las operaciones que se realizan son operaciones matriciales y dichas matrices tienen que estar en el dominio de la frecuencia. Por lo tanto hay que realizar los espectrogramas de las señales que están en el dominio del tiempo. El enventanado necesario puede ser de mayor o menor tamaño, el cual se divide en tamaño largo o tamaño corto. Dependiendo de la combinación que se haga de estos tamaños, se tiene el método Largocorto y el método Cortolargo. En ambos métodos, dependiendo de si se realizan los espectrogramas con ventana larga o con ventana corta, hay que realizar un proceso de umbralización en frecuencia o umbralización en el tiempo, respectivamente. Todas las etapas del algoritmo se explicarán en los siguientes apartados detalladamente.

4.2. Enventanado

Una de las partes más importantes del algoritmo que se va implementar es la longitud de la ventana que se aplica al realizar la transformada discreta de Fourier (FFT). Esto es así porque la longitud de la ventana afecta a la resolución espectral y de manera inversa a la resolución temporal. La resolución espectral se calcula de la siguiente manera:

$$Resolución\ espectral = f_s * \frac{4}{L} \tag{25}$$

Donde f_s es la frecuencia de muestreo y *L* es la longitud de la ventana. Esta ecuación se cumple para la ventana triangular, para a ventana de Hanning y para la ventana de Hamming. Sin embargo, si se utiliza una ventana rectangular la resolución espectral es la mitad de la ecuación 25. Por ejemplo para una longitud de ventana (no rectangular) de 1024 muestras se tiene una resolución espectral de:

$$Resolucion\ espectral = f_s \frac{4}{L} = 16000\ Hz \frac{4}{1024} = 62{,}5\ Hz \tag{26}$$

Esto quiere decir que en frecuencia se va a tener información de lo que ocurre cada 62,5 Hz, por tanto si hay componentes que estén separadas una frecuencia menor, se representarán como una única componente. La resolución en el tiempo es el tamaño de la ventana, que para el ejemplo de la ecuación 26 equivale a 64 ms. Lo que significa que solo se tendrán en cuenta las variaciones de señal cada 64 ms, por lo que si la señal varía en instantes de tiempo inferiores, no se apreciará en el espectrograma.

En la siguiente figura se pueden ver los 4 tipos de ventana más utilizados.

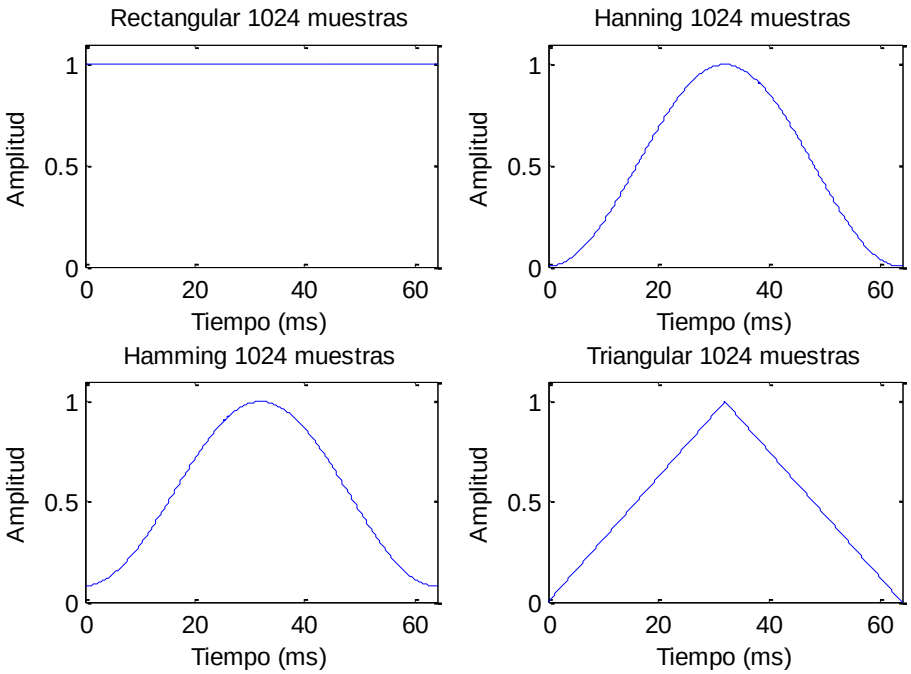

Figura 21: Los 4 tipos de ventana más utilizados (Rectangular, Hanning, Hamming y Triangular).

A continuación, en la figura 22 se muestra el módulo de la transformada de Fourier en decibelios de cada una de las ventanas. En esta figura se puede ver una característica fundamental que permite saber qué ventana es la que ofrece mejores prestaciones. Esta característica se conoce como grado de fugas y mide la diferencia en decibelios entre el lóbulo principal y el lóbulo secundario de la ventana. Fijándose en la figura 22 se observa que la ventana con un menor grado de fugas es la rectangular, cuya diferencia es de aproximadamente 13 dB. La siguiente ventana con menor grado de fugas es la triangular, con una diferencia de 26 dB aproximadamente. Después está la ventana de Hanning, con una diferencia de 32 dB aproximadamente. Y por último, está la ventana de Hamming que tiene el mayor grado de fugas, con 42 dB aproximadamente. Por eso se ha escogido esta ventana para realizar los espectrogramas del algoritmo propuesto, ya que cuanto mayor sea el grado de fugas, menor será la interferencia que experimenta la componente espectral que se está enventanando debido a otras componentes espectrales distintas.

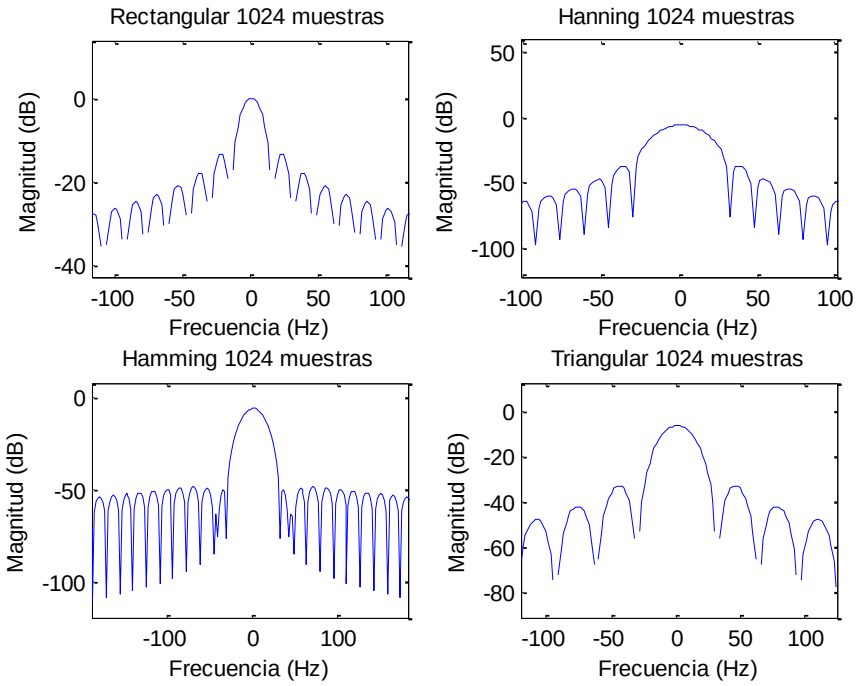

Figura 22: Módulos de las transformadas de Fourier de las 4 ventanas más utilizadas.

En las dos figuras siguientes (figuras 23 y 24) se muestra gráficamente lo que se ha demostrado numéricamente en la ecuación 26, es decir, para tamaños de ventana grandes la resolución espectral es mejor (muestras equiespaciadas una frecuencia menor), mientras que la resolución temporal es peor. Por el contrario si se utiliza una ventana más pequeña, la resolución espectral empeora (muestras equiespaciadas una frecuencia mayor), mientras que la resolución temporal mejora, teniéndose información de las variaciones temporales de la señal de forma precisa.

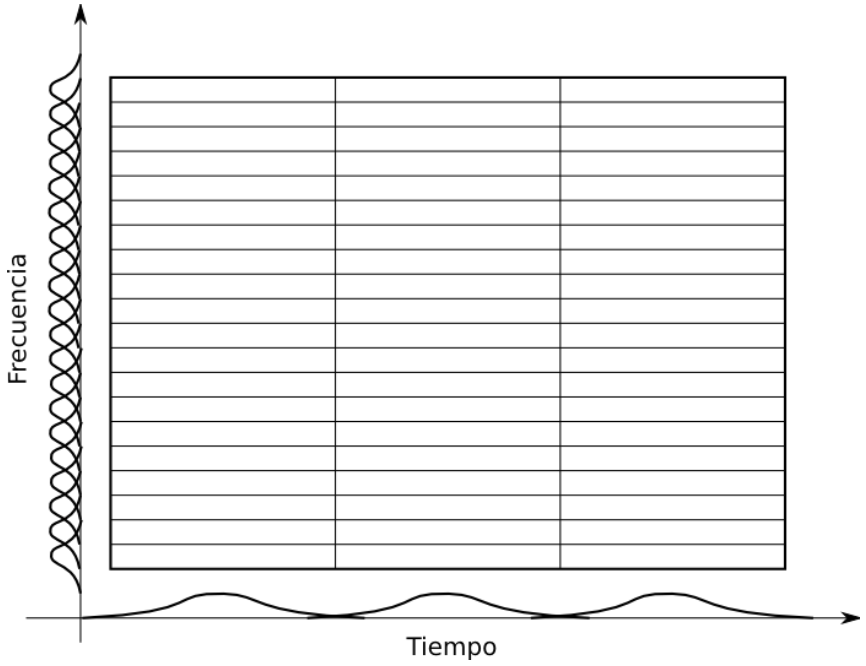

Figura 23: Comportamiento del espectrograma con tamaños de ventana grandes.

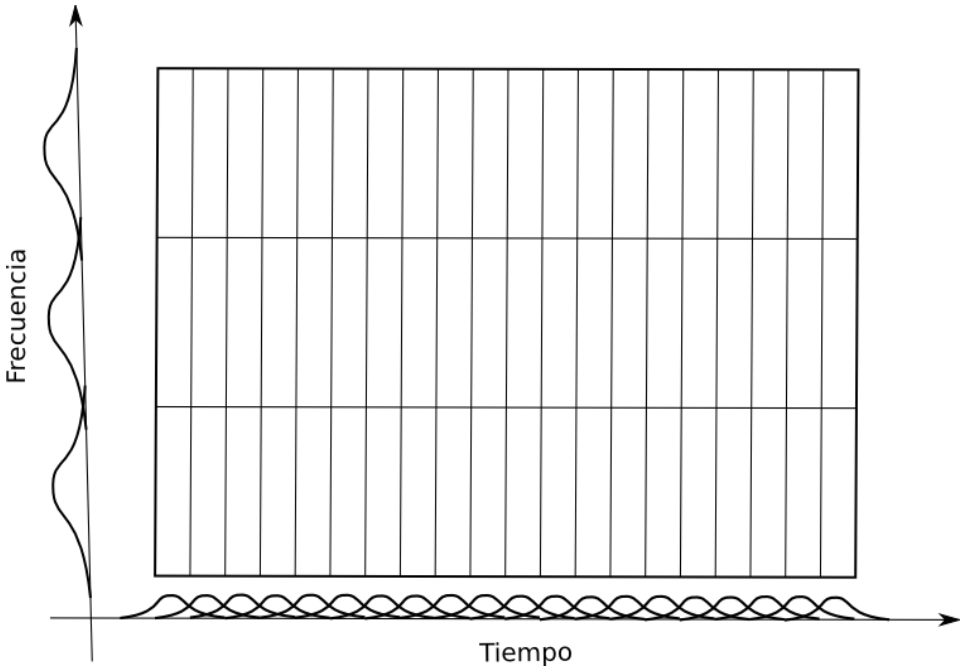

Figura 24: Comportamiento del espectrograma con tamaños de ventana pequeños.

Por tanto dependiendo de la característica que se quiera resaltar, se debe coger un tamaño de ventana u otro. Para mostrar las componentes armónicas hay que elegir una ventana con mejor resolución en frecuencia (tamaño de ventana grande), mientras que si se quiere mostrar como varía la señal en instantes muy cortos de tiempo hay que elegir una ventana con mejor resolución en el tiempo (tamaño de ventana pequeño).

A continuación se van a explicar las consecuencias de los distintos tamaños de enventanado, para ello se va a realizar espectrogramas con cada uno de los tamaños posibles de las ventanas, siempre utilizando la misma señal original (figura 25). Dicha señal está formada por aproximadamente 6 segundos de notas de piano (sonidos armónicos), 3 segundos de golpes de tambor (sonidos percusivos) y por 12 segundos de un fragmento de la canción “Hello” de Adele cantada a capela (sonidos vocales). La representación temporal de la señal se puede ver en la siguiente figura.

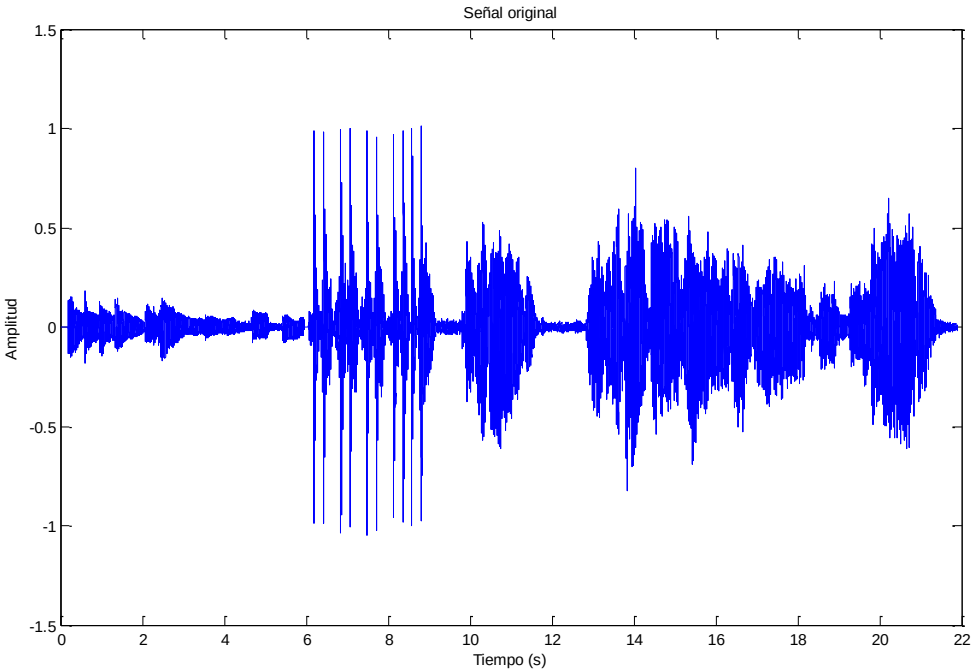

Figura 25: Señal original con la que se va a trabajar.

4.2.1. Ventana corta

Los tamaños de ventana que se han definido como cortos son 8 ms, 16 ms, 32 ms y 64 ms, que equivalen a 128 muestras, 256 muestras, 512 muestras y 1024 muestras, respectivamente.

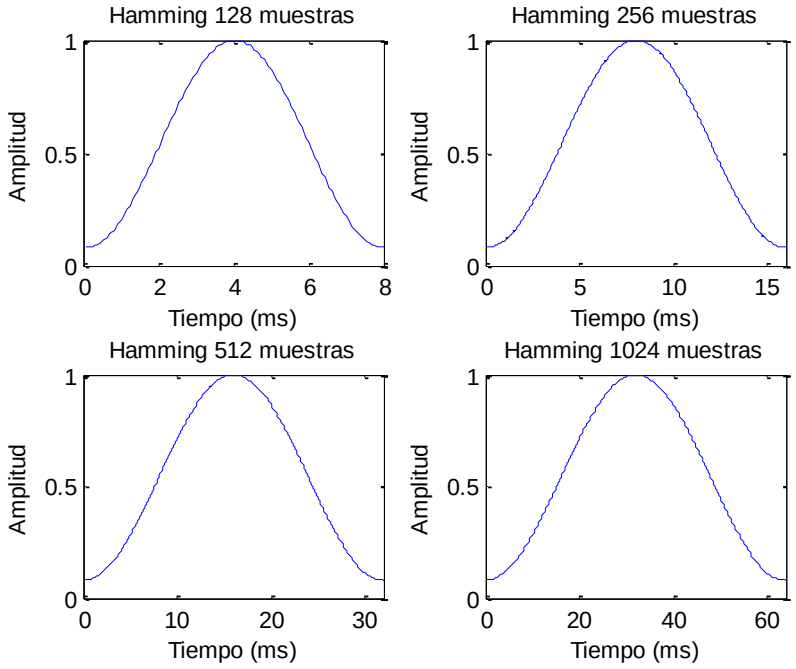

Figura 26: Representación de las ventanas cortas en el tiempo.

Como se puede ver en esta figura, la forma de las ventanas es la misma para los cuatro tamaños, en lo único que se diferencian entre sí es en el número de muestras con el que se representa cada una, y por tanto en el intervalo temporal que ocupan.

En la figura 27 se puede ver la representación espectral de cada una de las ventanas tras realizar la FFT con una longitud igual al doble del tamaño de cada ventana, para obtener una mejor interpolación entre sus puntos al realizar las gráficas. Como se puede observar, conforme va aumentando el tamaño de la ventana, la anchura del lóbulo principal va disminuyendo. Esto se debe a que se cumple la ecuación 25, en la que la anchura del lóbulo principal es inversamente proporcional a la longitud de la ventana.

En esta misma figura se puede ver que el grado de fugas para las cuatro ventanas es el mismo, aproximadamente unos 42 dB. Esto es así porque es una característica intrínseca de la ventana de Hamming y es invariante independientemente de su tamaño.

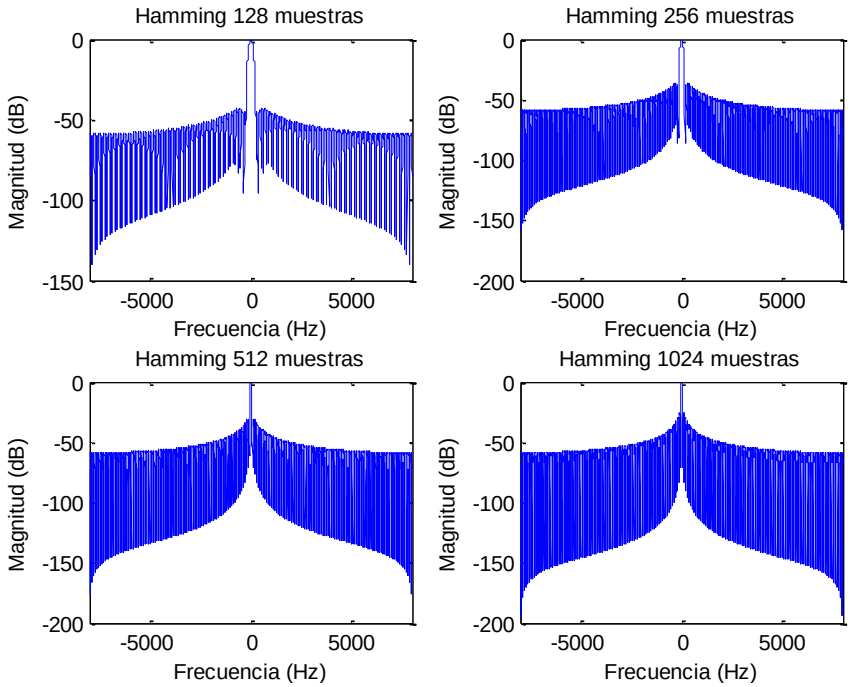

Figura 27: Representación del módulo de la FFT de cada ventana.

En la figura 28 se ha realizado un zoom de los lóbulos principales de cada ventana para mostrar la anchura que ocupa cada uno de ellos.

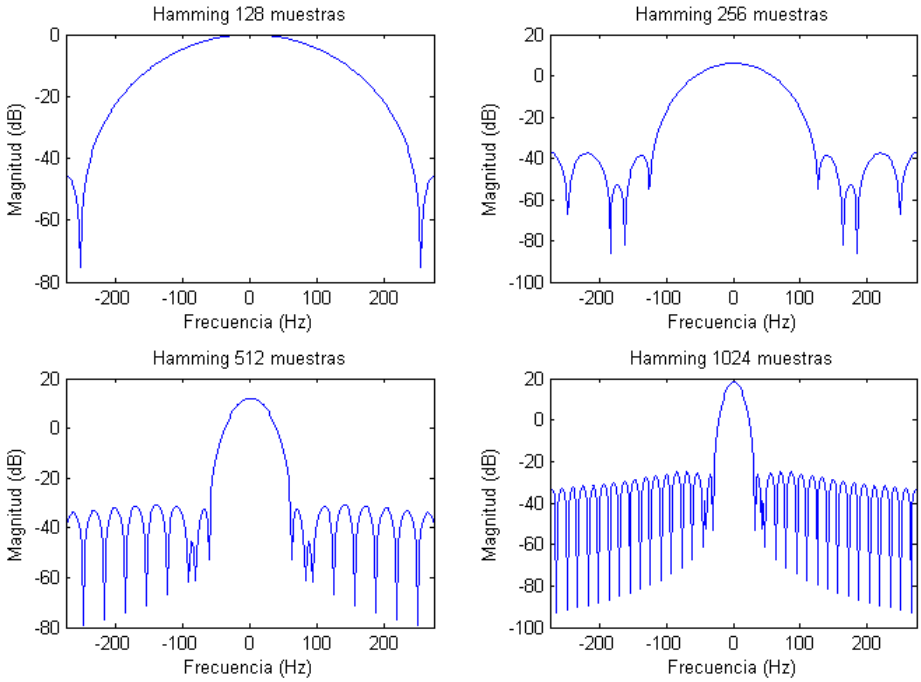

Figura 28: Zoom de los lóbulos principales de cada ventana.

Siguiendo la ecuación 25, la anchura del lóbulo principal para una longitud de 128 muestras es de 500 Hz, para 256 muestras es de 250 Hz, para 512 muestras es de 125 Hz y para 1024 muestras es de 62,5 Hz. y en la figura 59 se comprueba que esto es cierto.

Primero se va a realizar el espectrograma con la ventana más pequeña de las 8 que se han propuesto (tanto cortas como largas), la ventana que ocupa 8 ms, lo que equivale a 128 muestras. Como se ha explicado en el apartado anterior, para realizar el espectrograma se necesita además de la ventana y su tamaño, la frecuencia de muestreo, el solapamiento y la longitud de los segmentos en los que se divide la señal. La frecuencia de muestreo es de 16 KHz, el solapamiento es del 50% (la longitud de la ventana entre 2) y la longitud de los segmentos es la misma que la longitud de la ventana, ya que esta es una potencia de 2.

El resultado se puede ver en la figura 29, donde está representado el espectrograma de la señal original (figura 25) con una resolución en frecuencia de 500 Hz.

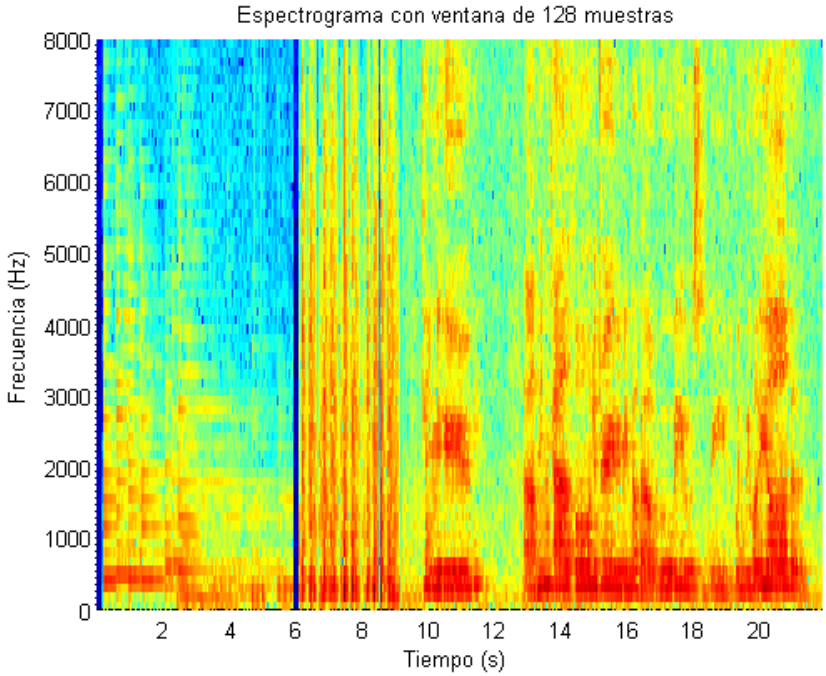

Figura 29: Espectrograma de la señal original con una ventana de 8 ms.

En el apartado de la introducción se han explicado las características de los sonidos armónicos y de los sonidos percusivos. Los sonidos armónicos muestran continuidad en el tiempo mientras que en frecuencia aparecen discontinuidades, debido a las componentes armónicas de su componente fundamental. Por el contrario, los sonidos percusivos se denominan sonidos de banda ancha, porque son continuos en frecuencia y además varían de forma muy rápida en el tiempo.

Estas características se pueden ver en la figura 29, donde al tener dicha ventana una resolución espectral muy mala, las componentes armónicas de los sonidos de piano no se aprecian con claridad. Por el contrario en la parte de los golpes de tambor se ve que en frecuencia aparecen de forma continua, pero se puede distinguir con claridad en que instante de tiempo se produce cada golpe.

Si se duplica la longitud de la ventana se tiene un tamaño de 256 muestras (16 ms), lo que equivale a una resolución en frecuencia de 250 Hz, la mitad que para la ventana anterior. El resultado se puede ver en la siguiente figura.

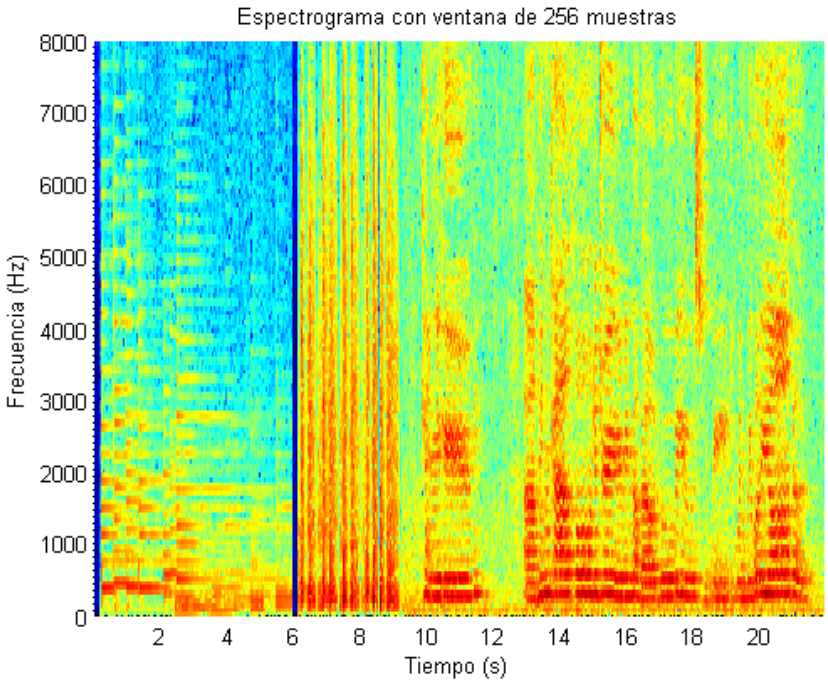

Figura 30: Espectrograma de la señal original con una ventana de 16 ms.

En este caso ya se puede ver cómo se van definiendo las componentes armónicas de las notas de piano, aunque todavía la resolución en frecuencia es bastante mala. En cuanto a los golpes de tambor, se siguen apreciando los instantes temporales en los que se producen.

Si se vuelve a duplicar el tamaño de la ventana, se alcanza un tamaño de 512 muestras (32 ms), lo que equivale a 125 Hz de resolución espectral. El espectrograma resultante se puede ver en la figura 31. En esta figura se observa que las componentes armónicas van apareciendo más claramente conforme mejora la resolución. Los golpes de tambor siguen mostrándose como señales continuas en frecuencia y todavía se pueden apreciar los instantes temporales en los que se producen cada uno.

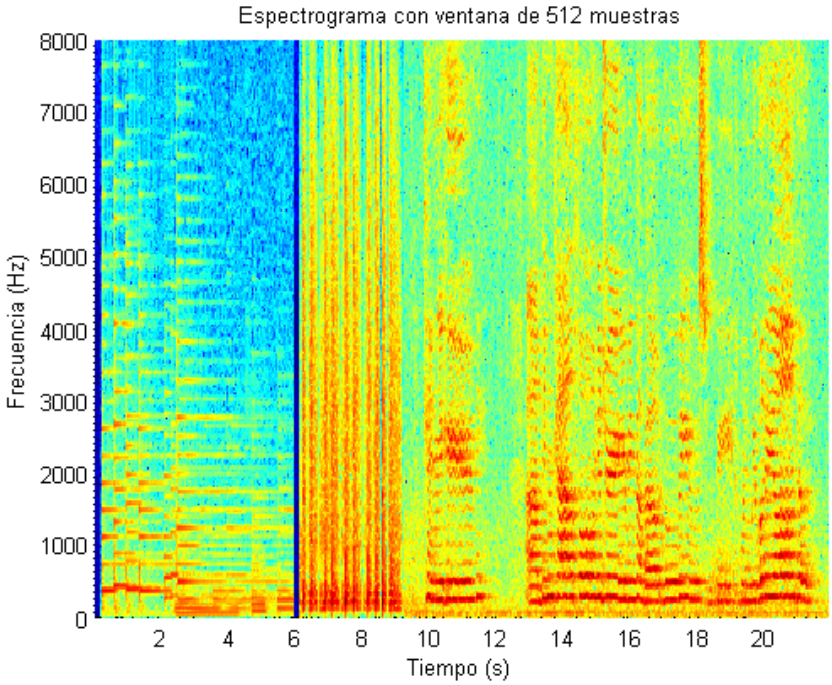

Figura 31: Espectrograma de la señal original con una ventana de 32 ms.

Si se vuelve a duplicar el tamaño de la ventana, se pasa a una longitud de 1024 muestras (64 ms). Teniendo en este caso una resolución espectral de 62,5 Hz. Esta ventana tiene la peculiaridad de que es común en los dos tipos de ventanas, tanto corta como larga, siendo el valor más alto de tamaño de ventana corta y el valor más pequeño de tamaño de ventana larga. Por tanto la figura 32 y su explicación serán válidas para ambos tipos de enventanado.

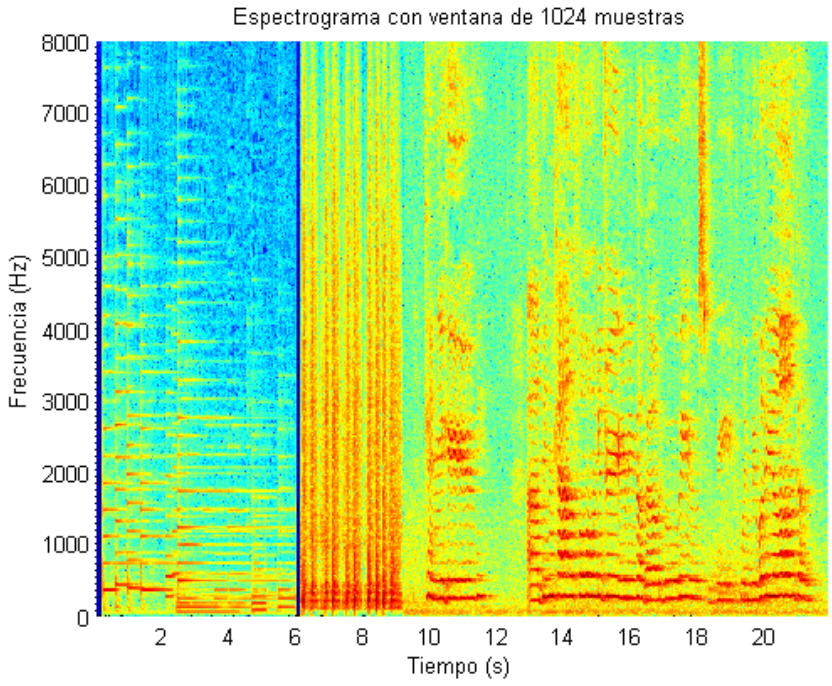

Figura 32: Espectrograma de la señal original con una ventana de 64 ms.

Nuevamente las componentes armónicas se aprecian cada vez de manera más clara y ocupan menos espacio en el eje de frecuencia, esto es porque la resolución cada vez es mejor. Los golpes de tambor siguen comportándose de la misma manera, aunque cada vez se van juntando más dichos golpes ya que la ventana se va haciendo más grande.

Después de ver y analizar los espectrogramas con ventana corta se llegan a las siguientes conclusiones sobre el comportamiento de la voz con ventanas que tienen baja resolución espectral.

- Una mala resolución en frecuencia implica que los canales en los que se divide el eje de frecuencia sean más anchos, lo que provoca que las fluctuaciones en la entonación de la voz no se aprecien.
- La energía de cada armónico es recogida por un único canal espectral, apareciendo continuos en el eje temporal como los sonidos armónicos.
- La energía de los fonemas sordos se extiende a lo largo de todo el eje de frecuencia, es decir, muestran continuidad en el eje de frecuencia, comportándose como los sonidos percusivos.
- Al realizar la separación para extraer los sonidos percusivos de la señal, también se toman como sonidos percusivos las componentes que pertenecen a los fonemas sordos de la voz cantada.
- Como se puede ver en el apartado de los resultados, dejar a la voz cantada sin sus fonemas sordos tiene una repercusión casi inapreciable.

4.2.2. Ventana larga

Ahora se va a proceder a realizar los espectrogramas con lo que denominan ventanas largas. Los tamaños son de 64 ms, 128 ms, 256 ms y 512 ms respectivamente. Esto equivale a 1024, 2048, 4096 y 8192 muestras respectivamente.

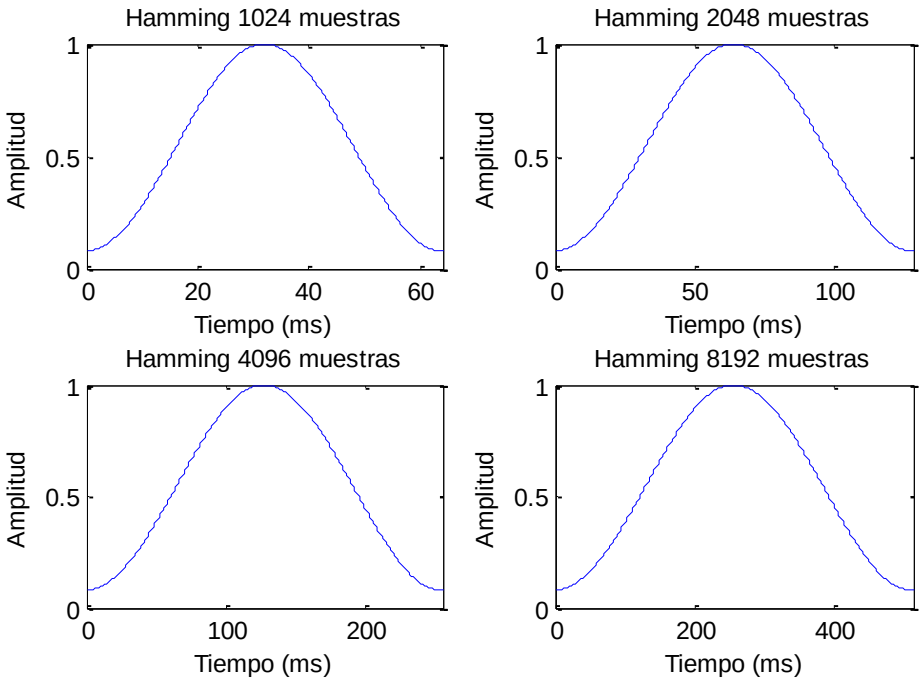

Figura 33: Representación de las ventanas largas en el tiempo.

Al igual que en la figura 26, la forma de las ventanas para los cuatro tipos de tamaños es la misma, la única diferencia entre ellas es su duración.

En la siguiente figura, se representa el módulo de la transformada rápida de Fourier de cada una de las ventanas y al igual que en la figura 27, se observa que el grado de fugas es aproximadamente el mismo para las cuatro y está en torno a los 42 dB. Ya que esa es una característica que permanece invariable independientemente del tamaño de la ventana.

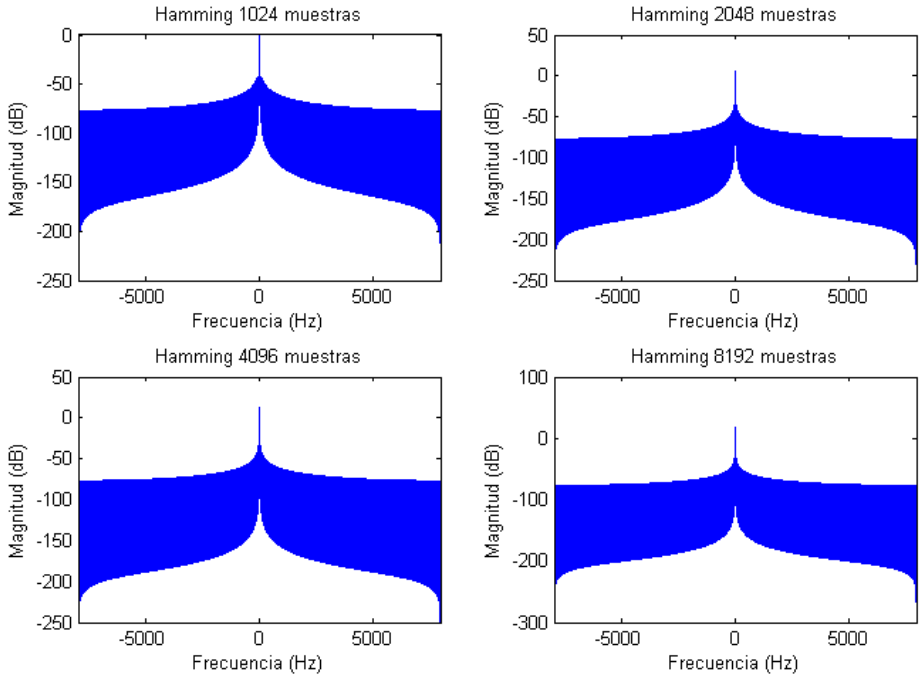

Figura 34: Representación del módulo de la FFT de cada ventana.

En la siguiente figura se va a realizar un zoom de los lóbulos principales de cada ventana para comprobar que se cumple lo que establece la ecuación 25. Según esta ecuación las anchuras de los lóbulos principales son 62,5 Hz, 31,25 Hz, 15,625 Hz y 7,8125 Hz, respectivamente.

Se puede comprobar que la ecuación 25 es totalmente cierta, ya que las anchuras de los lóbulos principales, es decir, la resolución espectral, son exactamente del mismo valor que las anchuras calculadas teóricamente.

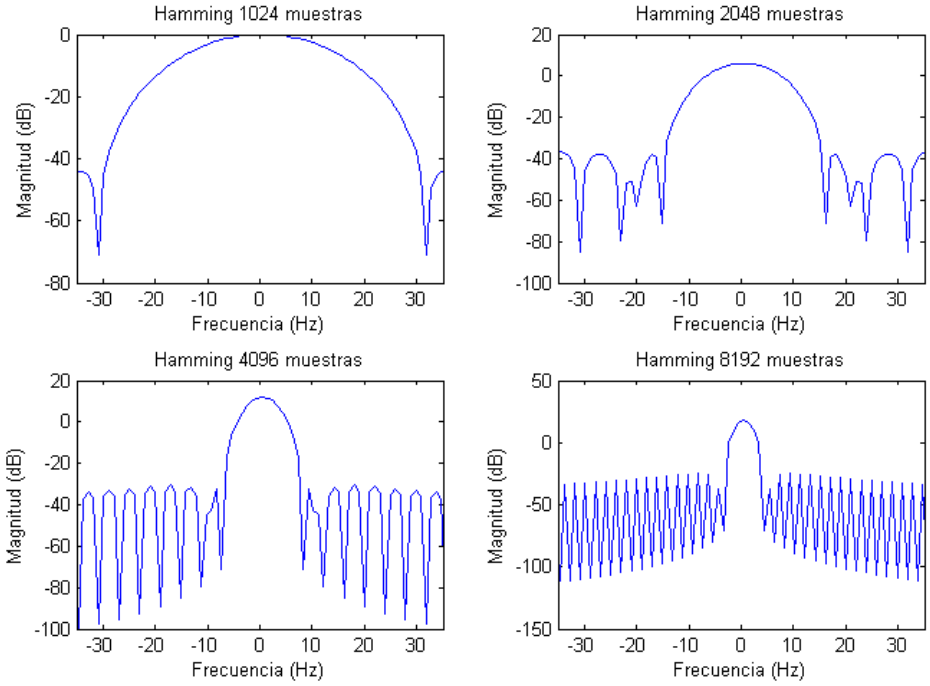

Figura 35: Zoom de los lóbulos principales de cada ventana.

Siguiendo la misma estructura que se ha seguido durante la explicación de la ventana corta, el primer espectrograma que habría que realizar sería con la ventana de 1024 muestras. Pero este espectrograma es común para ambos tipos de enventanado, y como ya se ha realizado en la figura 32, no se va a volver a repetir. Por tanto se va a realizar el espectrograma con el siguiente tamaño, 2048 muestras (128 ms).

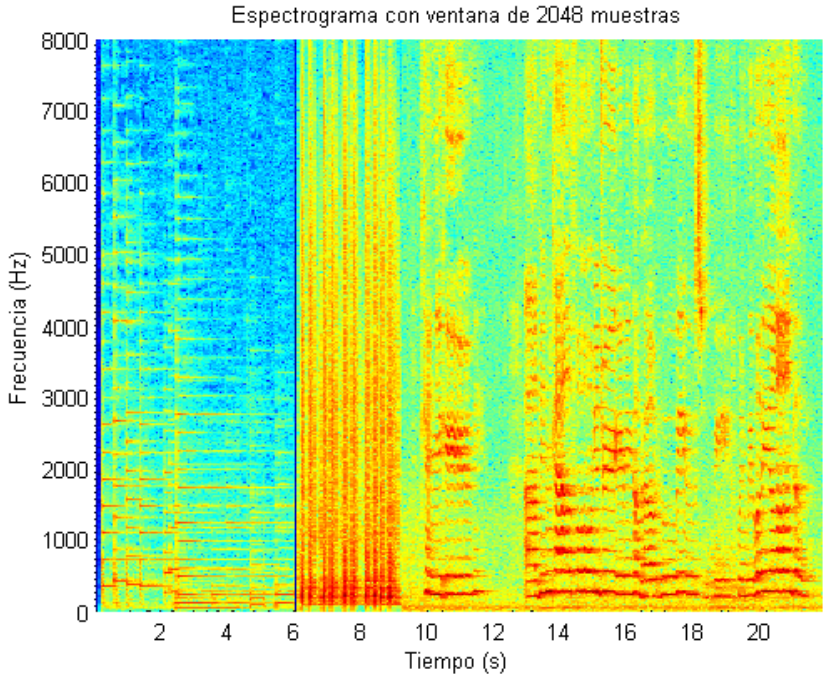

Figura 36: Espectrograma de la señal original con una ventana de 128 ms.

En este caso se tiene una resolución espectral de 31,25 Hz. En la figura anterior se observa como las componentes armónicas siguen siendo cada vez más estrechas, mientras que los golpes de tambor todavía se diferencian en el eje temporal.

A continuación se vuelve a duplicar el tamaño de la ventana hasta las 4096 muestras (256 ms). Esta ventana ofrece una resolución en frecuencia de 15,625 Hz. En la siguiente figura se ve el resultado de realizar dicho espectrograma.

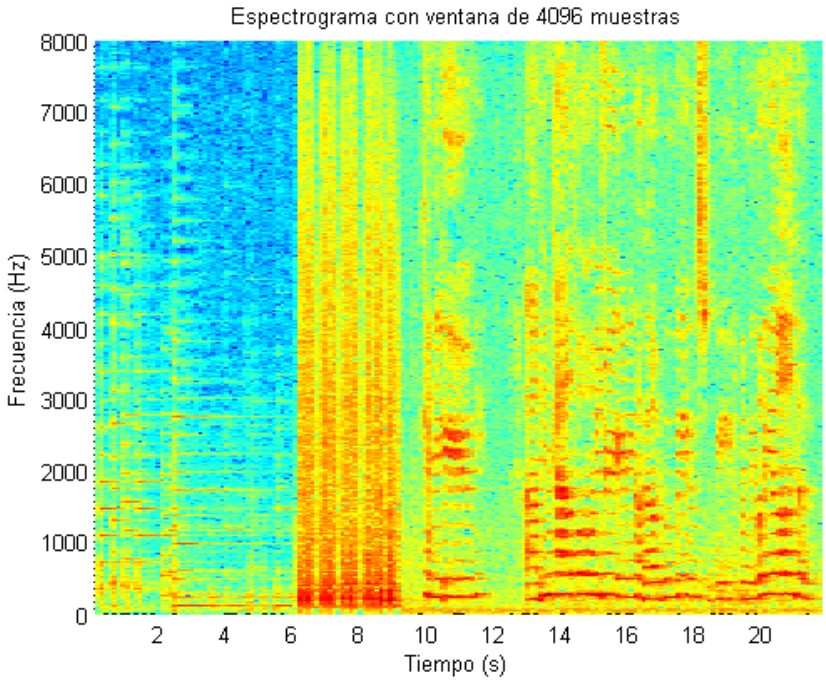

Figura 37: Espectrograma de la señal original con una ventana de 256 ms.

En este caso las componentes armónicas siguen su proceso de estrechamiento en frecuencia, mientras que los golpes de tambor sufren un cambio con respecto a los espectrogramas anteriores. En esta ocasión ya no se aprecian en qué instantes temporales se produce cada uno.

Por último, se realiza el espectrograma con el tamaño de ventana más grande de los 8 que se han utilizado, que es de 512 ms (8192 muestras). Con esta ventana se obtiene una resolución de 7,8125 Hz. En la siguiente figura se ve el resultado de realizar el espectrograma con dicha ventana.

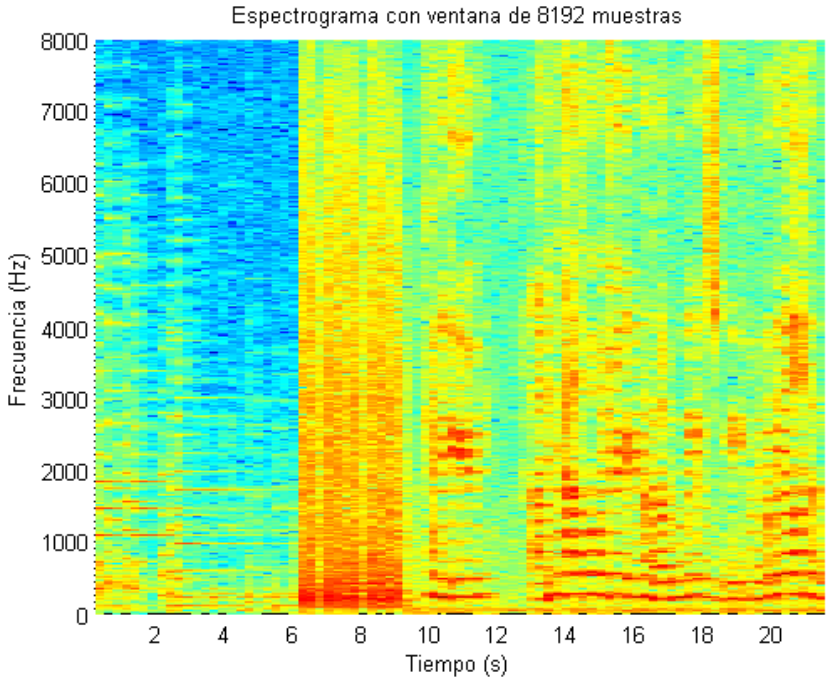

Figura 38: Espectrograma de la señal original con una ventana de 512 ms.

Como se puede ver, las componentes armónicas aparecen muy estrechas, mientras que siguen sin distinguirse los instantes temporales en los que se producen los golpes de tambor.

Tras de analizar los resultados de los espectrogramas realizados con ventana larga, se ha llegado a las siguientes conclusiones acerca del comportamiento de la voz cantada:

- La voz muestra rápidas fluctuaciones en la entonación, esto hace que la energía de cada segmento de voz se extienda a lo largo de los distintos canales en frecuencia. Por lo tanto no se comportan como los sonidos armónicos.
- La voz cantada puede contener fonemas sordos y como la resolución en el tiempo es mala, es muy probable que en la misma ventana se mezclen tanto fragmentos de voz con fonemas sonoros como fragmentos de voz con fonemas sordos. Esto también provoca que la energía se extienda a lo largo del eje de frecuencia.

4.3. Descripción del algoritmo

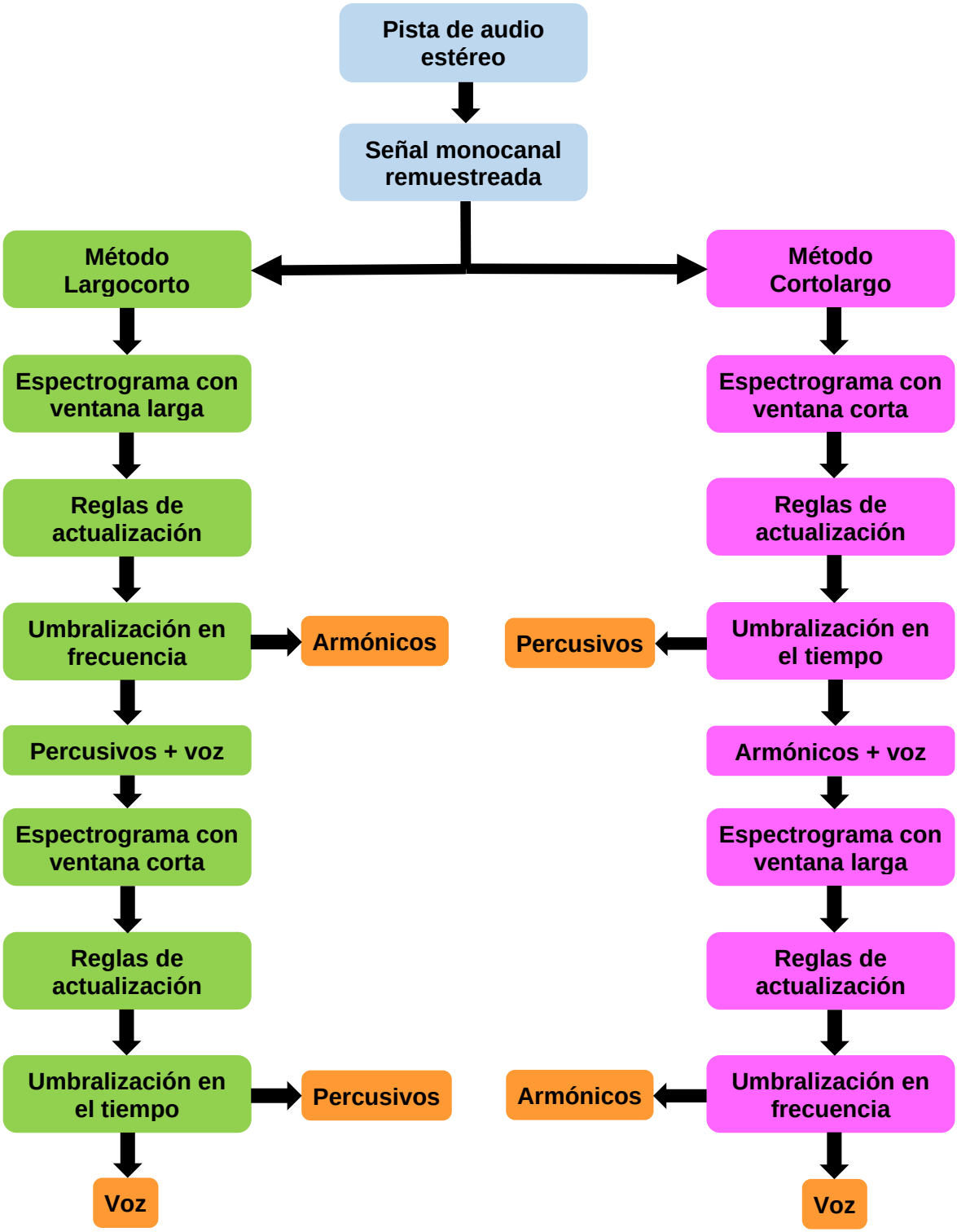

Figura 39: Esquema del algoritmo utilizado.

En la figura anterior se puede ver de forma esquemática todas las etapas que forman el algoritmo que se ha utilizado para la separación de una señal estéreo en sus

correspondientes tres pistas (sonidos armónicos, sonidos percusivos y sonidos producidos por la voz). Dicho algoritmo se basa en las técnicas NMF (que ya han sido explicadas en el apartado 3.3). A continuación se van a explicar cada una de las distintas etapas que se pueden ver en el esquema, detallando su funcionamiento y explicando todos sus parámetros:

Figura 40: Pista de audio estéreo.

1. El primer paso consiste en introducir en el sistema de separación una señal que sea monocanal, ya que como el nombre de este Trabajo Fin de Grado indica, se busca la separación de señales monoaurales. Pero en realidad las señales de la base de datos que se va a utilizar están en estéreo, esto que puede parecer un problema no lo es, ya que solo hay que realizar una tarea adicional, que es pasar la señal de estéreo a mono. Esta tarea se realiza mediante la suma de los dos canales que forman la señal estéreo y su posterior división entre 2. En uno de los canales se encuentra el acompañamiento musical, mientras que en el otro se encuentra la voz cantada sin ningún tipo de sonido instrumental, esta característica permite que se pueda medir la calidad de las separaciones de forma muy precisa con las medidas de calidad objetivas (SDR, SIR y SAR), que serán explicadas en el apartado 5.2. Si por el contrario no se dispusiera de la señal en estéreo, o más concretamente, si no se dispusiera de una pista en la que solo estuviera el acompañamiento musical y de otra pista en la que solo se encontrara la voz cantada, el proceso de separación sería el mismo, pero no se podrían obtener las medidas de calidad objetivas que demuestran con un número cómo de buena es la separación. Entonces dicha calidad solo podría ser comprobada subjetivamente mediante la escucha de cada una de las pistas generadas en el proceso de separación.

Figura 41: Señal monocanal remuestreada.

2. El segundo paso es el muestreo de la señal. Todas las canciones de la base de datos están muestreadas a 16 KHz, lo que quiere decir que en cada segundo de duración de la señal hay 16000 muestras de dicha señal. La razón por la que se ha utilizado este valor de frecuencia de muestreo es porque con este valor se aprecian las características espectrales y temporales propias de los sonidos armónicos percusivos y

vocales. La elección de una frecuencia de muestreo mayor no se traduciría en la aparición de características adicionales y lo único que ocurriría es que la eficiencia del algoritmo disminuiría. Entonces si la canción que se introduce al sistema no tiene esa frecuencia de muestreo, hay que remuestrearla para que tenga esa frecuencia, ya que todo el proceso de optimización va a ser realizado con canciones cuya frecuencia de muestreo es de 16000 Hz y un cambio en ese parámetro podría alterar los resultados.

3. Una vez que se tiene la señal en un solo canal y con una frecuencia de muestreo de 16 KHz, es el momento de comenzar el proceso de separación mediante las técnicas NMF. Para ello hay que distinguir entre los dos métodos propuestos, el método Largocorto y el método Cortolargo. Los dos métodos son iguales en los conceptos teóricos y en la implementación, en lo único que se diferencian es en el orden en el que se aplican los espectrogramas y por tanto, en el orden de extracción de las pistas musicales. En el método Largocorto primero se aplica un espectrograma con ventana larga (para extraer los sonidos armónicos) y luego se aplica un espectrograma con ventana corta (para extraer los sonidos percusivos), siendo la voz la señal que queda después de las extracciones. Por el contrario, en el método Cortolargo primero se aplica un espectrograma con ventana corta (para extraer los sonidos percusivos) y luego se aplica un espectrograma con ventana larga (para extraer los sonidos armónicos), siendo la voz la señal que queda después de las extracciones.

A continuación se van a explicar ambos métodos por separado, detallando todos los pasos que los forman:

4.3.1. Método Largocorto

Figura 42: Método Largocorto.

Este método se denomina así porque primero se le aplica a la señal un espectrograma con ventana larga y posteriormente un espectrograma con ventana corta. Al realizar el espectrograma con ventana larga los sonidos armónicos aparecen de forma discontinua en el eje espectral y muestran continuidad en el eje temporal, de la misma manera, los sonidos percusivos tienen continuidad en el eje espectral mientras que aparecen de forma discontinua en el eje temporal.

Figura 43: Espectrograma con ventana larga (Método Largocorto).

4. El primer paso es pasar la señal, que está en el dominio temporal, al dominio de la frecuencia, para ello se recurre a la transformada rápida de Fourier (FFT). Lo que se realiza es el espectrograma de la señal monocanal muestreada a 16 KHz. Para realizar el espectrograma hay que definir una serie de parámetros de los que depende el cálculo de dicha operación. Estos parámetros son:

4.1. **La frecuencia de muestreo:** como ya se ha explicado anteriormente, durante todo el algoritmo su valor es de 16 KHz.

4.2. **La longitud de la ventana:** es el tamaño de la ventana que se le aplica a los segmentos en los que se divide la señal. La ventana que se va a utilizar es la ventana de Hamming (figura 44). Como esta es la primera etapa del método Largocorto, la ventana utilizada se denomina ventana larga y puede tomar una longitud de 64 ms, de 128 ms, de 256 ms o de 512 ms. Como la frecuencia de muestreo es de 16000 Hz, es decir, 16000 muestras en cada segundo, las longitudes de la ventana larga en muestras pueden ser 1024, 2048, 4096 o 8192 muestras. Hay que tener en cuenta que todos los tamaños son potencias de 2, característica que se emplea en el cálculo de la longitud de los segmentos en los que se divide la señal.

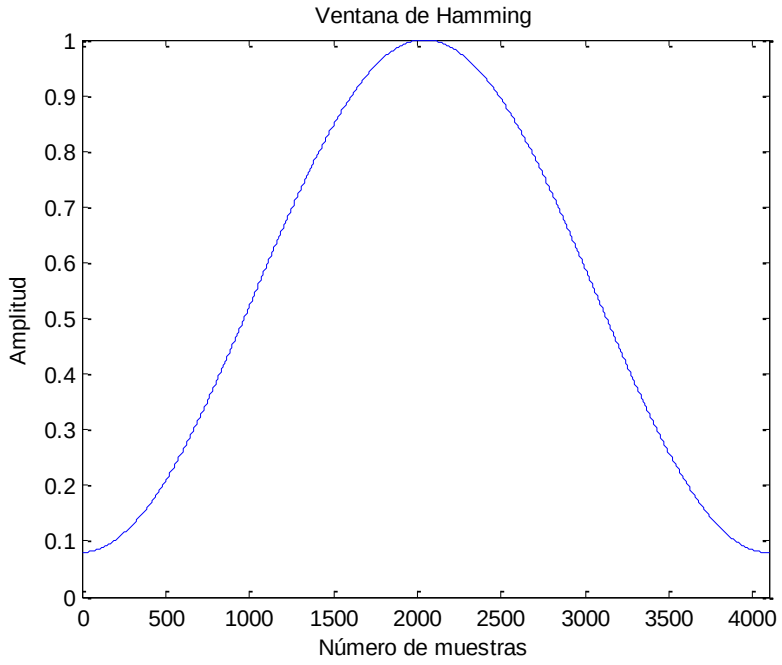

Figura 44: Ventana de Hamming de 4096 muestras de longitud.

4.3. El solapamiento: es el número de muestras que se solapan dos segmentos de señal consecutivos. Siempre va a ser la mitad del número de muestras que mide la ventana. En la siguiente figura se ve la forma en la que se solapan las ventanas.

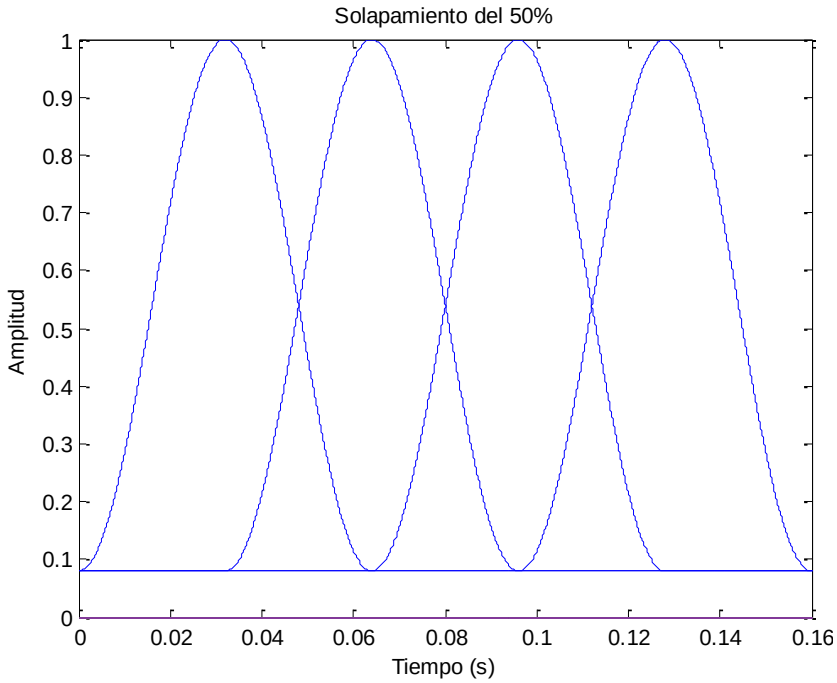

Figura 45: Solapamiento del 50% de las ventanas de Hamming.

La razón por la que se coge este valor de solapamiento es para que cuando se sumen las distintas ventanas el resultado siempre sea constante (figura 46). De esta forma es como si la señal estuviera enventanada por la función escalón.

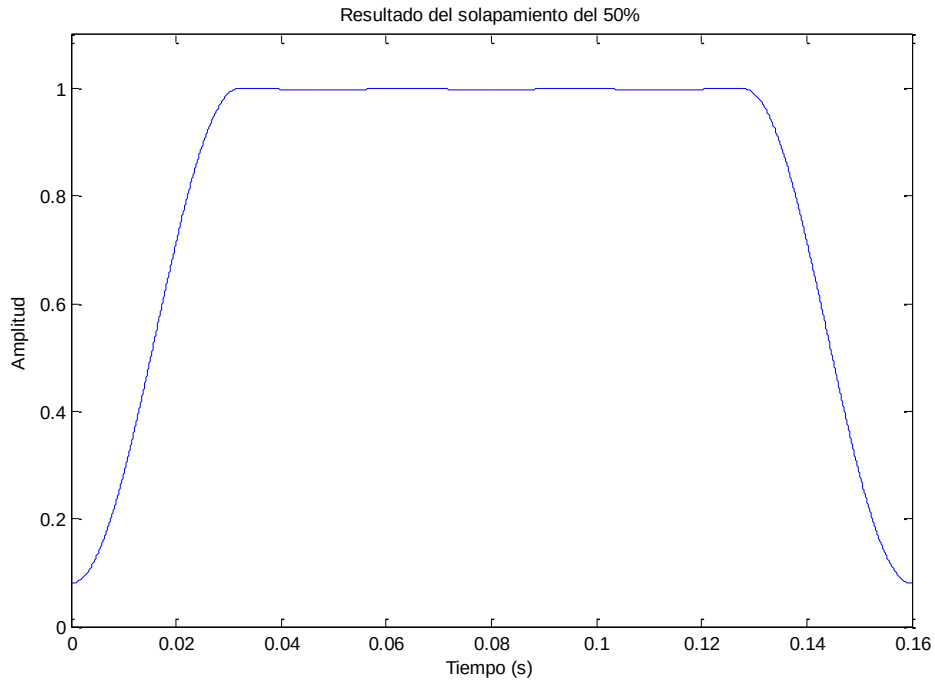

Figura 46: Suma de las ventanas de Hamming solapadas al 50%.

4.4. La longitud de los segmentos: es el tamaño de los segmentos de señal a los que se le aplica la transformada rápida de Fourier (FFT). Este tamaño es el resultado de elevar 2 al número que siendo potencia de 2, es igual o superior al tamaño de la ventana. Como los tamaños de las posibles ventanas son potencias de 2, la longitud de los segmentos siempre va a coincidir con la longitud de la ventana.

El resultado obtenido es el espectrograma de la señal original monocanal en forma de matriz, cuyas dimensiones serán FxT. A continuación se calcula el módulo y la fase de dicho espectrograma. Durante todo el proceso que queda hasta obtener las pistas separadas, todas las operaciones se realizarán con el módulo, mientras que la fase solo se tendrá en cuenta para pasar las señales al dominio temporal. En todas las representaciones gráficas de los espectrogramas que se van a realizar, lo que se representa es el módulo.

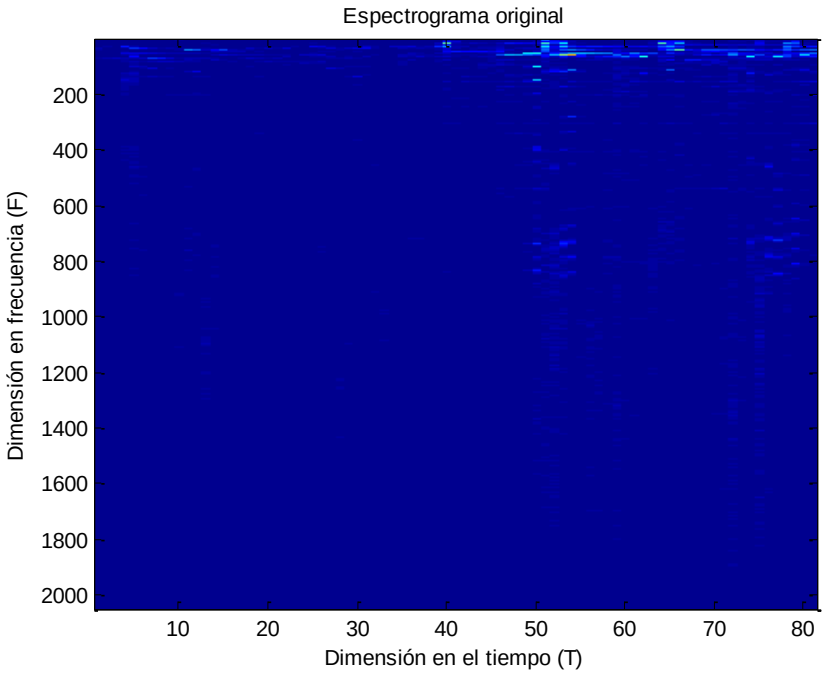

Figura 47: Espectrograma de la señal khair_1_04.

En la figura anterior se observa un ejemplo de la matriz resultante al realizar el espectrograma de la señal khair_1_04 con una ventana larga de 4096 muestras. Las dimensiones de la matriz son de 2049x81 (FxT).

5. Una vez que se tiene la señal en el dominio espectral, es el momento de definir las matrices de bases (W) y de activaciones (H). Para definirlas, se inicializan con valores aleatorios entre 0 y 1, los cuales siguen una distribución uniforme. Es en este momento donde entra en juego el número de bases, ya que la matriz W se genera con

unas dimensiones de F filas y B columnas, donde B es el número de bases. Del mismo modo la matriz H se genera con unas dimensiones de B filas y T columnas. De modo que se cumpla la ecuación 6.

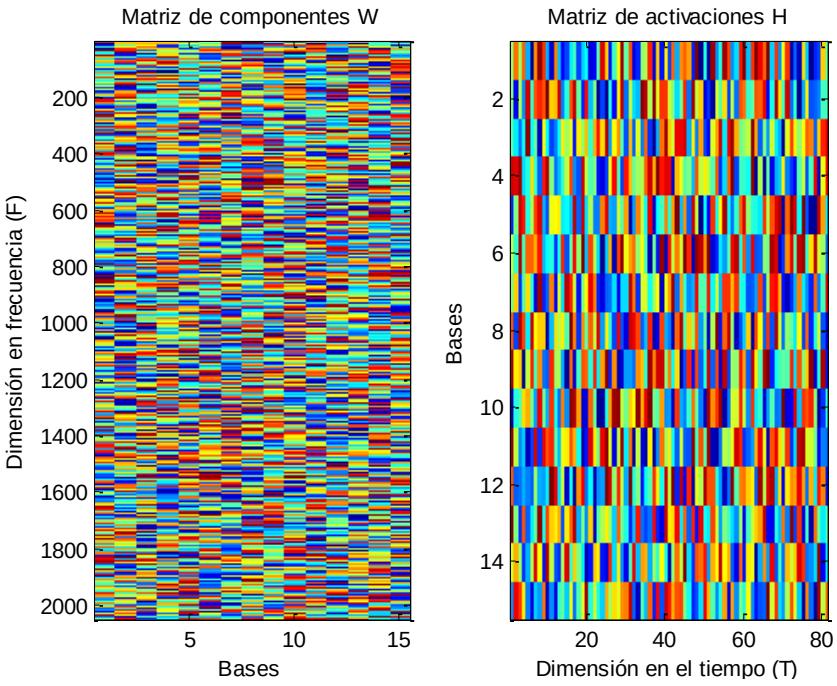

Figura 48: Matrices W y H inicializadas aleatoriamente.

Como se puede observar estas matrices no tienen ningún tipo de información de la señal original, la única característica común son sus dimensiones, ya que al multiplicar ambas matrices entre sí, se obtiene una matriz con las mismas dimensiones que el espectrograma original (figura 47). Como era de esperar la matriz W tiene unas de dimensiones de 2049x15 (FxB) y la matriz H mide 15x81 (BxT), ya que para este ejemplo se han utilizado 15 bases.

Reglas de actualización

Figura 49: Reglas de actualización del espectrograma con ventana larga (Método Largocorto).

6. Ahora es el momento de aplicar las reglas de actualización. Una vez inicializadas las matrices W y H, hay que aplicar las reglas de actualización de Kullback-Leibler (ecuaciones 20 y 21), calculando en cada actualización el valor de la función de coste de Kullback-Leibler (denominada divergencia), que mide la diferencia entre el espectrograma de la señal original y el espectrograma de la señal reconstruida como

resultado del producto de las matrices W y H (ecuación 8). La condición que se adopta como óptima para no realizar más actualizaciones es la siguiente:

$$condicion = \frac{abs\big(divergencia(i) - divergencia(i-1)\big)}{divergencia(i)} < 0.001 \qquad (27)$$

Donde abs es el valor absoluto, divergencia es la divergencia de Kullback-Leibler (ecuación 8) e i es el número de veces que se han aplicado las reglas de actualización, es decir, el número de iteraciones. Cuando se cumple la ecuación 27, la separación se da por buena y ya no se realizan más iteraciones. Para el cálculo de esta medida, se guardan todos los valores de la divergencia de cada iteración en un vector, porque así se puede comprobar de forma gráfica si la separación está convergiendo conforme aumenta el número de iteraciones. La ventaja de que la condición se calcule en función de la divergencia de cada separación, es que para cada canción que se vaya a separar y para cada parámetro que interviene en la ejecución del espectrograma, la separación convergirá dependiendo de los valores de su divergencia y no se realizarán un número de iteraciones fijo para cualquier separación. De esta forma además de asegurar que la separación se va a parecer en un alto grado a la señal original, también se garantiza que el proceso de actualización sea lo más eficiente posible.

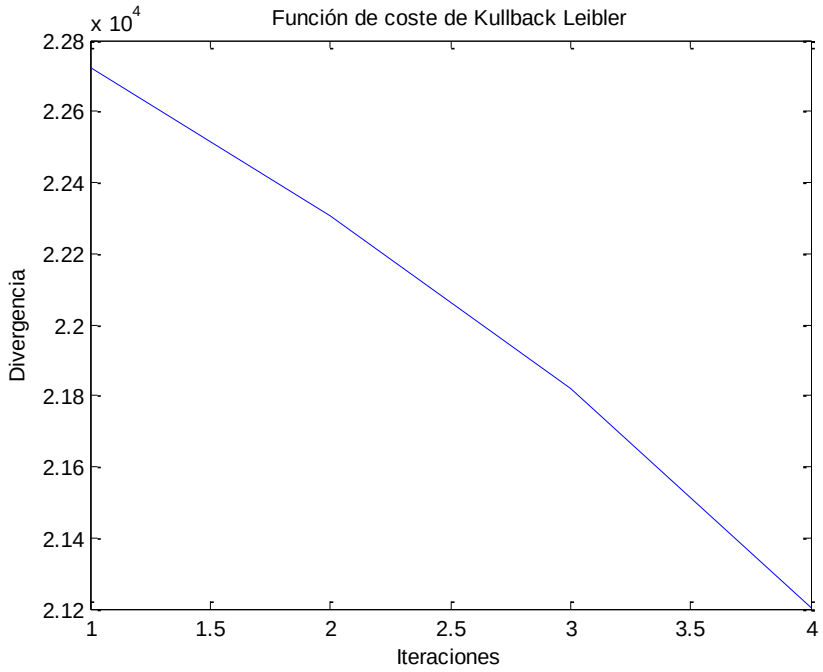

Figura 50: Divergencia de Kullback-Leibler con 4 iteraciones.

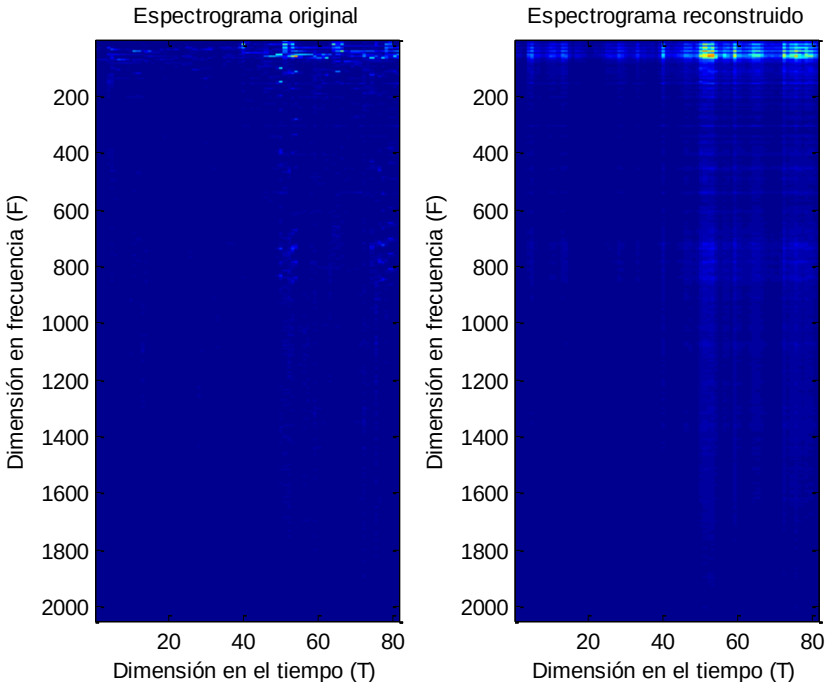

Figura 51: Comparativa entre el espectrograma de la señal original y el reconstruido.

Si el proceso de actualización se finaliza antes de que la divergencia sea constante, como pasa en la figura 50, el espectrograma de la señal original y el reconstruido se parecen muy sutilmente, como se observa en la figura 51. Por lo tanto se necesitan más iteraciones para que la divergencia converja y los espectrogramas sean iguales. En la figura 52 se han realizado 76 iteraciones y en esta ocasión el proceso de actualización sí ha finalizado, ya que la divergencia se llega a comportar de forma constante conforme pasan las iteraciones. Entonces al comparar ambos espectrogramas (figura 53) se ve como su parecido es muy grande.

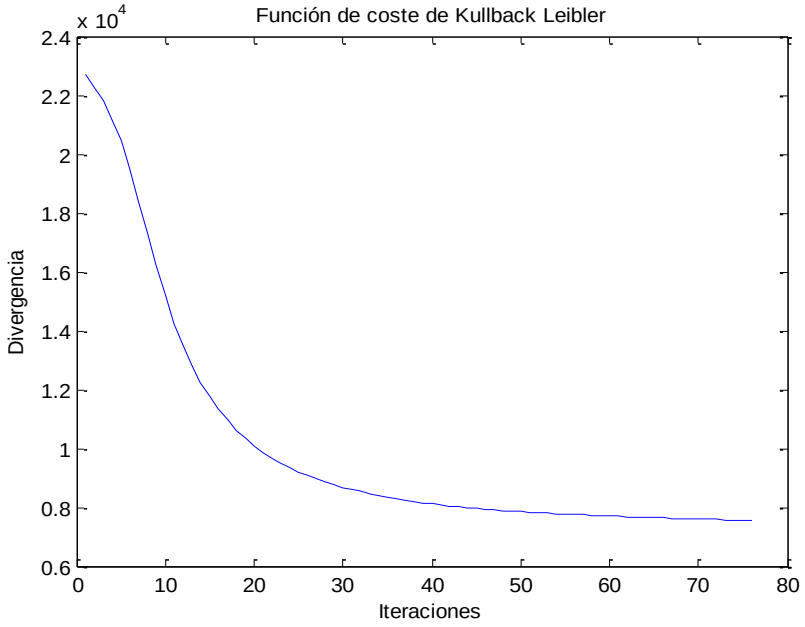

Figura 52: Divergencia de Kullback-Leibler con 76 iteraciones.

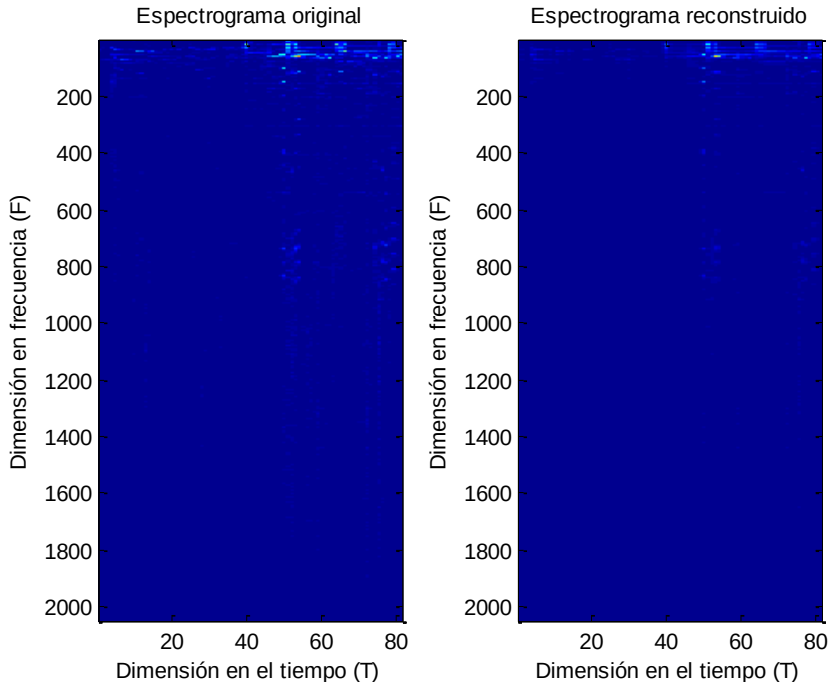

Figura 53: Espectrograma original y su reconstrucción.

Ahora las matrices de bases W y activaciones H sí dan información acerca de cómo está formada la señal original:

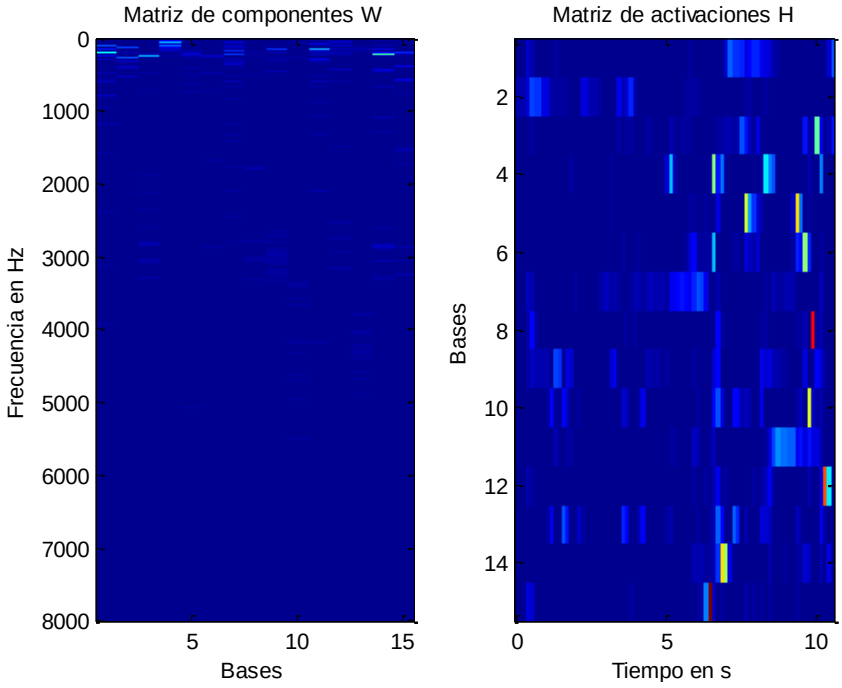

Figura 54: Matrices W y H después de 76 iteraciones.

A diferencia de la figura 48, esta vez sí se puede extraer gran cantidad de información de estas matrices. Si por ejemplo se hace un zoom a la base número 3 (figura 55), en la matriz W se observa que es una base cuya componente principal está

en torno a los 235 Hz, pero también tiene energía en torno a sus tres múltiplos posteriores (470 Hz, 705 Hz y 900 Hz). En la matriz H se observan los instantes temporales durante los cuales está activa la base número 11, estos instantes son aquellos intervalos en los que el color pasa de azul oscuro (el color predominante en ambas matrices, que quiere decir que no hay ni bases ni activaciones) a un color azul más claro.

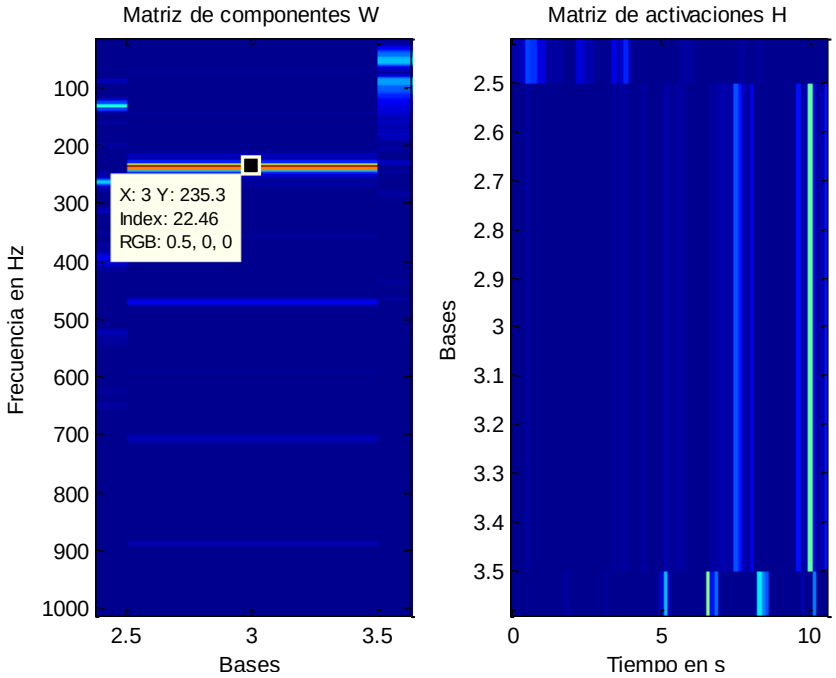

Figura 55: Zoom de la base número 3 en ambas matrices.

Umbralización en frecuencia

Figura 56: Umbralización en frecuencia (Método Largocorto).

7. Ahora es el momento del proceso de umbralización. Como se ha aplicado el espectrograma con ventana larga, los sonidos armónicos muestran discontinuidad espectral, por lo que se aplica el proceso de umbralización espectral, que permite saber qué bases se corresponden con los sonidos armónicos. Para cada una de las bases hay que calcular la siguiente medida:

$$d_s(X^b) = \frac{\sum_{f=2}^{F}\left(W_{f,b} - W_{f-1,b}\right)^2}{\sum_{f=1}^{F} W_{f,b}^2} > \theta_s \qquad (28)$$

Donde $X^b = W_{f,b} * H_{b,t}$ es el espectrograma de la base b con f desde 1 hasta F y con t desde 1 hasta T. Donde b va desde 1 hasta B, siendo B el número de bases de la matriz W y donde el denominador sirve para normalizar. Cuanto mayor es $d_s(X^b)$, más espectralmente discontinua es X^b. Si $d_s(X^b)$ es mayor que un umbral en frecuencia θ_s, X^b se considera que proviene de un instrumento armónico.

Una vez que se ha calculada la ecuación 28 para cada una de las bases, se calcula una nueva señal X' (percusivos más voz), en la que se eliminan las componentes armónicas de la señal original X. Esta operación se realiza mediante la siguiente ecuación:

$$X' = \max\left(0, X - \sum_{\substack{b=1,\ldots,B \\ d_s(X^b)>\theta_s}} X^b\right) \tag{29}$$

Donde 0 es una matriz con todos los elementos a 0 con las mismas dimensiones que X. Al calcular el máximo se asegura que la nueva señal X' (percusivos mas armónicos) no tenga elementos negativos.

Después de realizar la ecuación 29 se tiene calculada la señal formada por los sonidos percusivos más los sonidos vocales (X'). Por lo que, solo queda calcular la señal formada por todas las componentes armónicas, esta operación se realiza calculando el sumatorio que hay en la ecuación anterior, en el que se suman los productos matriciales de cada una de las bases consideradas como armónicas por sus correspondientes activaciones.

Este es el resultado de aplicar umbralización en frecuencia con un umbral de 0,4.

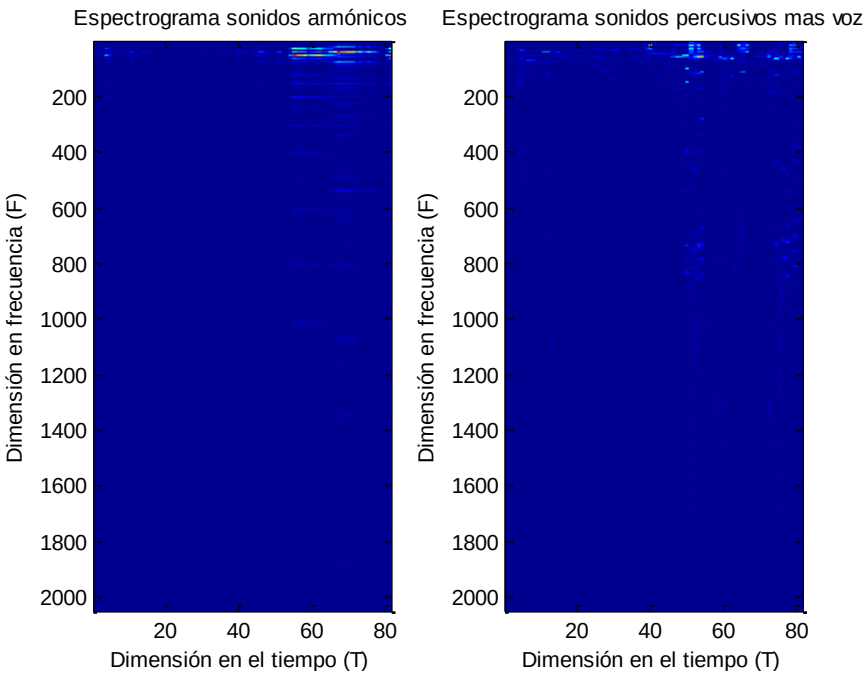

Figura 57: Resultado de la umbralización espectral.

Armónicos **Percusivos + voz**

Figura 58: Armónicos y Percusivos+voz en el dominio del tiempo (Método Largocorto).

8. El siguiente paso consiste en pasar las señales del domino espectral al dominio temporal, para ello se utiliza la transformada inversa de la transformada rápida de Fourier (IFFT). En ella intervienen los mismos parámetros que son necesarios para realizar el espectrograma (frecuencia de muestreo, longitud de la ventana, solapamiento y longitud de los segmentos). Además hay que tener en cuenta la fase del espectrograma original, esto se hace multiplicando los espectrogramas que se quieren pasar al dominio temporal por la fase del espectrograma original. Tras realizar este paso se tendrían dos pistas, la que se corresponde con los sonidos armónicos y la que se corresponde con los sonidos percusivos más los vocales. Los sonidos armónicos no necesitan ninguna forma de procesado adicional, por lo que ahora hay que centrarse en separar los sonidos percusivos de los sonidos producidos por la voz cantada.

Espectrograma con ventana corta

Figura 59: Espectrograma con ventana corta (Método Largocorto).

9. A continuación, comienza la segunda etapa del método Largocorto, en la que hay que pasar al dominio espectral la señal de los sonidos percusivos más los vocales utilizando la FFT. Para ello se necesitan los mismos parámetros que se han mencionado para el espectrograma de la señal original, la diferencia es que el tamaño de la ventana es diferente. Ahora se utiliza una ventana con menor número de muestras, por eso de denomina a esta etapa como separación con ventana corta. Los tamaños de las ventanas pueden ser de 8 ms, de 16 ms, de 32 ms o de 64 ms, que como la frecuencia de muestreo es de 16 KHz (al igual que en la etapa con ventana larga), se corresponden con ventanas de una longitud de 128, de 256, de 512 o de 1024 muestras, respectivamente. Los otros parámetros se calculan de la misma forma que se ha explicado en la etapa larga, pero hay que tener en cuenta que ahora el tamaño de la ventana ha cambiado, ahora es el de la ventana corta.

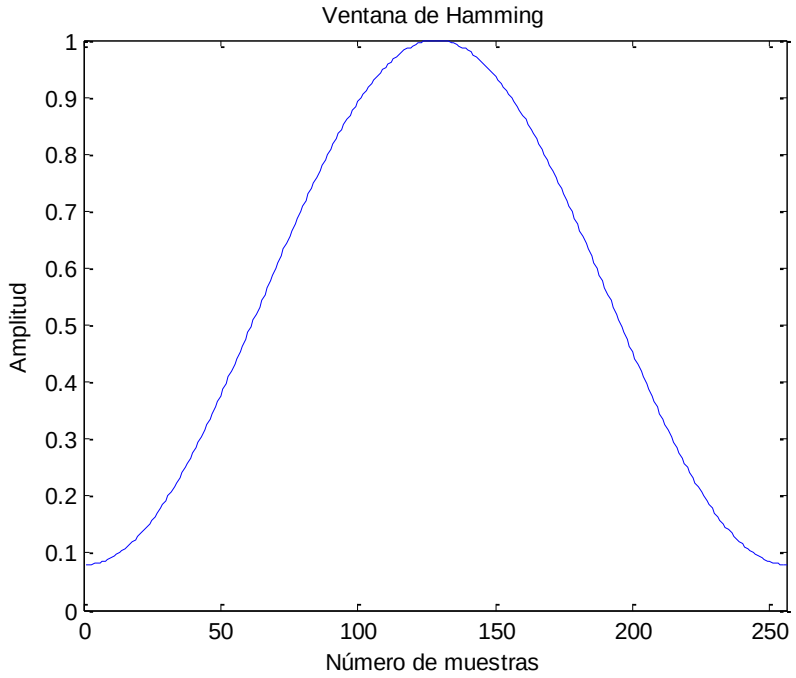

Figura 60: Ventana de Hamming con un tamaño de 256 muestras.

10. Una vez que se ha calculado el espectrograma, se tendrá una matriz cuyas dimensiones son $F \times T$, que no tendrán el mismo valor que las del espectrograma con ventana larga. Nuevamente se calcula el módulo y la fase de este espectrograma. Ahora es el momento de definir nuevamente las matrices W y H de forma aleatoria entre 0 y 1 siguiendo una distribución uniforme.

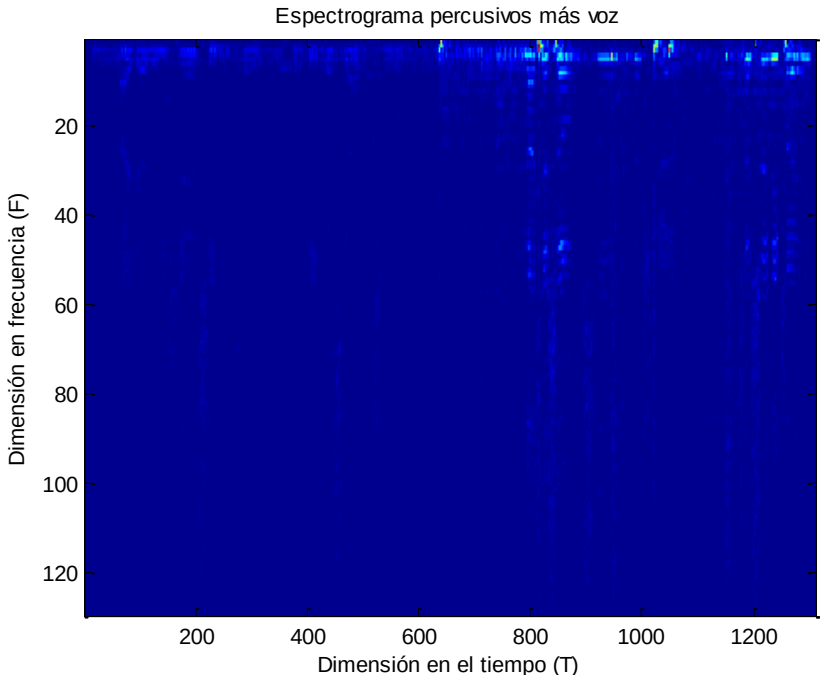

Figura 61: Espectrograma de los sonidos percusivos más la voz.

En la figura anterior se ve el espectrograma que se tiene al pasar la señal de los sonidos percusivos más la voz al dominio espectral, utilizando una ventana corta de 256 muestras. Como se puede observar, las dimensiones de esta matriz son distintas a las dimensiones de la matriz de la figura 47, en este caso la matriz es de 129x1315. Lo que quiere decir que las dimensiones de las matrices dependen del tamaño de ventana que se utilice.

En la siguiente figura se puede ver las matrices W y H inicializadas aleatoriamente. Al igual que en la figura 48, estas matrices no contienen ninguna información del comportamiento espectral de cada base ni del periodo de tiempo durante el cual permanecen activas. La única información que muestran son sus dimensiones, que son de 129x15 y de 15x1315, respectivamente, concordando con el tamaño del espectrograma de la figura 61.

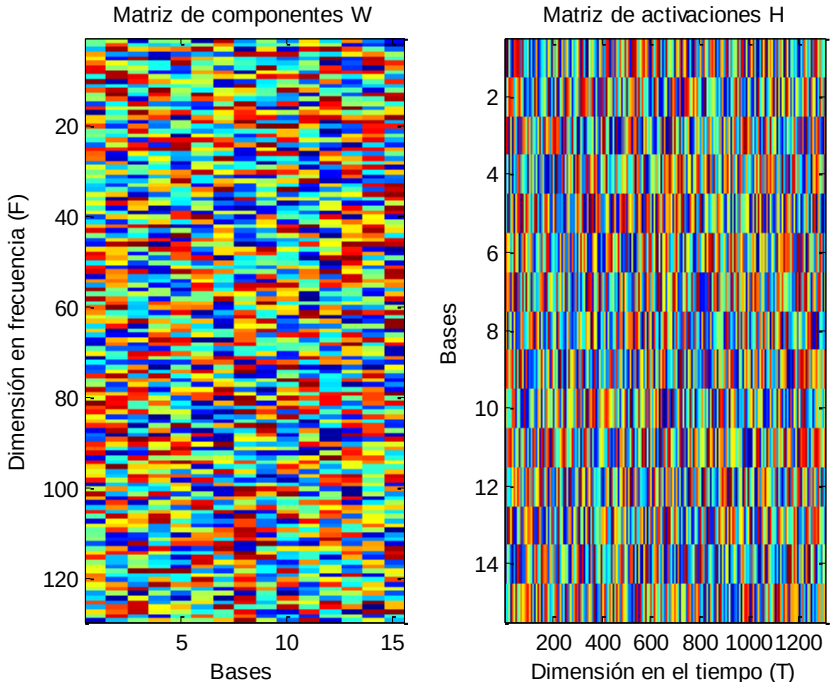

Figura 62: Matrices W y H inicializadas de forma aleatoria.

Reglas de actualización

Figura 63: Reglas de actualización del espectrograma con ventana corta (Método Largocorto).

11. Luego se ejecutan las reglas de actualización del mismo modo que en la etapa con ventana larga, calculándose la divergencia de Kullback-Leibler (ecuación 8) entre el espectrograma original (percusivos más voz) y el espectrograma resultante del producto entre W y H. Cuando se cumpla la ecuación 27 ($condición < 0{,}001$), la separación se da por buena y ya no se realizan más actualizaciones.

Tras realizar 82 iteraciones, la divergencia ha convergido, es decir, permanece constante (figura 64), dando como resultado las matrices W y H de la figura 65.

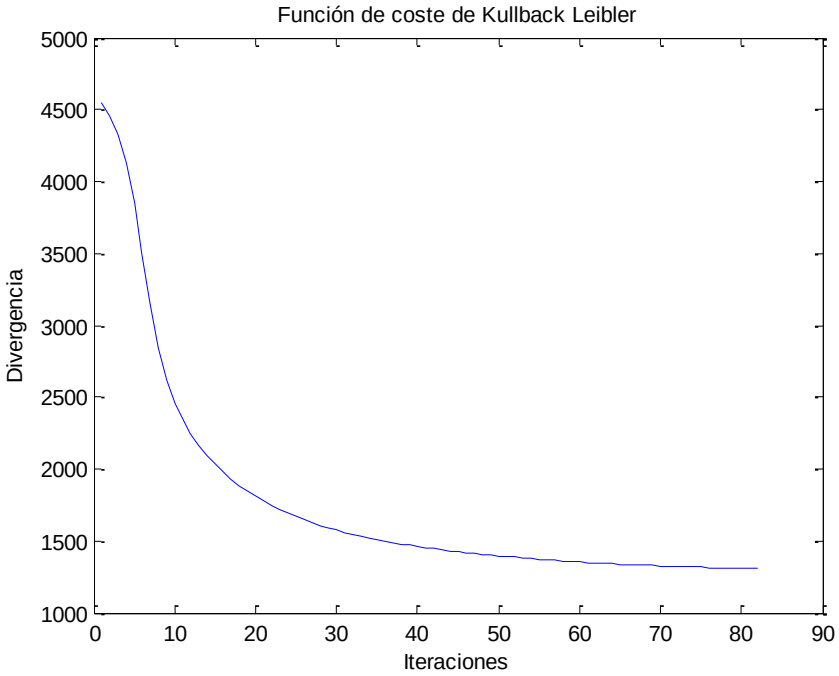

Figura 64: Divergencia de Kullback-Leibler con 82 iteraciones.

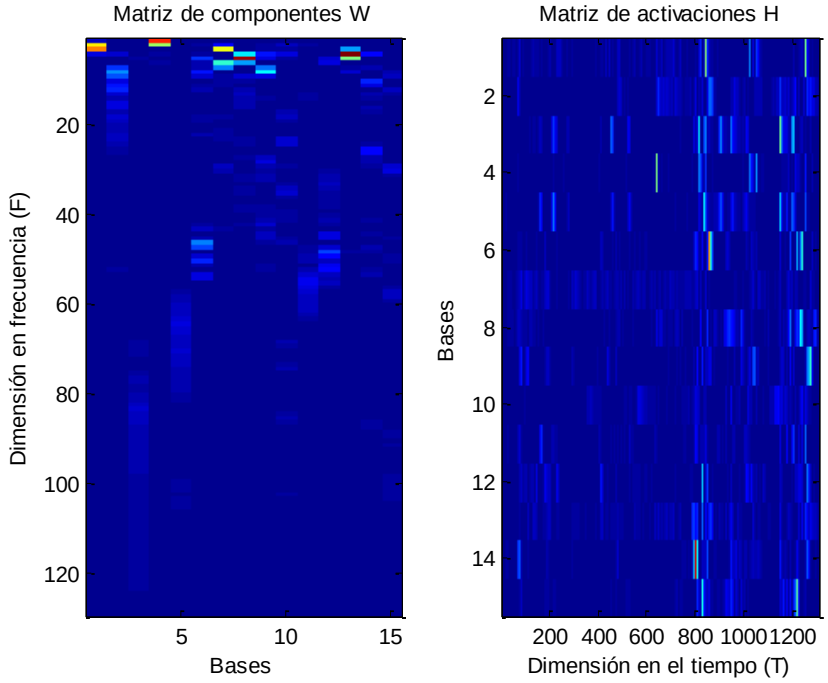

Figura 65: Matrices W y H después de 82 iteraciones.

A diferencia de las matrices de la figura 62, estas matrices sí dan información de cómo se comporta cada base a lo largo del eje de frecuencia (W) y en que instante de tiempo está activa cada una (H).

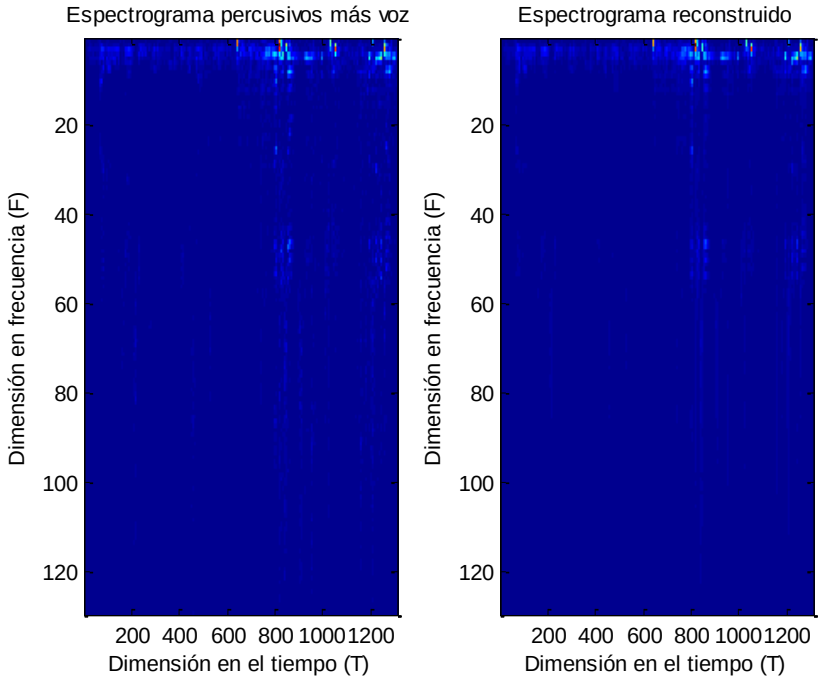

Figura 66: Espectrograma de los percusivos más la voz y su reconstrucción.

Multiplicando las matrices W y H actualizadas, se obtiene el espectrograma reconstruido de la figura anterior, que es igual que el espectrograma original (percusivos mas voz).

Umbralización en el tiempo

Figura 67: Umbralización en el tiempo (Método Largocorto).

12. Este paso sí es diferente a su homólogo en la etapa con ventana larga, ya que ahora se realiza el proceso de umbralización en el tiempo. Porque al realizar el espectrograma con ventana corta, los sonidos percusivos se comportan de manera continua en el eje de frecuencia mientras que aparecen discontinuidades en el eje temporal. Para saber las bases que se corresponden con los sonidos percusivos se calcula la siguiente ecuación:

$$d_t(X^b) = \frac{\sum_{t=2}^{T}\left(H_{b,t} - H_{b,t-1}\right)^2}{\sum_{t=1}^{T} W_{b,t}^2} > \theta_t \qquad (30)$$

Donde $X^b = W_{f,b} * H_{b,t}$ es el espectrograma de la base b con f desde 1 hasta F y con t desde 1 hasta T. Donde b va desde 1 hasta B, siendo B el número de bases de la matriz W. El denominador sirve para normalizar. Cuanto mayor es $d_t(X^b)$, más temporalmente discontinua es X^b. Si $d_t(X^b)$ es mayor que un umbral en el tiempo θ_t, X^b se considera que proviene de un instrumento percusivo.

Una vez que se ha calculada la ecuación 30 para cada una de las bases, se calcula una nueva señal X' (voz), en la que se eliminan las componentes percusivas de la señal formada por los sonidos percusivos más la voz (X). Esta operación se realiza mediante la siguiente ecuación:

$$X' = \max\left(0, X - \sum_{\substack{b=1,\dots,B \\ d_t(X^b)>\theta_t}} X^b\right) \tag{31}$$

Donde 0 es una matriz con todos los elementos a 0 con las mismas dimensiones que X. Al calcular el máximo se asegura que la nueva señal X' no tenga elementos negativos.

Después de realizar la ecuación 31 se tiene calculada la señal formada por los sonidos vocales (X'). Entonces, ya solo queda calcular la señal formada por todas las componentes percusivas, esta operación se realiza calculando el sumatorio que hay en la ecuación anterior, en el que se suman los productos matriciales de cada una de las bases consideradas como percusivas por sus correspondientes activaciones.

En la siguiente figura se muestra el resultado de aplicar la umbralización en el tiempo con un umbral de 0,2.

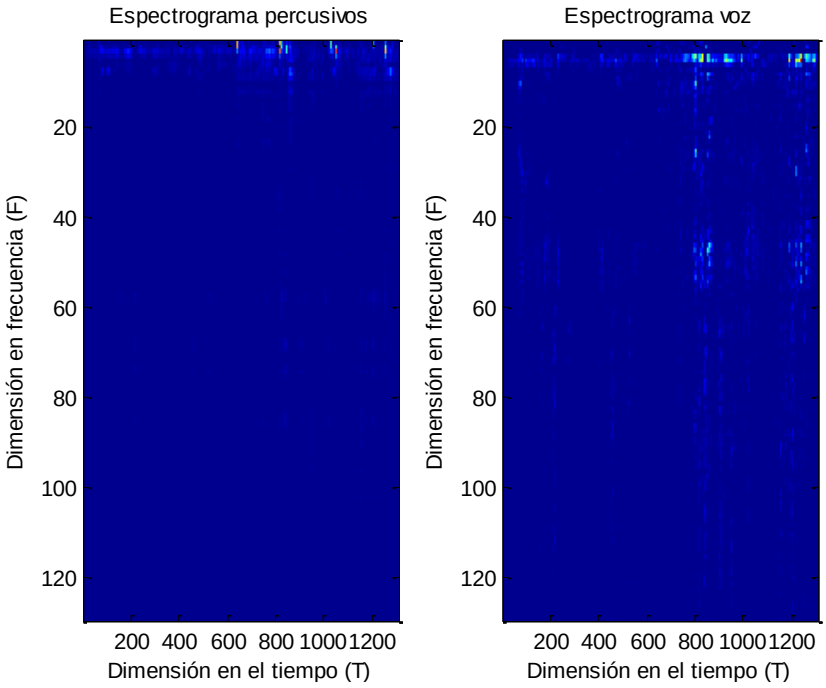

Figura 68: Resultado de la umbralización temporal.

Figura 69: Voz y sonidos percusivos en el dominio del tiempo (Método Largocorto).

13. El último paso es pasar las dos señales (percusivos y voz) al dominio temporal. Para ello se aplica la transformada inversa de la transformada rápida de Fourier (IFFT). Esta operación también requiere la multiplicación de los espectrogramas de las pistas separadas por la fase del espectrograma original (percusivos más voz), además de la frecuencia de muestreo, la longitud de la ventana, el solapamiento y la longitud de los segmentos.

14. Una vez que ya se tiene la señal original separada en las tres pistas musicales (sonidos armónicos, sonidos percusivos y voz), se puede medir la calidad de la separación mediante el cálculo de las medidas de calidad objetivas (SDR, SIR y SAR). Como en las canciones de la base de datos con la que se ha trabajado aparece todo el acompañamiento musical en un canal y la voz cantada sin música en el otro canal, no se puede medir la calidad de las tres pistas por separado. Por ello para realizar este proceso hay que sumar la pista de sonidos armónicos y la pista de sonidos percusivos, para poder compararla con la música original y por otro lado comparar la pista que contiene la voz separada con la voz original.

4.3.2. Método Cortolargo

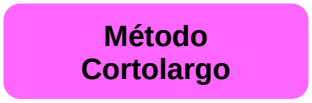

Figura 70: Método Cortolargo.

Como se puede prever, al igual que ocurre en el método anterior, este método tiene la nomenclatura de Cortolargo porque primero se aplica el espectrograma con ventana corta y posteriormente se aplica el espectrograma con ventana larga. Como ya se puede deducir del método anterior, en este caso primero se extraen los sonidos percusivos de la señal original y posteriormente se extraen los sonidos armónicos, siendo la voz la señal restante. A continuación se explican las diferentes fases:

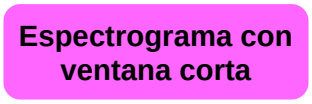

Figura 71: Espectrograma con ventana corta (Método Cortolargo).

4. El primer paso es pasar la señal original al dominio espectral, para ello se realiza el espectrograma aplicando la transforma rápida de Fourier (FFT). Los parámetros necesarios para su realización son los mismos que se han explicado anteriormente:

4.1. Una frecuencia de muestreo de 16 KHz.

4.2. Una longitud de ventana que se denomina ventana corta, que equivale a 128, o 256, o 512 o 1024 muestras.

4.3. Un solapamiento que es igual a la mitad de la longitud de la ventana (figura 72), para que cuando se sumen todas las ventanas el resultado siempre sea constante, tal y como se ve en la figura 73.

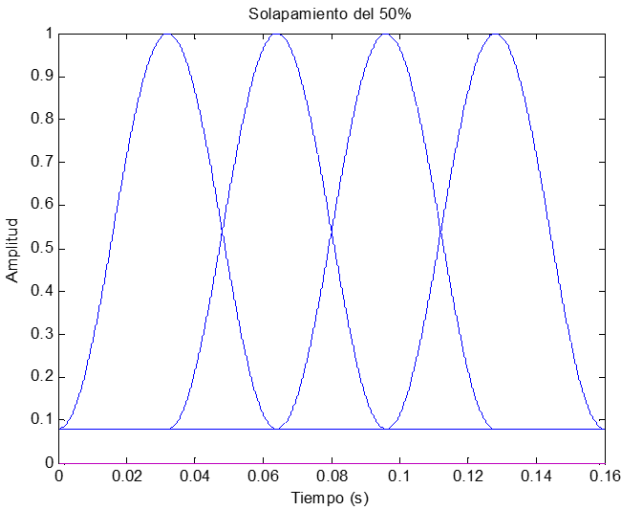

Figura 72: Ventanas de Hamming con un solapamiento del 50%.

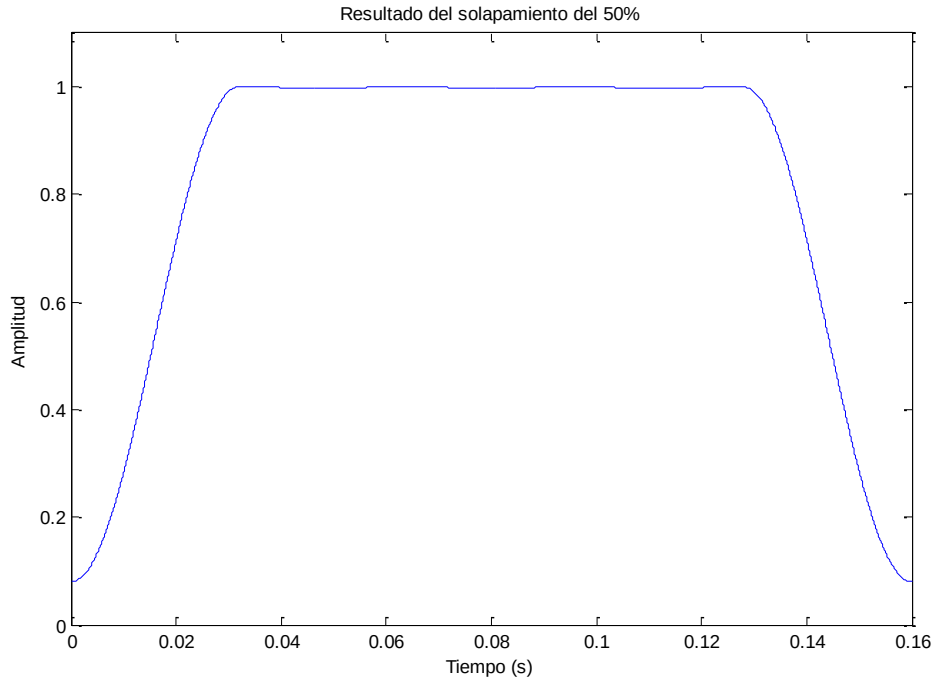

Figura 73: Suma de las ventanas de Hamming solapadas al 50%.

4.4. Un tamaño de los segmentos que es igual a la longitud de la ventana, ya que la condición que se pide es que el segmento tenga un tamaño que sea potencia de 2 mayor o igual a la longitud de la ventana. Por esta razón todas las longitudes posibles de la ventana corta son potencias de 2.

5. Del espectrograma resultante, que tendrá unas dimensiones de FxT, se calculan el módulo y la fase. Con el módulo se realizan las reglas de actualización y se calculan las divergencias, mientras que la fase se reserva para pasar las señales al dominio temporal. Al igual que sucede durante todo el método Largocorto, lo que se representa en las figuras son los módulos de los espectrogramas.

En la siguiente figura se muestra el espectrograma de la señal original con una ventana corta de 256 muestras. Como se puede ver, las dimensiones de este espectrograma son de 129x1315. No es casualidad que dichas dimensiones sean iguales que las del espectrograma de la figura 61. Esta coincidencia se debe a que la señal original es la misma y a que la ventana utilizada tiene el mismo tamaño, por tanto todos los parámetros que afectan a la realización de la FFT tienen también los mismos valores.

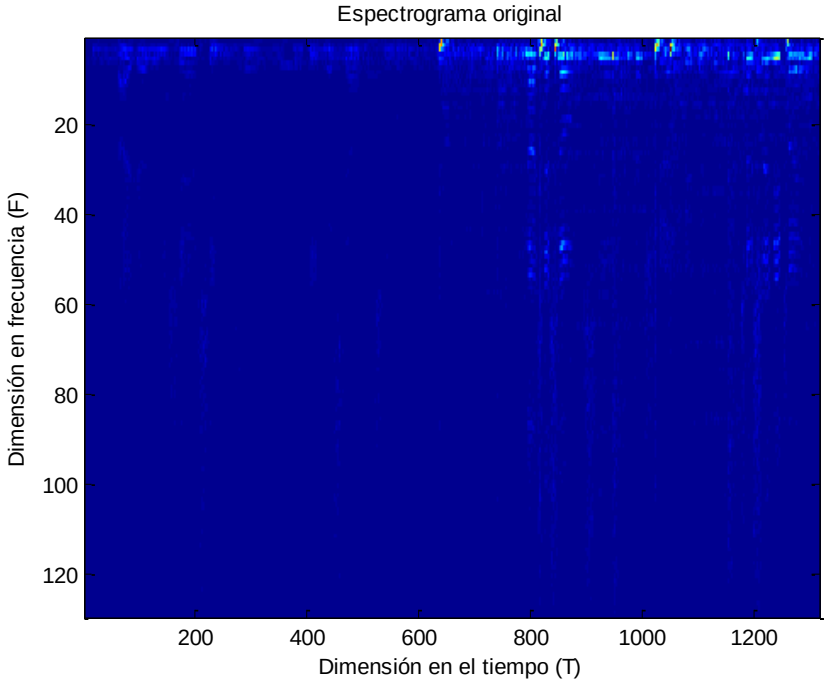

Figura 74: Espectrograma de la señal khair_1_04.

6. Se calculan las matrices W y H con valores aleatorios entre 0 y 1, siguiendo una distribución uniforme, las dimensiones serían FxB y BxT, respectivamente. De tal forma que se cumpla la ecuación 6.

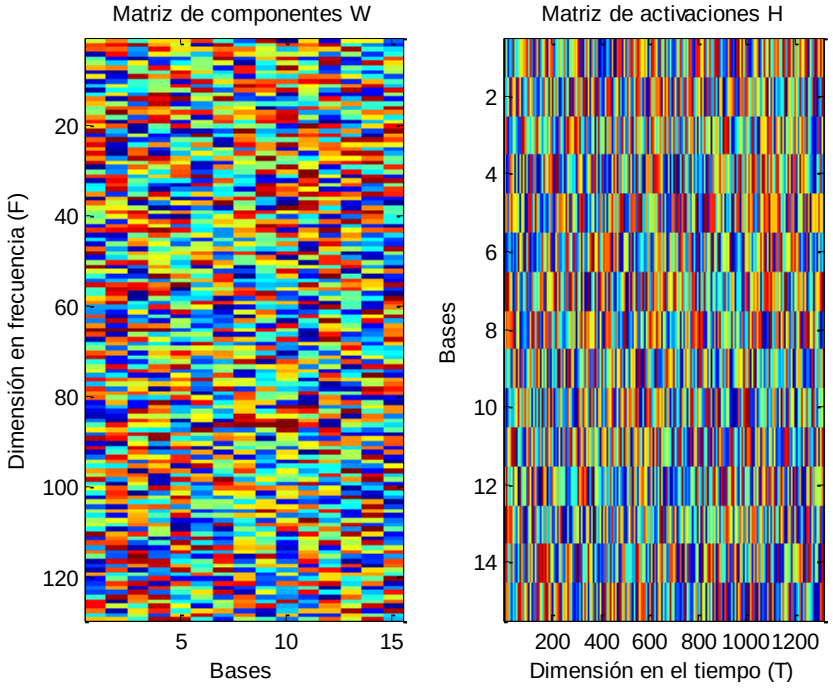

Figura 75: Matrices W y H inicializadas de forma aleatoria.

En la figura anterior se observan las matrices W y H inicializadas de forma aleatoria, las cuales solo dan información acerca de sus dimensiones FxB y BxT (129x15 y 15x1315). Siendo 15 el número de bases escogido para la realización del algoritmo.

Reglas de actualización

Figura 76: Reglas de actualización del espectrograma con ventana corta (Método Cortolargo).

7. Ahora comienzan a realizarse las reglas de actualización de Kullback-Leibler (ecuaciones 20 y 21), calculando para cada iteración la divergencia del mismo nombre (ecuación 8) y la condición que se adopta como óptima para no seguir realizando más iteraciones (ecuación 27).

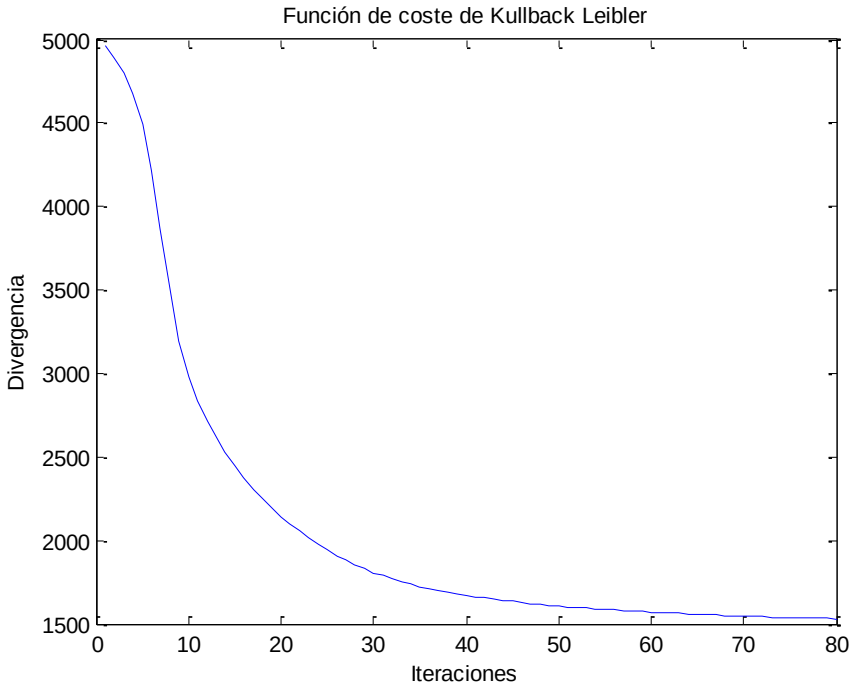

Figura 77: Divergencia de Kullback-Leibler con 80 iteraciones.

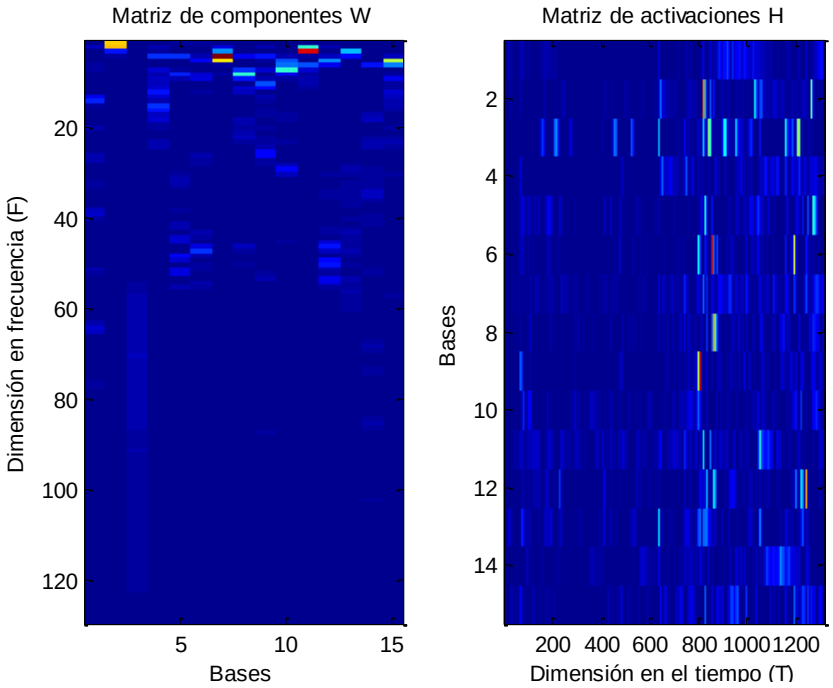

Figura 78: Matrices W y H después de 80 iteraciones.

Después de 80 iteraciones, la divergencia permanece constante (figura 77), por lo que se finaliza el proceso de actualización, obteniéndose las matrices W y H de la figura 78. A diferencia de las de la figura 75, estas matrices sí dan información de cómo se comporta cada base en frecuencia (W) y en qué instante de tiempo permanecen activas (H). El producto de ambas matrices da como resultado el espectrograma reconstruido de la figura siguiente, donde se observa que es similar al espectrograma original.

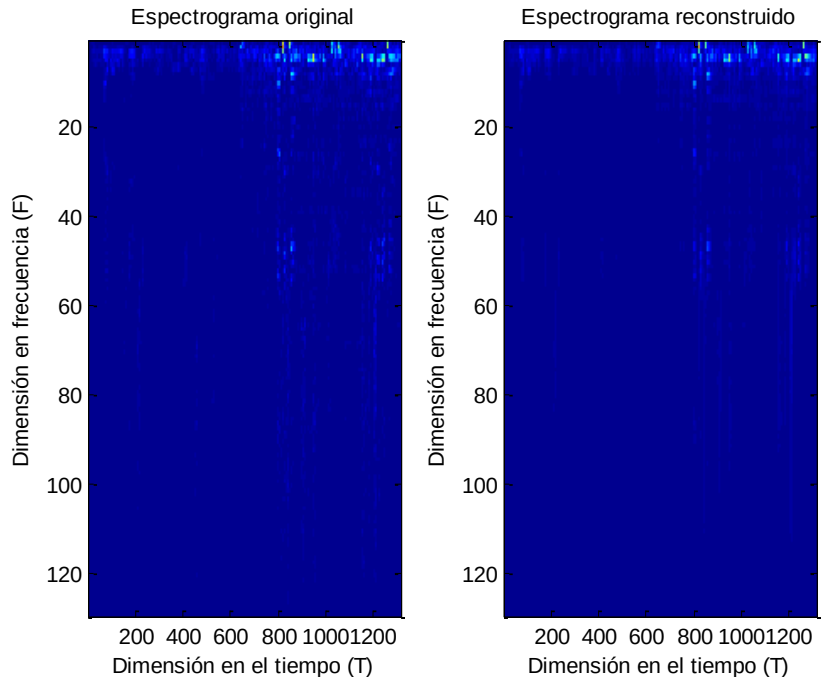

Figura 79: Espectrograma original y su reconstrucción.

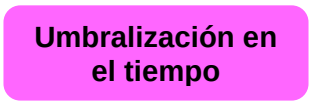

Figura 80: Umbralización en el tiempo (Método Cortolargo).

8. El siguiente paso es la umbralización, que como se ha utilizado una ventana corta para realizar el espectrograma, se realiza umbralización en el tiempo. Para saber las bases que se corresponden con los sonidos percusivos se calcula la ecuación 30.

Una vez que se ha calculado la ecuación 30 para cada una de las bases, se calcula una nueva señal X' (armónicos más voz) mediante la ecuación 31, en la que se eliminan las componentes percusivas de la señal original (X).

Entonces, ya solo queda calcular la señal formada por todas las componentes percusivas, esta operación se realiza calculando el sumatorio que hay en la ecuación 31, en el que se suman los productos matriciales de cada una de las bases consideradas como percusivas por sus correspondientes activaciones.

En la siguiente figura se muestra el resultado de aplicar la umbralización en el tiempo con un umbral de 0,2.

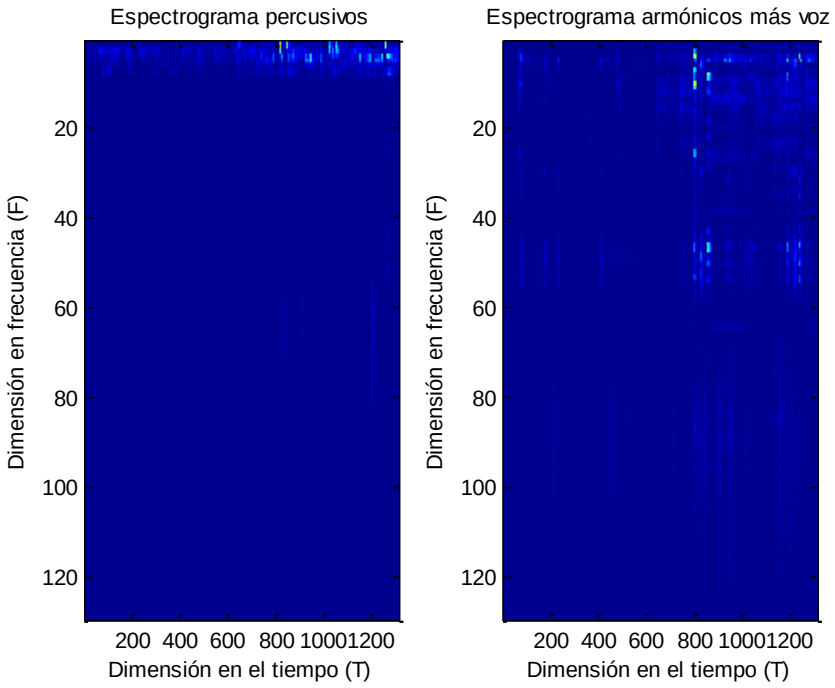

Figura 81: Resultado de la umbralización temporal.

Figura 82: Percusivos y Armonicos+voz en el dominio del tiempo (Método Cortolargo).

9. Ahora es el momento de pasar las señales al dominio temporal, para ello se realiza la transformada inversa de la transformada rápida de Fourier (IFFT) de los dos

espectrogramas que se han calculado. Para ello hay que multiplicar cada espectrograma por la fase del espectrograma de la señal original. También son necesarios los parámetros que intervienen en la (FFT), es decir, la frecuencia de muestreo, la longitud de la ventana, el solapamiento y el tamaño de los segmentos. Tras esto se obtienen dos pistas, una que solo contiene los sonidos percusivos y otra que contiene los sonidos armónicos y la voz. Evidentemente la siguiente etapa es separar esta última pista en sonidos armónicos y sonidos vocales.

Espectrograma con ventana larga

Figura 83: Espectrograma con ventana larga (Método Cortolargo).

10. Ahora comienza la segunda parte del método Cortolargo, en la que se realiza el espectrograma de la pista que contiene la mezcla de sonidos armónicos con la voz. En esta ocasión se utiliza la ventana larga, ya que lo que se pretende es extraer los sonidos armónicos de la señal. Al igual que en el método Largocorto, la ventana larga es de 1024, o de 2048, o de 4096, o de 8192 muestras. Además de la ventana, también son necesarios la frecuencia de muestreo, el solapamiento y el tamaño de los segmentos en los que se divide la señal.

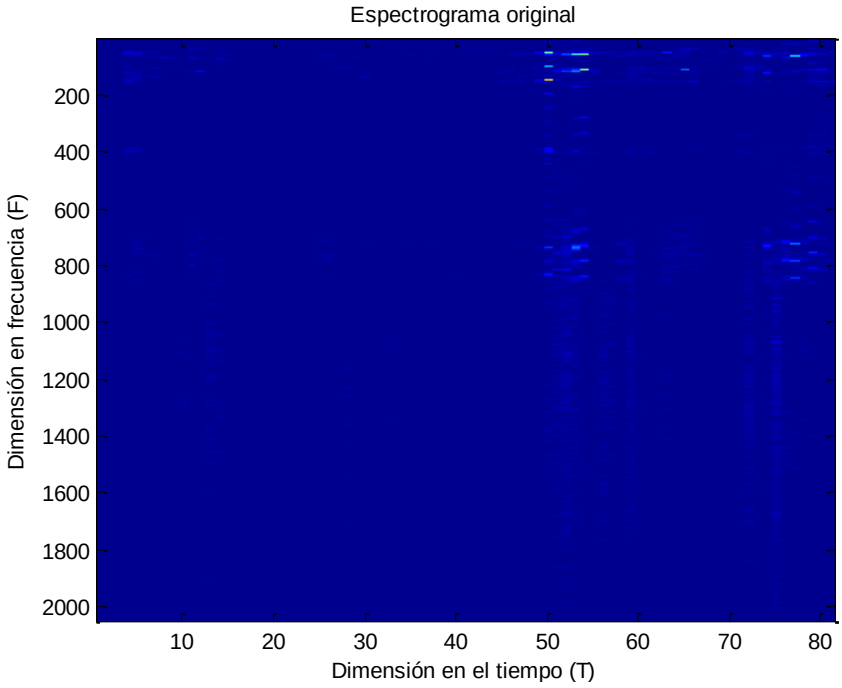

Figura 84: Espectrograma de los sonidos armónicos más la voz.

Este es el espectrograma de la señal formada por los sonidos armónicos más la voz. Se ha realizado con un tamaño de ventana de 4096 muestras. Sus dimensiones

son de 2049x81 (FxT), que son las mismas que las del espectrograma de la figura 47. Esto es así porque la señal original es la misma y porque se ha escogido el mismo tamaño de ventana.

11. Tras calcular el espectrograma, se calcula el módulo y la fase del mismo. El módulo se utilizará durante las reglas de actualización y para el cálculo de las divergencias, mientras que la fase se utilizará para pasar los espectrogramas al dominio temporal. También se inicializan las matrices W y H con elementos entre 0 y 1 siguiendo una distribución uniforme.

Al igual que en las figuras 48, 62 y 75, en la siguiente figura se muestran las matrices W y H inicializadas aleatoriamente, de las cuales solo son relevantes sus dimensiones 2049x15 y 15x81, respectivamente. Siendo 15 el número de bases utilizadas.

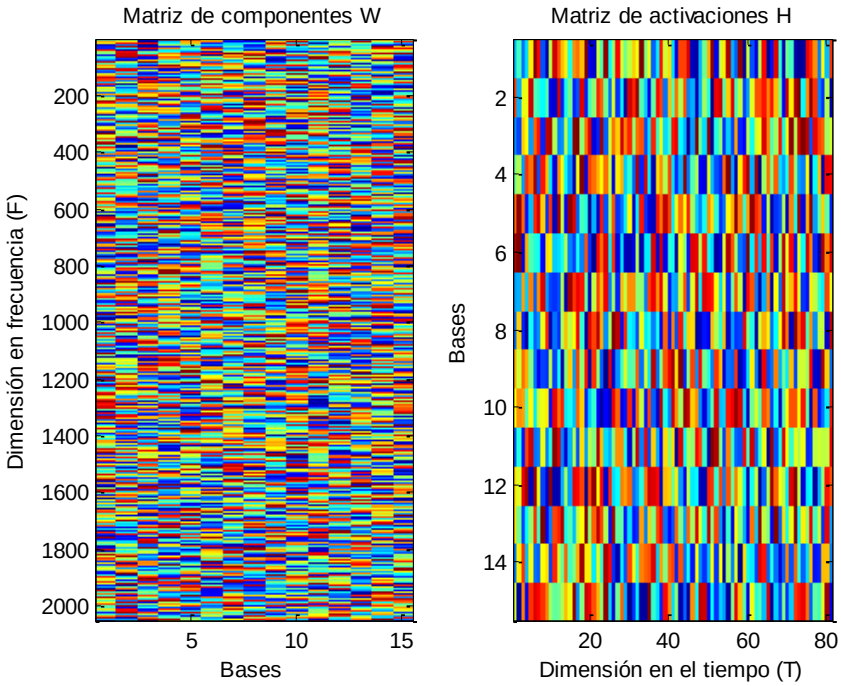

Figura 85: Matrices W y H inicializadas de forma aleatoria.

Reglas de actualización

Figura 86: Reglas de actualización del espectrograma con ventana larga (Método Cortolargo).

12. Es el momento de realizar las reglas de actualización de Kullback-Leibler (ecuaciones 20 y 21), calculándose la divergencia de Kullback-Leibler (ecuación 8) entre el espectrograma original (percusivos más voz) y el espectrograma resultante del producto entre W y H para cada iteración. Cuando se cumpla la ecuación 27

(condición<0,001), la separación se da por buena y ya no se realizan más actualizaciones.

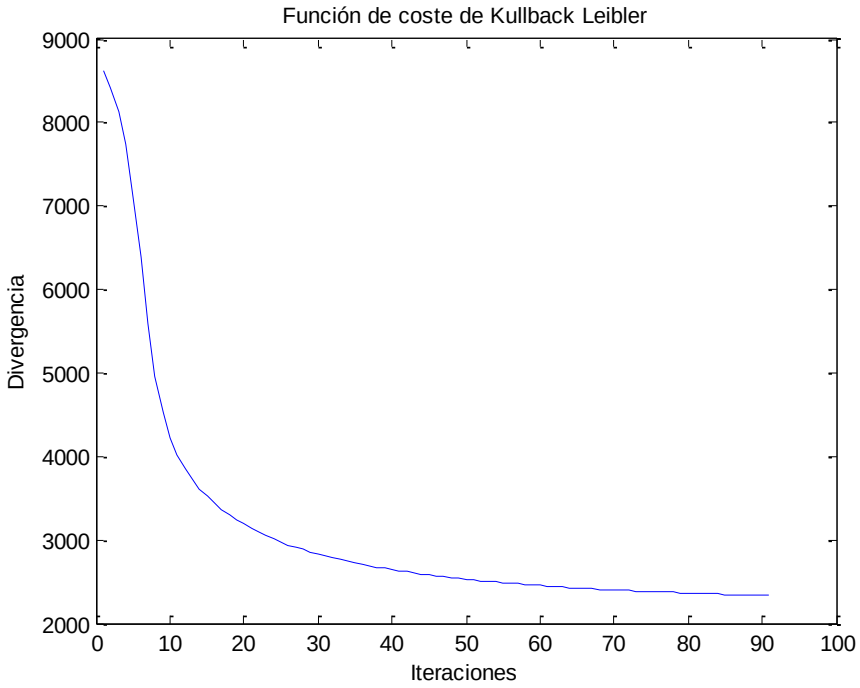

Figura 87: Divergencia de Kullback-Leibler con 91 iteraciones.

Después de realizar 91 iteraciones, la divergencia permanece constante, obteniéndose las siguientes matrices W y H (figura 88). En ellas se puede ver cómo se comportan en frecuencia las bases y durante que instantes de tiempo están activas.

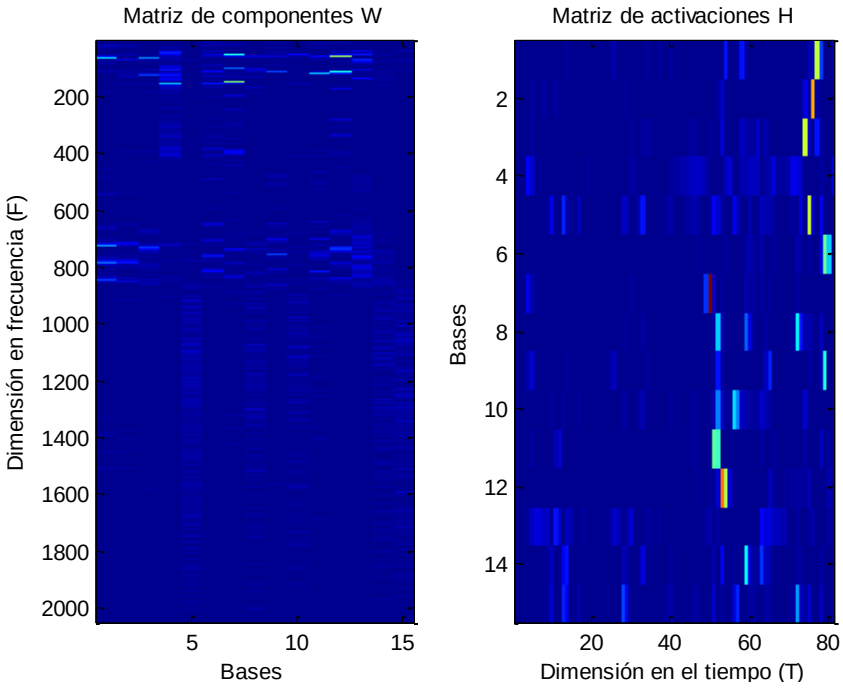

Figura 88: Matrices W y H después de 91 iteraciones.

El resultado del producto de las dos matrices anteriores da lugar al espectrograma reconstruido de la siguiente figura, que es similar al espectrograma original de los sonidos armónicos más la voz (figura 84):

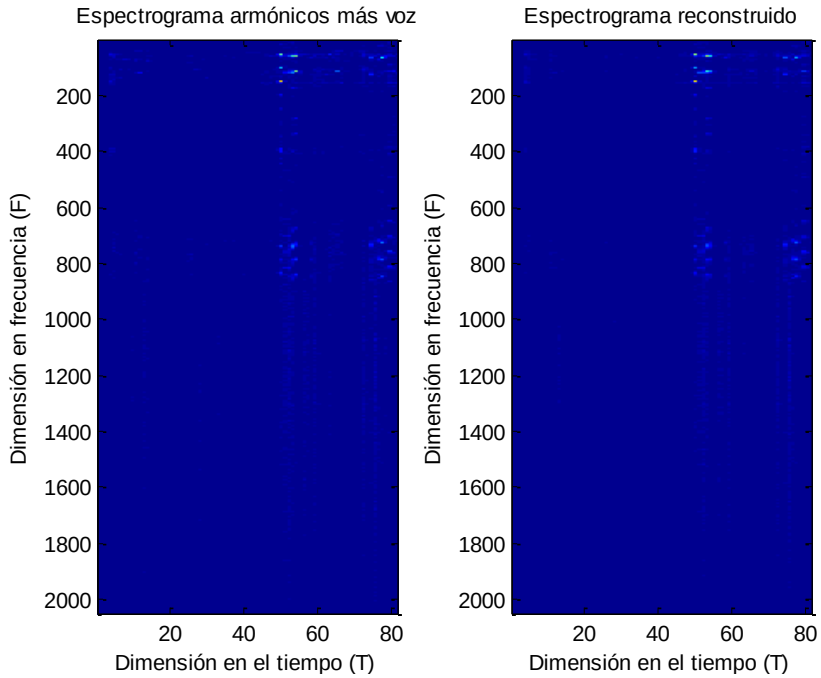

Figura 89: Espectrograma de los armónicos más la voz y su reconstrucción.

Umbralización en frecuencia

Figura 90: Umbralización en frecuencia (Método Cortolargo).

13. El siguiente paso es la umbralización, que como se ha utilizado una ventana larga para realizar el espectrograma, se realiza umbralización en frecuencia. Para saber las bases que se corresponden con los sonidos armónicos se calcula la ecuación 28.

Una vez que se ha calculado la ecuación 28 para cada una de las bases, se calcula una nueva señal X' (voz) mediante la ecuación 29, en la que se eliminan las componentes armónicas de la señal X (armónicos más voz). Esta operación se realiza mediante la siguiente ecuación:

Por lo que, solo queda calcular la señal formada por todas las componentes armónicas, esta operación se realiza calculando el sumatorio que hay en la ecuación 29, en el que se suman los productos matriciales de cada una de las bases consideradas como armónicas por sus correspondientes activaciones.

Este es el resultado de aplicar umbralización en frecuencia con un umbral de 0,4.

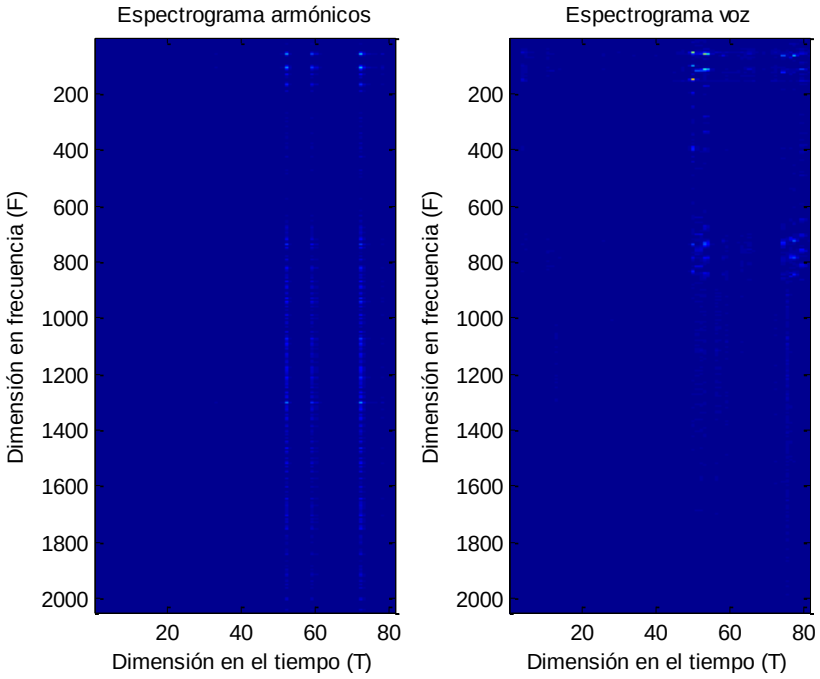

Figura 91: Resultado de la umbralización espectral.

Armónicos **Voz**

Figura 92: Armónicos y voz en el dominio del tiempo (Método Cortolargo).

14. Ya solo queda pasar los espectrogramas al dominio del tiempo, para lo que se realiza la transformada inversa de la transformada rápida de Fourier (IFFT), teniendo en cuenta que hay que multiplicar cada espectrograma por la fase del espectrograma original (armónicos más voz). Los parámetros necesarios son los mismos que para la FFT, es decir, la frecuencia de muestreo, la longitud de la ventana, el solapamiento y el tamaño de los segmentos de la señal.

15. Tras el paso anterior ya se tiene la señal original separada en tres pistas musicales (sonidos percusivos, sonidos armónicos y sonidos vocales). Ahora es el momento comprobar la calidad de la separación, para ello hay que calcular las medidas de calidad objetivas (SDR, SIR y SAR). La pista de los sonidos armónicos y la de los percusivos hay que sumarlas, porque no se dispone de los sonidos armónicos originales separados y de los sonidos percusivos originales separados, sino que se dispone del todo acompañamiento musical original mezclado. A la pista de la voz separada no hay que hacerle nada, ya que sí se dispone de la voz original sin música.

5. EVALUACION

5.1. Base de datos

No existe una base de datos estándar para evaluar los sistemas de separación de fuentes musicales. Sin embargo, hay un conjunto de bases de datos que son muy empleadas en este ámbito, su rigor está avalado y evaluado por la comunidad científica.

Las bases de datos se dividen en dos tipos [27], aquellas que están formadas por sonidos sintéticos y aquellas que están formadas por sonidos reales. El primer tipo son bases de datos formadas por señales cuyas notas han sido generadas por un sintetizador de ficheros simbólicos MIDI, los cuales son capaces de generar sonidos casi idénticos a los de un instrumento real. El segundo tipo son bases de datos de sonidos reales, que han sido generados por instrumentos musicales reales tocados por músicos profesionales. La principal característica de este tipo de bases de datos es que se tienen las particularidades de la interpretación del músico y del cantante, además de los rasgos acústicos de la sala y de los propios instrumentos. Sin ninguna duda es más conveniente trabajar con este tipo de bases ya que se asemejan altamente a un escenario real.

A su vez, las bases de datos de sonidos reales se dividen en dos tipos [27]. Un primer tipo que está formado por sonidos individuales, es decir, están compuestas por un conjunto de notas musicales tocadas de manera aislada por diferentes instrumentos, familias de instrumentos o estilos. Las notas se pueden encontrar en un orden aleatorio o siguiendo algún tipo de criterio (número de notas concurrentes, número de instrumentos), para así obtener distintos tipos de mezclas polifónicas. Además, las notas se encuentran aisladas, por lo que se puede analizar por separado cada nota del rango dinámico de un instrumento. Algunas de las bases de este tipo más utilizadas son the Musical Instrument Sound Database of Real World Computing (RWC) [39] o the McGill University's Master Samples (MUMS) [40].

El segundo tipo de bases de datos de sonidos reales son aquellas que están formadas por composiciones musicales. Este tipo de bases son muy utilizadas para la evaluación de los sistemas de separación de fuentes sonoras en escenas con diferentes interpretaciones (distintos instrumentos, distintas acústicas de sala, distinto nivel de polifonía…). Una base muy utilizada es the Classical Music Database and the Jazz Music Database de Real World Computing (RWC) [39].

Para este Trabajo Fin de Grado se va a utilizar la base de datos conocida como MIR-1K [25], que corresponde al grupo de bases de datos de sonidos reales formadas por composiciones musicales. El nombre de MIR-1K está formado por dos partes, MIR que

proviene de las siglas en ingles de “Recuperación de información multimedia” (Multimedia Information Retrieval), y 1K que significa 1000, esto es porque esta base de datos está formada por 1000 trozos de canciones o clips.

La MIR-1K ha sido diseñada para la investigación en la separación de la voz cantada. Estos 1000 clips tienen una duración de entre 4 (el más corto) y 13 segundos (el más largo), siendo la duración total de todos los clips de 133 minutos. Los clips provienen de 110 canciones de karaoke chinas estéreo, grabándose el acompañamiento musical en el canal izquierdo y la voz cantada en el derecho. Las canciones han sido seleccionadas de una base de datos aún mayor, formada por 5000 canciones de pop chino, las cuales han sido nuevamente interpretadas por 11 hombres y 8 mujeres, la mayoría de los cuales son amateur y no tienen una formación musical previa.

La nomenclatura de los clips tiene un orden y un sentido, por ejemplo en el clip abjones_1_01, abjones es el nombre del cantante, 1 es el número de canción interpretada por el cantante, y 01 es el número del clip. Entonces si se tiene el fichero amy_2_05, se puede saber que el nombre de la intérprete es amy, que es su segunda canción y que es el quinto fragmento de esa canción.

Además también se dispone de las canciones completas interpretadas por cada cantante y de la letra de cada canción leída por su correspondiente interprete.

5.2. Métricas

Para evaluar la calidad del algoritmo de separación es necesario ayudarse de un conjunto de medidas objetivas que permitirán obtener un valor numérico de dicha calidad de la separación. Estas medidas que se denominan SDR, SIR y SAR, son generadas por la función *bss_eval_images_nosort* que se encuentra en la toolbox de Matlab BSS_EVAL [26].

El objetivo de este método es descomponer el error total entre la señal estimada por el sistema $\hat{s}_j(t)$ y la señal real $s_j(t)$, en tres términos que se refieren a tres tipos de error:

$$\hat{s}_j(t) - s_j(t) = e_j^{interf}(t) + e_j^{noise}(t) + e_j^{artif}(t) \tag{32}$$

Donde $e_j^{interf}(t)$ es el término de error por las interferencias generadas por el resto de fuentes en la fuente j, $e_j^{noise}(t)$ es el error debido a la deformación del ruido de perturbación, y $e_j^{artif}(t)$ es el error debido a los artefactos creados por el algoritmo de separación, los cuales son generados de forma artificial.

Dependiendo de las relaciones entre los diferentes términos de la ecuación, se obtiene las medidas que se están buscando:

- Signal to Distorsion Ratio (SDR):

$$SDR = 10\log_{10}\left(\frac{\left\|s_j(t)\right\|^2}{\left\|e_j^{interf}(t) + e_j^{noise}(t) + e_j^{artif}(t)\right\|^2}\right) \tag{33}$$

Esta medida dice cómo se parece la forma de onda de la señal original a la de la fuente estimada en el proceso de separación. Será mayor cuanto más se parezcan ambas señales.

- Signal to Interference Ratio (SIR):

$$SIR = 10\log_{10}\left(\frac{\left\|s_j(t)\right\|^2}{\left\|e_j^{interf}(t)\right\|^2}\right) \tag{34}$$

Esta medida dice si se han introducido interferencias entre las fuentes, por ejemplo, que en la voz se hayan introducido algunas componentes de los sonidos armónicos. Será mayor cuantas menos interferencias haya.

- Signal to Artifacs Ratio (SAR):

$$SAR = 10\log_{10}\left(\frac{\left\|s_j(t) + e_j^{interf}(t) + e_j^{noise}(t)\right\|^2}{\left\|e_j^{artif}(t)\right\|^2}\right) \tag{35}$$

Esta medida muestra si se han generado muchos artefactos durante el proceso de separación. Será mayor cuantos menos artefactos se hayan generado.

5.3. Optimización

Para saber cuál es la combinación de parámetros con la que se obtiene los mejores resultados, se ha realizado un proceso de optimización hiperparamétrico del método NMF-KL-FT. Durante este proceso, se ha realizado un barrido de todos los parámetros posibles que afectan al algoritmo, es decir, las longitudes de las ventanas cortas y largas, el número de bases y los umbrales en frecuencia y en el tiempo. Además del orden en que se aplica el enventanado, es decir, si primero se aplica el espectrograma con ventana larga y después con ventana corta (método Largocorto), o si primero se aplica el espectrograma con ventana corta y posteriormente con ventana larga.

En cuanto al número de bases se han seleccionado 6 valores múltiplos de 5, que son 5, 10, 15, 20, 25 y 30 bases.

Como tamaños de ventana larga y de ventana corta se han utilizado aquellos que han sido explicados en el apartado 4.3, los cuales son de 64 ms, 128 ms, 256 ms y 512 ms para la ventana larga. Y de 8 ms, 16 ms, 32 ms y 64 ms para la ventana corta.

Como umbrales en frecuencia y umbrales en el tiempo se han utilizado el mismo rango de valores, estos son 0’1, 0’3, 0’5, 0’7 y 0’9.

Se han utilizado concretamente estos valores porque este Trabajo Fin de Grado se ha reproducido el algoritmo NMF-KL-FT [12] y en este algoritmo se han utilizado estos mismos valores, en cuanto a bases y tamaños de ventana se refiere. En cuanto a los valores de los umbrales se han utilizado la mitad, ya que en [12] estos valores van de 0,1 hasta 1 con saltos de 0,1, es decir 10 valores diferentes. La razón de este cambio se debe al tiempo de ejecución del proceso de optimización, ya que de esta forma se ha reducido a una cuarta parte.

Agrupando todo se obtiene el siguiente esquema:

- 2 métodos diferentes.
- 4 longitudes para la ventana larga.
- 4 longitudes para la ventana corta.
- 5 posibles umbrales en frecuencia.
- 5 posibles umbrales en el tiempo.
- 6 valores de bases.

Si se multiplican todos los posibles valores que pueden adoptar los parámetros, se obtiene que el número total de combinaciones es de 4800 (2*4*4*5*5*6), habiendo 2400 combinaciones para el método Largocorto y otras 2400 combinaciones para el método Cortolargo.

Este barrido de parámetros se ha realizado para 89 canciones, que son todas las que se corresponden con el intérprete “khair” y con la interprete “titon”, que son un hombre y una mujer, respectivamente.

Para cada fragmento de canción (por ejemplo khair_1_01) se ha creado una estructura denominada “supermatriz”, en la que se almacena toda la información necesaria para cada una de las distintas separaciones. La información almacenada para cada una de las distintas combinaciones de parámetros es la siguiente:

- **Ordenventana:** es el método utilizado, es decir, si es el método Largocorto o el Cortolargo.
- **VentanaL**: es el tamaño de la ventana larga y se introduce en número de muestras.
- **VentanaC**: es el tamaño de la ventana corta y se introduce en número de muestras.
- **Umbrals**: es el valor del umbral en frecuencia.
- **Umbralt**: es el valor del umbral en el tiempo.
- **Bases**: es el número de bases.
- **IteracionesLarga**: es el número de iteraciones realizadas durante las reglas de actualización utilizando el espectrograma con ventana larga.
- **IteracionCorta**: es el número de iteraciones realizadas durante las reglas de actualización utilizando el espectrograma con ventana corta.
- **DistorsionLarga**: es la divergencia de Kullback-Leibler para cada una de las iteraciones con ventana larga.
- **DistorsionCorta**: es la divergencia de Kullback-Leibler para cada una de las iteraciones con ventana corta.
- **SDR**: son las medidas de SDR para la música y para la voz.
- **SIR**: son las medidas de SIR para la música y para la voz.
- **SAR**: son las medidas de SAR para la música y para la voz.

Una vez que se ha realizado el barrido de parámetros para cada uno de los fragmentos de canciones, la información de cada una de las “supermatrices” se almacena en una estructura aún mayor, denominada “matriztotal”, en la que se añade la siguiente información:

- **Nombre:** El nombre del fragmento de la canción con el que se corresponde esa separación.

- **Contador:** es el número de la fila en la que está ubicada esa combinación en su "supermatriz" correspondiente.

Después de almacenar toda la información en la estructura "matriztotal", se obtiene una estructura con 427200 filas (4800 combinaciones de cada canción por 89 canciones) con todos sus correspondientes campos, anteriormente explicados. De cada una de las separaciones se guardan como ficheros de audio en una carpeta, que tiene el mismo nombre de la canción original, los sonidos armónicos, los sonidos percusivos, los sonidos vocales y los sonidos resultantes de la suma de los armónicos y los percusivos. Por lo tanto en cada una de las carpetas habrá 19202 archivos, que se corresponden con las 4 pistas separadas para cada combinación, además de la canción original y su correspondiente estructura "supermatriz".

Todo este proceso ha generado 430 GB de datos y el tiempo total de ejecución ha sido de 18 días y medio, tardando cada fragmento una media de 5 horas en ser separado con todas las combinaciones de parámetros.

Como los títulos de este Trabajo Fin de Grado y de [12] indican, lo que se busca es la separación de la voz cantada de las señales monoaurales, por tanto para decidir cuál es la mejor combinación de parámetros se van a tener en cuenta las medidas de calidad objetiva que relacionan la voz separada con la voz original. Dentro de las tres medidas explicadas en el apartado anterior, el SAR no se va a tener en cuenta para la obtención de la combinación óptima, ya que esta medida solo calcula la cantidad de artefactos que se han generado durante el proceso de separación. Por el contrario con el SDR y el SIR sí se puede calcular si la separación es buena, puesto que el SDR calcula cuánto se parece la voz separada a la voz original y el SIR calcula cuantas interferencias de los sonidos armónicos y de los sonidos percusivos se han colado en la voz separada.

En primer lugar se va a calcular cual es método con el que mejores resultados se obtienen, si el Largocorto o el Cortolargo. Así pues se va a realizar la media del SDR de las 89 canciones separadas para cada una de las posibles combinaciones de parámetros y para cada uno de los dos métodos propuestos.

En la siguiente figura están representadas cada una de las medias del SDR. Las primeras 2400 combinaciones se corresponden con el método Largocorto, mientras que las 2400 últimas se corresponden con el método Cortolargo. Tanto para el Largocorto como para el Cortolargo, la forma en la que se distribuyen los parámetros es la misma. Esta distribución es la siguiente:

- Primero se varían las bases, desde 5 hasta 30 con saltos de 5.
- Después se varían los umbrales temporales, desde 0,1 hasta 0,9 con saltos de 0,2.
- Luego se varían los umbrales en frecuencia, desde 0,1 hasta 0,9 con saltos de 0,2.

- Posteriormente se varía el tamaño de la ventana corta, desde los 8 ms hasta los 64 ms, pasando por 16 ms y 32 ms.
- Por último, se varía el tamaño de la ventana larga, desde los 64 ms hasta los 512 ms, pasando por 128 ms y 256 ms.

Entonces por ejemplo la combinación número 3 se corresponde con 15 bases, un umbral en el tiempo de 0,1, un umbral en frecuencia de 0,1, una tamaño de ventana corta de 8 ms y un tamaño de ventana larga de 64 ms. La combinación número 1000 se corresponde con 20 bases, 0,3 de umbral en el tiempo, 0,7 de umbral en frecuencia, 32 ms de tamaño de ventana corta y 128 ms de tamaño de ventana larga.

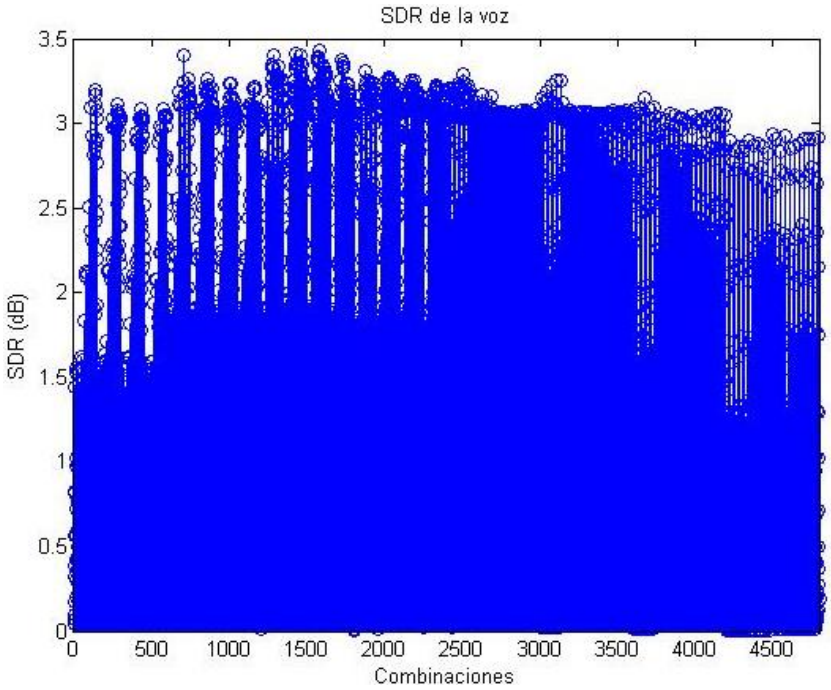

Figura 93: SDR de la voz.

En esta figura se puede observar como hay muchos valores medios de SDR de las primeras 2400 combinaciones (método Largocorto) que son superiores a los valores más altos de las 2400 combinaciones correspondientes al método Cortolargo. Los mejores resultados del método Largocorto están en torno a los 3,25 dB, habiendo picos en torno a los 3,4 dB. Mientras que para el método Cortolargo los mejores resultados están en torno a los 3,05 dB. Por tanto se llega a la conclusión de que el método Largocorto es el método con el que se obtienen mejores resultados.

Ahora se va a comprobar si este método también es el mejor en cuanto al SIR de la voz. En la figura 94 se ha representado la media del SIR para cada una de las combinaciones de parámetros de las 89 canciones separadas. Los valores del SIR de la voz para el método Largocorto (primeras 2400 combinaciones) son considerablemente

superiores a los valores del SIR para el método Cortolargo. Los valores del método Largocorto son mayoritariamente positivos, estando la mayoría concentrados en torno a los 2,75 dB y habiendo picos entre los 5 dB y los 6 dB. En cambio los mejores resultados del método Cortolargo están en torno a los 2 dB y además la mayoría de los valores son negativos, llegando incluso hasta los -7 dB. Por tanto se llega a la misma conclusión que en el análisis del SDR, con el método Largocorto se obtiene mejores resultados que con el método Cortolargo.

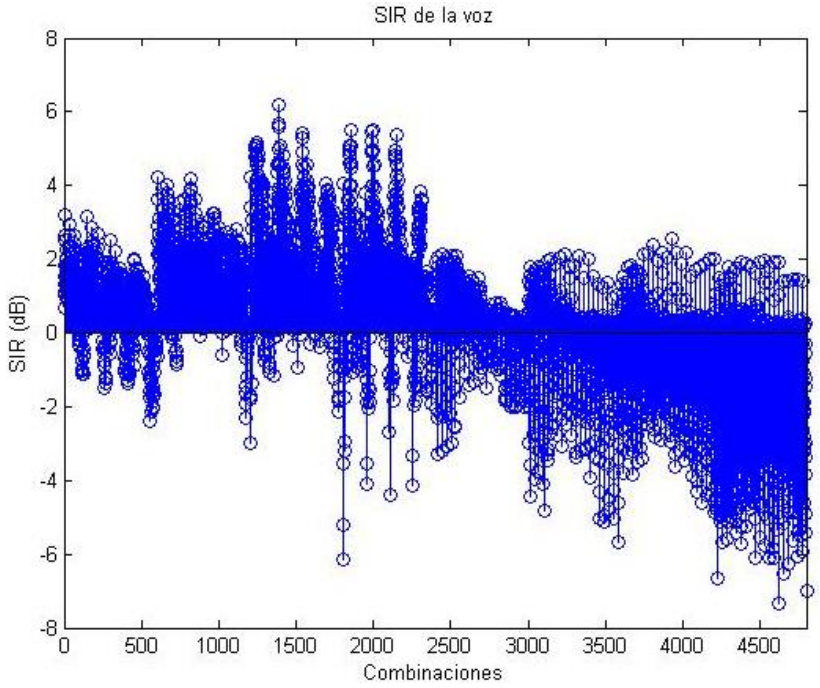

Figura 94: SIR de la voz.

La razón por la que con el método Largocorto se obtienen mejores resultados que con el método Cortolargo se debe a las características de la voz. Durante la separación con ventana larga en el método Largocorto, la voz se comporta de forma continua en frecuencia, como los sonidos percusivos, esto se debe a la existencia de los fonemas sordos. Por consiguiente, si se aplica el método Cortolargo, en la primera etapa (espectrograma con ventana corta) esos fonemas sordos se clasificarán como sonidos percusivos, por lo que la voz en la etapa con ventana larga no se comportará de forma continua en frecuencia y se clasificará como sonidos armónicos, produciéndose una separación errónea.

Una vez que se ha demostrado que el método Largocorto es con el que mejores resultados se obtienen, es el momento de calcular cuál es la combinación de parámetros con la que se obtiene la separación óptima.

Para demostrar cuál es la combinación óptima, primero se ha tenido en cuenta solo el SDR de la voz para el método Largocorto. En la figura 95 se ha representado el SDR de

la voz y se puede observar que la combinación con la que se obtiene un SDR de la voz mayor es la número 1581, cuyo valor es de 3,43 dB. Esta combinación se corresponde con 15 bases, una ventana larga de 4096 muestras, un umbral en frecuencia de 0,5, una ventana corta de 512 muestras y un umbral en el tiempo de 0,7.

El valor del SIR para esta combinación, la número 1581, es de 2,65 dB, que como se puede ver en la figura 96, no es el mejor valor ni tampoco está cercano a los valores más altos de SIR. Por tanto esta combinación de parámetros no se escoge como la óptima, ya que aunque se tenga un SDR máximo (es la combinación con la que más se parece la voz separada a la voz original), el SIR no es lo suficientemente bueno y por tanto habrá interferencias de los sonidos armónicos y los percusivos en la señal de voz.

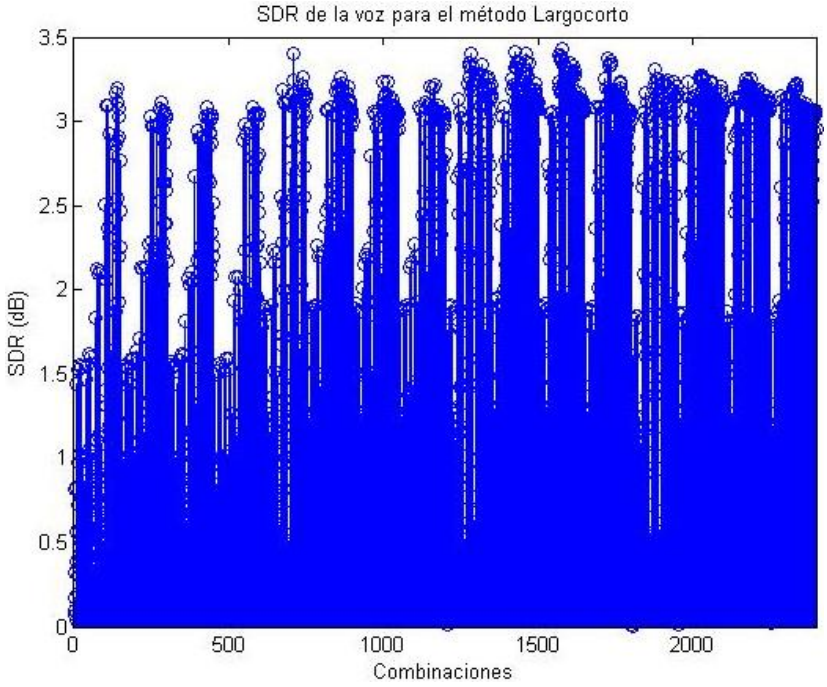

Figura 95: SDR de la voz para el método Largocorto.

Si por el contrario, se toma como combinación óptima aquella con la que se obtiene un SIR máximo, la combinación óptima sería la número 1392 con un SIR de 6,21 dB. Esta combinación se corresponde con 30 bases, una ventana larga de 4096 muestras, un umbral en frecuencia de 0,3, una ventana corta de 256 muestras y un umbral en el tiempo de 0,3. El valor de SDR para esta combinación es de 0,99 dB y al igual que en el caso anterior es un valor que no es cercano a los valores más grandes. Esta combinación también se descarta para ser la óptima, ya que aunque haya muy pocas interferencias de los sonidos armónicos y de los percusivos en la voz, la señal de voz resultante no se parecerá por completo a la voz original.

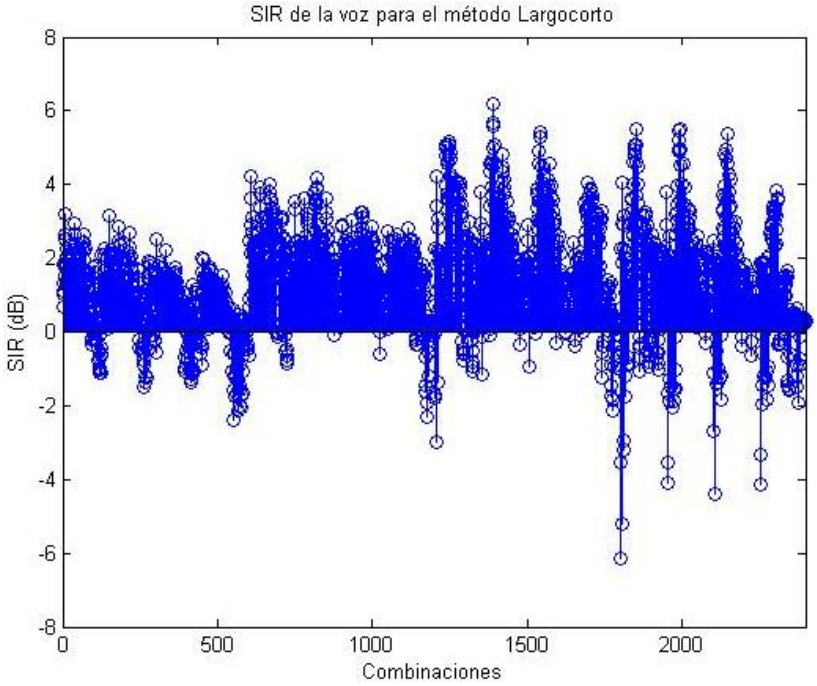

Figura 96: SIR de la voz para el método Largocorto.

La combinación de parámetros que se va a adoptar como óptima es aquella en la que la voz separada cumpla las siguientes características:

- Que tenga pocas interferencias debido a las componentes instrumentales (SIR alto).
- Que se parezca considerablemente a la voz original (SDR alto).

Para ello se va a escoger aquella combinación que tenga a la vez el mejor valor de SDR y de SIR. Esto se consigue realizando un barrido por las 2400 posibles combinaciones, tomando como referencia los valores mínimos de SDR y de SIR. Si para una combinación ambos valores son superiores a los tomados como referencia, esos valores pasan a ser los valores de referencia. Cuando finaliza el barrido se obtiene cuál es la combinación en la que ambos valores son máximos a la vez.

En este caso la combinación en la que tanto el SIR como el SDR son valores altos es la número 1388. Esta combinación se corresponde con 10 bases, una ventana larga de 4096 muestras, un umbral en frecuencia de 0,3, una ventana corta de 256 muestras y un umbral en el tiempo de 0,3. En este caso el valor de SDR es de 2,66 dB y el de SIR es de 4,65 dB. Como se puede observar en las dos figuras anteriores (figura 95 y figura 96), estos valores están 0,77 dB y 1,56 dB por debajo de los valores máximos, respectivamente.

Aunque los valores estén por debajo de los máximos, la calidad de la separación en su conjunto va a ser mejor si se realiza con esta combinación de parámetros que si se tiene en cuenta el SDR máximo o el SIR máximo.

5.4. Resultados

En este apartado se va a realizar la separación de los 911 fragmentos de canción de la base de datos MIR-1K que no se han utilizado durante el proceso de optimización, con la combinación de parámetros que se ha demostrado que es la óptima. Al igual que se ha hecho en el proceso de optimización, la información de cada una de las 911 separaciones se va a guardar en una estructura cuyos campos son los siguientes:

- Nombre: es el nombre del fragmento que se va a separar.

- Ordenventana: es el método utilizado para la separación, en este caso es el Largocorto.

- VentanaL: es el tamaño de la ventana que se ha utilizado para realizar el espectrograma durante la etapa larga, en este caso es de 4096 muestras (256 ms).

- VentanaC: es el tamaño de la ventana que se ha utilizado para realizar el espectrograma durante la etapa corta, en este caso es de 256 muestras (16 ms).

- Umbrals: es el umbral en frecuencia, en este caso es de 0,3.

- Umbralt: es el umbral en el tiempo, en este caso es de 0,3.

- Bases: es el número de bases que se van a utilizar, en este caso 10.

- IteracionLarga: es el número de veces que se han aplicado las reglas de actualización de Kullback-Leibler utilizando el espectrograma con ventana larga.

- IteracionCorta: es el número de veces que se han aplicado las reglas de actualización de Kullback-Leibler utilizando el espectrograma con ventana corta.

- DistorsionLarga: es el valor de la divergencia de Kullback-Leibler para cada una de las iteraciones de las reglas de actualización con ventana larga.

- DistorsionCorta: es el valor de la divergencia de Kullback-Leibler para cada una de las iteraciones de las reglas de actualización con ventana corta.

- SDR: es el valor de SDR tanto para la música como para la voz.

- SIR: es el valor de SIR tanto para la música como para la voz.

- SAR: es el valor de SAR tanto para la música como para la voz.

El proceso de separación se va a realizar con tres relaciones de amplitud entre la música original y la voz original distintas. Primero estando la voz original atenuada 5 dB con respecto a la música original (5.4.1), después siendo la relación entre ellas igual a 0 dB (5.4.2) y por último, estando la voz original amplificada 5 dB con respecto a la música original (5.4.3).

5.4.1. Separación a -5 dB

Ahora se va a realizar la separación de los 911 fragmentos, con la característica de que la voz original está atenuada 5 dB con respecto a la música original. En la siguiente figura se puede ver el diagrama de cajas de las medidas de la voz.

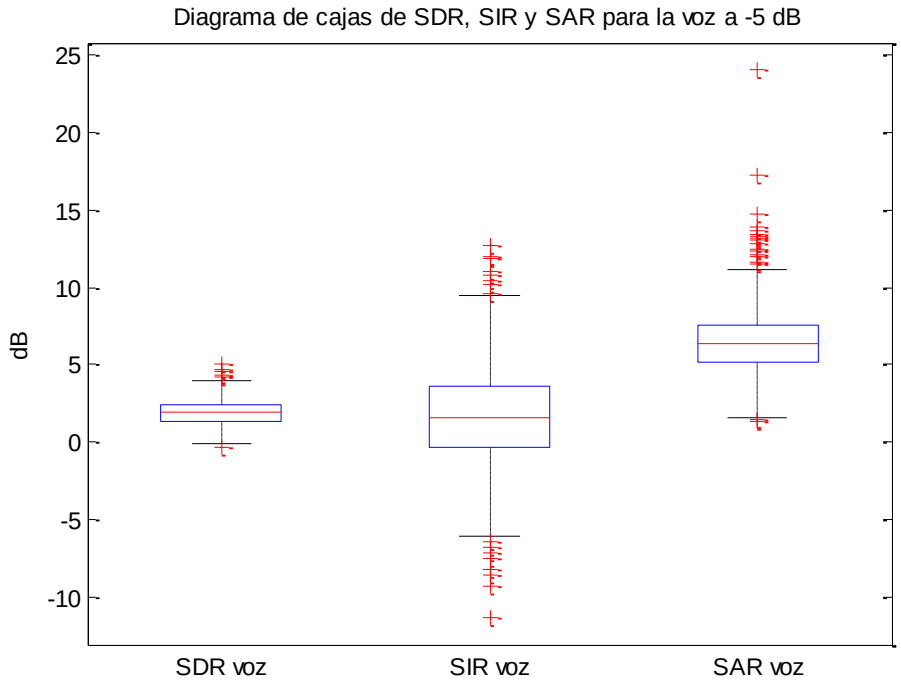

Figura 97: Diagrama de cajas de SDR, SIR y SAR para la voz a -5 dB.

El rango del SDR va desde los -0,3 dB hasta los 5,08 dB, teniendo una mediana de 1,88 dB. El SIR va desde los -11,3 dB hasta los 12,72 dB, siendo la mediana de 1,55 dB. La separación no introduce muchos artefactos ya que los valores del SAR son positivos, siendo su mediana de 6,41 dB.

Si se compara este diagrama con el diagrama de la figura 99 (voz a 0 dB), se puede ver que en conjunto, la calidad de la voz separada es menor cuando la voz original se atenúa con respecto a la música original. Esto es así porque las componentes de la música que se cuelan en la pista de la voz separada, aunque sean muy pocas, tienen más potencia que las componentes de la voz, por lo que la calidad de la voz separada baja conforme se atenúa la voz original con respecto a la música original.

En la siguiente figura se muestra el diagrama de cajas de las medidas de calidad objetivas de la música.

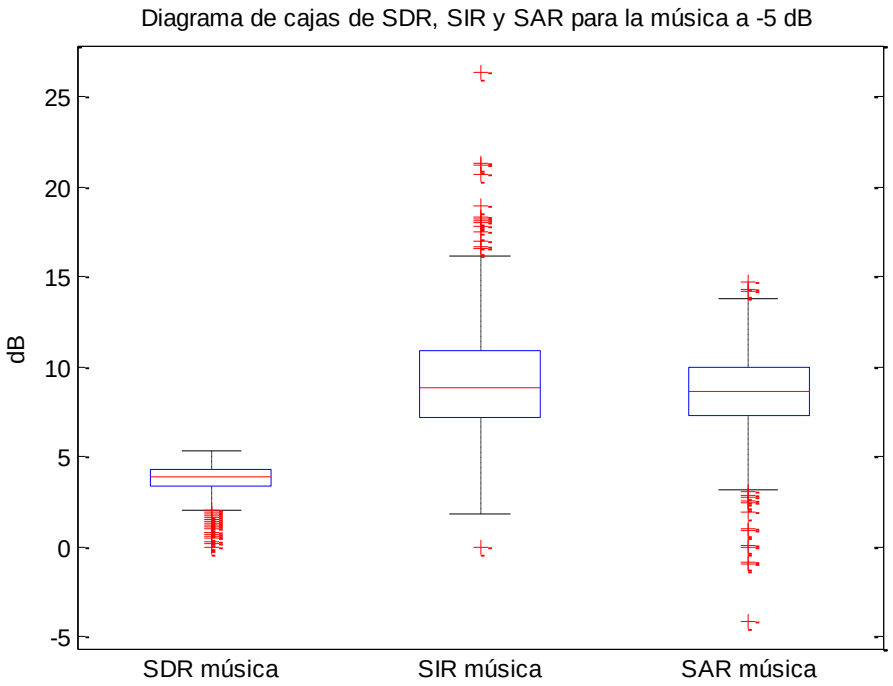

Figura 98: Diagrama de cajas de SDR, SIR y SAR para la música a -5 dB.

El SDR va desde los 0 dB hasta los 5,3 dB, teniendo una mediana de 3,86 dB. El SIR va desde los 0 dB hasta los 26,35 dB, con una mediana de 8,87 dB. El SAR tiene una mediana de 8,58 dB.

Si se compara con la figura 105 (música a 0 dB), se observa que al atenuar la voz original se obtiene una mejor separación de la música. En todas las pistas que contienen a la música separada se cuelan algunas componentes correspondientes a la voz, pero cuanto más atenuada esté la voz original con respecto a la música original, estas componentes serán menos influyentes y la calidad objetiva de la música separada será mejor.

5.4.2. Separación a 0 dB

En la siguiente figura (figura 99) está representado el diagrama de cajas del SDR, del SIR y del SAR de la voz separada. En él se puede observar que el rango que ocupa el SDR va desde los 0,3 dB hasta los 5,6 dB, teniendo una mediana de 3,07 dB. El rango de variación del SIR es considerablemente más grande, ya que va desde los -4,4 dB hasta los 16,5 dB, siendo la mediana de 5,83 dB. En cuanto al SAR, aunque no se haya tenido en cuenta para calcular cual era la combinación de parámetros con la que mejor resultados se obtenían, su rango va desde los 2,1 dB hasta los 24,3 dB, con una mediana de 9,13 dB. Los valores de SAR son todos positivos, lo que significa que el algoritmo utilizado no introduce demasiados artefactos que hagan que las señales se distorsionen.

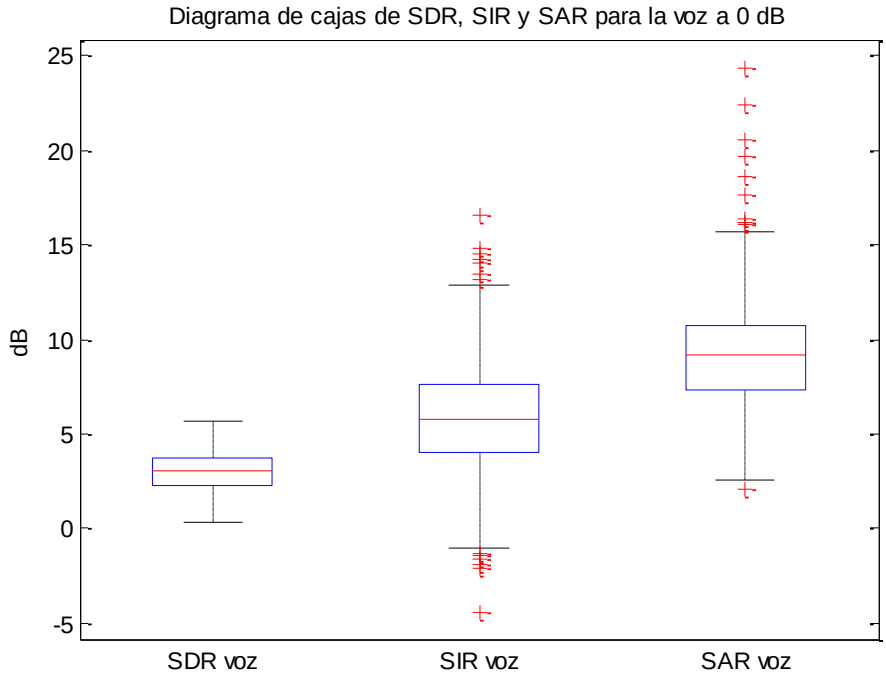

Figura 99: Diagrama de cajas del SDR, SIR y SAR para la voz a 0 dB.

Si observamos el SIR, se puede ver que hay 2 valores que destacan por encima del resto, uno por superior a los demás y otro por ser inferior a los demás. El valor más alto de SIR (16,57 dB) se corresponde con el fragmento “amy_13_04”, que como era de suponer es el fragmento que tiene también el mayor valor de SDR (5,56 dB). Este fragmento está interpretado por una mujer. Una de las características de este fragmento que hace que la calidad de la separación sea la mejor se puede ver en la siguiente figura. Esta característica es que la mayor parte de la voz está grabada con una amplitud ligeramente superior a la amplitud de la mayor parte del acompañamiento musical. Este hecho hace que en el proceso de separación, la voz se atenúe completamente en la pista de la música separada (sonidos armónicos más sonidos percusivos), y como consecuencia toda la voz aparezca en la pista de la voz separada, en la cual también aparecen componentes musicales, pero más atenuadas que las componentes de la voz.

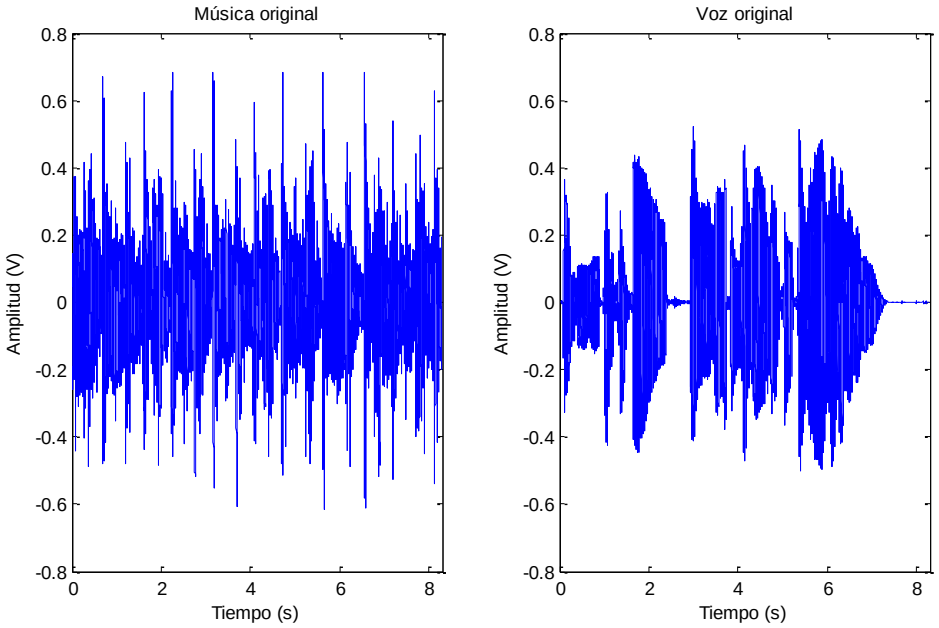

Figura 100: Amplitud de la música original y de la voz original del fragmento “amy_13_04”.

El valor más bajo de SIR es de -4,41 dB y se corresponde con el fragmento “bug_4_03”, que a su vez tiene un valor de SDR de la voz de 0,3 dB, siendo este el valor más bajo de SDR de todos los fragmentos separados. Este fragmento está interpretado por un hombre. Las razones por las que este fragmento es el que peor calidad tiene son las siguientes. Por un lado, la voz del intérprete en la canción original está grabada con una intensidad menor que la música (figura 101). Por lo que si se escucha la canción, lo que se aprecia es que la voz suena más atenuada que la música, pareciendo que la voz es lo que complementa a la música y no al revés, como sucede en la mayor parte de los otros fragmentos.

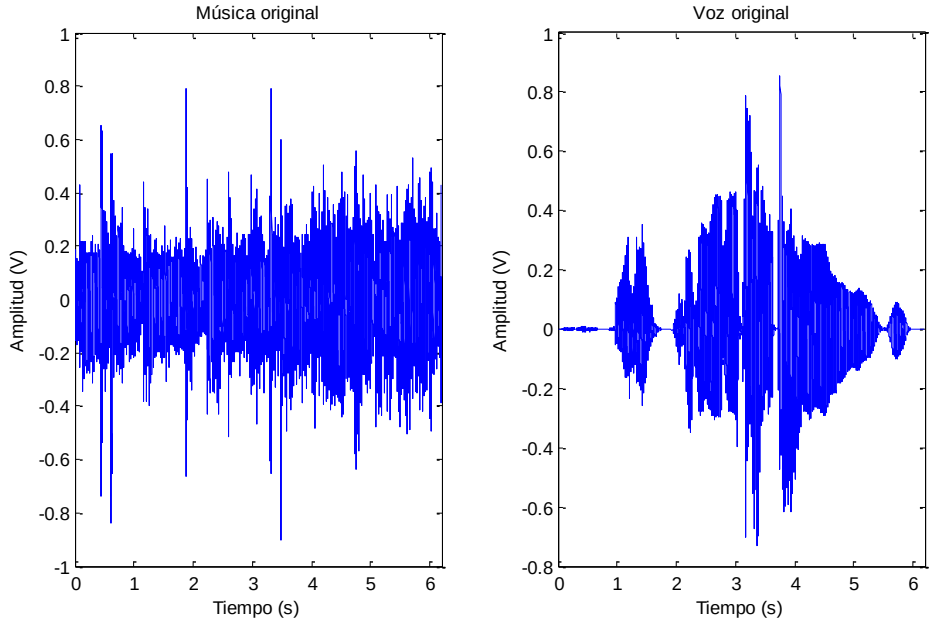

Figura 101: Amplitud de la música original y de la voz original del fragmento “bug_4_03”.

Pero la baja intensidad de la voz no es la causa principal, lo que afecta gravemente a la separación es que en la canción original, la voz del intérprete está acompañada por más voces realizando los coros. Pero estas voces no están grabadas en el canal derecho en la canción original, sino que están grabadas en el canal izquierdo con la música. Por tanto, cuando de calculan las medidas entre la pista de la voz separada y la voz original, los resultados son bajos, ya que en la pista de la voz separada el algoritmo además de la voz del interprete también ha introducido las voces de los coros.

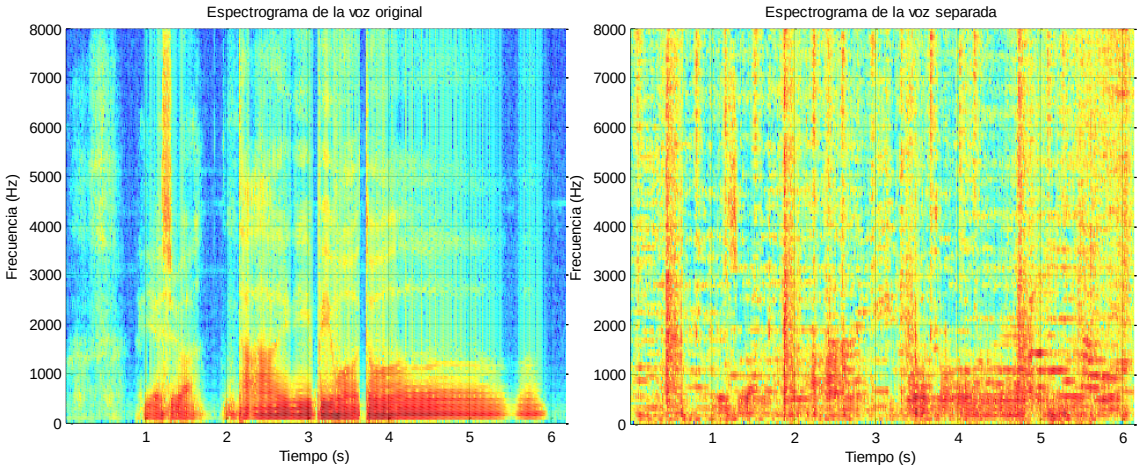

Figura 102: Espectrogramas de la voz original y de la separada del fragmento “bug_4_03”.

En la figura anterior se puede ver lo que se ha explicado, en el espectrograma de la voz original se ve que solo aparece la voz cantada por el intérprete, mientras que en el espectrograma de la voz separada aparte de que se cuelan componentes correspondientes a la música, también aparecen las componentes correspondientes a los coros, las cuales presentan una ligera periodicidad en frecuencia, por lo que también aparecen en la pista de los sonidos armónicos.

A continuación se va a realizar una separación de los fragmentos en dos grupos, un grupo donde estarán los fragmentos interpretados por los hombres y en otro grupo los fragmentos interpretados por mujeres. Esta tarea se va a efectuar para comprobar si se puede generalizar lo que anteriormente se ha demostrado, de que el fragmento con mejor calidad es el interpretado por una mujer y el de peor calidad es el interpretado por un hombre.

En la siguiente figura se puede ver el diagrama de cajas de las medidas de la voz de los fragmentos que han sido interpretados por hombres, los cuales son en total 546 fragmentos.

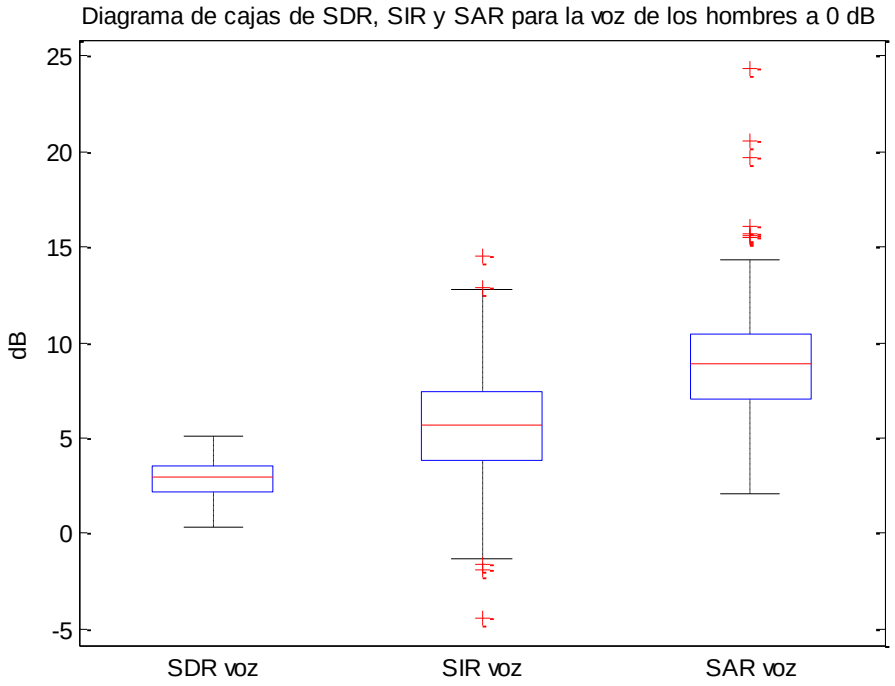

Figura 103: Diagrama de cajas del SDR, SIR y SAR para la voz de los hombres a 0 dB.

Como se puede observar, la mediana del SDR en la figura anterior es de 2,96 dB, teniendo un rango que va desde los 0,3 dB hasta los 5,08 dB. La mediana del SIR es de 5,65 dB, con un rango comprendido entre los -4,41 dB y los 14,58 dB. A continuación, se va a comparar con el diagrama de la voz de las mujeres, cuyos fragmentos son 365 en total.

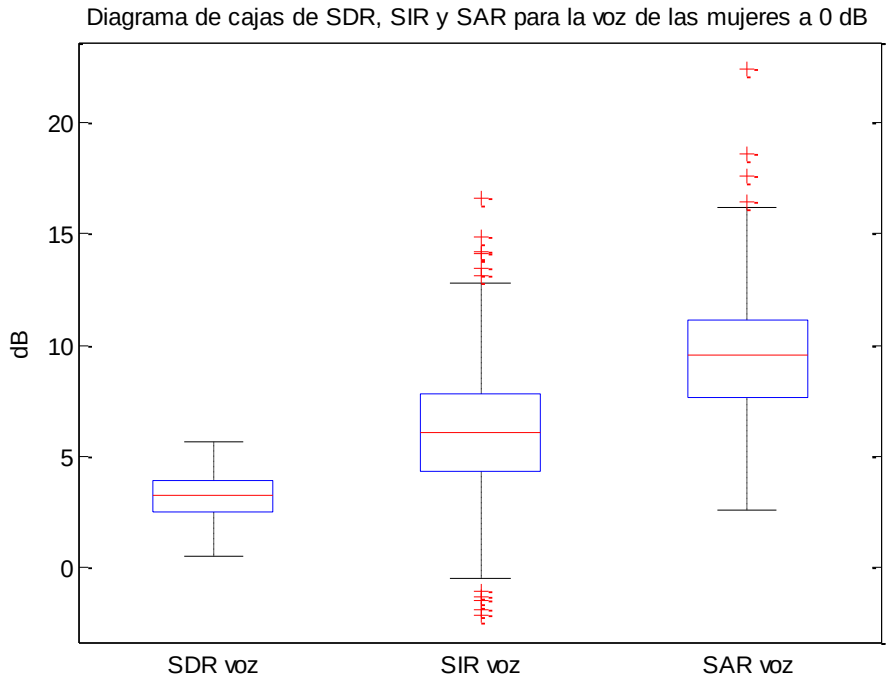

Figura 104: Diagrama de cajas del SDR, SIR y SAR para la voz de las mujeres a 0 dB.

La mediana del SDR es de 3,25 dB, con un rango que va desde los 0,55 dB a los 5,65 dB. La mediana del SIR es de 6,06 dB, con un rango comprendido entre los -2,12 dB y los 16,57 dB.

Comparando entre los resultados obtenidos en la figura 103 (fragmentos interpretados por hombres) y los resultados de la figura 104 (fragmentos interpretados por mujeres) se llega a la conclusión de que el algoritmo funciona mejor si la voz de las canciones está interpretada por una mujer que si está interpretada por un hombre.

En la siguiente figura se muestra el diagrama de cajas del SDR, SIR y SAR de la música separada de los 911 fragmentos.

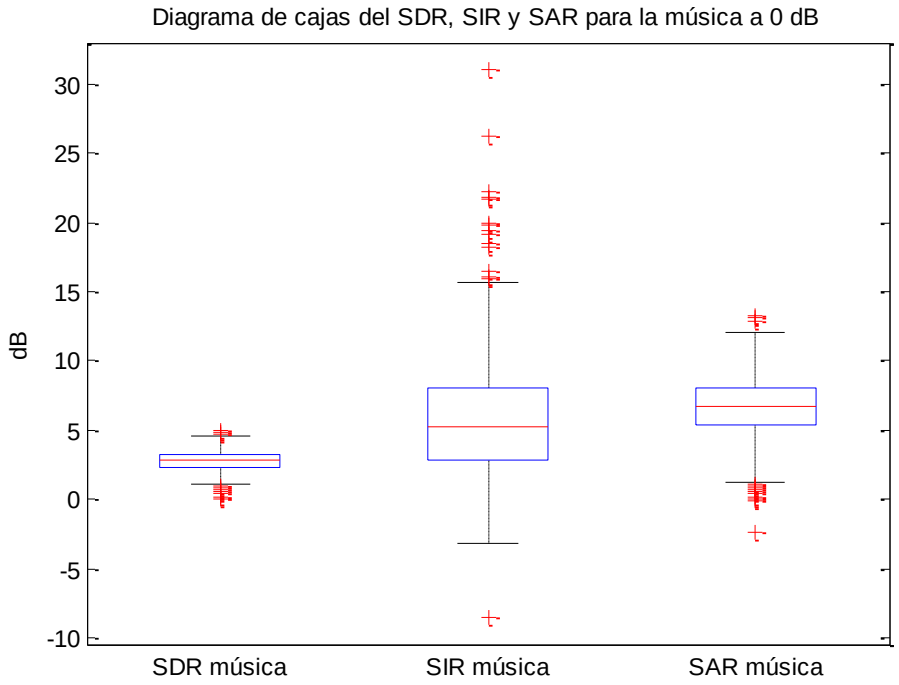

Figura 105: Diagrama de cajas del SDR, SIR y SAR para la música a 0 dB.

Se observa que el rango que ocupa el SDR es muy compacto, va desde los 0 dB hasta los 5 dB, teniendo una mediana de 2,87 dB. En cambio, el rango del SIR es 8 veces mayor, siendo su límite inferior los -8,49 dB y su límite superior los 31 dB, siendo su mediana de 5,25 dB. En cuanto al SAR, su rango va desde los -2,42 dB hasta los 13,24 dB, siendo su mediana de 6,77 dB.

En la figura se ve que hay un valor de SIR que destaca sobre el resto por ser extremadamente bajo. Este valor se corresponde con el fragmento “geniusturtle_5_03”, que tiene un valor de SDR de la música de 0,02 dB. La característica que hace que este fragmento tenga una mala calidad en la separación está en el acompañamiento musical.

Si se escucha la pista que contiene la música original y además se observa el espectrograma de la siguiente figura, se puede ver que hay un instrumento que no aparece

en ningún otro fragmento que tiene un comportamiento que no se corresponde ni con los sonidos armónicos estudiados ni con los sonidos percusivos.

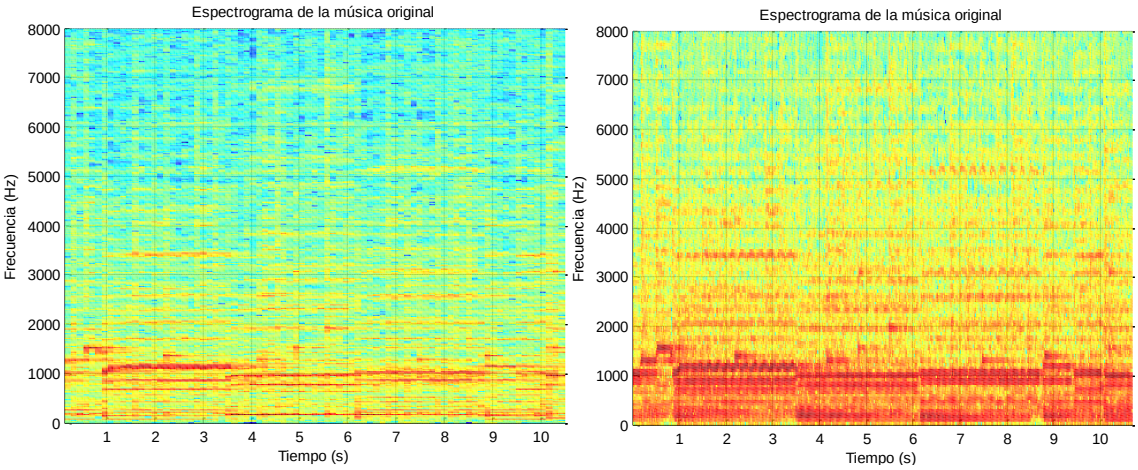

Figura 106: Espectrograma de la música original del fragmento “geniusturtle_5_03” con una ventana de 4096 muestras (izquierda) y de 256 muestras (derecha).

Al observar el espectrograma con ventana corta de 256 muestras, se puede ver como hay un instrumento cuyas componentes espectrales no son líneas horizontales continuas en el tiempo y discontinuas en frecuencia (sonidos armónicos), ni tampoco son líneas verticales continuas en frecuencia y discontinuas en el tiempo (sonidos percusivos). Como se puede ver en la figura 81, las componentes de dicho instrumento forman dientes de sierra, asemejándose al efecto conocido como vibrato, lo que significa que hay fluctuaciones en un rango muy pequeño de frecuencia a lo largo del tiempo. Un comportamiento que se asemeja más al comportamiento espectral de la voz, tal y como se puede ver en la figura 107.

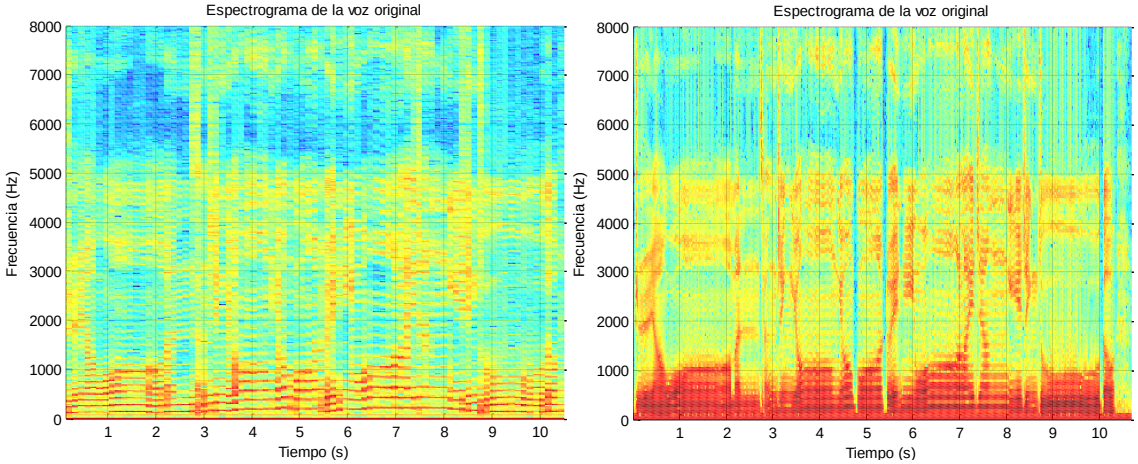

Figura 107: Espectrograma de la voz original del fragmento “geniusturtle_5_03” con una ventana de 4096 muestras (izquierda) y de 256 muestras (derecha).

Por lo tanto al realizar la separación de la canción, la pista de la música (figura 108) no contendrá las componentes que se identifican con este peculiar instrumento. Esta pista solo contendrá las componentes de los sonidos armónicos, ya que como se ha visto en la figura 106, en este fragmento no se disponen de sonidos producidos por instrumentos de percusión.

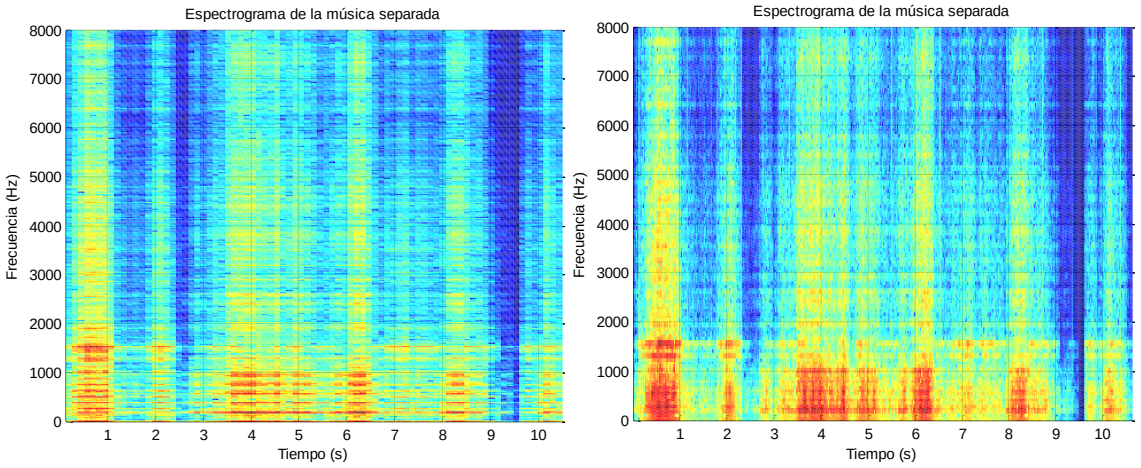

Figura 108: Espectrograma de la música separada del fragmento “geniusturtle_5_03” con una ventana de 4096 muestras (izquierda) y de 256 muestras (derecha).

Entonces las componentes del instrumento estarán en la pista de la voz separada, tal y como se puede ver en la siguiente figura.

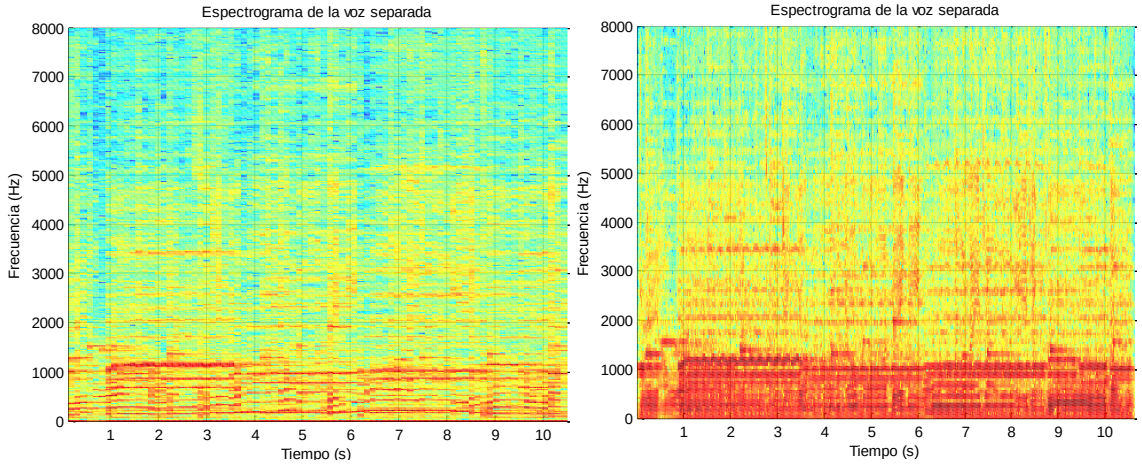

Figura 109: Espectrograma de la voz separada del fragmento “geniusturtle_5_03” con una ventana de 4096 muestras (izquierda) y de 256 muestras (derecha).

5.4.3. Separación a 5 dB

Por último, se va a realizar la separación de los 911 fragmentos con la característica de que la voz está amplificada 5 dB con respecto a la música original. En la siguiente figura se puede ver el diagrama de cajas de las medidas de la voz.

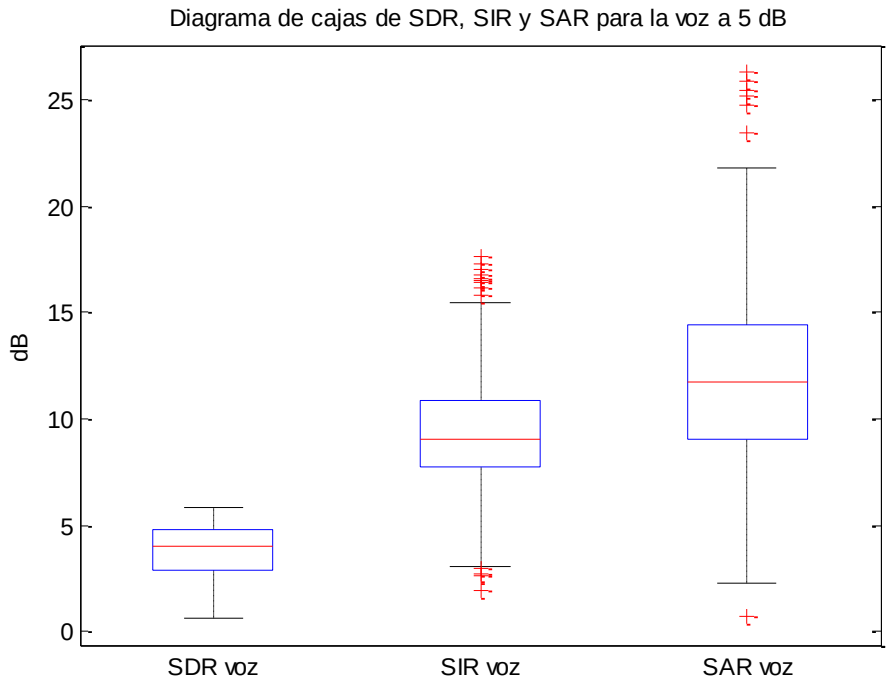

Figura 110: Diagrama de cajas de SDR, SIR y SAR para la voz a 5 dB.

Los valores del SDR van desde los 0,66 dB hasta los 5,83 dB, siendo la mediana de 4,01 dB. El rango del SIR va desde los 1,97 dB hasta los 17,66 dB, teniendo una mediana de 9,07 dB. El SAR va desde los 0,68 dB hasta los 26,29 dB, siendo la mediana de 11,75 dB. De nuevo los valores del SAR son positivos, teniendo una mediana de 11,75 dB.

Si se comparan estos resultados con los del diagrama de la figura 99 (voz a 0 dB), se observa que la calidad de la voz separada cuando la voz original está amplificada con respecto a la música original es mejor. Esto se debe a que cuanto mayor sea la potencia de la voz original con respecto a la música original, las componentes correspondientes a la música que se cuelan en la pista de la voz separada serán menos influyentes y la calidad de la voz será mejor.

A continuación se muestra el diagrama de cajas de las medidas de la música.

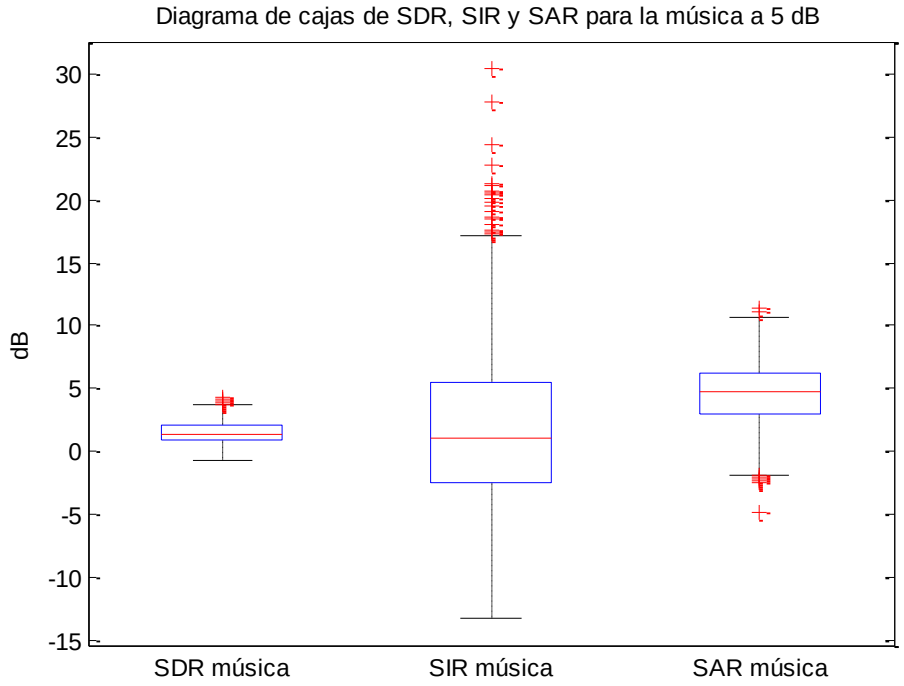

Figura 111: Diagrama de cajas de SDR, SIR y SAR para la música a 5 dB.

El SDR va desde los -0,69 dB hasta los 4,28 dB, siendo su mediana de 1,46 dB. El SIR va desde los -13,13 dB hasta los 30,34 dB, teniendo una mediana de 1,05 dB. En este caso el SAR tiene más valores negativos que el resto de diagramas que se han realizado, por lo que esta es la separación en la que más artefactos introduce el algoritmo, aunque su mediana es de 4,73 dB.

Al comparar este diagrama con el de la figura 105 (música a 0 dB), se observa que la calidad de la música, cuando la voz original esta amplificada con respecto a la música original, es peor. Esto se debe a que las componentes de la voz que se cuelan en la pista de la música separada, al estar amplificadas tienen mayor influencia, empeorando la calidad de la música separada.

5.5. Comparación con los resultados del método NMF-KL-FT [12]

En este apartado se van a comparar los resultados obtenidos, tras realizar el proceso de optimización y los procesos de separación con diferentes atenuaciones y amplificaciones, con los resultados que se muestran en [12].

En la siguiente tabla se pueden ver las combinaciones óptimas obtenidas tras el proceso de optimización en este Trabajo Fin de Grado y en [12].

	Trabajo Fin de Grado	NMF-KL-FT [12]
Método	Largocorto	Largocorto
Bases	10	15
Tamaño de ventana larga	256 ms	256 ms
Umbral en frecuencia	0,3	0,4
Tamaño de ventana corta	16 ms	16 ms
Umbral en el tiempo	0,3	0,2
SDR medio de la voz	2,66 dB	1,6 dB (etapa larga) 2,3 dB (etapa corta)
SIR medio de la voz	4,65 dB	_

Tabla 1: Comparativa de los parámetros óptimos y de las medidas de calidad objetivas de la voz entre el algoritmo de este Trabajo Fin de Grado y el algoritmo NMF-KL-FT [12].

La diferencia de los parámetros en ambos algoritmos se debe a que el criterio que han escogido los autores de NMF-KL-FT [12] para la selección de los parámetros óptimos solo se centra en el valor máximo del SDR de la voz, por eso el valor del SIR medio de la voz está vacío. Mientras que en este Trabajo Fin de Grado el criterio escogido es que el SDR y el SIR sean los más altos posibles a la vez.

Otra peculiaridad que tiene el algoritmo NMF-KL-FT [12], es que sus autores para la realización del barrido de parámetros no han realizado el método Largocorto y el método Cortolargo completos. Sino que han realizado la separación con la ventana larga y han calculado las medidas de calidad objetivas de la voz resultante. Y por otro lado han realizado la separación con la ventana corta y han calculado las medidas de la voz. Por esta razón hay dos valores del SDR de la voz en la figura anterior.

En la siguiente figura se muestran las medidas de calidad objetivas de la voz, obtenidas con el método NMF-KL-FT [12]. Las medidas que se corresponden con el método Largocorto son las de la columna cuyo nombre es “L+S”. Hay que tener en cuenta que en los diagramas de cajas que han publicado, han omitido los valores denominados “outliers”, que son los valores superiores e inferiores al bigote superior e inferior de cada diagrama, respectivamente. También hay que tener en cuenta que los valores numéricos que aparecen se corresponden con las medianas de cada una de las medidas. Por lo que en la tabla 2 se va a mostrar una tabla con los resultados obtenidos para la voz en este Trabajo Fin de Grado, siendo la mediana de cada medida su valor representativo.

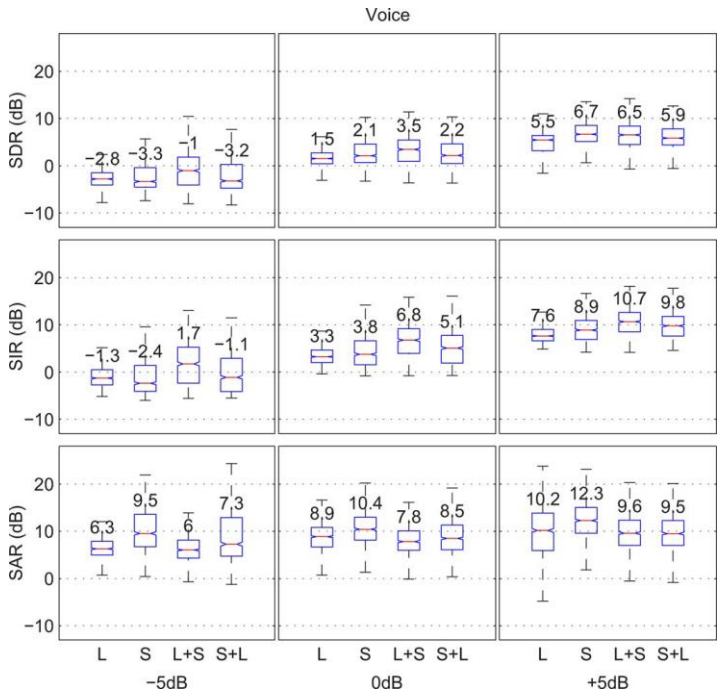

Figura 112: Medidas de calidad objetivas de la voz del método NMF-KL-FT [12].

	-5 dB	**0 dB**	**5 dB**
SDR	1,88 dB	3,07 dB	4,01 dB
SIR	1,55 dB	5,83 dB	9,07 dB
SAR	6,41 dB	9,3 dB	11,75 dB

Tabla 2: Medidas de calidad objetivas de la voz de este Trabajo Fin de Grado.

Después de comparar ambos conjuntos de medidas se llegan a las siguientes conclusiones:

- Cuando la voz original está atenuada 5 dB con respecto a la música original (-5 dB), los parámetros propuestos en este Trabajo Fin de Grado permiten obtener una mejor calidad de la voz que con los parámetros propuestos en NMF-KL-FT [12], ya que teniendo el mismo grado de interferencias debido a la música, las señales se parecen más ya que el SDR es mayor.
- Cuando la voz ni se atenúa ni se amplifica (0 dB), se obtienen mejores resultados con los parámetros en NMF-KL-FT [12] que con los de este Trabajo Fin de Grado, ya que las señales se parecen más a la original y hay menos interferencias por parte de la música.
- Por último, cuando la voz original está amplificada 5 dB con respecto a la música original (5dB), también se obtienen mejores resultados con el método propuesto en NMF-KL-FT [12].

- La característica que es unánime para las tres separaciones es que el algoritmo de este Trabajo Fin de Grado introduce menos artefactos en la señal de la voz separada que el algoritmo NMF-KL-FT [12].

Ahora se va a realizar para la música la misma comparativa que se ha hecho para la voz. Pero esta vez la columna en la que hay que fijarse es en la que tiene el nombre “P”.

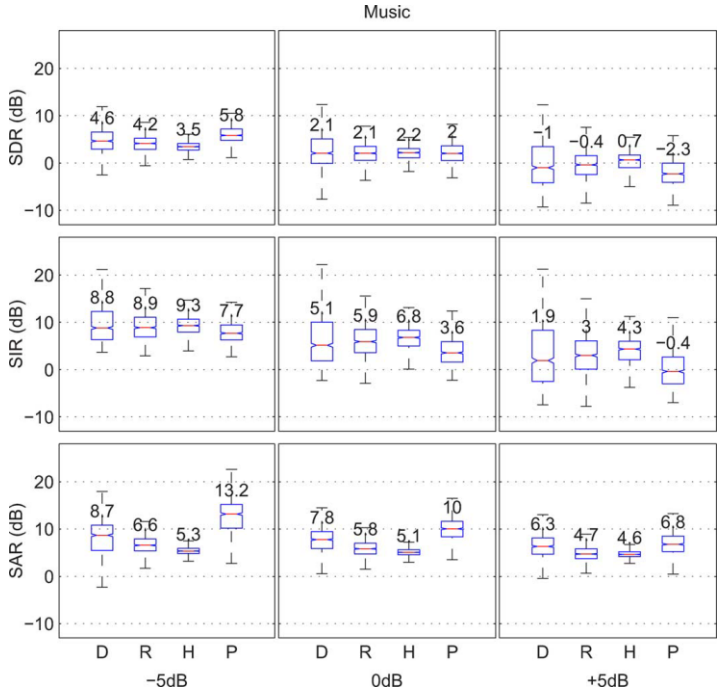

Figura 113: Medidas de calidad objetivas de la música del método NMF-KL-FT [12].

	-5 dB	**0 dB**	**5 dB**
SDR	3,86 dB	2,87 dB	1,46 dB
SIR	8,87 dB	5,25 dB	1,05 dB
SAR	8,58 dB	6,77 dB	4,73 dB

Tabla 3: Medidas de calidad objetivas de la música de este Trabajo Fin de Grado.

Después de comparar ambos conjuntos de medidas se llegan a las siguientes conclusiones:

- Cuando la voz original está atenuada 5 dB con respecto a la música original (-5 dB), se obtienen resultados opuestos con ambos métodos, ya que el algoritmo NMF-KL-FT [12] consigue que aun teniendo la señal de la música separada más interferencias que la señal que se obtiene con el algoritmo de este Trabajo Fin de Grado, esta señal se parece mucho más a la original.
- Cuando la voz ni se atenúa ni se amplifica (0 dB), se obtienen mejores resultados con los parámetros de este Trabajo Fin de Grado que con los del método NMF-

KL-FT [12], ya que las señales obtenidas se parecen más a la original y tienen menos interferencias por parte de la voz.

- Por último, cuando la voz original está amplificada 5 dB con respecto a la música original (5dB), también se obtienen mejores resultados con el método propuesto en este Trabajo Fin de Grado. La característica que es unánime para las tres separaciones es que el algoritmo de este Trabajo Fin de Grado introduce más artefactos en la señal de la música separada que el algoritmo del método NMF-KL-FT [12].

6. CONCLUSIONES

El objetivo fundamental de este Trabajo Fin de Grado era la separación de una señal de audio monocanal en tres pistas musicales distintas, es decir, en armónicos, percusivos y voz. Además de este objetivo, los objetivos específicos propuestos también han sido conseguidos.

Para ello se ha realizado un barrido de parámetros, tras el cual se ha demostrado que la combinación de parámetros óptima es la siguiente:

- **Método:** Largocorto.
- **Bases:** 10.
- **Tamaño de la ventana larga:** 256 ms.
- **Umbral en frecuencia:** 0,3.
- **Tamaño de la ventana corta:** 16 ms.
- **Umbral en el tiempo:** 0,3.

Después se han realizado tres procesos de separación. Tras realizar la separación a 0 dB se han llegado a las siguientes conclusiones:

- El algoritmo propuesto permite obtener mejores resultados si las canciones están interpretadas por mujeres que si están interpretadas por hombres.
- Cuando en las canciones aparte del acompañamiento musical y de la voz del intérprete, también aparecen voces en forma de coros, la calidad de la separación baja, ya que tienen un ligero comportamiento armónico que hace que algunas de sus componentes aparezcan en la señal de los sonidos armónicos.
- Si se dispone de instrumentos cuyo comportamiento no se asemeja ni al de los sonidos armónicos ni al de los sonidos percusivos, la calidad de la separación bajará ya que las componentes de este instrumento se introducirán en la señal de la voz separada.

Tras realizar la separación a -5 dB se ha llegado a la siguiente conclusión:

- Cuanto más atenuada esté la voz original con respecto a la música original, la calidad de la voz separada será peor, ya que las posibles componentes de la música que se cuelen en la voz tendrán mucha influencia, disminuyendo la calidad. Mientras que la calidad de la música será mejor, ya que las componentes de la voz que se cuelen en la música tendrán una escasa influencia en la medición de la calidad.

Tras realizar la separación a 5 dB se ha llegado a la siguiente conclusión:

- Cuanto más amplificada esté la voz original con respecto a la música original, la calidad de la voz separada será mejor, ya que las posibles componentes de la música que se cuelen en la voz tendrán una escasa influencia, lo que mejora la calidad. Mientras que la calidad de la música será peor, ya que las componentes de la voz que se cuelen en la música tendrán una gran influencia en la medición de la calidad, empeorándola.

7. LÍNEAS DE FUTURO

En este Trabajo Fin de Grado se ha buscado cual es la combinación de parámetros óptima para realizar la separación de una señal monocanal utilizando las técnicas NMF. Pero el conocimiento que se ha adquirido en la realización de este Trabajo Fin de Grado puede ser trasladado y ampliado a otras aplicaciones y escenarios.

Uno de estos escenarios puede ser la separación de señales multicanal, en la que se tiene un problema, ya que se añade la dimensión espacial, pero este a su vez da información espacial previa. Esto permite a priori saber la distribución de cada fuente, es decir, con qué micrófono se ha grabado cada fuente. Una característica de estas señales es que habitualmente cada micrófono no graba un solo instrumento, sino que cada fuente está formada por un grupo de instrumentos del mismo tipo. Debido a la información que se antes de realzar el algoritmo de separación, con estas señales se pueden conseguir resultados bastante sorprendentes.

Otra aplicación que está siendo muy demandada por la industria musical es el alineamiento entre la música y la partitura. En este campo todavía queda trabajo por realizar. Relacionado con dicha industria también está el acompañamiento automático a un instrumentista, que requiere de una síntesis previa de los instrumentos.

Una aplicación que está altamente relacionada con este Trabajo Fin de Grado consiste en la identificación de los instrumentos que forman las señales de armónicos y percusivos, y además poder mostrar las notas musicales de cada uno durante la melodía. Con la voz también se puede realizar el mismo proceso, consistente en la extracción de los fonemas y la representación de la letra de la canción.

Por último, las técnicas de separación, en concreto la que se ha utilizado en este Trabajo Fin de Grado (NMF), pueden ser utilizadas para mejorar la calidad de las señales. Por ejemplo, se podría separar la voz o la música del ruido, o por lo menos atenuarlo lo máximo posible. También se podría realizar la separación de distintas voces que aparecen de forma simultánea en una señal.

8. BIBLIOGRAFÍA

[1] Juan Sebastián Guevara Sanín. *Teoría de la música*. 2010

[2] Constantino Pérez Vega. *Sonido y Audición*. Universidad de Cantabria.

[3] A. Klapuri. *Sound onset detection by applying psychoacoustic knowledge*. IEEE International Conference on Acoustics, Speech and Signal Processing, 1999.

[4] A. Klapuri and M. Davy. *Signal Processing Methods for Music Transcription*. New York: Springer, 2006.

[5] A. Klapuri. *Signal Processing Methods for the Automatic Transcription of Music*. Tampere University of Technology, 2004.

[6] A. Bregman. Auditory Scene Analysis: *The Perceptual Organization of Sound*. MIT Press, 1990.

[7] Yipeng Li. *Monaural Musical Sound Separation*. The Ohio State University, 2008.

[8] M. Ryynänen and A. Klapuri. *Automatic transcription of melody, bass line, and chords in polyphonic music*. Computer Music Journal, 2008.

[9] M. D. Plumbley, S. A. Abdallah, J. P. Bello, M. E. Davies, G. Monti and M. B. Sandler. *Automatic Music Transcription and Audio Source Separation*. Cybernetics and Systems, 2002.

[10] T. Rossing. *The science of sound*. Addison Wesley, 1990.

[11] L. Regnier and G. Peeters. *Singing Voice Detection in Music Tracks Using Direct Voice Vibrato Detection*. IRCAM, 2009.

[12] Bilei Zhu, Wei Li, Ruijiang Li and Xiangyang Xue. *Multi-Stage Non-Negative Matrix Factorization for Monaural Singing Voice Separation*. IEEE Transactions on Audio, Speech and Language Processing, vol. 21, no. 10, October 2013.

[13] A. Ozerov and C. Févotte. *Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation*. IEEE Transactions on Audio, Speech and Language Processing, vol. 18, no. 3, March 2010.

[14] T. Virtanen. *Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria*. IEEE Transactions on Audio, Speech and Language Processing, March 2007.

[15] J. Fritsch and M.D. Plumbley. *Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis*. IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[16] C. Févotte, N. Bertin and J Durrieu. *Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis*. Neural Computation, 2009.

[17] Nishikawa, T. Abe, H. Saruwatari and H. Shikano. *Overdetermined blind source separation of real acoustic sounds based on multistage ICA using subarray processing*. Signal Processing and Information Technology, 2003. ISSPIT 2003. Proceedings of the 3rd IEEE International Symposium on.

[18] A. Hyvärinen and E. Oja. *Independent component analysis: algorithms and applications*. Neural networks, 13(4-5):411– 430, 2000.

[19] Z. Rafii and B. Pardo. *REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation*. IEEE Transactions on Audio, Speech and Language Processing, vol. 21, no. 1, January 2013.

[20] H. Laurberg. *Non-negative Matrix Factorization: Theory and Methods*. Aalborg University, 2008.

[21] D. D. Lee and H. S. Seung. *Learning the parts of objects by non-negative matrix factorization*. Nature, vol. 401, no. 6755, pp. 788-791, 1999.

[22] Youngmin Cho and Lawrence K. Saul. *Nonnegative Matrix Factorization for Semi-supervised Dimensionality Reduction*. University of California.

[23] D. D. Lee and H. S. Seung. *Algorithms for non-negative matrix factorization*. Adv. Neural Inf. Process. Syst., vol. 13, pp. 556–562, 2001.

[24] T. Virtanen. *Monaural Sound Source Separation by Non-negative Matrix Factorization With Temporal Continuity and Sparseness Criteria*. IEEE Transactions on Audio, Speech and Language Processing. (Vol: 15, issue: 3). March 2007.

[25] https://sites.google.com/site/unvoicedsoundseparation/mir-1k

[26] http://bass-db.gforge.inria.fr/bss_eval

[27] Francisco Javier Rodríguez Serrano. *Separación de fuentes sonoras en señales musicales*. Universidad de Jaén. 2014. ISBN: 978-84-8439-878-3.

[28] http://dle.rae.es/?id=Qsyp7LZ

[29] http://elrinconmusicaleduca.blogspot.com.es/2011/11/los-elementos-principales-que-nos.html

[30] http://solfamidas362.blogspot.com.es/2014/03/sesion-3-1-10-2013-con-francisco.html

[31] http://bibliotecadelfos.blogspot.com.es/2015/10/cualidades-del-sonido-duracion.html

[32] http://es.slideshare.net/Montacon/apresentao-sobre-ruido

[33] https://es.wikipedia.org/wiki/Sonoridad_(sicoac%C3%BAstica)

[34] https://gl.wikipedia.org/wiki/Curva_isof%C3%B3nica

[35] http://es.slideshare.net/JIMENAMARGARITAHERNANDEZ/el-aparato-fonador-48851245

[36] https://www.youtube.com/watch?v=vXJYrrLGNAo

[37] http://www.domingo-roman.net/vocales_esp_caract_acustica.html

[38] https://www.internationalphoneticassociation.org/content/ipa-pulmonic-consonants

[39] M. Goto, H Hashiguchi, T Nishimura, and R Oka. *RWC music database: Popular, classical, and jazz music database*. Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR), 2002.

[40] F Opolko and J Wapnick. *McGill University Master Samples*. McGill University, 1992.

ANEXO 1: MANUAL DE USUARIO

Para mostrar el algoritmo que se ha desarrollado en este Trabajo Fin de Grado, se ha realizado una interfaz gráfica (GUI) con la ayuda del software de MATLAB. Una de las ventajas de utilizar este programa es que tiene una gran cantidad de funciones desarrolladas, lo que supone acortar el tiempo de implementación requerido por el desarrollador de la interfaz.

Cuando se inicia la interfaz aparece la siguiente figura:

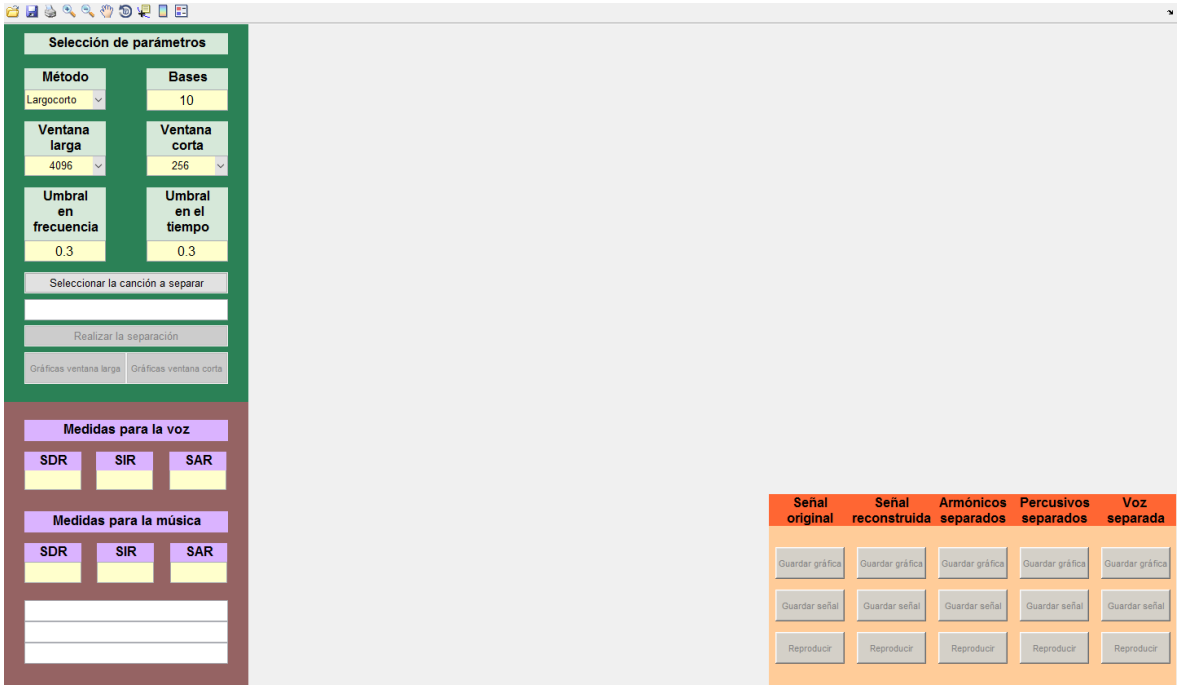

Figura 114: Inicio de la interfaz.

Como se puede ver hay cuatro zonas que se diferencia por sus colores:

1. Es la zona que tiene el fondo verde (figura 115), esta zona está reservada para seleccionar los parámetros con los cuales se quiere realizar la separación. Por defecto, al abrir la interfaz los parámetros seleccionados son los se han demostrado que realizan la separación óptima. Los valores “Bases”, “Umbral en frecuencia” y “Umbral en el tiempo” se introducen por teclado. Mientras que los valores “Método”, “Ventana larga” y “Ventana corta” no se pueden introducir por el teclado, sino que hay que seleccionarlos de los posibles valores propuestos, tal y como se puede ver en la figura 115. Aparte de la selección de los parámetros, en esta zona también se puede seleccionar la canción a la que se le quiere realizar la separación. También están los botones “Realizar la separación”, “Gráficas ventana larga” y “Gráficas ventana corta”, los cuales se explicarán más adelante.

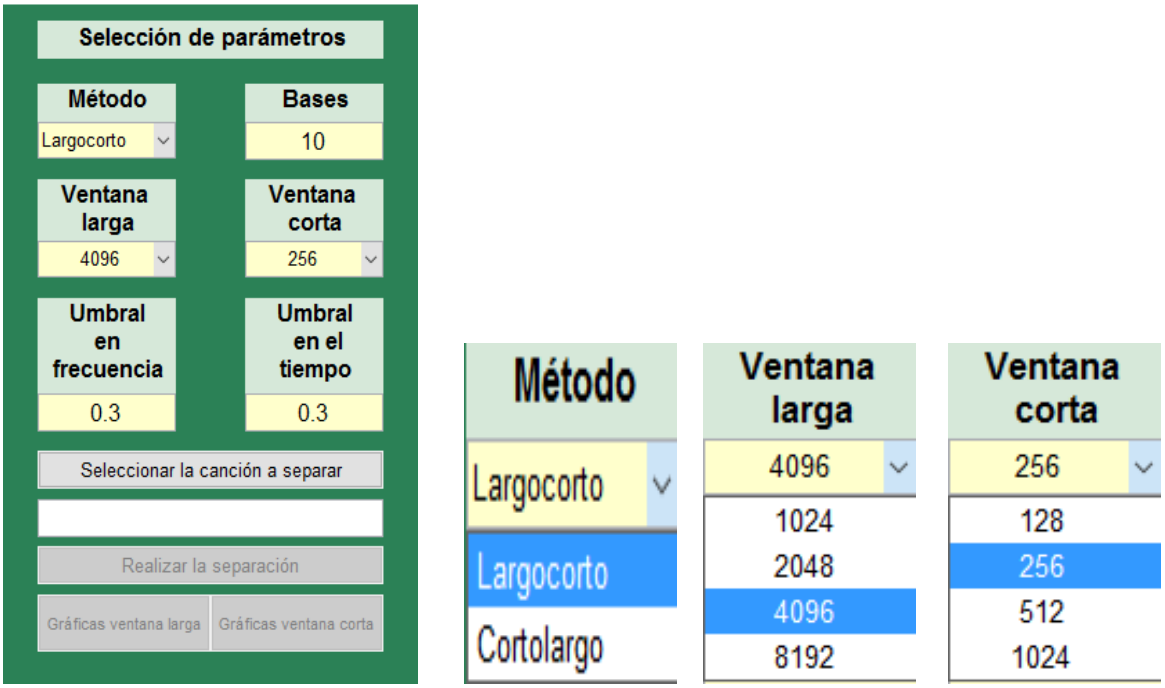

Figura 115: Sección para seleccionar los parámetros (imagen de la izquierda) y posibles valores de los parámetros Método, Ventana larga y Ventana corta (imágenes de la derecha).

2. Es la zona que tiene el fondo morado (figura 116), en esta zona se muestran las medidas de calidad objetivas tanto de la música como de la voz, una vez que se ha realizado la separación. Los tres recuadros con el fondo blanco que se encuentran en la parte inferior, sirven para mostrar los parámetros con los que se ha realizado la separación.

Figura 116: Sección para mostrar SDR, SIR y SAR.

3. Es la zona que tiene el fondo naranja (figura 117), en esta zona se pueden reproducir y guardar todas las pistas que se han generado durante la separación.

También se pueden guardar las gráficas (que se representan en la zona 4) que se generan cuando se pulsan los botones “Gráficas ventana larga” o “Gráficas ventana corta”.

Figura 117: Zona con los botones de guardar y reproducir.

4. Esta zona es la que está de color gris en la figura 114, esta zona está reservada para mostrar los distintos espectrogramas de cada una de las señales generadas en la separación. Las gráficas solo aparecen cuando se pulsan los botones “Gráficas ventana larga” o “Gráficas ventana corta”.

Como se puede ver en la figura 114, al abrir la interfaz solo se pueden seleccionar los parámetros y la canción que se quiere separar, el resto de botones están inactivos. Si se pulsa el botón “Seleccionar la canción a separar” aparece la ventana que se muestra en la figura 118. En esta ventana se selecciona la canción a separar. Una vez que se ha seleccionado la canción, su nombre aparece en el rectángulo de fondo blanco que hay debajo de dicho botón (figura 119) y además se habilita el botón “Realizar la separación”.

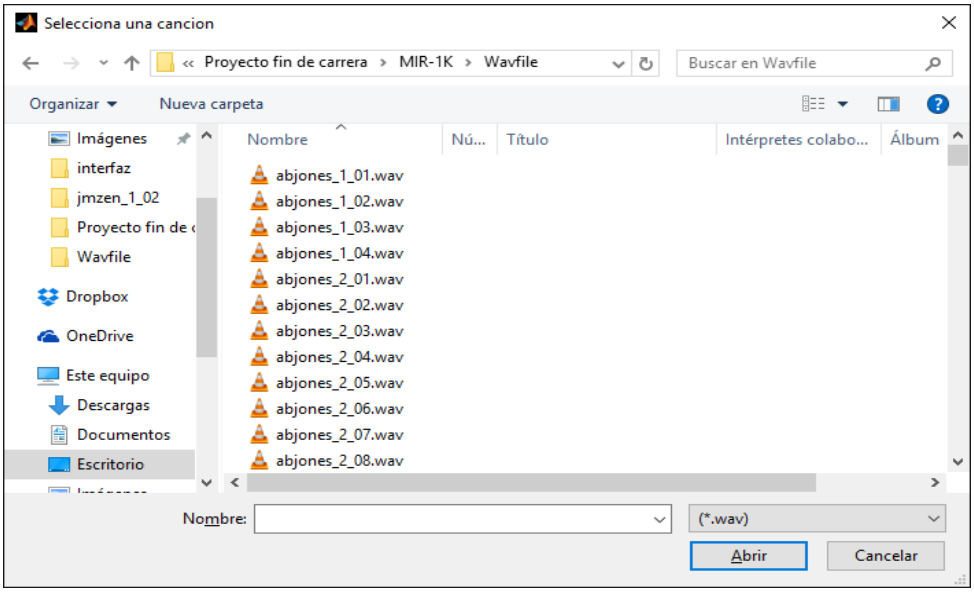

Figura 118: Ventana para seleccionar la canción a separar.

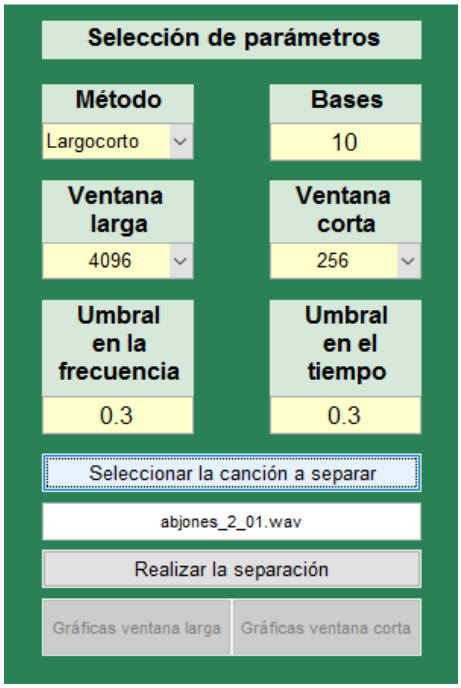

Figura 119: Zona 1 después de seleccionar la canción.

Al presionar el botón “Realizar la separación”, en la zona 2 aparecen se muestran los valores de SDR, SIR y SAR tanto para la música como para la voz (figura 120). En la parte inferior de esta misma zona también se muestran los parámetros con los que se ha realizado la separación. En la zona 1 se habilitan los botones “Gráficas ventana larga” y “Gráficas ventana corta”. Y por último en la zona 3 se habilitan los botones “Reproducir” y “Guardar señal” para cada una de las pistas generadas durante la separación.

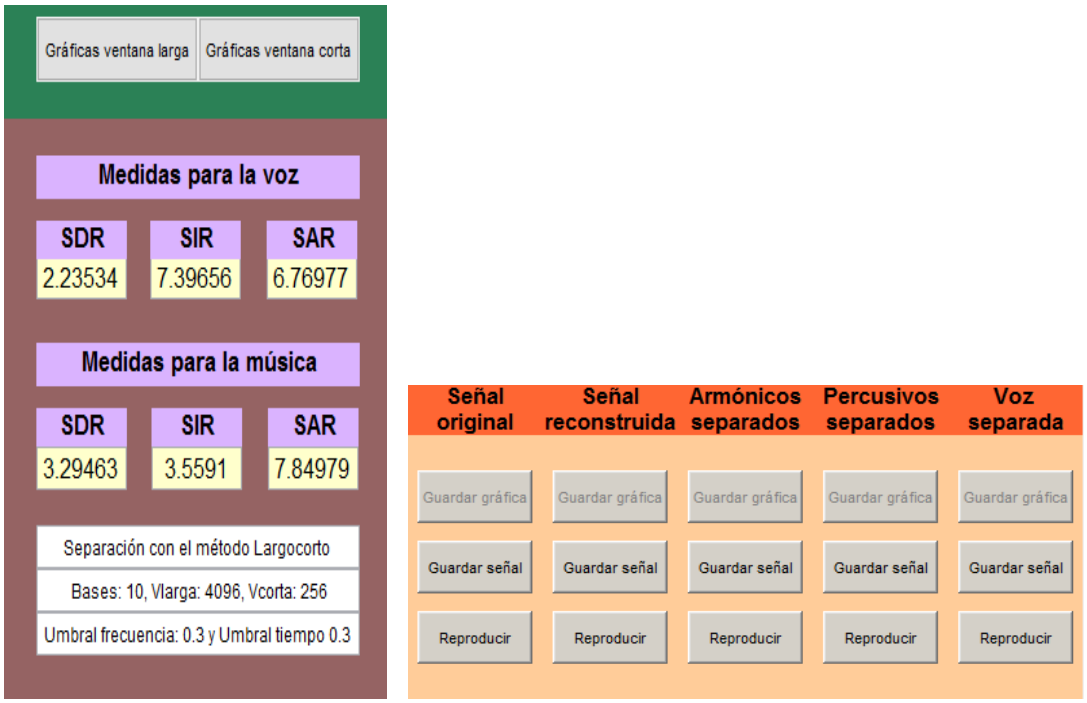

Figura 120: Zonas 1, 2 y 3 después de pulsar el botón “Realizar la separación”.

Si se pulsa el botón “Reproducir” de cualquiera de las señales, dicha señal se escuchará. Si se quiere guardar la señal, lo que hay que hacer es pulsar el botón “Guardar señal” y se abre la ventana de la figura 121. En esta ventana se puede guardar la señal con el nombre y en el directorio que se quiera.

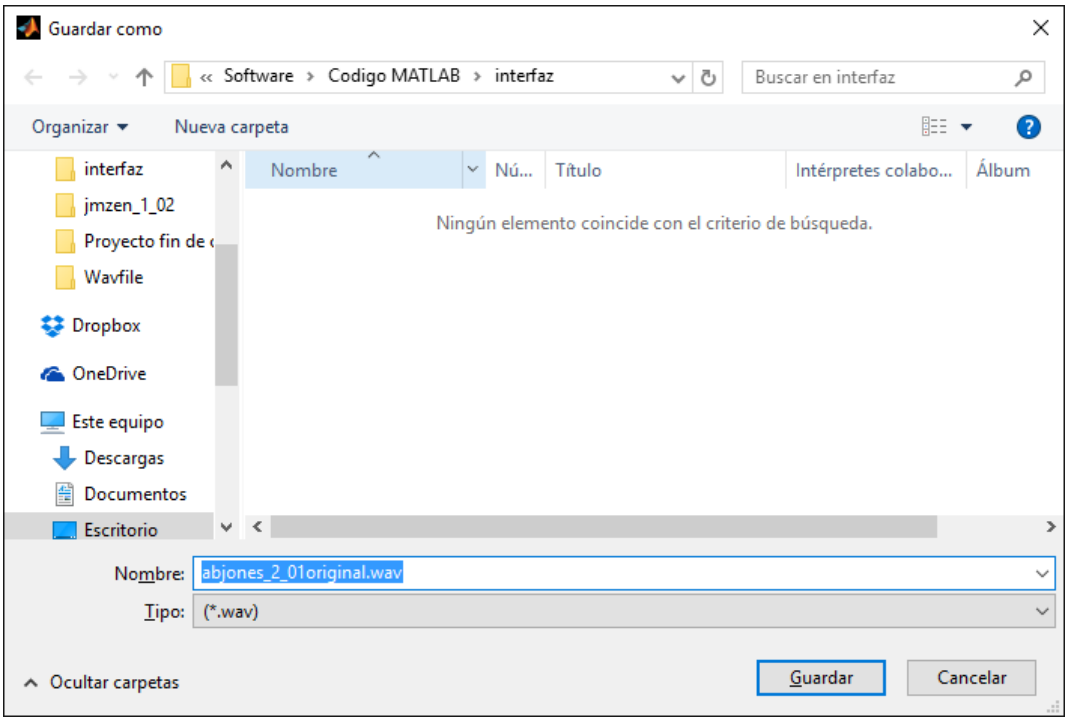

Figura 121: Ventana que se abre para guardar las señales de audio.

Para finalizar, si se pulsa el botón “Gráficas ventana larga”, en la zona 4 aparecen los espectrogramas con ventana larga de las señales generadas durante la separación (figura 122). Tras pulsar dicho botón se habilitan los botones “Guardar gráfica”, lo cuales al pulsarlos permiten guardar cada uno de los espectrogramas generados (figura 123).

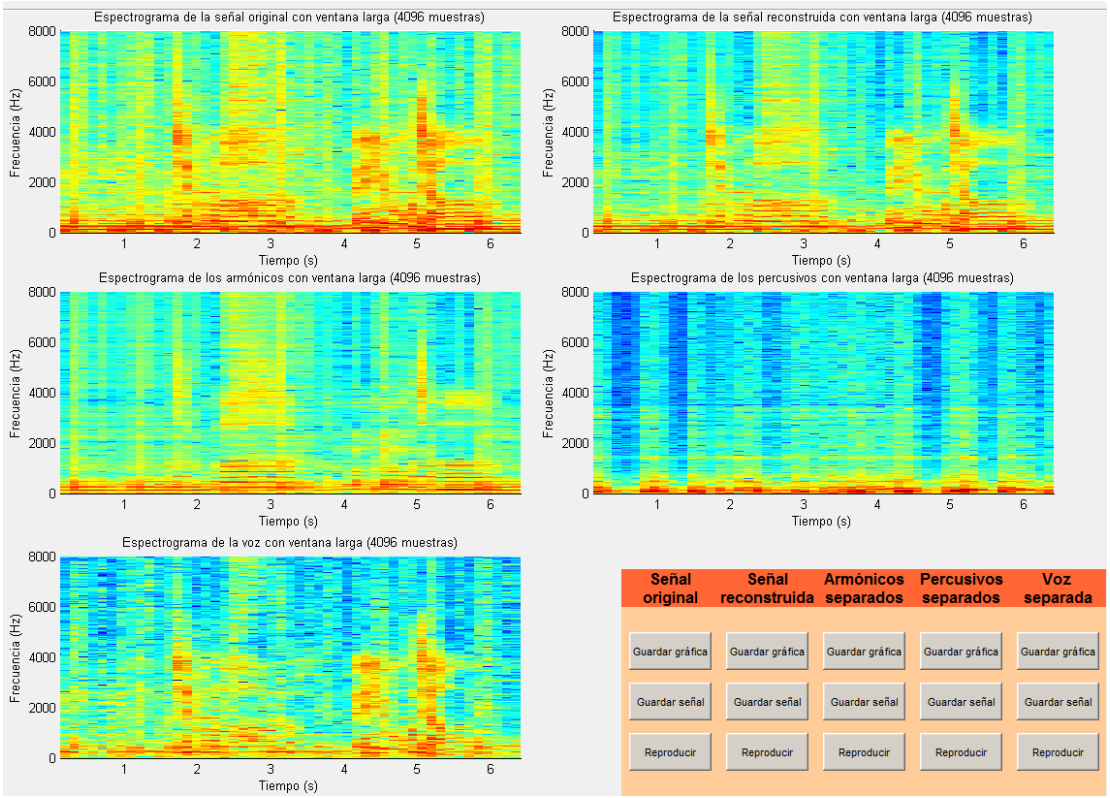

Figura 122: Zonas 3 y 4 después de pulsar “Gráficas ventana larga”.

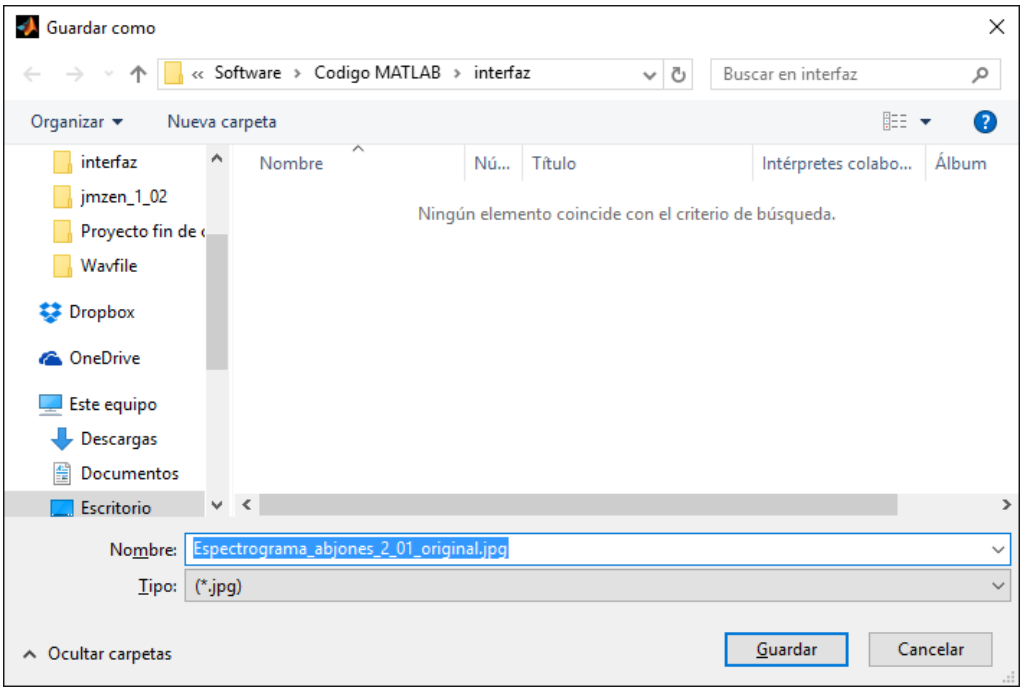

Figura 123: Ventana generada al pulsar el botón “Guardar gráfica”.

De la misma forma que ocurre al pulsar el botón “Gráficas ventana larga”, al pulsar el botón “Gráficas ventana corta” también aparecen en la zona 4 los espectrogramas con

ventana corta de las de las señales generadas durante la separación (figura 124). Y de nuevo aparecen habilitados los botones “Guardar gráfica”.

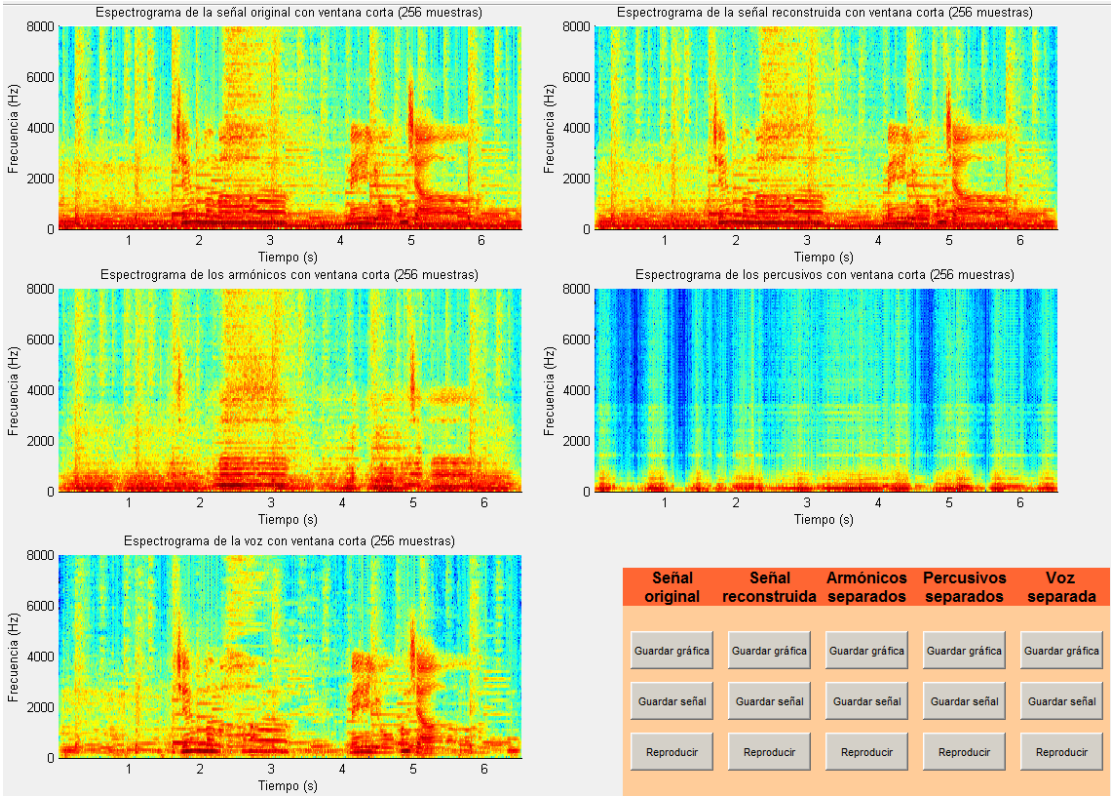

Figura 124: Zonas 3 y 4 después de pulsar “Gráficas ventana larga”.

Una característica distintiva de esta interfaz, es que aunque la separación se haya realizado con unos tamaños de ventana determinados, los espectrogramas se pueden representar con cualquier tamaño de ventana. Pero hay que tener en cuenta que las señales que se representan son el resultado de una separación con los parámetros que aparecen en la parte inferior de la zona 2. Esta característica se puede ver en las figuras 125 y 126. En la figura 125 aparecen los espectrogramas con una ventana larga de 1024 muestras, pero las señales a las que se les ha realizado el espectrograma han sido el resultado de realizar el proceso de separación con un tamaño de ventana larga de 4096 muestras, tal y como se puede ver en la figura 126.

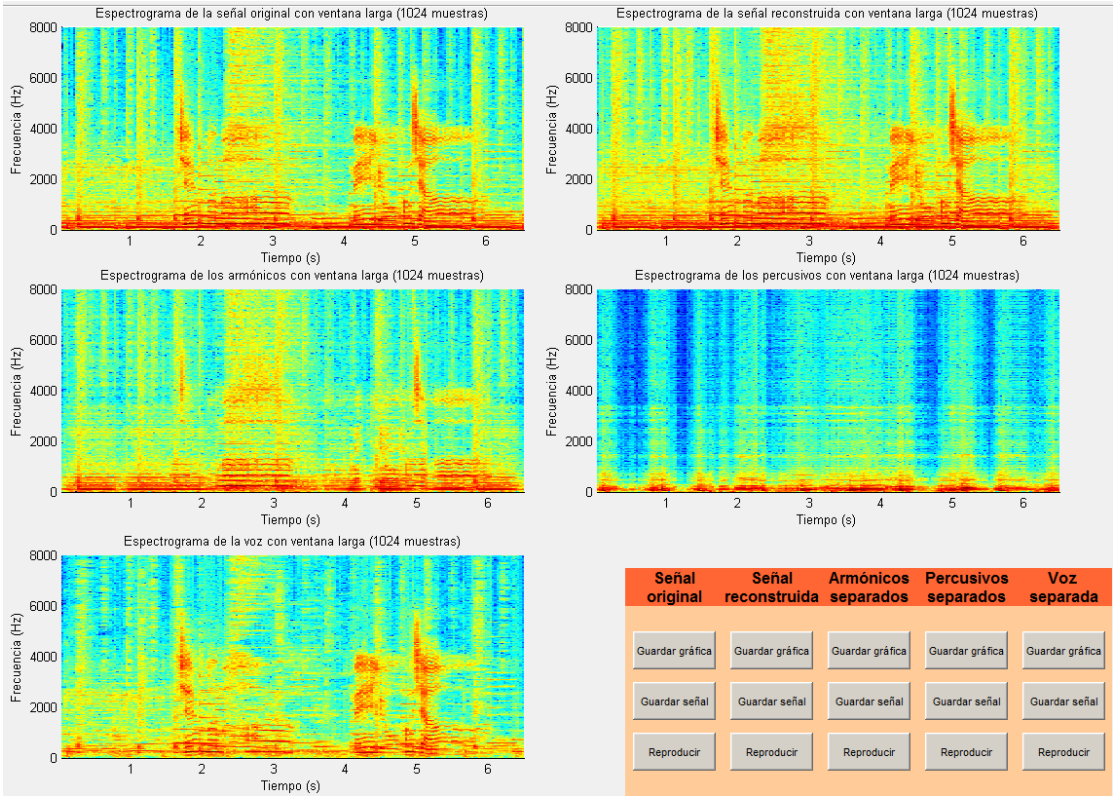

Figura 125: Espectrogramas con una ventana larga de 1024 muestras.

Figura 126: Parámetros de la separación.

ANEXO 2: INFORMACIÓN ADICIONAL

En el anexo cuyo nombre es “información adicional” está estructurado de la siguiente forma:

- **BASE DE DATOS MIR-1K:** esta carpeta contiene la base de datos que se ha utilizado, que como ya se ha explicado en el apartado 5.1, su nombre es MIR-1K.
- **CÓDIGO DE MATLAB:** esta carpeta contiene todo el código de MATLAB que ha sido necesario implementar.
- **PROCESO DE OPTIMIZACIÓN:** esta carpeta contiene información relativa al proceso de optimización (apartado 5.3).
- **SEPARACIONES CON LOS PARÁMETROS ÓPTIMOS:** esta carpeta contiene información relativa a las diferentes separaciones realizadas con los parámetros óptimos (apartado 5.4).

1. BASE DE DATOS MIR-1K

Esta carpeta está formada a su vez por dos carpetas más:

- **khair_titon:** contiene los fragmentos que se han utilizado para realizar el proceso de optimización, concretamente son los fragmentos interpretados por “khair” y por “titon”.
- **Wavfile_sin_khair_titon:** contiene los fragmentos que se han utilizado para realizar las diferentes separaciones con los parámetros óptimos, es decir, todos los fragmentos de la base de datos excepto los interpretados por “khair” y “titon”.

2. CÓDIGO DE MATLAB

Esta carpeta está formada a su vez por tres carpetas más:

- **INTERFAZ:** esta carpeta contiene las funciones que han sido necesarias implementar para realizar la interfaz de usuario.
- **PROCESO DE OPTIMIZACIÓN:** esta carpeta contiene todas las funciones que han sido necesarias implementar para realizar la etapa de optimización. Además contiene otra carpeta denominada “GRÁFICAS”, en la que están implementadas las funciones necesarias para extraer y representar los resultados de SDR, SIR y SAR.
- **SEPARACIÓN CON PARÁMETROS ÓPTIMOS:** esta carpeta contiene todas las funciones que han sido necesarias implementar para realizar las separaciones con los parámetros óptimos. Además contiene otra carpeta denominada “GRÁFICAS”,

en la que están implementadas las funciones necesarias para extraer y representar los resultados de SDR, SIR y SAR.

3. PROCESO DE OPTIMIZACIÓN

Esta carpeta contiene la estructura “matriztrotal”, que recoge toda la información generada durante la etapa de optimización. También están recogidas en forma de vectores cada una de las medidas de calidad objetivas (SDR, SIR y SAR) para el método Largocorto y Cortolargo, tanto para la música como para la voz.

4. SEPARACIONES CON LOS PARÁMETROS ÓPTIMOS

Esta carpeta está formada a su vez por tres carpetas más:

- **-5 dB:** esta carpeta contiene la estructura “matrizoptima”, que recoge toda la información generada durante la etapa de separación a -5 dB (apartado 5.4.1). También contiene los vectores de cada una de las medidas de calidad objetivas (SDR, SIR y SAR), tanto para la música como para la voz. Además hay dos carpetas más en las que hay ejemplos de separaciones de alta calidad (SEPARACIONES DE ALTA CALIDAD) y separaciones de baja calidad (SEPARACIONES DE BAJA CALIDAD). A su vez en cada una de estas carpetas hay dos carpetas más:
  - **SEGÚN LAS MEDIDAS DE LA MÚSICA:** en esta carpeta hay ejemplos de separaciones de alta (SEPARACIONES DE ALTA CALIDAD) o baja (SEPARACIONES DE BAJA CALIDAD) calidad, teniendo en cuenta únicamente loas medidas de calidad objetivas de la música.
  - **SEGÚN LAS MEDIDAS DE LA VOZ:** en esta carpeta hay ejemplos de separaciones de alta (SEPARACIONES DE ALTA CALIDAD) o baja (SEPARACIONES DE BAJA CALIDAD) calidad, teniendo en cuenta únicamente loas medidas de calidad objetivas de la voz.

- **0 dB:** esta carpeta contiene la estructura “matrizoptima”, que recoge toda la información generada durante la etapa de separación a 0 dB (apartado 5.4.2). También contiene los vectores de cada una de las medidas de calidad objetivas (SDR, SIR y SAR), tanto para la música como para la voz. Además hay dos carpetas más en las que hay ejemplos de separaciones de alta calidad (SEPARACIONES DE ALTA CALIDAD) y separaciones de baja calidad

(SEPARACIONES DE BAJA CALIDAD). A su vez en cada una de estas carpetas hay dos carpetas más:

- **SEGÚN LAS MEDIDAS DE LA MÚSICA:** en esta carpeta hay ejemplos de separaciones de alta (SEPARACIONES DE ALTA CALIDAD) o baja (SEPARACIONES DE BAJA CALIDAD) calidad, teniendo en cuenta únicamente loas medidas de calidad objetivas de la música.
- **SEGÚN LAS MEDIDAS DE LA VOZ:** en esta carpeta hay ejemplos de separaciones de alta (SEPARACIONES DE ALTA CALIDAD) o baja (SEPARACIONES DE BAJA CALIDAD) calidad, teniendo en cuenta únicamente loas medidas de calidad objetivas de la voz.

- **5 dB:** esta carpeta contiene la estructura “matrizoptima”, que recoge toda la información generada durante la etapa de separación a 5 dB (apartado 5.4.3). También contiene los vectores de cada una de las medidas de calidad objetivas (SDR, SIR y SAR), tanto para la música como para la voz. Además hay dos carpetas más en las que hay ejemplos de separaciones de alta calidad (SEPARACIONES DE ALTA CALIDAD) y separaciones de baja calidad (SEPARACIONES DE BAJA CALIDAD). A su vez en cada una de estas carpetas hay dos carpetas más:
  - **SEGÚN LAS MEDIDAS DE LA MÚSICA:** en esta carpeta hay ejemplos de separaciones de alta (SEPARACIONES DE ALTA CALIDAD) o baja (SEPARACIONES DE BAJA CALIDAD) calidad, teniendo en cuenta únicamente loas medidas de calidad objetivas de la música.
  - **SEGÚN LAS MEDIDAS DE LA VOZ:** en esta carpeta hay ejemplos de separaciones de alta (SEPARACIONES DE ALTA CALIDAD) o baja (SEPARACIONES DE BAJA CALIDAD) calidad, teniendo en cuenta únicamente loas medidas de calidad objetivas de la voz.